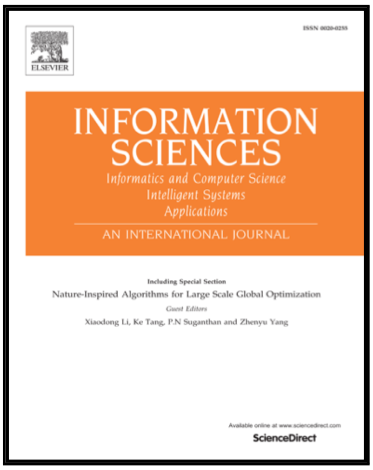

ISSN 0020-0255
ELSEVIER
INFORMATION SCIENCES
Informatics and Computer Science
Intelligent Systems
Applications
AN INTERNATIONAL JOURNAL
Including Special Section
Nature-Inspired Algorithms for Large Scale Global Optimization
Guest Editors
Xiaodong Li, Ke Tang, P.N Suganthan and Zhenyu Yang
Available online at www.sciencedirect.com
ScienceDirect

Incorporating Product Description to Sentiment Topic Models for Improved Aspect-based Sentiment Analysis

Reinald Kim Amplayo, Seanie Lee, Min Song

PII: S0020-0255(18)30349-9
DOI: 10.1016/j.ins.2018.04.079
Reference: INS 13622

To appear in: *Information Sciences*

Received date: 8 August 2017
Revised date: 15 January 2018
Accepted date: 30 April 2018

Please cite this article as: Reinald Kim Amplayo, Seanie Lee, Min Song, Incorporating Product Description to Sentiment Topic Models for Improved Aspect-based Sentiment Analysis, *Information Sciences* (2018), doi: 10.1016/j.ins.2018.04.079

Incorporating Product Description to Sentiment Topic Models for Improved Aspect-based Sentiment Analysis

Reinald Kim Amplayo[a], Seanie Lee[a], Min Song[a,*]

[a] *Yonsei University, Seoul, Korea*

Abstract

Sentiment topic models are used as unsupervised methods to solve the specific problems of the general aspect-based sentiment analysis (ABSA) problem. One of the main problems of the technique is its substandard aspect term extraction, which leads to difficulties in aspect label determination. This paper is focused on improving the aspect term extraction of topic models by incorporating product descriptions to the current state-of-the-art sentiment topic model, Aspect Sentiment Unification Model (ASUM). We present two models that extend from ASUM differently to leverage on the information found in the product description: Seller-aided Aspect-based Sentiment Model (SA-ASM) and Seller-aided Product-based Sentiment Model (SA-PSM). SA-ASM has its topic distribution inside the review while SA-PSM has its topic distribution inside the product description. Based on experiments conducted to reviews of laptops and mobile phones, results show that SA-ASM performs better in micro-level problems such as sentiment classification and aspect assignment and SA-PSM performs better in macro-level problems like aspect category detection. Both models achieve better performances compared to current topic modeling methods for the ABSA problem.

Keywords: aspect-based sentiment analysis, aspect extraction, topic models, product description

[*] Corresponding author

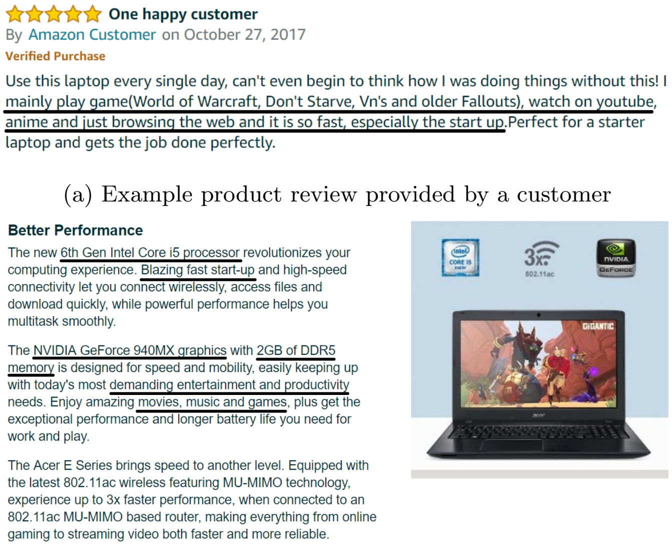
Email addresses: rktamplayo@yonsei.ac.kr (Reinald Kim Amplayo), lsnfamily02@naver.com (Seanie Lee), min.song@yonsei.ac.kr (Min Song)

1. Introduction

The web is flooded with review data that customers write about the products and services they paid for. It is an important task to automatically extract in these large volumes of literature information useful to both customers and sellers. This task requires work beyond a simple sentiment analysis to judge whether a review is positive or negative for a product. In one review a customer wrote, several aspects of a product may be presented, and the user may have different opinions regarding each of the presented aspects. Hence, a more fine-grained sentiment analysis should be considered. In most literature [23, 25, 24], this set of wide range of sub-problems is collectively called **aspect-based sentiment analysis** (ABSA).

In this paper, we consider three sub-problems of ABSA: *sentiment classification*, *aspect assignment*, and *aspect category detection*. Sentiment classification is a problem where one is given a text and is asked to assign a sentiment to it. The sentiment can be many things; it can be a polarity (i.e. positive, negative), it can be a rate (i.e. 80% positive), or it can be an emotion (i.e. anger, happiness, etc.). In this paper, we focus on the classification of sentiment based on polarity. Aspect assignment is a problem where one is given a set of words and is asked to assign a representative term, an aspect, that is assumed to be related to the product. For example, the set of words "battery charge charger long last" is assigned an aspect "battery". Lastly, aspect term prediction is a problem that is reverse with aspect assignment. The problem states that given an aspect, determine the list of terms that are related to it. For example, the aspect "memory" can have terms "sd card ram gb" but cannot have terms like "battery" and "audio".

There have been lots of efforts done in solving these three sub-problems. One approach is to use topic models, specifically extensions to the widely known Latent Dirichlet Allocation (LDA) topic model [4]. Currently, there are two kinds of state-of-the-art topic models. The first sentiment topic model introduced in the literature is the Joint Sentiment Topic Model [17]. The second senti-



★★★★★ One happy customer
By Amazon Customer on October 27, 2017
Verified Purchase

Use this laptop every single day, can't even begin to think how I was doing things without this! I mainly play game(World of Warcraft, Don't Starve, Vn's and older Fallouts), watch on youtube, anime and just browsing the web and it is so fast, especially the start up.Perfect for a starter laptop and gets the job done perfectly.

(a) Example product review provided by a customer

Better Performance

The new 6th Gen Intel Core i5 processor revolutionizes your computing experience. Blazing fast start-up and high-speed connectivity let you connect wirelessly, access files and download quickly, while powerful performance helps you multitask smoothly.

The NVIDIA GeForce 940MX graphics with 2GB of DDR5 memory is designed for speed and mobility, easily keeping up with today's most demanding entertainment and productivity needs. Enjoy amazing movies, music and games, plus get the exceptional performance and longer battery life you need for work and play.

The Acer E Series brings speed to another level. Equipped with the latest 802.11ac wireless featuring MU-MIMO technology, experience up to 3x faster performance, when connected to an 802.11ac MU-MIMO based router, making everything from online gaming to streaming video both faster and more reliable.

(b) Example product description provided by the manufacturer

Figure 1: An example of a product with information widely available in Amazon

ment topic model is an extension of the previous model, Aspect and Sentiment Unification Model (ASUM) [14]. However, due to its mediocre performance in extracting aspects, subsequent works [6, 7, 3, 34] have focused on improving the aspect extraction part of the model.

The main contribution of this paper is an extension of the current state-of-the-art topic models. Specifically, we consider the addition of the product descriptions provided by the seller to the sentiment topic models as another dimension to improve the aspect extraction functionality of the model. When sellers post a product to sell to the customers, they usually provide product descriptions which describe the functionalities of the product in detail. The contribution is from the observation that product description contains a clearer and more correct aspect terms for the product. Normally, sentiment classification models only use the customer reviews, as shown in Figure 1a, to build a classification model. In this paper, we attempt to incorporate product descriptions, as shown in Figure 1b, to the sentiment topic models.

The intuition behind the improvement of incorporating product description can be demonstrated by looking at the example product review and description in Figure 1. The review provided by the happy customer shows limited aspect terms that indicates the aspect the customer likes is the performance of the product. On the other hand, the product description provided by the manufacturer is more detailed and includes many aspect terms that describes the performance of the product. Therefore, incorporating the product description helps in improving the performance of sentiment topic models.

Specifically, we present two models that incorporate product description: Seller-aided Aspect-based Sentiment Model (SA-ASM) and Seller-aided Product-based Sentiment Model (SA-PSM). The difference between both models is the location of the topic distribution, with SA-ASM placing it inside the review plate and SA-PSM placing it inside the product description plate. Both models are applied to the tasks mentioned above. We compare our proposed models to ASUM [14] and JAST [34], and our experiments using two different datasets show that the models perform better compared to other models in the following:

- **Sentiment classification**: SA-ASM performs better than other topic models using both datasets. This also holds true both when the dataset used is a small dataset (i.e. only the boundary sentiment labels, that is review scores 1 and 5, are used) and when the dataset used is a large dataset (i.e. non-boundary sentiment labels, that is review scores 2 and 4, are also considered).
- **Aspect assignment**: SA-ASM generally performs better compared to other topic models in terms of diversity, specificity, and agreeability using both datasets.
- **Aspect term extraction**: SA-PSM performs better compared to other topic models in terms of precision and coherence.

The rest of the paper is organized as follows. Section 2 discusses the related work on aspect-based sentiment analysis, specifically on topic models. We provide brief explanations on the base models used in this paper in Section

3. Section 4 explains in detail our proposed models to incorporate product description to sentiment topic models. Section 5 describes briefly the experimental setup and reports the results from experiments on multiple subtasks of aspect-based sentiment analysis. Finally, we conclude in Section 6 providing possible future works and improvements.

2. Related work

In this section, we briefly describe related work in unsupervised methods for aspect-based sentiment analysis, specifically on topic models that motivated the methodology proposed in this paper.

2.1. Aspect-based sentiment analysis

After its introduction, aspect-based sentiment analysis (ABSA) has attracted a lot of researchers in the machine learning and natural language processing community. There are two kinds of techniques that have been introduced: supervised and unsupervised techniques. Since our proposed models are unsupervised techniques, we focus on unsupervised techniques. One common technique is to separate the problem into two parts: aspect extraction and sentiment analysis. [2] used a generalized method to learn multi-word aspects and employed a set of heuristic rules to take into account the influence of the opinion words. They then pruned the list of aspects to remove incorrect aspects. The extracted aspects are then used in conjunction with a sentiment classifier. [16] used pattern-based bootstrapping to extract product aspects and opinion words. They used prevalence and reliability to assess both patterns and features. These are then clustered and finally fed to a sentence sentiment strength calculator which calculates sentiment strength using a scoring formula. Both [5] and [13] used Latent Dirichlet Allocation topic model to extract aspects from reviews. They differ in their technique on classifying the sentiment, the former using graph-based propagation method constructed using adjectives extracted from the reviews and the latter using different kinds of sentiment lexicons and doing a naive word search.

Recent literature also provides combined unsupervised techniques for the ABSA problem. [30] used a feature-based heuristic that utilizes a SentiWordNet-based method with different linguistic feature selections composed of adjectives, adverbs, verbs and n-grams. SentiWordNet was also used in a method to compute the document-level sentiment for each review. [8] used word2vec to solve both aspect extraction and sentiment classification. They built knowledge graphs that connects aspect terms and opinion words to each other, with edge weights assigned using the terms' cosine similarity difference based on the word2vec model. They calculated the PageRank value of all the terms in the knowledge graph and extracted the most important aspect terms. Then they used the word2vec model to compare all the terms in the sentence with a set of positive and negative terms to calculate the polarity of the sentence. [28] made use of discourse units found in reviews. They first separate texts into discourse units and segment them based on the aspects. The segmented aspects are then aggregated to calculate the polarity score for each aspects. However, approaches based on topic models are proven to perform better [17, 14]. These methods are explored in the next section.

Other recent methods to ABSA include neural network- and deep learning-based methods such as recursive neural conditional random fields [36], recursive neural networks [21], restricted Boltzmann machines [33], hierarchical bidirectional LSTMs [27], and attention-based LSTMS [37]. Deep learning methods perform well on sentiment classification especially when domain adaptation is necessary [9]. However, these methods are purely supervised methods. Our model does not require labelled review dataset and just needs a small sentiment lexicon for training. Thus, we cannot directly compare these methods to our model.

2.2. Topic models for ABSA

A more successful combined unsupervised technique is by using topic models to infer both sentiment and aspects. It has been shown that incorporating topic information to a sentiment analysis model improves the performance of the model [20]. [19] used a mixture model called Topic Sentiment Mixture (TSM) model that uses a hidden Markov model structure to extract topic life cycles and sentiment dynamics. They separately created language models for topics and sentiments. The model has been extended by [35] to analyze sentiments of online product reviews. Another sentiment topic model that follows TSM is Multi-Aspect Sentiment (MAS) topic model [31] in which the model accepts as input a review as well as a set of predefined aspects and the review's corresponding sentiment score. Their goal is to divide the full review into phrases regarding specific aspects and cluster them based on the aspects. Following the previous attempts to create a sentiment topic model is the Joint Sentiment-Topic (JST) model [17]. The intuition behind JST model is simple; they assumed that a word has both a topic and a sentiment and the topic of the word depends on the intended sentiment of the word. This resulted to the addition of a sentiment random variable which depends on a multinomial distribution drawn from a Dirichlet hyperparameter. JST model is immediately extended by Aspect Sentiment Unification Model (ASUM) [14]. The extension is based on the assumption that within a sentence in a review, there is only one sentiment pertaining to at most one aspect. This resulted into adding a sentence dimension into the JST model. It is also good to note that both JST and ASUM use a small set of sentiment seed words to guide the sentiment distribution of the model. However, the method in which they insert the guidance of the seed words is different; JST only utilizes the lists during initialization while ASUM incorporates the lists to the topic-word Dirichlet hyperparameter completely until the end of the training.

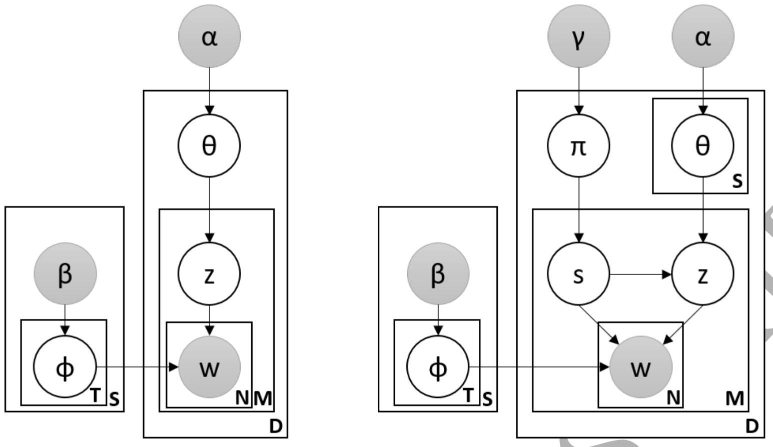
In spite the success of JST and ASUM, more recent literature is still trying to improve the aspect extraction part of the model in many ways. In [6, 7], they used the notion of *must-sets* and *cannot-sets* in which it restricts two terms to be in the same set or to be in different sets, respectively. This notion is used to extend JST and improve the extraction of aspect terms. [3] extended LDA by introducing a Markov chain that makes subsequent words more likely to be in the same topic. They also introduced a new random variable to detect multiword aspects. This model, however, does not take note of the sentiment label of a word. The most recent literature on sentiment topic models used a combination of LDA and hidden Markov models [26] and an extension of JST by applying maximum entropy and lifelong machine learning [34]. The former was extended to explicitly capture topic coherence and sentiment consistency and to accurately extract latent aspects and corresponding sentiment polarities. The latter introduced a unified topic model called Joint Aspect-based Sentiment Topic (JAST) model that can be used for multiple ABSA problems. They also improved their own model by applying lifelong machine learning in order to make use of datasets from other domains. Recent models for aspect-based sentiment analysis limit their context to the reviews, which may only contain limited informative aspect terms for inference.

Our proposed models are most similar to [22], where they incorporate product specifications to jointly model product reviews and specifications. They used information from a structured list of specifications of the product to improve the extraction of aspects. However, acquisition of a structured product specification is a very expensive task. In our study, we incorporate product descriptions, which are mostly unstructured and cheap since sellers almost always provide these information for the customers.

3. SLDA and ASUM

In this section, we discuss the two main topic models that helped form the base of our proposed models, the two topic models proposed in [14], Sentence-LDA (SLDA) and Aspect Sentiment Unification Model (ASUM).

The basic topic model, Latent Dirichlet Allocation (LDA) topic model [4], assumes a bag-of-words representation, which makes the model simple. This means that it only considers naive co-occurrence between two words and does not take note of other factors in text such as parts-of-speech, grammar, and proximity. Although the former two does not really affect a lot when extracting aspects, words about an aspect usually co-occur within close proximity to each



(a) Sentence-LDA

(b) ASUM

Figure 2: Graphical representation of SLDA and ASUM. Meanings of the notations are described in Table 1.

other. Sentence-LDA (SLDA) lifts this constraint by using a full sentence as a window for co-occurrence. This constrains the words within a sentence to co-occur with words in other sentence, although both of them are still part of the same document. SLDA is shown as graphical models in Figure 2a. As shown, the simple extension can be done by adding an inner plate that bundle up words in one sentence. By doing this, extraction of aspects improves over a simple LDA topic model.

The generative process of SLDA is as follows:

1. For every aspect z, draw a word distribution $\phi \sim Dirichlet(\beta)$
2. For each review d,
   (a) Draw the review's aspect distribution $\theta \sim Dirichlet(\alpha)$
   (b) For each sentence,
      i. Choose an aspect $z \sim Multinomial(\theta)$
      ii. Generate words $w \sim Multinomial(\phi)$

Based on SLDA, the Joint Sentiment Topic (JST) model [17] is extended into Aspect Sentiment Unification Model (ASUM). ASUM follows a generative

Table 1: Meanings of the notations used in the models.

D	Number of products
R	Number of reviews per product
M, M'	Number of sentences of each product description (of each review)
N, N'	Number of words of each sentence of a product description (of a review)
S	Number of sentiments
T	Number of topics (i.e. aspects)
w, x	word in a product description (in a review)
y, z	aspect in a product description (in a review)
s	sentiment
θ	multinomial distribution over aspects
ϕ	multinomial distribution over words
π	multinomial distribution over sentiments
α	Dirichlet prior over θ
β	Dirichlet prior over ϕ
γ	Dirichlet prior over π

process that is similar to the following way of writing a review. First, a reviewer first decides to write a review about a general topic (e.g. laptop) and his/her corresponding sentiment rating about it (e.g. 70% happy). Then, he/she decides which aspects about a general topic make him/her feel his/her sentiment (e.g. 90% happy about the memory, 30% unhappy about the battery, etc.). Lastly, he/she thinks of the sentences that he/she will write and their corresponding sentiment and aspect. Figure 2b shows the graphical representation of the described model.

The generative process of ASUM is as follows:

1. For every pair of sentiment s and aspect z, draw a word distribution $\phi \sim Dirichlet(\beta)$
2. For each review d,
    (a) Draw the review's sentiment distribution $\pi \sim Dirichlet(\gamma)$
    (b) For each sentiment s, draw an aspect distribution $\theta \sim Dirichlet(\alpha)$

(c) For each sentence,

i. Choose a sentiment $s \sim Multinomial(\pi)$

ii. Choose an aspect $z \sim Multinomial(\theta)$

iii. Generate words $w \sim Multinomial(\phi)$

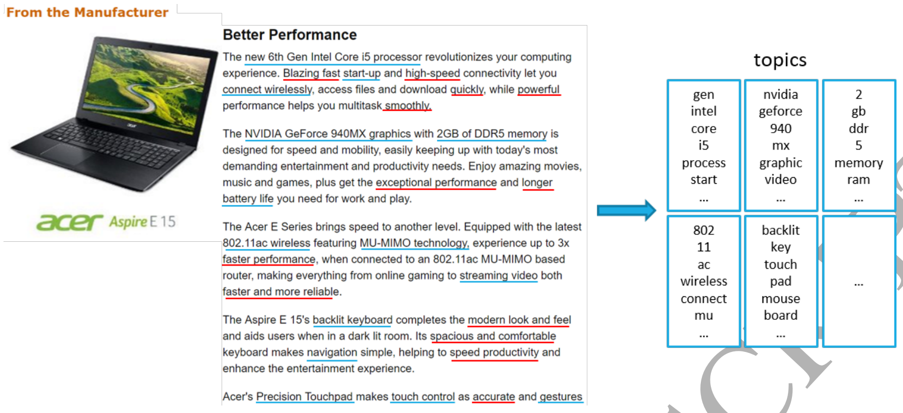
4. Incorporating product description

In this paper, we extend the generative process of ASUM by including the events that happen before a reviewer writes a review. Before a reviewer writes a review, he/she buys the product and tries to use it. While using it, he/she checks whether the product is good based on its several aspects. After examining the product in detail, the reviewer starts to write the review as explained in the generative process of ASUM. In this section, we show two kinds of methods in order to model the generative process as described above.

Since we cannot directly model the product usage of the reviewer, we need to find another method that similarly corresponds to the product usage of the reviewer. We posit that the product description provided by the seller can model the product usage of the reviewer similarly. This is due to the fact that the product description is assumed to list all the aspects found in the product.

Incorporating product descriptions help in increasing the performance on all subtasks of ABSA since it provides better aspect term extraction. Product descriptions contain expertly written descriptions about the product (e.g. laptop) and its aspect (e.g. performance, battery, etc.). These descriptions may not be found in reviews because most of the reviewers are ordinary users who may not know how to describe the aspects of the products properly. Based on this observation, our models improve over previous models which only use user reviews as their context.

The main problem in incorporating product descriptions is that the information written in the text is biased towards the positive sentiment. Sellers use the product description not only to inform the consumers on product specifications (i.e. aspects) but also to appeal to the consumers why they should buy the



From the Manufacturer

Better Performance

The new 6th Gen Intel Core i5 processor revolutionizes your computing experience. Blazing fast start-up and high-speed connectivity let you connect wirelessly, access files and download quickly, while powerful performance helps you multitask smoothly.

The NVIDIA GeForce 940MX graphics with 2GB of DDR5 memory is designed for speed and mobility, easily keeping up with today's most demanding entertainment and productivity needs. Enjoy amazing movies, music and games, plus get the exceptional performance and longer battery life you need for work and play.

The Acer E Series brings speed to another level. Equipped with the latest 802.11ac wireless featuring MU-MIMO technology, experience up to 3x faster performance, when connected to an 802.11ac MU-MIMO based router, making everything from online gaming to streaming video both faster and more reliable.

The Aspire E 15's backlit keyboard completes the modern look and feel and aids users when in a dark lit room. Its spacious and comfortable keyboard makes navigation simple, helping to speed productivity and enhance the entertainment experience.

Acer's Precision Touchpad makes touch control as accurate and gestures

acer Aspire E 15
topics
gen intel core i5 process start …
nvidia geforce 940 mx graphic video …
2 gb ddr 5 memory ram …
802 11 ac wireless connect mu …
backlit key touch pad mouse board …
…

Figure 3: Aspect extraction from product description

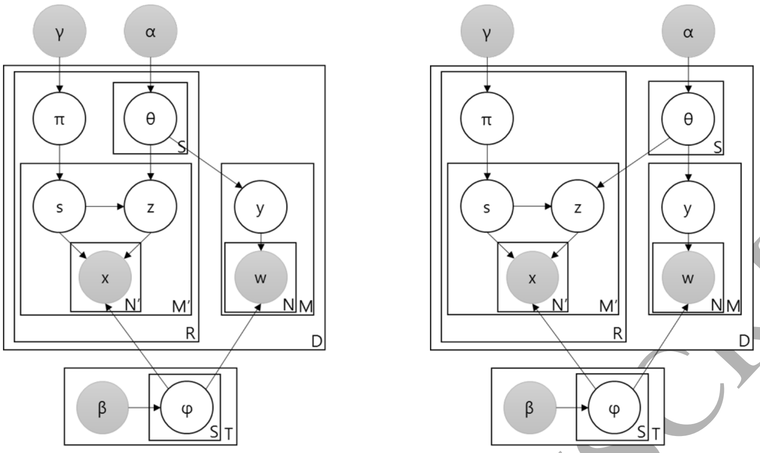
product. An example of a product description and possible extracted aspects is shown in Figure 3. Because of this, it is with necessity that the modeling of the product description should be without the sentiment label. To this end, we use the intuition of SLDA to model the product description and of ASUM to model the product reviews.

We introduce two sentiment topic models that incorporates product description: Seller-aided Aspect-based Sentiment Model (SA-ASM) and Seller-aided Product-based Sentiment Model (SA-PSM). The graphical representation of the models are shown in Figure 4.

4.1. Aspect based: SA-ASM

One way to combine both product reviews and product descriptions in a topic model can be explained informally as follows. After the reviewer bought the product, the reviewer tries to use the product with the purpose of examining it in detail based on its aspects. The reviewer looks at each small detail regarding each aspect of the product, which is usually found in the product description. Then, the reviewer writes the product review through the same process as described above.

We call this model Seller-aided Aspect-based Sentiment Model (SA-ASM). The model is shown graphically in Figure 4a. As described above, in the model,



(a) Seller-aided Aspect-based Sentiment Model

(b) Seller-aided Product-based Sentiment Model

Figure 4: Graphical representation of SA-ASM and SA-PSM. Meanings of the notations are described in Table 1.

the review looks directly to the sentences of the product description to seek help for extracting aspects. The model is expected to have a good performance on specific tasks such as sentiment classification and aspect assignment, because of the location of the topic distribution θ being inside the review plate.

Formally, the generative process of the topic model is as follows:

1. For every pair of sentiment s and aspect z, draw a word distribution $\phi \sim Dirichlet(\beta)$
2. For each product d,
    (a) For each review r,
        i. Draw the document's sentiment distribution $\pi \sim Dirichlet(\gamma)$
        ii. For each sentiment s, draw its aspect distribution $\theta \sim Dirichlet(\alpha)$
        iii. For each sentence in the review m',
            A. Choose a sentiment $s \sim Multinomial(\pi)$
            B. Choose an aspect $y \sim Multinomial(\theta)$
            C. Generate words $w \sim Multinomial(\phi)$

(b) For each sentence in the product description m,

i. Choose an aspect $y \sim Multinomial(\theta)$

ii. Generate words $w \sim Multinomial(\phi)$

We use collapsed Gibbs sampling [10] to estimate the latent variables ϕ, θ, and π. At each step of the Markov chain, we separately choose values for the sentiment and aspect of the review and the aspect of the product description. The sentiment and aspect of the ith sentence of a review are chosen using the conditional probability

$$\Pr(s_i = j, z_i = k | s_{-i}, z_{-i}, x) \propto \frac{C^{DRS}_{drj} + \gamma_j}{\sum_{j'=1}^{S}(C^{DRS}_{drj'} + \gamma_{j'})} \frac{C^{DRST}_{drjk} + \alpha_{r_k}}{\sum_{k'=1}^{T}(C^{DRST}_{drjk'} + \alpha_{r_{k'}})}$$
$$\frac{\Gamma(\sum_{w'=1}^{W}(C^{STW}_{jkw'} + \beta_{jw'}))}{\Gamma(\sum_{w'=1}^{W}(C^{STW}_{jkw'} + \beta_{jw'}) + m_i^{(r)})} \prod_{w'=1}^{W} \frac{\Gamma(C^{STW}_{jkw'} + \beta_{jw'} + m_{iw}^{(r)})}{\Gamma(C^{STW}_{jkw'} + \beta_{jw'})} \quad (1)$$

where C variables are counts over documents, sentiments, aspects, and/or words, $m_i^{(r)}$ is the number of sentences in review r, and $m_{iw}^{(r)}$ is the number of words in the ith sentence in review r. Γ is the digamma function.

The aspect of the ith sentence of a product description is chosen using the conditional probability

$$\Pr(y_i = k | y_{-i}, w) \propto \sum_{r=1}^{R} \sum_{s=1}^{S} \frac{C^{DRST}_{drsk} + \alpha_{r_k}}{\sum_{k'=1}^{T} C^{DRST}_{drsk'} + \alpha_{r_{k'}}}$$
$$\sum_{s=1}^{S} \left[\frac{\Gamma(\sum_{w'=1}^{W}(C^{STW}_{skw'} + \beta_{sw'}))}{\Gamma(\sum_{w'=1}^{W}(C^{STW}_{skw'} + \beta_{sw'}) + m_i^{(p)})} \prod_{w'=1}^{W} \frac{\Gamma(C^{STW}_{skw'} + \beta_{sw'} + m_{iw}^{(p)})}{\Gamma(C^{STW}_{skw'} + \beta_{sw'})} \right] \quad (2)$$

where $m_i^{(r)}$ is the number of sentences in product description p, and $m_{iw}^{(r)}$ is the number of words in the ith sentence in production description p.

The approximate probability of sentiment j in document d and review r is

$$\frac{C^{DRS}_{drj} + \gamma_j}{\sum_{j'=1}^{S}(C^{DRS}_{drj'} + \gamma_{j'})} \quad (3)$$

The approximate probability of aspect k in document d, review r, and sentiment j is

$$\frac{C_{drjk}^{DRST} + \alpha_{r_k}}{\sum_{k'=1}^{T}(C_{drj'k'}^{DRST} + \alpha_{r_{k'}})} \tag{4}$$

Finally, the approximate probability of word w with sentiment j and aspect k is

$$\frac{C_{jkw}^{STW} + \beta jw}{\sum_{w'=1}^{W}(C_{jkw'}^{STW} + \beta_{jw'})} \tag{5}$$

4.2. Product based: SA-PSM

Another way to combine both product reviews and product descriptions in a topic model can be done as follows. After the reviewer bought the product, the reviewer uses the product without explicitly looking into specific aspects of the product. After using it for quite some time, he/she might have discovered several sentiments about specific aspects of the product. The reviewer then writes a product review through the same process as described above.

We call this model Seller-aided Product-based Sentiment Model (SA-PSM). The model is shown graphically in Figure 4b. As described above, in the model, the review looks to the product description as a whole to seek help for extracting aspects. The model is expected to have a good performance on general tasks such as aspect term extraction, because of the location of the topic distribution θ being outside the review plate but inside the product plate.

Formally, the generative process of the topic model is as follows:

1. For every pair of sentiment s and aspect z, draw a word distribution $\phi \sim Dirichlet(\beta)$
2. For each product d,
    (a) For each sentiment s, draw its aspect distribution $\theta \sim Dirichlet(\alpha)$
    (b) For each review r,
        i. Draw the document's sentiment distribution $\pi \sim Dirichlet(\gamma)$
        ii. For each sentence in the review m',
            A. Choose a sentiment $s \sim Multinomial(\pi)$

B. Choose an aspect $y \sim Multinomial(\theta)$

C. Generate words $w \sim Multinomial(\phi)$

(c) For each sentence in the product description m,

i. Choose an aspect $y \sim Multinomial(\theta)$

ii. Generate words $w \sim Multinomial(\phi)$

We also use collapsed Gibbs sampling to infer the latent variables ϕ, θ, and π in SA-PSM. At each step of the Markov chain, we separately choose values for the sentiment and aspect of the review and the aspect of the product description. The sentiment and aspect of the ith sentence of a review are chosen using the conditional probability

$$\Pr(s_i = j, z_i = k | s_{-i}, z_{-i}, x) \propto \frac{C^{DRS}_{drj} + \gamma_j}{\sum_{j'=1}^{S}(C^{DRS}_{drj'} + \gamma_{j'})} \frac{C^{DRT}_{drk} + \alpha_{r_k}}{\sum_{k'=1}^{T}(C^{DRT}_{drk'} + \alpha_{r_{k'}})}$$
$$\frac{\Gamma(\sum_{w'=1}^{W}(C^{STW}_{jkw'} + \beta_{jw'}))}{\Gamma(\sum_{w'=1}^{W}(C^{STW}_{jkw'} + \beta_{jw'}) + m_i^{(r)})} \prod_{w'=1}^{W} \frac{\Gamma(C^{STW}_{jkw'} + \beta_{jw'} + m_{iw}^{(r)})}{\Gamma(C^{STW}_{jkw'} + \beta_{jw'})} \quad (6)$$

and the aspect of the ith sentence of a product description is chosen using the conditional probability

$$\Pr(y_i = k | y_{-i}, w) \propto \sum_{s=1}^{S} \frac{C^{DST}_{dsk} + \alpha_{d_k}}{\sum_{k'=1}^{T} C^{DST}_{dsk'} + \alpha_{d_{k'}}}$$
$$\sum_{s=1}^{S} \left[\frac{\Gamma(\sum_{w'=1}^{W}(C^{STW}_{skw'} + \beta_{sw'}))}{\Gamma(\sum_{w'=1}^{W}(C^{STW}_{skw'} + \beta_{sw'}) + m_i^{(p)})} \prod_{w'=1}^{W} \frac{\Gamma(C^{STW}_{skw'} + \beta_{sw'} + m_{iw}^{(p)})}{\Gamma(C^{STW}_{skw'} + \beta_{sw'})} \right] \quad (7)$$

The approximate probability of sentiment j in document d and review r, and word w with sentiment j and aspect k are shown in Equation 3 and Equation 5, respectively. However, the approximate probability of aspect k in document d with sentiment j is

$$\frac{C^{DST}_{djk} + \alpha_{d_k}}{\sum_{k'=1}^{T}(C^{DST}_{dj'k'} + \alpha_{d_{k'}})} \quad (8)$$

5. Experiments

In this section, we present several experiments done using the models proposed and compared to recent and similar topic models.

5.1. Experimental setting

5.1.1. Datasets

We use two different set of reviews gathered from Amazon. One dataset, called Laptops is a collection of reviews regarding laptops and notebooks. We use a simple query "laptop" to search for the products. Another dataset, called Cellphones is a collection of reviews regarding cellular phones. We use a simple query "cellphone" to search for the products. Each product contains one product description and multiple reviews. Each review contains the review text as well as the review score ranging from 1 (negative) to 5 (positive).

The statistics of the datasets are shown in Table 2. The Laptops dataset contains more reviews compared to the Cellphones dataset. One interesting fact to note is the average number of sentences and tokens in a product description and in a review. In the Cellphones dataset, the average number of sentences in a product description is approximately four times the average number of sentences in a review. Similarly, in both datasets, the average number of tokens in a sentence of a product description is a lot larger compared to that of a review. This indicates that the product description contains lots of terms which can denote the elaborative nature of the descriptions in product descriptions. It is also good to note that the datasets are dirty and not cleaned; there are spelling errors, fake reviews, etc. and we assume that since this is a mere minority of the dataset, the model places them with low probability.

Preprocessing is also done before doing the experiments by delimiting tokens using space and lemmatizing and separating the text into sentences using Stanford CoreNLP [18]. We also remove sentences with words over 50 since this has a greater possibility of having multiple aspects. We also use a standard stopword list to remove stopwords. Since negation is a very important aspect in sentiment analysis, we also do negation processing simply by attaching "not" to words that are negated using simple regular expression rules (e.g. "not bad" turns to "not_bad").

Sentiment seed words are also used to guide the inference of the sentiment

Table 2: Statistics of the datasets

Dataset	LAPTOPS	CELLPHONES
Total Number of products	121	164
Total Number of reviews	447,781	109,939
Average number of reviews in product	3700	670
Average number of sentences in product description	16.2	20.1
Average number of tokens in a sentence in product description	27.6	12.5
Average number of sentences in review	17.2	4.5
Average number of tokens in a sentence in review	6.8	5.5
Percentage of reviews with one star	0.11	0.11
Percentage of reviews with at most one star	0.16	0.16
Percentage of reviews with at least four stars	0.77	0.76
Percentage of reviews with five stars	0.57	0.59

labels of the words. We use PARADIGM+ as described in [14]. PARADIGM+ is an extension of the original sentiment oriental paradigm words from [32], which contains seven positive and seven negative words. PARADIGM+ is extended by carefully choosing affective words and general evaluative words. The words are carefully chosen as to not include specific evaluative and affective words because they are assumed to be unknown before training. The list of seed words are listed in Table 3.

5.1.2. Baselines

We compare our proposed models to two other topic models: the current state-of-the-art **Aspect Sentiment Unification Model (ASUM)** [14] and one of the recent topic models **Joint Aspect-based Sentiment Topic (JAST)** model [34]. ASUM is already discussed above in Section 3. JAST model is created to holistically model four types of sub-problems of the ABSA problem. One of the unique properties of this topic model is the separation of the word distribution into general opinion word distribution, aspect-specific

Table 3: Paradigm+ sentiment seed words

Positive seeds	good nice excellent positive fortunate correct superior amazing attractive awesome best comfortable enjoy fantastic favorite fun glad great happy impressive love perfect recommend satisfied thank worth
Negative seeds	bad nasty poor negative unfortunate wrong inferior annoy complain disappointed hate junk mess dislike unworthy problem regret sorry terrible trouble unacceptable upset waste worst worthless not_good not_like not_recommend not_worth

term distribution and aspect-specific opinion word distribution. Another difference is that JAST needs a large set of opinion lexicons. Therefore, we utilize the opinion lexicon [12], as utilized in the original paper. We do not use JST [17] as a baseline because it has been shown empirically that it performs worse than ASUM [14] and JAST [34].

5.1.3. Parameter settings

We set the number of iterations to 2000, including the burn-in period. The hyperparameters are set to $\alpha = 0.1$, $\beta = 0.01$, and $\gamma = 1$. The number of topics are set to 15 and the number of sentiments are set to 2 (i.e. positive and negative). The parameters are set following the empirical evaluations of JST [17] and ASUM [14].

Since topic models in general are based on the word frequencies of the corpus. That is, if a word occurs a lot in the corpus, it is with high probability that it will appear at the top of the word distribution over aspects. One major example is the term "laptop" for the Laptops dataset. This will appear almost always with high probability on all the word distribution, because it often appears in a review or in a product description. To answer this problem, we compute the term scores for all the terms for each word distribution over aspects. The term score, as used by [14], gives a lower score to the words common across various word distribution and higher score to the words that occur exceptionally often in one aspect. One can think of the term score as a weighted tf-idf score where the documents are the word distributions.

Another thing that we bring into consideration is the separation of the word distribution in the JAST model. To accommodate a fair comparison between the four models, we also divide ASUM's and our proposed models' word distribution into three using the opinion lexicon [12], which is used as the seed lists of JAST.

5.2. Sentiment classification

The most basic task of ABSA is the sentiment classification task, where one is given a text (i.e. review) and classifies it as having a positive or a negative sentiment. Sentiment topic models can easily classify the sentiment of the reviews in an unsupervised manner because they model the sentiment distribution (i.e. π in our models) of the reviews. Although our proposed models are extended from ASUM to improve aspect extraction, not sentiment classification, the accuracy should be similar to that of the ASUM.

One major problem is dataset imbalance; there are too many positive reviews compared to negative reviews. In order to solve this problem, we do multi-fold evaluations from the dataset by sampling multiple sub-datasets with equal number of positive and negative reviews. We then average the evaluation results of each sample. This provides an approximate to the real performance of the models. It is to note that the ranking between the performance of pairs of models do not change even if this remedy is not done, although the difference between two performance scores increases.

For each dataset, we consider two kinds of sub-datasets. One sub-dataset, called Large considers reviews with scores greater than 3 as positive and less than 3 as negative. Another sub-dataset, called Small considers reviews with scores greater than 4 as positive and less than 2 as negative. For each model, we calculate its accuracy and precision, recall, and f1 scores for each of the sentiments (e.g. positive and negative). We then calculate the macro f1 score based on the positive and negative precision and recall.

Results are summarized in the Table 4. Interestingly, JAST achieves a very poor performance compared to the other three models. This may be because of the fact that JAST was not created for sentiment classification, but rather for aspect precision. However, it is hard to say that the topic model is holistic if its performance of the sentiment classification is poor. Meanwhile, contrary to the expectation that the results of the proposed models do not differ with the results of ASUM, there is a slight yet clear difference between the three models. Out of the three, SA-PSM achieves the lowest performance. This is because SA-PSM is built with the θ topic distribution outside the review layer, which means that SA-PSM is built for more general problems. Between SA-ASM and ASUM, the former achieves the better performance. Although both have the same ASUM component for the reviews, the addition of the product description, and thus the presumed improvement on aspect extraction, may have helped the performance of the sentiment classification. This can be explained through an example, where the "fast" in "draining fast" and "fast performance" have different sentiments.

5.3. Aspect assignment

From this section, we verify the improvement in aspect extraction of the proposed models over ASUM and JAST. We start with the aspect assignment problem, where one is given a list of words (i.e. word distribution over aspects) and assigns it an aspect label. Because the problem is specific to a word distribution, we expect that SA-ASM performs better compared to other topic models.

Unfortunately, all the models only can produce a word distribution, not the aspect label. Thus, we can only check how easy a person can assign the word distribution an aspect label. We predefined the aspects that may be found on reviews and product description of each dataset. We also include three domain unspecific labels: *general* for list of words related to the domain but is hard to assign a label, *other* for list of words related to the domain and is specific but is not in the predefined labels, and *none* for list of words unrelated to the domain. The aspect labels we predefined are shown in Table 5. Using these aspect labels,

Table 4: Sentiment classification results. Best results have 95% statistically significant improvement over the other results based on McNemar's test. These results are bold-faced.

(a) Laptops dataset

Model	Small sub-dataset				Large sub-dataset			
	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
JAST	53.86	75.76	53.88	62.97	53.05	75.05	53.06	62.17
ASUM	83.26	83.28	83.28	83.28	79.94	80.00	79.95	79.98
SA-ASM	**84.26**	**84.30**	**84.25**	**85.28**	**81.27**	**81.30**	**81.26**	**81.28**
SA-PSM	75.26	75.60	75.25	75.42	72.79	73.25	72.78	73.01

(b) Cellphones dataset

Model	Small sub-dataset				Large sub-dataset			
	Acc	Prec	Rec	F1	Acc	Prec	Rec	F1
JAST	59.86	76.68	59.78	67.25	58.87	76.20	58.88	66.43
ASUM	84.81	84.83	84.82	84.83	81.76	81.78	81.78	81.78
SA-ASM	**85.73**	**85.85**	**85.72**	**85.78**	**82.48**	**82.62**	**82.47**	**82.54**
SA-PSM	84.43	84.55	84.43	84.49	81.41	81.55	81.40	81.47

Table 5: Predefined aspect labels for annotation

Dataset	Aspect labels
LAPTOPS	general, other, none, audio, battery, connection, design, display, external, function, hardware, installation, memory, model, processing, software, value
CELLPHONES	general, other, none, accessory, audio, battery, camera, connection, design, display, memory, processor, software, utility, value

we ask two annotators to assign aspects to the word lists. We note that the list of words used in this evaluation is the aspect-specific term distribution of each topic model.

Using the annotated word lists, we evaluate the topic models using three metrics: **diversity**, **specificity**, and **agreeability**.

5.3.1. Diversity

Diversity measures how different the annotated labels are. The measure gives a higher value if the number of distinct annotated labels is also high, considering the number of word distributions annotated as a specific label. We use the Shannon diversity index to measure the diversity. The measure is given as follows:

$$\mathbf{H} = -\sum_{a=1}^{A} \frac{C_a}{N} \log\left(\frac{C_a}{N}\right) \tag{9}$$

where A is the list of aspect labels, C_a is the number of word distributions annotated as label a, and $N = T * S * 2 = 60$ is the total number of word distributions.

5.3.2. Specificity

Specificity measures how particular the annotated labels are. The measure gives a higher value if the number of distinct annotated labels are high, not considering the number of word distributions annotated as a specific label. We use a simple metric for the specificity measure as follows:

$$S = \frac{N - N(general, other, none)}{N} \tag{10}$$

where $N(general, other, none)$ is the number of annotations using the labels "general", "other", or "none".

5.3.3. Agreeability

Agreeability measures the agreement between annotators with regards to their annotated labels. The measure gives a higher value if annotations of annotators are the same. We use Cohen's kappa coefficient for multiple categories [15], defined as follows:

$$\kappa_0 = \frac{\bar{P} - P_e}{1 - P_e} \frac{1 - \bar{P}}{N * (1 - P_e)} \tag{11}$$

where $\bar{P}$ is the relative observed agreement among annotators and P_e is the hypothetical probability of chance agreement.

Results are shown in Table 6. Overall, JAST still performs the worst in terms of diversity, specificity, and agreeability. SA-ASM performs the best in terms of specificity and agreeability. For the Laptops dataset, SA-ASM also performs the best in terms of diversity.

5.4. Aspect term prediction

Up to this point, the models are evaluated using specific problems such as sentiment classification and aspect assignment. In this section, we compare the models using a general problem: aspect term prediction. Aspect term prediction is a problem where one is given an aspect (e.g. battery) and predicts the possible aspect terms (e.g. charge, last, minutes, etc.). Because the problem looks at the dataset as a whole, we expect that SA-PSM performs better than the other topic models.

The assigned labels in the last section is used as the input labels in this section. Only those labels that are agreed by the annotators are used as labels.

Table 6: Aspect assignment results. **H** refers to the Shannon diversity index, S is the specificity measure, and κ_0 is the Cohen's kappa coefficient for multiple categories.

(a) Laptops dataset

Model	**H**	S	κ_0
JAST	0.468	0.250	0.796
ASUM	1.045	0.617	0.935
SA-ASM	**1.224**	**0.750**	**0.941**
SA-PSM	0.891	0.633	0.853

(b) Cellphones dataset

Model	**H**	S	κ_0
JAST	0.136	0.100	0.800
ASUM	**1.116**	0.600	0.923
SA-ASM	1.015	**0.633**	**0.926**
SA-PSM	0.948	0.583	0.848

Note that we also use the aspect-specific term distribution as the output list. We evaluate the models by its precision and cohesiveness.

5.4.1. Aspect term precision

505 We evaluate the precision of the output list of a topic model using the following method. First, an annotator is presented the top 10 words from the output list and decides how many words are related to the input label. Then, we calculate the precision in two ranks: precision@5, in which only the first five terms are considered and precision@10, in which all the ten terms are considered.

510 The results of the experiment are shown in Table 7. As shown in the table, SA-PSM performs the best out of the four topic models. This result holds both when only the first 5 terms are considered and when the first 10 terms are considered for evaluation.

Table 7: Aspect term precision results.

(a) Laptops dataset

Model	p@5	p@10
JAST	0.450	0.450
ASUM	0.653	0.597
SA-ASM	0.544	0.523
SA-PSM	**0.667**	**0.623**

(b) Cellphones dataset

Model	p@5	p@10
JAST	0.114	0.114
ASUM	0.723	0.708
SA-ASM	0.743	0.693
SA-PSM	**0.750**	**0.725**

5.4.2. Term intrusions

Term intrusions are also used to evaluate topic models for its semantic coherence and meaningful topic generation [26]. Term intrusions measures both the coherence of the terms in the list and the exclusiveness of the terms to a specific list. Term intrusions are used as evaluation method using the following method. First, an annotator is presented seven terms: the top five terms of the current list, an **intra-topic intrusive term**, a term also in the current list but has a very low ranking, and an **inter-topic intrusive term**, a term extracted from the top five terms of another arbitrary list. Next, the annotator selects two words that do not belong to the group. These two words are compared to the original intra-topic and inter-topic intrusive terms. We measure the recall of both intra-topic intrusive terms and inter-topic intrusive terms. Intra-topic intrusive term recall measures the coherence of the terms in the list, while inter-topic intrusive term recall measures the exclusiveness of the terms to a specific list. For a combined comparison, we calculate the average of the two recalls.

The results of the experiment are shown in Table 8. As shown in the table,

Table 8: Term intrusion results

(a) LAPTOPS dataset

Model	Intra	Inter	Average
JAST	0.35	0.30	0.33
ASUM	0.67	0.27	0.47
SA-ASM	0.63	0.26	0.45
SA-PSM	0.67	0.37	**0.52**

(b) CELLPHONES dataset

Model	Intra	Inter	Average
JAST	0.38	0.27	0.34
ASUM	0.93	0.18	0.56
SA-ASM	0.83	0.25	0.54
SA-PSM	0.83	0.33	**0.58**

530 SA-PSM performs the best out of the four topic models overall. In the LAPTOPS dataset, SA-PSM performs the best in intra-topic and inter-topic intrusive term recall. However, in the CELLPHONES dataset, ASUM achieves a higher intra-topic intrusive term recall. But this is paired up with a very low inter-topic intrusive term recall, which in turn results to a lower average intrusive term 535 recall. The combination of high precision and high intrusive term recall makes SA-PSM the better model for general ABSA problems.

5.5. Examples

Lastly, we show examples of labelled top 10 terms extracted by each model, Table 9 for JAST, Table 10 for ASUM, Table 11 for SA-ASM, and Table 12 for 540 SA-PSM. Term lists that have asterisks (*) are term lists that did not have an agreeing annotation (i.e. only one annotator labelled the list as such). Terms that are colored red and italicized are terms that are not related to the aspect label or terms that do not correspond to the correct sentiment. As shown in the tables, JAST has term distributions that are not related to the domain and

Table 9: Aspect terms extracted by JAST. (+) and (-) correspond to the sentiment (i.e. positive and negative) of the aspect. Term lists with asterisk (*) are lists that did not have an agreeing annotation.

none (+)	function (+)	*general (-)*	hardware* (-)
dosent	*hp*	*laptop*	*chromebook*
custmer	*laptop*	*price*	*use*
wasent	want	*microsoft*	flip
tryed	college	*buy*	screen
luckely	stuff	*con*	tablet
execpt	online	*color*	*google*
dirve	buy	*processor*	keyboard
900x3c	*brand*	*web*	*chrome*
dictation	function	*drive*	*os*
not_study	fullest	*cd*	battery

distributions that are not specific. The first table, for example, contains terms that do not correspond to a single aspect; most of these terms are words that are spelled incorrectly. Since the datasets contain a lot of words that are not spelled correctly and words that are not English, JAST provides term distributions exclusively for those terms. This makes most of the distributions useless and not interpretable to the users. SA-ASM has two term lists with asterisks. This means that given only the first 10 terms of the term distribution, annotating an aspect label to the distribution is difficult. As for ASUM and SA-PSM, both generated similar number of term distributions with no asterisks. However, as shown in the figure, ASUM generated more terms that are not related to the aspect labels (colored red and italicized), as annotated by annotators. In contrary, the term distributions extracted by SA-PSM have significantly lesser unrelated terms.

5.6. Discussion

Previous sections show quantitatively and empirically that our models show superior performance over previous state-of-the-art models on several tasks. In

Table 10: Aspect terms extracted by ASUM

battery (+)	function (+)	memory (-)	hardware (-)
battery	*laptop*	remove	trackpad
hour	buy	crapware	*minute*
life	recommend	ram	keyboard
5	who	*sticker*	screen
last	*hp*	windows	process
laptop	great	ssd	really
see	use	gb	*use*
update	purchase	32	*ad*
hold	*heat*	*space*	*plus*
hand	amazing	*like*	audio

Table 11: Aspect terms extracted by SA-ASM

battery (+)*	function (+)	memory (-)	hardware (-)*
battery	need	ssd	*hour*
life	web	drive	integrate
hour	use	2	*brightness*
fan	*machine*	5	battery
last	word	4	microphone
no	video	samsung	*say*
time	browsing	inch	point
use	budget	ram	audio
long	*laptop*	*ticket*	*price*
state	college	8gb	*put*

Table 12: Aspect terms extracted by SA-PSM

battery (+)	function (+)	memory (-)	hardware (-)
battery	use	ssd	*pretty*
hour	browsing	install	hp
10	college	sata	keyboard
last	photo	problem	line
5	price	samsung	responsive
life	need	drive	bad
second	web	hard	screen
log	budget	8gb	speaker
in	great	ram	*ticket*
charge	perfect	inch	touch

this section, we explain intuitively how our models achieve improvements over the base model (i.e. ASUM) and further discuss the advantages and limitations of the models.

The major advantages of SA-ASM and SA-PSM over ASUM can be explained in two-folds. First, the models leverage on the information found in the product description. Product descriptions contain more precise descriptions of the different aspects of the product [22]. The use of product descriptions contributes greatly in the improvement of the performance of the models. Second, both models effectively incorporate the texts found in the product description, thus consequently effectively supplements the inference of the aspect distribution θ. This is done by (a) generating all the words in the same sentence at the same time during the inference of the aspect distribution, which has been proven to greatly improve the aspect extraction part of the model [14], and (b) connecting this generative process to the aspect distribution θ and the word distribution ϕ of the original generative process of ASUM.

Both models have several limitations that are orthogonal to each other. As also seen empirically in the previous sections, SA-ASM is better in sentiment classification and aspect assignment subtasks, while SA-PSM is better in aspect term prediction subtask. However, SA-ASM does not perform as good as SA-PSM in predicting aspect terms, and SA-PSM also does not perform as good as SA-ASM in classifying sentiments of reviews and assigning aspect names. This can be explained through the placement of the aspect distribution θ inside the model; θ is generated specifically for each review in the SA-ASM, and it is generated generally for each product in the SA-PSM. The difference in the specificity of θ on both models makes them perform differently on different tasks. Since both sentiment classification and aspect assignment subtasks only require a specific part of the dataset, the SA-ASM performs better on these subtasks. On the other hand, since it is required to look at the full dataset to solve the aspect term extraction subtask, the SA-PSM performs better on this task.

6. Conclusion

In this paper, we proposed two topic models extended from the state-of-the-art sentiment topic model ASUM to improve the extracting of aspects by incorporating product description into the topic model. In the SA-ASM, the review looks directly to the sentences of the product description to seek help for extracting aspects. This makes the model to work well on specific problems in Aspect-based Sentiment Analysis (ABSA) such as sentiment classification and aspect assignment. In the SA-PSM, the review looks to the product description as a whole to seek help for extracting aspects. This makes the model work well on general problems in ABSA such as aspect term prediction.

We presented several experiments using two different datasets, comparing our models to ASUM and JAST, and showed that our hypotheses regarding the models hold. In the case of sentiment classification, SA-ASM performed better both when using a small and a large dataset. In the case of assigning aspect labels, SA-ASM generally performed better in terms of diversity, specificity and agreeability. Lastly, in the case of aspect term prediction, SA-PSM generally performed better in terms of precision and semantic coherence.

There are several possibilities to improve both SA-ASM and SA-PSM. In general, topic models are easily extensible using other metadata. In the case of sentiment analysis for online reviews, aside from the product description, similar and related products which are provided by e-commerce websites based on the visits of customers to other products. Categories are also good metadata information that can be inserted easily in both SA-ASM and SA-PSM. Lastly, e-commerce websites also provide question and answer (Q&A) section for customers and manufacturers to communicate informally by giving and answering questions. This is also a good way to specifically extract out aspects based on user needs. Another possible improvement is to combine both models into one. In this way, the advantages of both models in both micro-level and macro-level problems in ABSA are saved. Finally, images found in the product description and, also possibly, in the review text may also contain informative features to improve the sentiment prediction [38, 39] and, more likely, the aspect extraction [11, 29] part of the model.

For future work, we plan to use our models to other aspect-based sentiment analysis subtasks such as sentiment quantification, latent aspect rating analysis, review summarization, among others. We can also further our analysis by using different seed sets, more datasets from other domains such as forums and editorials, and emotions instead of sentiments. It is also interesting to extend the model to work on other problems such as word sense disambiguation [1] and sentiment summarization [40]. We also share our datasets and models[1] to encourage the research community for future research.

Acknowledgment

This work was supported by the Ministry of Education of the Republic of Korea and the National Research Foundation of Korea (NRF-2015S1A3A2046711).

[1] Link provided after paper acceptance.

References

References

[1] Diego R Amancio, Osvaldo N Oliveira Jr, and Luciano da F Costa. Unveiling the relationship between complex networks metrics and word senses. *EPL (Europhysics Letters)*, 98(1):18002, 2012.

[2] Ayoub Bagheri, Mohamad Saraee, and Franciska De Jong. Care more about customers: unsupervised domain-independent aspect detection for sentiment analysis of customer reviews. *Knowledge-Based Systems*, 52:201–213, 2013.

[3] Ayoub Bagheri, Mohamad Saraee, and Franciska De Jong. Adm-lda: An aspect detection model based on topic modelling using the structure of review sentences. *Journal of Information Science*, 40(5):621–636, 2014.

[4] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[5] Samuel Brody and Noemie 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] Zhiyuan Chen, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. Exploiting domain knowledge in aspect extraction. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1655–1667, 2013.

[7] Zhiyuan Chen, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. Leveraging multi-domain prior knowledge in topic models. In *Proceedings of the International Joint Conference on Artificial Intelligence*, 2013.

[8] Aitor Garcıa-Pablos, Montse Cuadros, and German Rigau. V3: Unsupervised aspect based sentiment analysis for semeval-2015 task 12. *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, page 714, 2015.

[9] Xavier Glorot, Antoine Bordes, and Yoshua Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 513–520, 2011.

[10] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences*, 101(suppl 1):5228–5235, 2004.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] Minqing Hu and Bing 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.

[13] Salud M Jiménez-Zafra, M Teresa Martín-Valdivia, Eugenio Martínez-Cámara, and L Alfonso Ureña-López. Combining resources to improve unsupervised sentiment analysis at aspect-level. *Journal of Information Science*, page 0165551515593686, 2015.

[14] Yohan Jo and Alice 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.

[15] Helena Chmura Kraemer. Extension of the kappa coefficient. *Biometrics*, pages 207–216, 1980.

[16] Yan Li, Zhen Qin, Weiran Xu, and Jun Guo. A holistic model of mining product aspects and associated sentiments from online reviews. *Multimedia Tools and Applications*, 74(23):10177–10194, 2015.

[17] Chenghua Lin and Yulan 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.

[18] Christopher D Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. The stanford corenlp natural language processing toolkit. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (System Demonstrations)*, pages 55–60, 2014.

[19] Qiaozhu Mei, Xu Ling, Matthew Wondra, Hang Su, and ChengXiang 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.

[20] Tony Mullen and Nigel Collier. Incorporating topic information into sentiment analysis models. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 25. Association for Computational Linguistics, 2004.

[21] Thien Hai Nguyen and Kiyoaki Shirai. Phrasernn: Phrase recursive neural network for aspect-based sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 2509–2514, 2015.

[22] Dae Hoon Park, C Zhai, and Lifan Guo. Speclda: Modeling product reviews and specifications to generate augmented specifications. In *Proceedings of the 2015 SIAM International Conference on Data Mining. SIAM*. SIAM, 2015.

[23] Ioannis Pavlopoulos and Ιωάννης Παυλόπουλος. Aspect based sentiment analysis. *Athens University of Economics and Business*, 2014.

[24] Maria Pontiki, Dimitrios Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. Semeval-2015 task 12: Aspect based sentiment

analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, 2015.

[25] Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014)*, pages 27–35. Citeseer, 2014.

[26] Md Mustafizur Rahman and Hongning Wang. Hidden topic sentiment model. In *Proceedings of the 25th International Conference on World Wide Web*, pages 155–165. International World Wide Web Conferences Steering Committee, 2016.

[27] Sebastian Ruder, Parsa Ghaffari, and John G Breslin. A hierarchical model of reviews for aspect-based sentiment analysis. *arXiv preprint arXiv:1609.02745*, 2016.

[28] Nuttapong Sanglerdsinlapachai, Anon Plangprasopchok, and Ekawit Nantajeewarawat. Exploring linguistic structure for aspect-based sentiment analysis. *Maejo International Journal of Science and Technology*, 10(2):142, 2016.

[29] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[30] Vivek Kumar Singh, Rajesh Piryani, A Uddin, and P Waila. Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification. In *Automation, Computing, Communication, Control and Compressed Sensing (iMac4s), 2013 International Multi-Conference on*, pages 712–717. IEEE, 2013.

[31] Ivan Titov and Ryan T McDonald. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, volume 8, pages 308–316. Citeseer, 2008.

[32] Peter D Turney and Michael L Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346, 2003.

[33] Linlin Wang, Kang Liu, Zhu Cao, Jun Zhao, and Gerard de Melo. Sentiment-aspect extraction based on restricted boltzmann machines. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 616–625, 2015.

[34] Shuai Wang, Zhiyuan Chen, and Bing Liu. Mining aspect-specific opinion using a holistic lifelong topic model. In *Proceedings of the 25th International Conference on World Wide Web*, pages 167–176. International World Wide Web Conferences Steering Committee, 2016.

[35] Wei Wang. Sentiment analysis of online product reviews with semi-supervised topic sentiment mixture model. In *Fuzzy Systems and Knowledge Discovery (FSKD), 2010 Seventh International Conference on*, volume 5, pages 2385–2389. IEEE, 2010.

[36] Wenya Wang, Sinno Jialin Pan, Daniel Dahlmeier, and Xiaokui Xiao. Recursive neural conditional random fields for aspect-based sentiment analysis. *arXiv preprint arXiv:1603.06679*, 2016.

[37] Yequan Wang, Minlie Huang, Li Zhao, and Xiaoyan Zhu. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2016.

[38] Sicheng Zhao, Yue Gao, Xiaolei Jiang, Hongxun Yao, Tat-Seng Chua, and Xiaoshuai Sun. Exploring principles-of-art features for image emotion recognition. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 47–56. ACM, 2014.

[39] Sicheng Zhao, Hongxun Yao, Yue Gao, Rongrong Ji, and Guiguang Ding. Continuous probability distribution prediction of image emotions

via multitask shared sparse regression. *IEEE Transactions on Multimedia*, 19(3):632–645, 2017.

[40] Jingbo Zhu, Muhua Zhu, Huizhen Wang, and Benjamin K Tsou. Aspect-based sentence segmentation for sentiment summarization. In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 65–72. ACM, 2009.