SentiLDA — An Effective and Scalable Approach to Mine Opinions of Consumer Reviews by Utilizing Both Structured and Unstructured Data

Fan Liu(✉) and Ningning Wu

Information Science, University of Arkansas at Little Rock,
2801 S. University Ave., Little Rock, AR 72204, USA
{fxliu,nxwu}@ualr.edu

Abstract. With the help of Internet and Web technologies, more and more consumers tend to seek opinions online before making purchase decisions. However, with the ever-increasing volume of user generated reviews, people are overwhelmed with the amount of data they have. Thus there is a great need for a system that can summarize the reviews and produce a set of aspects being mentioned in the reviews together with the pros/cons being expressed to them. To address the need, this paper proposes a new probabilistic topic model, SentiLDA, for mining reviews (unstructured data) and their ratings (structured data) jointly to detect the product/service aspects and their corresponding positive and negative opinions simultaneously. A key feature of SentiLDA is that it is capable of mining positive and negative sub-topics under the same aspect without the need of sentiment seed words. Experiment results show that the performance of SentiLDA outperforms the other related state-of-the-art models in detecting product/service aspects and their corresponding sentiments in reviews.

Keywords: Opinion mining · Topic modeling · Big data

1 Introduction

The Internet has greatly changed our shopping experiences over the last decade. Coming to the big data era, user-generated content (UGC) becomes a rich information source on the Internet. Customer reviews can be found from all kinds of social media, from internet tycoons like amazon.com to personal blogs. Many online review platforms allow users to submit reviews in free text format to comment on pros and cons of a product or service, and to give numerical ratings on the overall satisfaction level. These reviews can help people seek information before making shopping decisions but they also bring problems to consumers. A survey shows 32 % of internet users have been confused by information they have found online during their shopping; 30 % have felt overwhelmed by the amount of information they found online [6]. It is impossible for a shopper to digest the huge volume of reviews available online without any post-processing of the data. Thus, some people just rely on the ratings, leave the

S. Madria and T. Hara (Eds.): DaWaK 2016, LNCS 9829, pp. 82–96, 2016.
DOI: 10.1007/978-3-319-43946-4_6

actual reviews aside. But ratings cannot tell it all. Besides, there is not a golden standard to guide how people give ratings. For example, someone may give a 5-star with minor defects, but others give a 5 only when they are 100 % satisfied. Thus, there are studies trying to provide an adjusted rating for the reviews [17]. However, most consumers need more than just a score. They actually need a system, which can digest big volume of reviews and produce a set of aspects being mentioned in the reviews together with the pros/cons being expressed to the aspects. Therefore, techniques for review summarization and integration are in great demand for the improvement of online shopping experience.

Many studies have been carried out to address the problem of review summarization. Some solely detect the aspects in reviews [7, 13], while others separate the opinions from the facts [20]. But they are still far from obtaining the opinion orientations associated with the aspects. Predicting the sentiments of the words requires extra knowledge about the words, and sometimes even the knowledge about the aspects being discussed. To conquer this difficulty, researchers developed different topic models incorporating prior information from a set of seed words with general deterministic sentiment orientations [8, 9, 18].

Only one study [2] tried to jointly model ratings and reviews of the movies. It applied an approach based on collaborative filtering and topic modeling, which had some limitations on the dataset due to the nature of collaborative filtering. To the best of our knowledge, no previous research has developed an approach to predict the sentiment orientations of the words in reviews by incorporating ratings information solely based on topic modeling.

The major contribution of this research is to propose a new probabilistic topic model, SentiLDA, for mining reviews (unstructured data) and their ratings (structured data) jointly to detect the product/service aspects and their corresponding positive and negative opinions simultaneously without using seed words. A key feature of SentiLDA is that it is capable of mining positive sub-topics and negative sub-topics under the same aspect topic without prior information of the sentiment seed words. Since it does not rely on domain knowledge, SentiLDA is general and can be applied to similar problems with unstructured text data and structured numerical data.

The second contribution is we implement SentiLDA in Spark [12] following the MapReduce paradigm by using variational inference. It takes the advantage of the parallel distributed in-memory computing environment to scale up and speed up the model inference. Experiment results show SentiLDA outperforms the other related state-of-the-art models in detecting product/service aspects and their corresponding sentiments in reviews.

In addition, the implementation avoids the scalability issue of the traditional Gibbs sampling technique, and thus makes it very suitable for big data analysis in distributed environment. SentiLDA could be used to support other research based on domain specific sentiment words, such as review rating prediction, opinions summarization and Integration.

The remaining of the paper is organized as follows. Section 2 introduces the related works. The proposed model is described in Sect. 3, followed by a brief explanation of the implementation in Sect. 4. Experiment setup, results and analysis are presented in Sect. 5. Finally, Sect. 6 concludes the paper and suggests the future work.

2 Related Work

Early studies in sentiment prediction mainly depended on using WordNet [3] or pointwise mutual information (PMI) [14] to determine the sentiment of a word. However, this approach has difficulties in predicting domain-specific sentiment words. Certain words may be positive in a domain, but negative in another domain, e.g. "big" is good for cell phone screen size, but bad for battery size of the cell phone.

Opinions are always expressed to objects. In order to perform review summarization and integration, it is desirable to know both the sentiment orientation of a word and what aspect it is talking about. Many researchers have been trying to solve this ultimate problem by using topic models. Tying-JST [9], TSM [11], ASUM [8], and JAS [18], are popular models proposed for this objective. Tying-JST modifies LDA by adding one variable to control the sentiment orientations of the words in the reviews. The sentiment variable is drawn from a document level sentiment distribution determined by a Dirichlet distribution. The approach of TSM is similar as Tying-JST. In additional, it introduces a background words variable in the model. ASUM adds some constraints on the basis of Tying-JST. It assumes the words from the same sentence are of the same topic and sentiment. JAS brings in more variables to control the subjectivities of the words and the sentiments of the subjective words. It also assumes each sentence in the review has two sentence-level sentiment distributions for opinion and fact respectively. The models mentioned above all have their drawbacks. Tying-JST and TSM extract topics-sentiments solely based on words co-occurrences, which loses the locality information of the words in the reviews. ASUM restricts the sentiments of the words in the same sentence to be the same, which is not held in many reviews. Rather than discovering T topics with positive and negative words separated, ASUM discovers T positive topics and T negative topics, which requires further post work to perform review summarization and integration. JAS introduces many latent variables, which increase the computational complexity of the model inference. Moreover, the generative process described by JAS is not intuitive as it assumes two sentiment distributions for each sentence in a review. Last but not least, in order to distinguish facts from opinions, and positive sentiment from negative sentiment, all these models heavily rely on a good set of sentiment seed words which is not always easy to obtain.

Review rating has been studied in some research recently. But most of them are trying to do rating prediction or justification based on review context [4, 16]. JMARS [2] is the closest one to our study. It also models aspects, ratings and sentiments jointly on movie reviews. The approach is based on collaborative filtering and topic modeling. In order to perform collaborative filtering, it requires each user writes more than one reviews on different movies. Then the model is able to construct user's expectations and movie's properties. For the reason of privacy protection and data accessibility, it is not easy to obtain multiple product reviews from the same user. Therefore, it prevents JMARS from being applied to the problem we are trying to solve in this research. Furthermore, product reviews are often shorter than movie reviews, the locality information of words are of great importance, but JAMRS fails to model it.

To the best of our knowledge, this is the first study to exploit review rating to predict the aspect-specific sentiment orientations of the words in reviews by solely using topic modeling and without using sentiment seed words.

3 Method

3.1 SentiLDA

In order to extract topics/aspects and their associated sentiment opinion bearing words in the reviews respectively, we proposed a probabilistic graphical model, which follows a hierarchical topic-vocabulary structure shown in Fig. 1. It is a two-level vocabulary hierarchy. The root node V stands for the whole vocabulary of the corpus. The vocabulary is virtually split into K child nodes in the first level, where each node contains all words for one topic/aspect of the product. In the second level each topic/aspect node is divided into three leaf nodes, which are the neutral, positive and negative words of the corresponding topic/aspect respectively. By definition, the neutral words mean purely descriptive ones that do not express any opinion, such as "hotel", "room", and "restaurant" etc. The positive and negative words stand for the ones that convey sentimental opinions, such as "excellent", "terrible", and so on. Negations will be detected and handled appropriately by using Stanford CoreNLP NLPT [10]. These words are not necessarily constrained to be adjectives and adverbs. Nouns and verbs can also bear sentiments, e.g., "noise" is negative and "recommend" express a positive opinion.

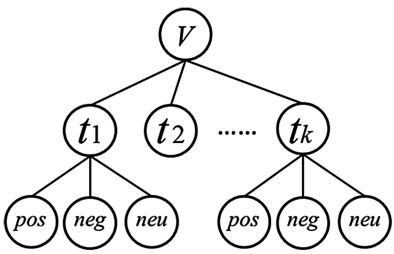

Fig. 1. Hierarchy of the topic-vocabulary structure

In LDA [1], a bag-of-words assumption is proposed, where the relative position of each individual word is neglected. Words located close to each other in the documents can be assigned to totally unrelated topics, which are inappropriate in many scenarios, especially when the model is applied to short documents like reviews. Thus, we made a stricter assumption that words co-occurring in the same sentence must be of the same topic. This assumption is similar to ASUM, but it has two main differences: (1) there is only one topic distribution for each review, compared to a positive and a negative distribution respectively in ASUM; (2) unlike ASUM, the sentiment orientations of the words in the same sentence are not constrained to be the same. Actually it is quite

common that two sides of a coin are discussed in the same sentence of a review, e.g., "The restaurant in the hotel was great but fairly expensive." "Great" and "expensive" are two opposite sentiments of the topic restaurant in hotel reviews. The assumption in SentiLDA is more intuitive and logical, because people tend to discuss issues in a review topic by topic. For each topic there would be opinions of both positive and negative side, rather than setting a sentimental orientation first and then choose a topic to write. Moreover, this assumption also reduces the complexity of the model by introducing only one topic distribution instead of two. (3) We observe many narrative sentences in reviews, e.g. "I spend the Xmas with my family at hotel ABC this year", which express no sentiment, but just a fact. In ASUM, it has to be either positive or negative. But in SentiLDA, it could be neutral. Based on our observation and experiment results the proposed model outperforms ASUM. The graphical representation of SentiLDA is shown in Fig. 2. There are D reviews in the whole corpus, where each review consists of $\boldsymbol{M_d}$ sentences, and there are $\boldsymbol{N_{d,m}}$ words in each sentence. The details of the model are described below.

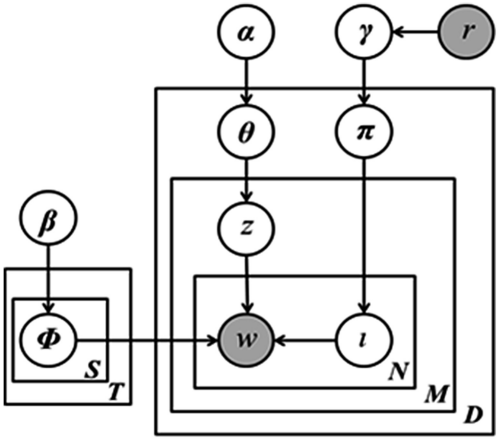

Fig. 2. Graph model representation of SentiLDA

Before actually writing any reviews, first draw three word distributions $\Phi_{t,s} \sim$ Dirichlet (β) for each topic t, in which s corresponds to neutral (facts-topic), positive, and negative sentiment topic respectively. When a reviewer writes a review d, the generative process for each word in a review is as following.

1. Draw a topic distribution $\boldsymbol{\theta_d} \sim$ Dirichlet $(\boldsymbol{\alpha})$ for the review
2. Draw a sentiment distribution $\boldsymbol{\pi_d} \sim$ Dirichlet (γ) for the review
3. For each sentence m in the review d,
   (a) Choose a topic $t \sim$ Multinomial $(\boldsymbol{\theta_d})$
   (b) For each token n in the sentence m,
      (i) Choose a sentiment label $\iota_n \sim$ Multinomial $(\boldsymbol{\pi_d})$
      (ii) Choose a word $w \sim$ Multinomial (Φ_{t,ι_n}).

In the SentiLDA, $\boldsymbol{\pi}$ plays an important role in assigning a sentiment label to each word in the document, and it is generated from a Dirichlet distribution with a hyper parameter $\boldsymbol{\gamma}$. For ease of use, it is suggested empirically to use a symmetric hyper parameter for a Dirichlet distribution; however, a symmetric $\boldsymbol{\gamma}$ means a random sentiment distribution in the proposed model [15]. Without a guidance of the overall sentiment distribution of the review, it is impossible to effectively separate the words into different sentimental orientations, because all the words are clustered solely based on co-occurrences. Fortunately, besides the unstructured text in the reviews, there is also a numerical overall rating being accompanied with the reviews in most of the online review platforms. It can be exploited to provide a clue of the sentiment distribution of the review. However, shown by previous study [17], the review ratings are inconsistent among different users, different review platforms. Therefore, generating a sentiment distribution $\boldsymbol{\pi}$ just based on the absolute value of the rating is not appropriate. But the ratings could be a very good clue for setting a prior $\boldsymbol{\gamma}$ for a Dirichlet distribution, which generates a sentiment distribution $\boldsymbol{\pi}$. Then, the value of $\boldsymbol{\pi}$ could be further optimized in the parameters inference. Based on this assumption, SentiLDA is proposed with a variable r. Note that r is in a shadowed node, which means it is an observed value. In the case of modeling reviews, it is the review overall rating provided by the review writer. It determines the value of the prior $\boldsymbol{\gamma}$, where $\gamma \in \Gamma$, and Γ is a set of possible priors corresponding to different ratings.

3.2 Model Inference

The key to solving the problem is to infer the latent variables in the proposed SentiLDA model. In practice, the latent variables are derived by maximizing the log-likelihood of the observed data. Given the hyper parameters $\boldsymbol{\alpha}, \boldsymbol{\gamma}$, and word distributions over topics $\boldsymbol{\Phi}$, the joint distribution of the latent topic distribution $\boldsymbol{\theta}$, sentiment distribution $\boldsymbol{\pi}$, topic assignments $\boldsymbol{z}$, sentiment assignments $\boldsymbol{\iota}$, and observed words w is given by,

$$p(\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{z}, \boldsymbol{\iota}, \boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\Phi}) = p(\boldsymbol{\theta}|\boldsymbol{\alpha})p(\boldsymbol{\pi}|\boldsymbol{\gamma}) \prod_{m=1}^{M} \left\{ p(z_m|\boldsymbol{\theta}) \prod_{n=1}^{N_m} [p(\iota_n|\boldsymbol{\pi_m})p(w_n|z_m, \iota_n, \boldsymbol{\Phi})] \right\} \tag{4.1}$$

If we integrate and sum over all the latent variables, then the marginal distribution of a review is obtained. After taking product of the marginal probability of every single review, we can obtain the likelihood of the whole set of reviews,

$$\begin{aligned} p(\boldsymbol{D}|\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\Phi}) = \prod_{d=1}^{D} \iint p(\boldsymbol{\theta_d}|\boldsymbol{\alpha})p(\boldsymbol{\pi_d}|\boldsymbol{\gamma}) \Big(\prod_{m=1}^{M} \sum_{z_m}^{T} p(z_m|\boldsymbol{\theta_d}) \\ \times \prod_{n=1}^{N_m} \sum_{\iota_n}^{S} p(\iota_n|\boldsymbol{\pi_d})p(w_n|z_m, \iota_n, \boldsymbol{\Phi}) \Big) d\boldsymbol{\pi_d} d\boldsymbol{\theta_d} \end{aligned} \tag{4.2}$$

There are two ways to find the values of the latent variables to maximize the probability of generating such a corpus, a collapsed Gibbs sampler based on Markov chain Monte Carlo (MCMC) and an inference technique based on variational methods. Due to its simplicity to be understood and implemented, the collapsed Gibbs sampler [5]

dominates the research community in solving latent variables inference problem. However, it has several limitations that prevent it from being applied to big data scenario [19]. Therefore, we use variational method as an alternative technique to solve the variable inference problem in our proposed model. Compared to Gibbs sampling, variational inference has the following advantages: (1) there is clear convergence criterion for variational inference; (2) it does not require a shared state during each iteration; (3) it takes less number of iterations, usually 20 to 40, to converge, and thus reduces the communication overhead; and (4) it is able to optimize the hyper parameters due to its statistical nature.

Variational Inference. By introducing variational parameters $\boldsymbol{\delta}, \boldsymbol{\lambda}, \boldsymbol{\varepsilon}$, and $\boldsymbol{\eta}$, shown in Fig. 3, the dependencies between $\boldsymbol{\theta}$ and z, $\boldsymbol{\pi}$ and l are dropped. A family of distribution $q(\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{z}, \boldsymbol{\iota})$ on the latent variables is obtained. It is used to approximate the true posterior distribution of the latent variables in the proposed model. By minimizing the difference, Kullback-Leibler (KL) divergence, between these two distributions, the optimal values of variational parameters $\boldsymbol{\delta}, \boldsymbol{\lambda}, \boldsymbol{\varepsilon}$, and $\boldsymbol{\eta}$ can be derived.

$$q(\boldsymbol{\theta}, \boldsymbol{\pi}, \boldsymbol{z}, \boldsymbol{\iota}) = q(\boldsymbol{\theta}|\boldsymbol{\delta}) q(\boldsymbol{\pi}|\boldsymbol{\lambda}) \prod_{m=1}^{M} \left[q(z_m|\boldsymbol{\varepsilon_m}) \prod_{n=1}^{N_m} q(l_n|\boldsymbol{\eta_n}) \right] \tag{4.3}$$

Minimizing the KL divergence between the variational distribution and true posterior distribution of the latent variables is equivalent to maximizing the evidence lower bound (ELBO) of the corpus.

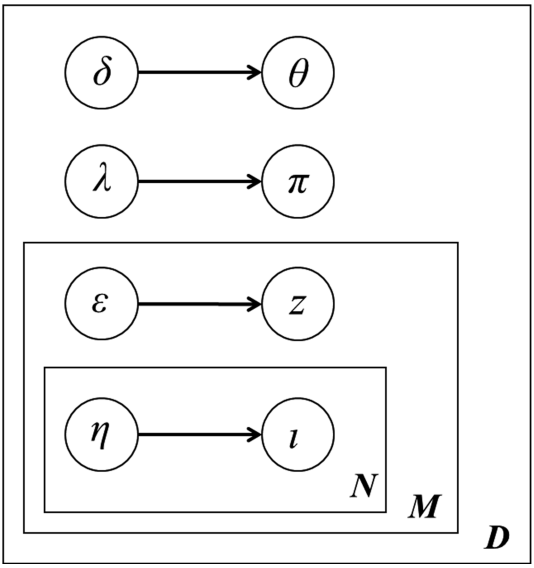

Fig. 3. Graphical Model Representation of the variational approximation of the posterior in SentiLDA

$$\mathcal{L} = \mathrm{E}_q[\log p(\boldsymbol{\theta}, \boldsymbol{\pi}, z, \boldsymbol{\imath}, \boldsymbol{w}|\boldsymbol{\alpha}, \boldsymbol{\gamma}, \boldsymbol{\Phi})] - \mathrm{E}_q[q(\boldsymbol{\theta}, \boldsymbol{\pi}, z, \boldsymbol{\imath})] \tag{4.4}$$

Therefore, the updating equations for the variational parameters are obtained,

$$\delta_i = \alpha_i + \sum\nolimits_{m=1}^{M} \varepsilon_{m,i} \tag{4.5}$$

$$\lambda_s = r_s + \sum\nolimits_{m=1}^{M} \sum\nolimits_{n=1}^{N_m} \eta_{m,n,s} \tag{4.6}$$

$$\varepsilon_{m,i} \propto \prod\nolimits_{n=1}^{N_m} \prod\nolimits_{s}^{S} \Phi_{i,s,w_n}^{\eta_{m,n,s}} \exp\left[\Psi(\delta_i) - \Psi\left(\sum\nolimits_{i=1}^{T} \delta_i\right)\right] \tag{4.7}$$

$$\eta_{m,n,s} \propto \prod\nolimits_{i=1}^{T} \Phi_{i,s,w_n}^{\varepsilon_{m,i}} \exp\left[\Psi(\lambda_s) - \Psi\left(\sum\nolimits_{s}^{S} \lambda_s\right)\right] \tag{4.8}$$

The word distribution over topics is derived as follow,

$$\Phi_{i,s,j} \propto \sum\nolimits_{d=1}^{D} \sum\nolimits_{m=1}^{M} \sum\nolimits_{n=1}^{N_m} \varepsilon_{d,m,i} * \eta_{d,m,n,s} * w_{d,m,n}^{j} \tag{4.9}$$

Incorporating POS Information. In the decoupled variational model $\boldsymbol{\eta}$ are the multinomial distributions for the sentiment labels of each word respectively. They will start from some random initial values, and to be updated iteratively during the inference process. However, a fully random initialization of these values may not be sufficient to separate the words of different sentiments effectively. Thus, we exploit the part of speech (POS) information from the reviews to initialize $\boldsymbol{\eta}$, which has never been applied by previous researches. We use Stanford CoreNLP NLPT [10] to tag the reviews for the POS information. If the POS tag of a word is a noun or verb, $\boldsymbol{\eta}$ is initialized to a higher value to neutral, 0.6 in our experiment, and equal values to positive and negative, 0.2 in our experiment. Otherwise, it is initialized to a lower value to neutral (0.2), and 0.4 for both positive and negative.

4 Implementation

Spark is a popular big data processing engine that supports parallel distributed in-memory computing. Every document is independent to each other in the inference procedure, thus it adapts to the paradigm of MapReduce in Spark seamlessly. And the iterative model inference procedure requires the same set of data being processed many times. Spark's ability of doing in-memory computing could reduce the cost of I/O traffic of reading in the data significantly. Instead of reading in the data multiple times in Hadoop MapReduce, it only needs to read in the data once in Spark. Therefore, we implement the proposed model in Spark. The inference procedure can be implemented in two stages, map stage and reduce stage.

4.1 Map Stage: Document Level

There is a set of variational parameters $\boldsymbol{\delta}$, $\boldsymbol{\lambda}$, $\boldsymbol{\varepsilon}$, and $\boldsymbol{\eta}$ for each document. There is no dependency between sets of variational parameters of different documents. Thus, all the documents can be processed parallelly. Equations (4.5)–(4.8), and document level (4.9) are implemented in the map stage. The map stage emits document level word distribution for topics $\boldsymbol{\Phi}'$ after processing each document.

4.2 Reduce Stage: Corpus Level Aggregation

Word distribution for topics $\boldsymbol{\Phi}$ is a global variable. In order to update it, document level word distribution for topics $\boldsymbol{\Phi}'$ has to be aggregated at the corpus level according to different key values, which consists of a topic index, a sentiment index and a word index. Corpus level Eq. (4.9) is implemented in the reduce stage.

5 Experiments

5.1 Data Set

We crawled a set of reviews covering 36 major hotels on the Strip in Las Vegas, USA from 4 websites: expedia.com, hotels.com, orbitz.com, and tripadvisor.com. It contains all the reviews in English and their numerical ratings from each source. We choose hotel reviews because it contains many different aspects, each has a lot domain specific sentiment words. SentiLDA is proposed to solve problems of such characteristics, but not limited to hotel reviews. All reviews were preprocessed through a pipeline consists of tokenization, sentences splitting, lemmatization, and POS tagging by using Stanford CoreNLP Toolkit [10]. Negations were also detected by Stanford NLP Toolkit. Words modified by negations were added with a prefix of "not_". Punctuations and stop-words were removed. Only nouns, verbs, adjectives and adverbs that carry actual meanings were kept. In order to reduce the sparsity of the vocabulary, we further removed words that appear less than 10 times in the corpus since they barely convey meaningful information. All the ratings are in the scale of 1 to 5, and are integers only. Table 1 shows the statistics of the resulting dataset in the experiment.

Table 1. Statistics of the corpus

Rating	# of reviews	# of sentences	# of words
1	25,490	210,719	1,429,640
2	40,000	311,828	2,093,742
3	80,213	575,936	3,814,216
4	157,389	1,026,331	6,679,301
5	185,268	1,097,737	6,959,797
Total	488,360	3,222,551	20,976,696

5.2 Experiment Setting

In this experiment, SentiLDA is compared with ASUM [8], Tying-JST [9] and JAS [18]. All of them are popular models in review aspect/sentiment discovery. However, in the original paper of the above models, they were all inferred by Gibbs sampling. In order make them run faster, we migrate them to Spark by using variational inference as well. We set the number of topics $\boldsymbol{T}$ for all the models to 35, since it could discover all the major features and has the least number of uninterpretable features. Hyper parameter α_i is set to 2 for each aspect in the Dirichlet distribution. $\Phi_{i,s,j}$ is randomly initialized and normalized for all words in an aspect-sentiment distribution. Since the proposed model exploits the review ratings to indicate the prior of sentiment distribution in a review, there is a set of $\boldsymbol{\gamma}$ corresponding to different ratings. Table 2 shows the different configurations of $\boldsymbol{\gamma}$. All the other models do not exploit review rating information and use the default symmetric setting of $\boldsymbol{\gamma}$ instead. The sum of the elements in all configurations of $\boldsymbol{\gamma}$ is kept to be 1.

Table 2. Different γ settings by different ratings

Rating	Neutral	Positive	Negative
1	0.75	0.02	0.23
2	0.75	0.07	0.18
3	0.75	0.12	0.13
4	0.75	0.18	0.07
5	0.75	0.23	0.02
Symmetric	0.33	0.33	0.33

5.3 Qualitative Analysis

Tables 3, 4 and 5 show the top 20 words of each sentiment for the customer service aspect obtained from the proposed SentiLDA model. For comparison purpose, the top words of the sentiments for the same aspects derived from JAS and ASUM are also shown in the tables. We have also compared with Tying-JST. Since its result is similar to JAS, we do not include it here due the length limit of this paper. ASUM doesn't extract neutral sentiment words directly, but we can extract them by looking for the common words in positive and negative sentiment.

Table 3. Top 20 words of customer service obtained by SentiLDA for each sentiment

neutral	staff, friendly, helpful, hotel, service, room, great, desk, nice, clean, front, check, good, always, courteous, excellent, pleasant, housekeeping, stay, extremely
positive	concierge, professional, attentive, greet, welcome, make, name, spa, smile, warm, special, reception, level, feel, doorman, gracious, hotel_3, outstanding, impeccable, efficient
negative	rude, customer, unfriendly, unhelpful, attitude, not_helpful, poor, manager, horrible, terrible, bad, management, lack, not_friendly, less, not_care, unprofessional, dirty, worst, employee

Table 4. Top 20 words of customer service obtained by JAS for each sentiment

neutral	make, hotel, stay, help, feel, go, staff, check, need, time, ask, more, get, room, guest, take, way, question, treat, say
positive	staff, service, friendly, helpful, great, hotel, room, nice, excellent, good, customer, concierge, housekeeping, always, professional, pleasant, courteous, polite, best, wonderful
negative	desk, staff, front, rude, customer, people, hotel, employee, manager, work, check, attitude, management, person, kind, guest, poor, speak, extremely, member

Table 5. Top 20 words of customer service obtained by ASUM for each sentiment

positive	staff, friendly, helpful, hotel, clean, nice, room, great, courteous, service, pleasant, extremely, polite, professional, always, casino, check, stay, accommodate, well
negative	Service, hotel, staff, room, customer, poor, top, rude, notch, bad, experience, food, horrible, restaurant, terrible, lack, overall, housekeeping, cleanliness, casino

The comparison shows SentiLDA model captures most of the neutral words that are discovered by JAS and ASUM, such as "staff", "front", "desk", "service", "room", "check", "housekeeping", "stay" in customer service aspect. SentiLDA discovers more aspect-specific sentiment words than both JAS and ASUM, such as "attentive", "greet", "welcome", "smile", "warm", "gracious", "outstanding", "impeccable", and "efficient" in the positive side, "unfriendly", "unhelpful", "not_friendly", "not_care", "not_helpful", and "unprofessional" in the negative side.

Most of the sentiment words discovered by JAS and ASUM are from sentiment seed words set, or closely related to them. Because sentiment seed words are also the words used frequently by people, such as "great", "excellent", "rude", "horrible", they tend to dominate the high possibility words in an aspect sentiment. However, instead of relying on sentiment seed words, SentiLDA exploits review ratings as sentiment distribution prior. It evens out the frequent words to all sentiments according to the sentiment prior. Thus, aspect-specific sentiment words have higher probabilities in the correct sentiments.

Furthermore, SentiLDA is capable of detecting non-adjective sentiment bearing words. Table 6 shows the top 20 words of each sentiment for the aspect of bathroom. Words like "slipper", "robe", "toiletries" are all neutral if mentioned not in the bathroom. But when people talk about amenity in hotel bathrooms, they are definitely good to have. Thus they convey positive sentiment is this situation. On the contrast, the

Table 6. Top 20 words of bathroom aspect obtained by SentiLDA for each sentiment

neutral	room, shower, bathroom, water, bed, floor, towel, clean, day, get, tub, dirty, toilet, hair, carpet, sink, sheet, leave, stain, take
positive	slipper, robe, toiletries, chocolate, lotion, kit, amenity, gel, provide, product, bath, body, shave, cotton, toothbrush, razor, spa, cream, polish, steam
negative	stain, dirty, filthy, carpet, blood, sheet, bug, mold, black, cover, look, disgusting, dirt, wall, gross, break, notice, nasty, foot, find

existence of “stain”, “mold”, “bug”, and even “blood” in the bathroom is obviously a negative sign. Models solely rely on sentiment seed words, such as ASUM, JAS and JST are not able to discover this kind of sentiment aspects.

5.4 Quantitative Analysis

Convergence Test. We study the convergence speed of different models on the training data set. The iterative updates process stops when the improvement of log likelihood is less than 0.01 %. We test on the proposed SentiLDA, ASUM, Tying-JST, and JAS. Figure 4 shows the result. From the plot, we can observe that SentiLDA achieve slightly better log likelihood than ASUM. All of them are much better than the other two. Regarding to the iterations take to convergence, SentiLDA and ASUM are quite similar with 33, 37 respectively. JAS only takes 18 iterations to converge, but with a much worse log likelihood. It may be caused by being trapped in a local optimal. Tying-JST takes the most iteration (78) to converge to the worst result of them. One of the possible reasons would due to the number of variables of Tying-JST. It has a topic variable and sentiment variable for each word in the review. Thus, it needs more iterations to update them all to a stationary state. Another possible reason for the low log likelihood may due to the lack of constraints of Tying-JST. The words in a review can be of any topics and sentiments, the combinations of the values of the variables are much larger than the other models, and then results in low log likelihood.

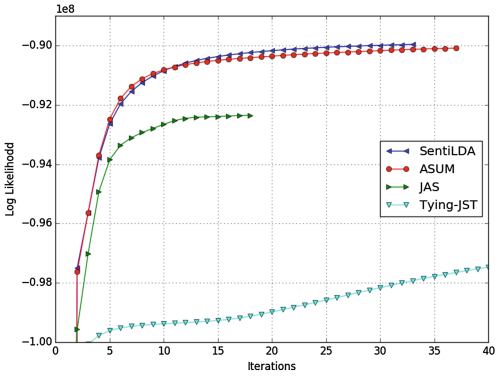

Fig. 4. Comparison of Log Likelihood convergence

Perplexity Test. The ability to predict unseen data is another important metric to evaluate the fitness of a topic model. We divided the data set in this experiment into two parts. We use two thirds of the data as a training set to train a model, and then use remaining one thirds of data as a held-out test set to evaluate the training models. The perplexity of the held-out test set is computed for comparison. The perplexity is calculated by take the inverse of the geometric mean per-word likelihood. A lower

perplexity indicates better generalization ability of a model. Table 7 shows the results of the perplexity comparison. SentiLDA obtained the best predictive perplexity on held-out data set. It indicates that the proposed SentiLDA is not only good in extracting aspects-specific sentiment topics from seen data, but also performs well on unseen data. More complex models, such as JAS, and Tying-JST, tend to suffer from over-fitting problem.

Table 7. Perplexities of the held-out test set by applying different models

Model	SentiLDA	ASUM	JAS	JST
Perplexity	646.99	664.40	760.94	991.83

5.5 Result Analysis

From the qualitative analysis and quantitative analysis, it shows the proposed SentiLDA outperforms the other state-of-art opinion mining topic models. It takes less iterations to converge to a higher log likelihood on the training data, and performs better generalization ability in unseen testing data. SentiLDA is capable of discovering aspect specific sentiment words without using sentiment seed words. However, some general sentiment words, such as “nice”, “excellent”, “great” appear in neutral side after being modeled by SentiLDA. The possible reason might be without the hard constraints of sentiment seed words, some common sentiment strong words that frequently used by people will be detected as common fact(neutral) words by SentiLDA. But the bottom line is they have never been detected as the opposite sentiment.

6 Conclusion and Future Work

In this paper, we studied the problem of mining aspect specific sentiments from unstructured text data and structured numerical data by exploiting the numerical review rating as a prior for the sentiment distribution in the unstructured review. In specific, we defined a novel problem of mining opinions from reviews and ratings without relying on sentiment seed words and proposed an effective and scalable approach to solve this problem. The experiment results show SentiLDA outperforms the other state-of-art topic models in discovering aspect specific sentiment words, converging faster to a higher log likelihood, and better predicting unseen data.

There are some interesting future directions of this study. First, we have not tried to optimize the hyper parameters according to different ratings. It would be interesting to study how the performance of the model would improve if the hyper parameters are optimized. Second, sentiment seed words are not incorporated in this model, in the future we would like to study how we can incorporate the seed words in the model to improve the performance.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
2. Diao, Q., Qiu, M., Wu, C.Y., Smola, A.J., Jiang, J., Wang, C.: 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, pp. 193–202. ACM, August 2014
3. Fellbaum, C.: WordNet. Blackwell Publishing Ltd. (1998)
4. Ganu, G., Elhadad, N., Marian, A.: Beyond the stars: improving rating predictions using review text content. In: WebDB, vol. 9, pp. 1–6, June 2009
5. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proc. Natl. Acad. Sci. **101**(suppl. 1), 5228–5235 (2004)
6. Horrigan, J.A.: Online shopping. In: Pew Internet & American Life Project Report (2008)
7. Hu, M., Liu, B.: Mining opinion features in customer reviews. In: AAAI, vol. 4, No. 4, pp. 755–760, July 2004
8. Jo, Y., Oh, A.H.: Aspect and sentiment unification model for online review analysis. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, pp. 815–824. ACM, February 2011
9. Lin, C., He, Y.: Joint sentiment/topic model for sentiment analysis. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, pp. 375–384. ACM, November 2009
10. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, pp. 55–60, June 2014
11. Mei, Q., Ling, X., Wondra, M., Su, H., Zhai, C.: Topic sentiment mixture: modeling facets and opinions in weblogs. In: Proceedings of the 16th International Conference on World Wide Web, pp. 171–180. ACM, May 2007
12. http://spark.apache.org/
13. Titov, I., McDonald, R.: Modeling online reviews with multi-grain topic models. In: Proceedings of the 17th International Conference on World Wide Web, pp. 111–120. ACM, April 2008
14. Turney, P.D.: Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, pp. 417–424. Association for Computational Linguistics, July 2002
15. Wallach, H.M., Mimno, D.M., McCallum, A.: Rethinking LDA: why priors matter. In: Advances in Neural Information Processing Systems, pp. 1973–1981 (2009)
16. Wang, H., Lu, Y., Zhai, C.: 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, pp. 783–792. ACM, July 2010
17. Wu, N., Liu, F., Zhang, J.: A study on consistency of cross-site online reviews. In: The 10th IEEE International Conference on Pervasive, Intelligence and Computing, December 2013
18. Xu, X., Tan, S., Liu, Y., Cheng, X., Lin, Z.: Towards jointly extracting aspects and aspect-specific sentiment knowledge. In: Proceedings of the 21st ACM International Conference on Information and Knowledge Management, pp. 1895–1899. ACM, October 2012

19. Zhai, K., Boyd-Graber, J., Asadi, N., Alkhouja, M.L.: Mr. LDA: a flexible large scale topic modeling package using variational inference in mapreduce. In: Proceedings of the 21st International Conference on World Wide Web, pp. 879–888. ACM, April 2012
20. Zhai, Z., Liu, B., Xu, H., Jia, P.: Clustering product features for opinion mining. In: Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, pp. 347–354. ACM, February 2011