On the Design of LDA Models for Aspect-based Opinion Mining

Samaneh Moghaddam
School of Computing Science
Simon Fraser University
Burnaby, BC, Canada
sam39@cs.sfu.ca

Martin Ester
School of Computing Science
Simon Fraser University
Burnaby, BC, Canada
ester@cs.sfu.ca

ABSTRACT

Aspect-based opinion mining, which aims to extract aspects and their corresponding ratings from customers reviews, provides very useful information for customers to make purchase decisions. In the past few years several probabilistic graphical models have been proposed to address this problem, most of them based on Latent Dirichlet Allocation (LDA). While these models have a lot in common, there are some characteristics that distinguish them from each other. These fundamental differences correspond to major decisions that have been made in the design of the LDA models. While research papers typically claim that a new model outperforms the existing ones, there is normally no "one-size-fits-all" model. In this paper, we present a set of design guidelines for aspect-based opinion mining by discussing a series of increasingly sophisticated LDA models. We argue that these models represent the essence of the major published methods and allow us to distinguish the impact of various design decisions. We conduct extensive experiments on a very large real life dataset from Epinions.com (500K reviews) and compare the performance of different models in terms of the likelihood of the held-out test set and in terms of the accuracy of aspect identification and rating prediction.

Categories and Subject Descriptors

I.7.0 [**Document and Text Processing**]: General; G.3 [**Mathematics of Computing**]: Probability and Statistics—*statistical computing, multivariate statistics*

General Terms

Algorithms, Design, Experimentation

Keywords

aspect-based opinion mining, latent dirichlet allocation, variational methods, aspect identification, rating prediction

1. INTRODUCTION

Other people's opinions have always been an important piece of information during the decision-making process when buying a new product. Today people like to make their opinions available to strangers via the Internet. As a result, the Web has become an excellent source for gathering consumer opinions. As online commerce activity continues to grow, the role of online reviews is expected to become increasingly important [18]. Epinions,Amazon, and Cnet,are examples of the most important Web resources containing such opinions.

However, reading all product reviews to make a good decision is a time-consuming job. Reading different and possibly even contradictory opinions written by different reviewers may even make customers more confused. While some of the reviewing websites ask reviewers to express an overall rating (as stars) for the reviewed item, focusing on just overall ratings will not be sufficient for a user to make decision. These needs have inspired a new line of research on mining online product reviews, which is *aspect-based opinion mining.* Aspect-based opinion mining aims to extract major aspects of a product and to predict the rating of each aspect from the product reviews. Aspects are attributes or components of products (e.g., 'LCD', 'battery life', etc. for a digital camera) and ratings are the intended interpretation of user satisfaction in terms of numerical values. Aspects and ratings clearly provide more detailed information than the overall rating and help users to make better decisions [23]. Aspect-based opinion mining is not only very useful for potential customers to know the opinions of other users before they use a service or purchase a product, but also crucial for businesses to find consumer opinions on their products and services [16]. These techniques can also provide valuable sources of information for other applications, e.g. advertising methods, recommendation systems, etc.

In the last decade several probabilistic models have been proposed to address the problem of aspect-based opinion mining (e.g., [26, 30, 2, 29, 10, 19, 6, 15, 15]). Most of these models are based on Latent Dirichlet Allocation (LDA) which is typically used in topic modeling. While these models have a lot in common, there are some characteristics that distinguish them from each other. In the following we point out some of the main distinctive features:

- Modeling words using one latent variable vs. having separate latent variables for aspects and ratings.
- Modeling all words of the reviews vs. modeling only opinion phrases.
- Modeling the dependency between aspects and ratings vs. modeling them independently.

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.
Copyright 2012 ACM 978-1-4503-1156-4/12/10 ...$10.00.

- Using only review texts as input vs. also using additional input data, e.g. a review's overall rating.

These features correspond to major decisions that must be made in the design of an LDA model. While research papers typically claim that a new model outperforms the existing ones and demonstrate this with experiments on some data sets, there is normally no "one-size-fits-all" model. However, in the literature it is normally not shown why and in which scenarios a model performs better than another one. For example, let us consider a proposed model A generating all words which considers the dependency between aspects and their ratings and takes the overall rating of reviews as an additional observed variable. The comparison of model A with the basic LDA model on a dataset with a large number of reviews per product shows better performance of model A. The problem with this experimental evaluation is that it does not reveal whether the better performance of model A is due to the dependency assumption between aspects and ratings or due to the additional observed data. It is also not clear whether model A still performs better for another dataset containing products with few reviews.

We argue that the best choice for some design decision may depend on other design decisions and on the content and the size of the dataset. So, in this paper we do not propose yet another LDA model for the problem of aspect-based opinion mining, but we present design guidelines for such models. To derive these guidelines, we discuss a series of increasingly sophisticated probabilistic graphical models based on LDA. We start with the basic LDA model and then gradually extend the model by adding latent and observed variables as well as dependencies. The discussed models are as follows:

- LDA: The basic LDA model proposed in [1] which learns general topics of reviews using all words of the training reviews.

- S-LDA: An extension of LDA where the model learns both aspects and ratings from reviews.

- D-LDA: Extension of S-LDA considering the dependency between aspects and their ratings.

- PLDA: The basic LDA model which learns general topics of reviews from opinion phrases.

- S-PLDA: An extension of PLDA where the model learns both aspects and ratings from phrases.

- D-PLDA: An LDA-based model learning aspects and their corresponding ratings from opinion phrases while considering the dependency between the aspects and ratings.

We argue that these six models represent the essence of the major published methods and allow us to tease apart the impact of various design decisions. For example, the comparison of S-LDA and S-PLDA against D-LDA and D-PLDA shows whether the dependency of aspects and ratings improves the performance, independent from the type of observed data (all words or opinion phrases).

Since there is no benchmark dataset for the problem of aspect-based opinion mining, current works have been evaluated on different data sets. However, one dataset may have products with only few reviews, while the products of another dataset may have been selected to have at least a few hundred reviews. We crawled the well-known reviewing website, Epinions.com, and built a very large dataset containing 505,978 reviews about 94,792 products from 257 different product categories. We made the dataset publicly available for research purposes[1]. We evaluate the performance of the six models in terms of the likelihood of a held-out test set. To measure the impact of the training set size, we perform experiments for different subsets of products with different numbers of reviews. We also evaluate the accuracy of the models in aspect identification and rating prediction (precision, recall, and mean squared error) on a labeled dataset and find a strong correlation between model perplexity and accuracy.

As a novel technical contribution, we present a method for extracting opinion phrases based on grammatical relations provided by a dependency parser. This technique promises to generate opinion phrases more accurately than current methods that consider only syntactic properties such as the proximity of words.

The remainder of the paper is organized as follows. The next section is devoted to related work. Section 3 introduces the problem statement and discusses our contributions. Section 4 presents different LDA models for the considered problem. Section 5 describes the inference and estimation techniques for the presented models. The proposed technique for extracting opinion phrases is discussed in 6. In Section 7 we report the results of our experimental evaluation. Finally, Section 8 concludes the paper with a summary of our design guidelines and the discussion of future work.

2. RELATED WORK

The problem of aspect-based opinion mining has recently attracted increasing attention. As a result, there are several lines of related works on this problem. Most of the early works on this problem are feature-based approaches (e.g., [8, 17, 22]). These approaches normally apply some constraints on high-frequency noun phrases to identify product aspects. As a result they usually produce too many non-aspects and miss low-frequency aspects [6]. In addition, feature-based approaches require the manual tuning of various parameters which makes them hard to port to another dataset [21].

Latent variable models overcome these limitations by automatically learning the model parameters from the data. Some of the proposed models are based on Hidden Markov Model (HMM) [28, 10], Conditional Random Field (CRF) [13, 3], and Latent Semantic Association [19, 6, 5]. However, most of the current models are based on Latent Dirichlet Allocation (LDA), e.g., [26, 25, 30, 2, 11, 29, 27, 12, 14, 15, 7]. Since the focus of our paper is LDA-based models, in the following we will only discuss the most recent and most important LDA-based models presented in the literature for the problem of aspect-based opinion mining.

Titov et al. [26, 25] propose a model, based on LDA, for extracting aspects from reviews. They consider two types of topics for each review: global topics, which correspond to global properties of the product (e.g., product brand) and local topics which are related to the product aspects. Their proposed model assumes that each corpus has a certain dis-

[1] http://www.sfu.ca/~sam39/Datasets/EpinionsReviews/

tribution of global topics and each review has a distribution of local topics. Each word of a review is generated by sampling a topic from one of these distributions. An indicator variable is used to select the type of topic for each word.

An LDA-based model for jointly identifying aspects and sentiments is proposed in [30]. The model assumes that all of the words in a sentence are generated from a single topic. It also considers different word distributions for aspects, sentiments, and background words and leverages Part-Of-Speech (POS) tags of words (using indicator variables) to separate them. The authors of [2, 11] also assume that all words in a single sentence are generated from one topic and apply LDA at the sentence level to extract product aspects. The model presented in [11] is further extended to extract sentiments related to each aspect. In this model, each review has a distribution over sentiments and each sentiment has a distribution over aspects. For each sentence of a review, the model first draws a sentiment from the sentiment distribution and then an aspect is sampled conditioned on the selected sentiment. Each words of the sentence is then generated based on the selected aspect and sentiment.

The authors of [27] propose an LDA-based model to jointly identify aspects, their ratings, and the weight placed on each aspect by the reviewer. As input the model takes all words of reviews about a product and the overall ratings assigned by reviewers to that product. Each word of a review is generated by sampling an aspect from the learned distribution. The word and the aspect together generate the rating of the aspect. Finally, the aspect weights are sampled from a normal distribution and the overall rating of review is generated based on the weighted sum of all the aspect ratings.

The authors of [12] propose an LDA-based model for identifying aspects, sentiments, and their ratings. The model assumes that each review has a distribution over aspects and another distribution over sentiments. To generate each word of a review, an aspect and a sentiment are first chosen from the corresponding distributions. Then a word is generated conditioned on both aspect and sentiment. The rating of each sentiment is also computed using a normal linear model learned by the overall rating of review. The model further captures the dependencies between the syntactic class of aspects and sentiments. The syntactic class of a word is chosen based on the POS tag of that word and the POS tag of the previous word.

Zhan's model [29] is defined on opinion phrases (see Section 3.1 for definition). They apply a set of POS patterns on the review text to extract nouns and related adjectives as opinion phrases. Then the basic LDA model is applied at the sentence level to extract topics from opinion phrases. Each topic of this model is a pair of head term and modifier. The model presented in [21] also learns from opinion phrases, which are extracted from frequent nouns and adjectives. This model assumes the dependency between aspects and ratings. To generate each opinion phrase, an aspect is first chosen from a Dirichlet distribution. Then a rating is selected conditioned on the aspect. Finally, a head term and a modifier are generated based on the selected aspect and rating, respectively.

There are also some works [14, 15, 7] proposing LDA-based models for extracting topics and their polarity from documents. While these works did not talk about aspects and rating specifically, their model can be applied on reviews for identifying aspects and their ratings. The generative assumption of the model proposed in [14] is as follows: To generate a word of a document, a topic is first selected and the word polarity (positive or negative) is then sampled depending on the chosen topic. Finally, a word is generated conditioned on both topic and polarity. The authors further extend this model by considering the dependency of word polarity on the local context. For example, if two words are connected using 'and', their ratings are considered the same.

The authors of [15, 7] assume that each document has a specific distribution over the polarity of words and there is a specific topic distribution for each polarity (positive and negative) which is independent from the documents (global topic distribution). To generate a word, the proposed model first chooses the polarity of word form the document specific polarity distribution. Then it selects a topic from the global topic distribution conditioned on the selected polarity. Finally, a word is generated conditioned on the selected topic and polarity.

3. PROBLEM DEFINITION

In this section we first define the basic terminologies we will use in this paper. Then we formulate the problem statement and discuss our contributions.

3.1 Problem Statement

Review: A review contains a sequence of words describing opinions of reviewer regarding a specific item (e.g. product or service).

Aspect: An aspect (also called product feature [8]) is an attribute or component of the product that has been commented on in a review, e.g. 'battery life', 'zoom', 'shutter lag', etc. for a digital camera.

Sentiment: Sentiment is a linguistic term which refers to the direction in which a concept or opinion is interpreted [16]. We use sentiment in a more specific sense as an opinion about an aspect which is usually expressed by an adjective. For example, 'great' is a sentiment for the aspect 'picture quality' in the sentence "It has great picture quality".

Opinion Phrase: An opinion phrase $< h, m >$ is a pair of head term h and modifier m [19]. Usually the head term is a candidate aspect, and the modifier is a corresponding sentiment expresses some opinion towards the aspect, e.g. <LCD, blurry>, <screen, inaccurate>, etc.

Rating: A sentiment can be classified in n-level orientation scale. Sentiment orientation is an intended interpretation of the user satisfaction in terms of numerical values [21]. Polarity is a two-level orientation scale. In this scale a sentiment is either positive or negative. Most of the reviewing websites use five-level orientations, presented by stars in the range from 1 to 5 which is called rating.

Aspect-based Opinion Mining: Given a set of reviews for product P, the task is to identify the k major aspects of P and to predict the rating of each aspect for P.

3.2 Our Contributions

As discussed in the introduction section, our goal is to present a set of design guidelines for LDA models for learning aspects and their ratings from reviews. We focus on the following design choices:

- Is it better to have separate latent variables for aspects and ratings?

- Is it better to assume dependency between ratings and aspects?
- Is it better to learn from bag-of-words or preprocess the reviews and learn from opinion phrases?
- Which preprocessing technique for extracting opinion phrases works best?
- Does the answer to the above questions differ for products with few reviews and products with many reviews?

We start our investigation with the basic LDA model [1] and then gradually extend the model by considering different probabilistic assumptions. In summary we discuss five LDA-based models with various underlying assumptions for our problem: *S-LDA* extends LDA by assuming the review is generated by a set of aspects and their ratings. *D-LDA* adds the dependency between aspects and their ratings. *PLDA* learns one latent variable from opinion phrases. *S-PLDA* learns both aspects and their corresponding ratings from opinion phrases. Finally, *D-PLDA* learns aspects and ratings from opinion phrases while considering the dependency between the generated aspects and ratings.

We also present a novel technique for extracting opinion phrases from reviews based on dependency parsing. Different from current preprocessing methods which are mainly based on syntactic properties our technique is based on the semantic relationships between words. We conduct extensive experiments on a real life dataset from Epinions.com, and based on the results we propose a set of design guidelines for LDA models for aspect-based opinion mining.

4. LDA MODELS FOR ASPECT-BASED OPINION MINING

In this section, we first briefly discuss the basic LDA model for the problem of aspect-based opinion mining. Then we introduce a series of models based on LDA making different probabilistic assumptions.

4.1 LDA

Latent Dirichlet Allocation (LDA) is a generative probabilistic model of a corpus [1]. The basic idea is that documents are represented as mixtures over latent topics where topics are associated with a distribution over the words of the vocabulary. Figure 1 shows the graphical model of this model. Following the standard graphical model formalism, nodes represent random variables and edges indicate possible dependency. Shaded nodes are observed random variables and unshaded nodes are latent random variables. Finally, a box around groups of random variables is a 'plate' which denotes replication. The outer plate represents reviews and the inner plate represents words. D and N denote the number of product reviews and the number of words in each review, respectively. In our domain, LDA assumes the following generative process:

1. Sample $\theta \sim Dir(\alpha)$.
2. For each word w_n, $n \in \{1, 2, ..., N\}$
   (a) Sample a topic $z_n \sim Mult(\theta)$
   (b) Sample a word $w_n \sim P(w_n|z_n, \beta)$, a multinomial distribution conditioned on the topic z_n.

Translating this process into a joint probability distribution results in the following expression:

$$P(\boldsymbol{z}, \boldsymbol{w}, \theta|\alpha, \beta) = P(\theta|\alpha) \prod_{n=1}^{N} [P(z_n|\theta)P(w_n|z_n, \beta)] \quad (1)$$

The key inference problem is to compute the posterior distribution of the latent variables given a review:

$$P(\boldsymbol{z}, \theta|\boldsymbol{w}, \alpha, \beta) = \frac{P(\boldsymbol{z}, \boldsymbol{w}, \theta|\alpha, \beta)}{P(\boldsymbol{w}|\alpha, \beta)} \quad (2)$$

Since this distribution is intractable to compute (due to the coupling between θ and β), variational inference is used to approximate this distribution [1]. We will explain the inference and estimation techniques in Section 5.

Some of the current works [2, 30, 26, 25] apply this model on reviews to extract topics as product aspects. In [2] and [30], the model has been used at the sentence level. The authors of [30] further improve the model by considering different word distributions for aspects, ratings, and background words. In [26, 25] the basic LDA model is improved by considering different topic distributions for local and global topics and is applied at the document level.

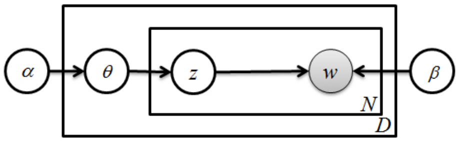


Figure 1: LDA

4.2 S-LDA

The second model replaces the one latent variable for topics by two separate variables for aspects and ratings. We call this model Separate-LDA (S-LDA). For every aspect/rating pair, θ contains the probability of generating that combination of aspect and rating. The variable θ is sampled once per review. After sampling θ, the latent variables a_n and r_n are sampled independently (conditional independency), and then a word w_n is sampled conditioned on the sampled aspect and rating (Figure 2). The joint probability distribution of this model is as follows:

$$P(\boldsymbol{a}, \boldsymbol{r}, \boldsymbol{w}, \theta|\alpha, \beta) = P(\theta|\alpha) \prod_{n=1}^{N} [P(a_n|\theta)P(r_n|\theta)P(w_n|a_n, r_n, \beta)] \quad (3)$$

A model similar to S-LDA has been proposed in [12], learning two Dirichlet distributions (one for aspects and one for sentiments) per review. While that model further considers the syntactic dependency between words and is more an LDA-HMM model, the generation of words conditioned on both aspects and ratings is the same as the generative process of S-LDA.

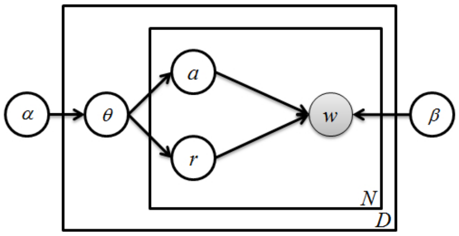


Figure 2: S-LDA

4.3 D-LDA

Dependency-LDA (D-LDA) also models the dependency between the latent aspects and ratings while learning from a bag-of-words model of reviews (Figure 3). The joint probability distribution of D-LDA considers this dependency:

$$P(\boldsymbol{a}, \boldsymbol{r}, \boldsymbol{w}, \theta|\alpha, \beta, \omega) = \\ P(\theta|\alpha) \prod_{n=1}^{N} [P(a_n|\theta)P(r_n|a_n, \omega)P(w_n|a_n, r_n, \beta)] \quad (4)$$

Several models similar to D-LDA have been proposed in the literature [14, 15, 7, 27, 11]. The model presented in [14] further considers the dependency of the rating from the local context. The model proposed in [15] (further extended in [7]) assumes the dependency of the selected aspect from the sampled rating, i.e., the opposite direction of the dependency. The authors of [11] also make the same assumption as [15] and apply the model at the sentence level to extract aspects and ratings.

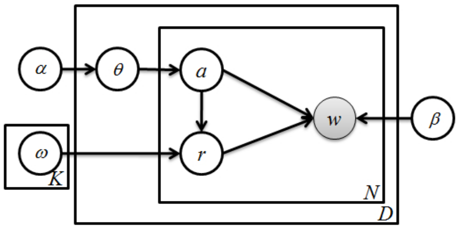


Figure 3: D-LDA

4.4 PLDA

While LDA assumes a bag-of-words model, Phrases-LDA (PLDA) assumes a bag-of-phrases model of product reviews (Figure 4). A review is preprocessed into a bag-of-opinion-phrases $< h_n, m_n >$ which leads to two observed variables h_n (head term) and m_n (modifier). Translating this process into a joint probability distribution results in the expression:

$$P(\boldsymbol{z}, \boldsymbol{h}, \boldsymbol{m}, \theta|\alpha, \beta, \pi) = \\ P(\theta|\alpha) \prod_{n=1}^{N} [P(z_n|\theta)P(h_n, m_n|z_n, \beta, \pi)] \quad (5)$$

In [29] a similar model is applied on opinion phrases to extract topics from reviews. While the generation of opinion phrases in [29] is the same as the generative process of PLDA, their model further assumes that each sentence of the review is related to only one topic.

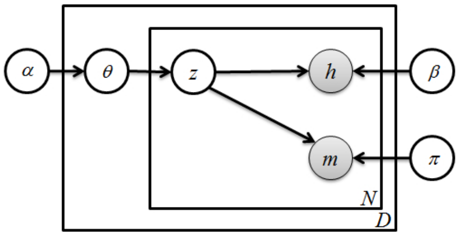


Figure 4: PLDA

4.5 S-PLDA

Compared to PLDA, the S-PLDA model introduces a separate rating variable which is conditionally independent from the aspect. In this model a review is assumed to be generated by first choosing a value of θ, and then repeatedly sampling N aspects and ratings as well as opinion phrases $< h_n, m_n >$ conditioned on the chosen value of θ. Similar to S-LDA, θ represents the aspect/rating pairs and for every pair, α contains the probability of generating that combination of aspect and rating (Figure 5). The joint probability distribution of S-PLDA is as follows:

$$P(\boldsymbol{a}, \boldsymbol{r}, \boldsymbol{h}, \boldsymbol{m}, \theta|\alpha, \beta, \pi) = \\ P(\theta|\alpha) \prod_{n=1}^{N} [P(a_n|\theta)P(r_n|\theta)P(h_n|a_n, \beta)P(m_n|r_n, \pi)] \quad (6)$$

$P(h_n|a_n, \beta)$ and $P(m_n|r_n, \pi)$ are multinomial distributions conditioned on the aspect a_n and rating r_n, respectively. The same model is presented in [21] as a comparison partner.

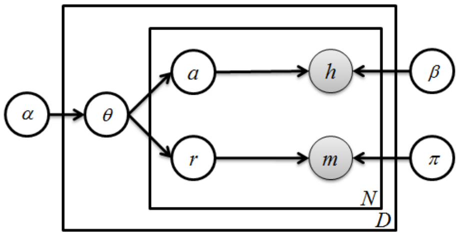


Figure 5: S-PLDA

4.6 D-PLDA

Similar to the step from S-LDA to D-LDA, compared to S-PLDA, the D-PLDA model adds the dependency between ratings and aspects. There are various options for dependencies between the two latent variables and the two observed variables. We assume that modifiers depend on the aspect and the rating. To illustrate the importance of this dependency, consider the following opinion phrases: 'low LCD resolution' and 'low price'. The modifier 'low' expresses a negative opinion for the head term 'LCD resolution', while it is a positive opinion for 'price'. On the other hand, we assume that the rating of an aspect does not affect the choice of a head term for that aspect.

D-PLDA can be viewed as generative process that first generates an aspect and subsequently generates its rating. In particular, for generating an opinion phrase, this model first generates an aspect a_n from an LDA model. Then it generates a rating r_n conditioned on the sampled aspect a_n. Finally, a head term h_n is drawn conditioned on a_n and a modifier m_n is generated conditioned on both the aspect a_n and rating r_n (Figure 6). D-PLDA specifies the following joint distribution:

$$P(\boldsymbol{a}, \boldsymbol{r}, \boldsymbol{h}, \boldsymbol{m}, \theta|\alpha, \omega, \beta, \pi) = P(\theta|\alpha) \\ \prod_{n=1}^{N} [P(a_n|\theta)P(r_n|a_n, \omega)P(h_n|a_n, \beta)P(m_n|a_n, r_n, \pi)] \quad (7)$$

In [21] a similar model is presented. D-PLDA generates the modifier conditioned on both aspect and rating, but in [21] only the selected rating generates the modifier.

5. INFERENCE AND ESTIMATION

Computing the posterior distribution of the latent variables for the LDA models is intractable. Blei et al. [1] proposed to obtain a tractable lower bound by modifying the

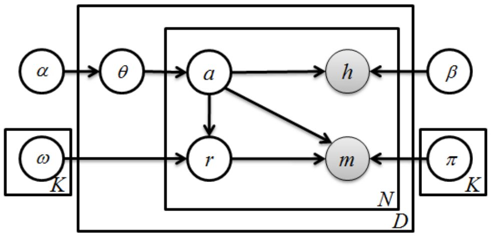


Figure 6: D-PLDA

graphical model through considering a variational Dirichlet parameter for generating θ and a variational multinomial parameter for generating each latent variable. In a good approximation, the KL-divergence between the variational distribution and the true posterior will be minimum. So, by setting the derivative of the KL-divergence with respect to variational parameters equal to zero, the update equations can be obtained. Using Variational Estimation-Maximization (EM) technique [1], a lower bound on the posterior probability can be obtained.

Regarding the computational complexity, each iteration of variational inference for the basic LDA requires $O(Nk)$ operations [1] where k is the number of topics. According to the variational inference algorithms, S-LDA and D-LDA require $O(5Nk)$, PLDA and S-PLDA require $O(2Nk)$, and D-PLDA require $O(6Nk)$ operations for each iteration. As stated in [1], the number of iterations required for a single document is on the order of the number of words in the document. This means that the total number of operations for the LDA models is roughly on the order of $O(N^2k)$.

When working with conditional distributions, over-fitting is always a serious problem [1]. A new review is very likely to contain words that did not appear in any of the reviews in a training corpus. Maximum likelihood estimate of the model parameters assign zero probability to such words, and so zero probability to new reviews. Smoothing is a standard approach to dealing with this problem [1]. We smooth all the parameters which depend on the observed data by assigning positive probability to all vocabulary terms whether or not they are observed in the training set.

Finally, when estimating model parameters using maximum likelihood estimation, it is possible to increase the likelihood by adding parameters, which may however result in over-fitting. Since our goal is to compare the average performance of different models, we perform our experiments for different values of k. Note that, before applying the models on reviews, we first apply the Porter Stemmer algorithm [24] and then remove stop words using a standard lists of stop words[2].

6. EXTRACTION OF OPINION PHRASES

Bag-of-words is a popular representation of documents in text processing. In the area of aspect-based opinion mining most of the current works [2, 30, 26, 25, 12, 14, 15, 7, 27, 11] adopt this representation of reviews. However, it is not clear whether representing a review as a bag of words is sufficient for this problem. Some of the recent works [29, 21] propose to preprocess the product reviews to extract opinion phrases and present LDA models that generate only opinion phrases. These works typically use some simple parsing techniques to extract pairs of frequent noun and nearest adjective, or use POS patterns, e.g. "[adjective] [noun]", "[noun] [verb] [adjective]", etc.

In this section, we present a novel method for extracting opinion phrases based on the Stanford Dependency parser [20], which is a parser widely used in the area of text mining. A dependency parser determines the semantic relationships between words and promises to generate opinion phrases more accurately than methods that consider only the proximity of words. Dependency parsers provide a simple description of the grammatical relationships in a sentence. In the following, we briefly explain the grammatical relations [20] we use:

- Adjectival complement (*acomp*): An adjectival phrase which functions as the complement, e.g., "The auto-mode works amazing" parsed to 'acomp(works, amazing)'.
- Adjectival modifier (*amod*): An adjectival phrase that serves to modify the meaning of a noun phrase, e.g., "It has a wide screen" parsed to 'amod(screen, wide)';
- "And" conjunct (*conj_and*): A relation between two elements connected by the coordinating conjunction "and", e.g., "The LCD is small and blurry" parsed to 'conj_and(small, blurry)'.
- Copula (*cop*): A relation between the complement of a copular verb and the copular verb, e.g., "The batteries are ok" parsed to 'cop(ok, are)'.
- Direct object (*dobj*): A noun phrase which is the object of the verb, e.g., "I like the auto-focus" parsed to 'dobj(like, auto-focus)'.
- Negation modifier (*neg*): A relation between a negation word and the word it modifies, e.g., "The shutter lag is'n fast" parsed to 'neg(fast, n't)'.
- Noun compound modifier (*nn*): A noun that serves to modify the head noun, e.g., "The shutter lag is'n fast" parsed to 'nn(lag, shutter)'.
- Nominal subject (*nsubj*): A noun phrase which is the syntactic subject of a clause, e.g., "The zoom is disappointing" parsed to 'nsubj(disappointing, zoom)'.

We employ these grammatical relations to define a set of dependency patterns for extracting opinion phrases. In the following we list the extraction patterns (N indicates a noun, A an adjective, V a verb, h a head term, m a modifier, and $< h, m >$ an opinion phrase). Table 1 shows some examples of how these patterns are used for extracting opinion phrases.

1. $\mathit{amod}(N, A) \rightarrow < N, A >$
2. $\mathit{acomp}(V, A) + \mathit{nsubj}(V, N) \rightarrow < N, A >$
3. $\mathit{cop}(A, V) + \mathit{nsubj}(A, N) \rightarrow < N, A >$
4. $\mathit{dobj}(V, N) + \mathit{nsubj}(V, N') \rightarrow < N, V >$
5. $< h_1, m > + \mathit{conj\_and}(h_1, h_2) \rightarrow < h_2, m >$
6. $< h, m_1 > + \mathit{conj\_and}(m_1, m_2) \rightarrow < h, m_2 >$
7. $< h, m > + \mathit{neg}(m, \mathit{not}) \rightarrow < h, \mathit{not} + m >$
8. $< h, m > + \mathit{nn}(h, N) \rightarrow < N + h, m >$
9. $< h, m > + \mathit{nn}(N, h) \rightarrow < h + N, m >$

[2] `http://ir.dcs.gla.ac.uk/resources/linguistic_utils/stop_words`

Table 1: Dependency patterns for extracting opinion Phrases

Sentence	Dependency Relations	Patrn.	Opinion Phrases
This camera has great zoom and resolution.	amod(zoom,great), conj_and(zoom,resolution)	1, 5	<zoom,great>, <resolution,great>
It comes with small and rechargeable batteries.	amod(batteries,rechargeable), conj_and(small,rechargeable)	1, 6	<batteries,rechargeable>, <batteries,small>
The camera case looks nice.	acomp(looks,nice),nsubj(looks,case),nn(case,camera)	2, 8	<camera case,nice>
I love the picture quality.	dobj(love,picture),nsubj(love,I),nn(quality,picture)	4, 9	<picture quality,love>
The screen is wide and clear.	cop(wide,is),nsubj(wide,screen),conj_and(wide,clear)	3, 6	<screen,wide>, <screen,clear>
The battery life is not long.	cop(long,is),nsubj(long,life),nn(life,battery),neg(long,not)	3,8,7	<battery life,not long>

7. EXPERIMENTS

As discussed in the introduction, there is still no benchmark dataset for the problem of aspect-based opinion mining. Since our goal is to compare the performance of different LDA models for products with different numbers of reviews, we have built a very large real life dataset by crawling Epinions.com. In the next subsections, we first briefly describe our dataset and then present a qualitative and quantitative evaluation of the LDA models. For qualitative analysis we compare the top words obtained by different models. For the quantitative analysis we measure the performance of the models in terms of test set likelihood.

7.1 Dataset

We built a crawler to extract product reviews from the well-known reviewing website Epinions.com. The dataset contains 505,978 reviews about 94,792 products from 257 product categories (e.g., camcorder, cellular phone, digital camera, Mp3 player, etc.). Note that, in aspect-based opinion mining, since both aspects and ratings are product-specific, one model is learnt per product. Out of 94,792 products, 49,324 products have only one review which makes it impossible to train and test, so these products were removed.

Table 2: Statistics of the Dataset

Subset	#Products	#Rev./Product
$1 < \#Rev. <= 10$	36,166	3
$10 < \#Rev. <= 50$	7,886	19
$50 < \#Rev. <= 100$	869	67
$100 < \#Rev. <= 200$	368	137
$200 < \#Rev.$	179	341

To the best of our knowledge, the impact of the size of the training dataset has not been evaluated in the literature on aspect-based opinion mining. To do so, we define five subsets of products with different numbers of reviews. The first subset contains products with at least 2 and at most 10 reviews. We also consider subsets of products with more than 10 and less than 50 reviews, between 50 and 100 reviews, from 100 to 200 reviews, and more than 200 reviews. For each subset of products, Table 2 shows the number of products in that subset and the average number of reviews per product. In the following sections we present evaluation results for each subset of products as the average of the results for the products of that subset.

7.2 Qualitative Evaluation

To perform a qualitative evaluation, we select a product from the digital camera category that has 166 reviews. Table 3 shows the top (most probable) words/phrases extracted by different models for this product. We also compare the performance of models learning from phrases (PLDA, S-PLDA, and D-PLDA) using different preprocessing techniques:

- Frequent noun technique: Pairs of frequent nouns and nearest adjectives [21].
- POS patterns: Pairs of nouns and adjectives extracted using POS patterns [29].
- Dependency patterns (introduced in Section 6)

By comparing the top words of the first three models, we observe that the extracted words are almost the same, i.e., S-LDA and D-LDA do not perform better than the basic LDA. Comparing the extracted phrases using frequent noun technique and POS patterns, we see that the frequent noun technique is not that promising since there are some frequent phrases which are not relevant (e.g., <time, hard> and <pictur, mani>) and there are lots of opinion phrases which are not frequent (e.g., <pictur, good> and <displai, nice>). It is also shown that often the extracted phrases based on dependency patterns are more informative, e.g., <qualiti, amaz>, <view find, love>, and <photo qualiti, high>.

7.3 Quantitative Evaluation

If the necessary ground truth is available, the performance of a model can be evaluated by measures such as accuracy, precision and recall. However, in large data sets such as our Epinions dataset ground truth is typically not available. In such cases, a standard approach for the evaluation of graphical models is comparing the likelihoods of a held-out test set. We hold out 10% of the reviews for testing purposes and use the remaining 90% to train the model. As is standard for LDA models [1, 29, 21], we computed the *perplexity* of the held-out test set for all models for various numbers of aspects, $k = \{5, 10, 15, 20, 25\}$. Since the relative results are similar for different values of k, we choose $k = 15$ for our discussion. The perplexity is monotonically decreasing in the likelihood of the test data, and a lower perplexity score indicates better performance. More formally, for a test set of N reviews, the perplexity is defined as [1]:

$$perplexity(R_{test}) = exp\{-\frac{\sum_{d=1}^{D} \log P(\boldsymbol{w}_d)}{\sum_{d=1}^{D} N_d}\} \qquad (8)$$

Table 3: Top words/phrases extracted by different LDA Models

Prep.	Model	Top words/phrases extracted for a digital camera (stemmed)
N/A	LDA	good, pictur, digit, resolut, set, disk, great, time, shot, featur
	S-LDA	featur, zoom, disk, good, pictur, set, shot, resolut, camera, floppi
	D-LDA	bright, time, resolut, good, disk, set, great, digit, shot, camera
Freq. nouns	PLDA	<pictur,mani>,<resolut,high>,<time,hard>,<camera,digit>,<disk,floppi>
	S-PLDA	<time,hard>,<drive,floppi>,<resolut,mani>,<camera,digit>,<featur,good>
	D-PLDA	<resolut,high>,<camera,digit>,<drive,floppi>,<usb,easi>,<disk,hard>
POS patrn.	PLDA	<usag,normal>,<price,high>,<drive,floppi>,<pictur,good>,<featur,sever>
	S-PLDA	<pictur,good>,<effect,special>,<displai,nice>,<life,long>,<printer,great>
	D-PLDA	<batteri,dead>,<resolut,mani>,<camera,digit>,<pictur,good>,<featur,offer>
Dep. patrn.	PLDA	<displai,nice>,<zoom,optic>,<effect,mani>,<pictur,good>,<reolut,high>
	S-PLDA	<resolut,high>,<qualiti,amaz>,<printer,compat>,<displai,nice>,<zoom,optic>
	D-PLDA	<photo qualiti,high>,<zoom,optic>,<storag capac,unlimit>,<printer,compat>,<viewfind,love>

For models learning from opinion phrases, $P(\boldsymbol{w}_d)$ is being replaced by $P(\boldsymbol{h}_d, \boldsymbol{m}_d)$. While perplexity is a well-established measure for comparing LDA models, it is not clear how perplexity relates to the accuracy of aspect identification and rating prediction. To this end, we use a well-known public dataset with ground truth used in [8, 9, 4] (with 314 reviews of 5 products) to analyze the correlation between model perplexity and accuracy of these tasks. Table 4 shows the average precision and recall of aspect identification, the Mean Squared Error (MSE) of rating prediction and the perplexity of different models on this dataset averaged over all products.

Table 4: Evaluation on Labeled Dataset

Model	Precision	Recall	MSE	Perplexity
S-LDA	0.54	0.51	1.25	813.11
LDA	0.54	0.52	1.22	795.72
D-LDA	0.58	0.55	1.18	748.26
PLDA	0.81	0.73	0.96	587.82
S-PLDA	0.83	0.73	0.93	335.02
D-PLDA	0.87	0.78	0.85	131.80

We observe a strong correlation of the perplexity and precision, recall, and MSE. All three accuracy measures improve monotonically with improving (decreasing) perplexity. Extrapolating these results to the much larger Epinions dataset, for which we have no ground truth, we argue that the perplexity results reported in our quantitative evaluation provide a good indication of the relative accuracy of aspect identification and rating prediction for the compared models.

7.3.1 Comparing Preprocessing Techniques

We compare the perplexity of models learning from opinion phrases for different preprocessing techniques. As the baselines we use pairs of one frequent noun and one adjective [21] and phrases generated from POS patterns [29].

Which preprocessing technique for extracting opinion phrases works best? To answer this question, Table 5 presents the average perplexity of different models using different preprocessing techniques. The results indicate that using POS patterns for extracting opinion phrases is more effective than the frequent noun technique. However, it is unclear whether this better performance is due the infrequency of some of the opinion phrases or because of the inaccuracy of extracted frequent phrases. Table 5 also demonstrates that our proposed technique based on dependency parsing clearly and consistently outperforms the other preprocessing techniques for all subsets of products. This confirms our hypothesis that exploiting the semantic relationship between words pays off for extracting opinion phrases.

7.3.2 Evaluation of LDA Models

In this section we compare the performance of the discussed models to answer the key design questions stated in the introduction. Figure 7(a) shows the perplexity of all models for different subsets of products. The perplexity of the models learning from opinion phrases are given for preprocessing using dependency patterns, which performs best according to Table 5. Figure 7 compares only products with at least 10 reviews, similar to the literature. At the end of this section, we also discuss the models' performance for products with less than 10 reviews.

Is it better to have separate latent variables for aspects and ratings? To answer this question, we compare the performance of LDA vs. S-LDA (Figure 7(b)), and also PLDA vs. S-PLDA (Figure 7(c)). It turns out that having separate latent variables for aspects and ratings cannot improve the performance of a model having only one observed variable (i.e., using the bag-of-words model). As a result, the performance of LDA and S-LDA are almost the same. However, S-PLDA outperforms PLDA thanks to the separate aspect and rating variables generating head terms and modifiers, respectively.

Is it better to assume dependency between ratings and aspects? Here we compare the perplexity of S-LDA vs. D-LDA (Figure 7(d)) and also S-PLDA vs. D-PLDA (Figure 7(e)). The higher perplexity of D-LDA compared to S-LDA demonstrates that increasing model complexity while keeping the observed data fixed decreases the performance of the model. However, the comparison of D-PLDA and S-PLDA indicates that assuming the dependency between ratings and aspects improves the performance of a model that generates opinion phrases.

Is it better to learn from bag-of-words or preprocess the reviews and learn from opinion phrases? To address this question, we compare three pairs of models: LDA vs. PLDA (Figure 7(f)), S-LDA vs. S-PLDA (Figures 7(g)), and D-LDA vs. D-PLDA (Figure 7(h)). The only difference between these pairs of models is generation of words vs. generation of opinion phrases. Except for products with more

Table 5: Perplexity of the LDA models using different preprocessing techniques

Preprocessing	Frequent nouns			POS patterns			Dependency patterns		
Subset of Products	PLDA	S-PLDA	D-PLDA	PLDA	S-PLDA	D-PLDA	PLDA	S-PLDA	D-PLDA
$1 < \#Rev. <= 10$	11834.99	11824.35	11740.95	7735.32	7679.10	7650.67	5463.80	5422.24	5413.66
$10 < \#Rev. <= 50$	6724.61	6129.57	5829.33	4573.20	4124.06	3847.92	3438.19	2937.36	1975.35
$50 < \#Rev. <= 100$	3024.59	2243.22	1882.91	2734.58	1912.88	1381.86	1998.42	1514.71	592.40
$100 < \#Rev. <= 200$	1885.52	1511.47	660.96	1753.39	1328.77	411.94	1481.00	1038.25	164.19
$200 < \#Rev.$	1337.93	1165.88	406.27	1301.00	1066.89	353.73	1284.19	879.29	142.37

than 200 reviews, LDA performs better than PLDA, showing that extracting opinion phrases does not help if only one latent variable is used to generate both head terms and modifiers. Comparing S-LDA and S-PLDA, we observe that for products with more than 50 reviews S-PLDA achieves a lower perplexity (better performance), while for products with fewer reviews it performs poorer. Finally, D-PLDA performs significantly better than D-LDA in all of the product subsets, showing the advantage of learning from opinion phrases rather than bag-of-words when modeling aspects, ratings, and their dependency.

Does the answer to the above questions differ for products with few reviews and products with many reviews? Here, we discuss the performance of the LDA models for products with less than 10 reviews. Since products of this subset have only 3 reviews on average (see Table 2), learning a proper model is very difficult. As Table 6 shows, the basic LDA model outperforms the more complex models for these products, demonstrating that neither preprocessing nor model complexity can improve the performance of LDA. Thus, the best design choices are very different for products with many reviews and for those with few reviews.

Table 6: Perplexity of products with $\#Rev. <= 10$

LDA	S-LDA	D-LDA	PLDA	S-PLDA	D-PLDA
4413.6	4567.2	4729.3	5463.8	5422.2	5413.6

8. CONCLUSION

Aspect-based opinion mining, which aims to extract aspects and their corresponding ratings from online product reviews, provides very useful information for users to make their purchase decision. LDA-based models are considered to be state-of-the-art for aspect-based opinion mining. Realizing that there is no "one-size-fits-all" model that always outperforms all other models, in this paper we developed a set of design guidelines. We conducted extensive experiments on a very large real life dataset from Epinions.com (500K reviews) and compared the performance of different models in terms of the likelihood of the held-out test set. Based on our experimental results, we formulate the following guidelines for the design of LDA models for aspect-based opinion mining:

- When learning from bag-of-words, having separate latent variables for aspects and ratings cannot improve the performance of a model. However, when learning from opinion phrases, it does help to consider two latent variables for generating head terms and modifiers.

- When learning from opinion phrases and having separate latent variables for aspect and rating, assuming their dependency improves the performance.

- When separate latent variables are assumed for aspects and ratings, using preprocessing techniques can improve the performance.

- Using dependency patterns consistently achieves the best performance for extracting opinion phrases.

- For products with few reviews, the basic LDA model outperforms the more complex models. For products with many reviews, the model learning aspects and ratings from opinion phrases with dependency assumption performs best.

This paper suggests several directions for future research, such as investigating different factors for improving the model performance for products with only few reviews. Another direction is exploring the impact of different additional input sources (e.g. review's overall rating) on the performance of the models.

9. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 2003.

[2] S. Brody and N. Elhadad. An unsupervised aspect-sentiment model for online reviews. In *HLT '10.*

[3] Y. Choi and C. Cardie. Hierarchical sequential learning for extracting opinions and their attributes. In *ACL '10.*

[4] X. Ding, B. Liu, and P. S. Yu. A holistic lexicon-based approach to opinion mining. In *WSDM '08.*

[5] H. Guo, H. Zhu, Z. Guo, and Z. Su. Domain customization for aspect-oriented opinion analysis with multi-level latent sentiment clues. In *CIKM '11.*

[6] H. Guo, H. Zhu, Z. Guo, X. Zhang, and Z. Su. Product feature categorization with multilevel latent semantic association. In *CIKM '09.*

[7] Y. He, C. Lin, and H. Alani. Automatically extracting polarity-bearing topics for cross-domain sentiment classification. In *HLT '11.*

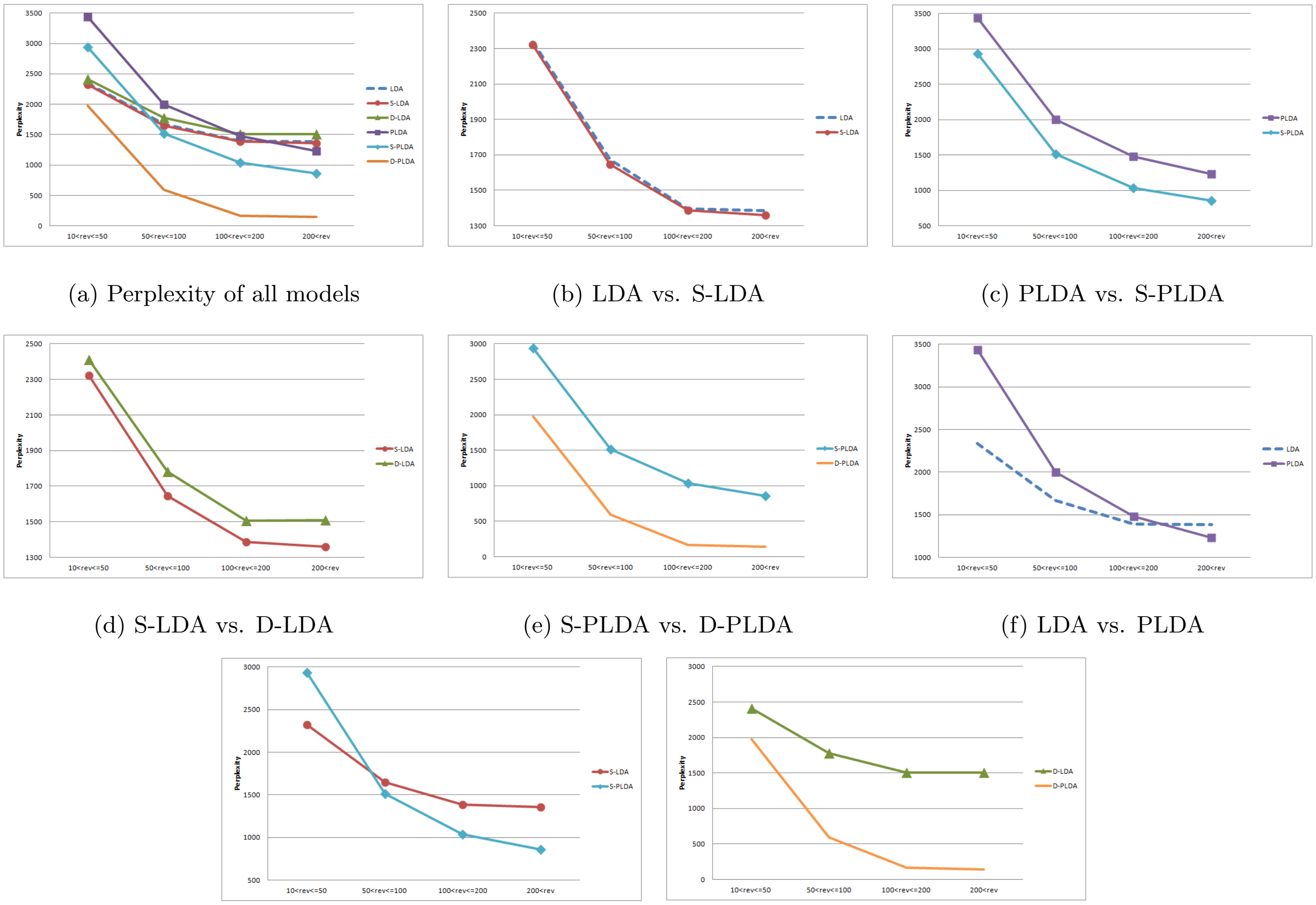
[8] M. Hu and B. Liu. Mining and summarizing customer reviews. In *KDD '04.*

[9] M. Hu and B. Liu. Mining opinion features in customer reviews. In *AAAI'04.*

[10] W. Jin, H. H. Ho, and R. K. Srihari. Opinionminer: a novel machine learning system for web opinion mining and extraction. In *KDD '09.*

[11] Y. Jo and A. H. Oh. Aspect and sentiment unification model for online review analysis. In *WSDM '11.*

[12] H. Lakkaraju, C. Bhattacharyya, I. Bhattacharya, and S. Merugu. Exploiting coherence for the simultaneous discovery of latent facets and associated sentiments. In *SDM '11.*



(a) Perplexity of all models

(b) LDA vs. S-LDA

(c) PLDA vs. S-PLDA

(d) S-LDA vs. D-LDA

(e) S-PLDA vs. D-PLDA

(f) LDA vs. PLDA

(g) S-LDA vs. S-PLDA

(h) D-LDA vs. D-PLDA

Figure 7: Perplexity comparisons for different subsets of products

[13] F. Li, C. Han, M. Huang, X. Zhu, Y.-J. Xia, S. Zhang, and H. Yu. Structure-aware review mining and summarization. In *COLING '10*.

[14] F. Li, M. Huang, and X. Zhu. Sentiment analysis with global topics and local dependency. In *AAAI '10*.

[15] C. Lin and Y. He. Joint sentiment/topic model for sentiment analysis. In *CIKM '09*.

[16] B. Liu. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, chapter 11. Springer, 2007.

[17] B. Liu, M. Hu, and J. Cheng. Opinion observer: analyzing and comparing opinions on the web. In *WWW '05*.

[18] Y. Lu, P. Tsaparas, A. Ntoulas, and L. Polanyi. Exploiting social context for review quality prediction. In *WWW '10*.

[19] Y. Lu, C. Zhai, and N. Sundaresan. Rated aspect summarization of short comments. In *WWW '09*.

[20] B. M. Marie-Catherine de Marneffe and C. D. Manning. Generating typed dependency parses from phrase structure parses. In *LREC '06*.

[21] S. Moghaddam and M. Ester. ILDA: interdependent LDA model for learning latent aspects and their ratings from online product reviews. In *SIGIR '11*.

[22] S. Moghaddam and M. Ester. Opinion Digger: an unsupervised opinion miner from unstructured product reviews. In *CIKM '10*.

[23] S. Moghaddam, M. Jamali, and M. Ester. ETF: Extended Tensor Factorization model for personalizing prediction of review helpfulness. In *WSDM '12*.

[24] M. F. Porter. An algorithm for suffix stripping. *Readings in information retrieval*, 1997.

[25] I. Titov and R. McDonald. A joint model of text and aspect ratings for sentiment summarization. In *ACL-HLT '08*.

[26] I. Titov and R. McDonald. Modeling online reviews with multi-grain topic models. In *WWW '08*.

[27] H. Wang, Y. Lu, and C. Zhai. Latent aspect rating analysis without aspect keyword supervision. In *KDD '11*.

[28] T.-L. Wong, L. Bing, and W. Lam. Normalizing web product attributes and discovering domain ontology with minimal effort. In *WSDM '11*.

[29] T.-J. Zhan and C.-H. Li. Semantic dependent word pairs generative model for fine-grained product feature mining. In *PAKDD '11*.

[30] W. X. Zhao, J. Jiang, H. Yan, and X. Li. Jointly modeling aspects and opinions with a maxent-lda hybrid. In *EMNLP '10*.