Multi-aspect Sentiment Analysis with Topic Models

Bin Lu*†‡, Myle Ott†§, Claire Cardie† and Benjamin Tsou*‡
*Department of Chinese, Translation and Linguistics, City University of Hong Kong, Hong Kong
†Department of Computer Science, Cornell University, Ithaca, NY, USA
‡Research Centre on Linguistics and Language Information Sciences, Hong Kong Institute of Education, Hong Kong
lubin2010@gmail.com, {myleott,cardie}@cs.cornell.edu, btsou99@gmail.com

***Abstract*—We investigate the efficacy of topic model based approaches to two multi-aspect sentiment analysis tasks: *multi-aspect sentence labeling* and *multi-aspect rating prediction*. For sentence labeling, we propose a weakly-supervised approach that utilizes only minimal prior knowledge—in the form of seed words—to enforce a direct correspondence between topics and aspects. This correspondence is used to label sentences with performance that approaches a fully supervised baseline. For multi-aspect rating prediction, we find that overall ratings can be used in conjunction with our sentence labelings to achieve reasonable performance compared to a fully supervised baseline. When gold-standard aspect-ratings are available, we find that topic model based features can be used to improve unsophisticated supervised baseline performance, in agreement with previous multi-aspect rating prediction work. This improvement is diminished, however, when topic model features are paired with a more competitive supervised baseline—a finding not acknowledged in previous work.**

***Keywords*-multi-aspect sentiment analysis; topic modeling;**

I. Introduction

The ever-increasing popularity of websites that feature user-generated opinions (e.g., TripAdvisor.com and Yelp.com) has led to an abundance of customer reviews that are often too numerous for a user to read. Consequently, there is a growing need for systems that are able to automatically extract, evaluate and present opinions in ways that are both helpful and easy for a user to interpret.

Early approaches to this problem [1]–[4] have focused on determining either the overall polarity (i.e., positive or negative) or the sentiment rating (e.g., one-to-five stars) of a review. However, only considering coarse overall ratings fails to adequately represent the multiple potential dimensions on which an entity can be reviewed. For example, while the following review from OpenTable.com might express an overall sentiment rating of 3-stars, it additionally expresses a positive opinion toward the restaurant's food, as well as negative opinions toward the restaurant's ambiance and service:

> "The food was very good, but it took over half an hour to be seated, ... and the service was terrible. The room was very noisy and cold wind blew in from a curtain next to our table. Desserts were very good, but because of [the] poor service, I'm not sure we'll ever go back!"

Looking beyond just overall ratings is important for users, too, because they are likely to differ in how much value they ascribe to each of these distinct aspects. For example, while a gourmand may forgive a restaurant's poor ambiance, they may be uncompromising when it comes to food quality. Accordingly, a new branch of sentiment analysis has emerged, called MULTI-ASPECT SENTIMENT ANALYSIS, that aims to take into account these various, potentially related aspects often discussed within a single review.

Recently, several topic modeling approaches based on Latent Dirichlet Allocation (LDA) [5] have been proposed for multi-aspect sentiment analysis tasks [6]–[8]. These approaches use variations of LDA to uncover latent topics in a document collection, with the hopes that these topics will correspond to rateable aspects for the entity under review.

In this work, we investigate the role of several unsupervised and weakly supervised topic modeling approaches to two popular multi-aspect sentiment analysis tasks: (1) MULTI-ASPECT SENTENCE LABELING, where each sentence in a review is labeled according to the aspects it discusses (see Section III-A), and (2) MULTI-ASPECT RATING PREDICTION, the goal of which is to predict implicit aspect-specific star ratings for each review (see Section III-B).

For *multi-aspect sentence labeling*, we propose a weakly supervised topic modeling approach (see Section III-A1) that uses minimal prior knowledge in the form of seed words to encourage a correspondence between topics and ratable aspects. We find that these models generally perform quite well (see Section VI-A), and that the best of these models performs comparably to a supervised approach.

For *multi-aspect rating prediction*, we consider two settings. In the first, we assume that aspect-ratings are unavailable, but find (in Section VI-B) that by leveraging overall ratings in conjunction with our *multi-aspect sentence labeling* approach, we can produce significant improvements over an aspect-blind baseline. In our second setting, we use gold-standard aspect-ratings to train supervised classifiers both with and without topic model based features. We find (in Section VI-C) that these additional features improve performance over an online supervised baseline (Perceptron Rank). However, this improvement is diminished when a more competitive supervised baseline is used instead (Support-Vector Regression)—a finding not previously acknowledged.

For both tasks, we examine and compare four types of topic models (see Section IV): LDA, Local LDA [6], Multi-Grain LDA [7], and Segmented Topic Models (STM)—a recently proposed [9] topic model that, to date, has not been applied to sentiment analysis tasks.

Lastly, we perform our experiments using three datasets (see Section V-A) from two domains (hotel and restaurant reviews). Specifically, we evaluate our data coming from CitySearch, OpenTable, and TripAdvisor.

II. Related Work

While sentiment analysis has been studied extensively for some time [10], most approaches have focused on document-level overall sentiment. Recently, there has been a growing interest in sentiment analysis at finer levels of granularity, and specifically approaches that take into account the multi-aspect nature of many sentiment analysis tasks.

A. Multi-aspect Sentiment Analysis

Early multi-aspect work focused on creating aspect-based review summaries using mined product features [11]–[13]. More recent work [14], [15] has also began modeling implicit aspects. For example, [16] develop an aspect-based review summarization system that extracts and aggregates aspects and their corresponding sentiments.

Recent work has also began to look at multi-aspect rating prediction. [17] present the Good Grief algorithm, which jointly learns ranking models for individual aspects using an online Perceptron Rank (PRank) [18] algorithm. [19] and [20] bootstrap aspect terms with seed words for unsupervised multi-aspect opinion polling and probabilistic rating regression, respectively. [21] integrate a document-level HMM model to improve both multi-aspect rating prediction and aspect-based sentiment summarization.

B. Multi-aspect Topic Models

While early generative approaches to sentimenent analysis tasks focused only on latent topics [22]–[24], recently work has begun to additionally model multiple aspects present in a single document. For example, [7] present Multi-grain LDA (MG-LDA), in which review-specific elements and ratable aspects are modeled by global and local topics, respectively. [6] introduce Local-LDA, a sentence-level LDA that discovers ratable aspects in reviews. [8] present MaxEnt-LDA, a maximum entropy hybrid model that discovers both aspects and aspect-specific opinion words.

However, the mapping between topics and aspects in these models is still largely implicit, which can be burdensome when working with different parameterizations or datasets. [25] integrate ground-truth aspect ratings into MG-LDA to force topics to correlate directly with aspects. However, their approach requires gold-standard aspect ratings. In contrast, in this work we both consider settings in which aspect ratings are available (see Section III-B), and settings in which they are unavailable (see Section III-A).

III. Multi-aspect Sentiment Analysis Tasks

A. Multi-aspect Sentence Labeling

The first phase of multi-aspect sentiment analysis is aspect identification and mention extraction. This step identifies the relevant aspects for a rated entity and extracts all textual mentions associated with those aspects [25].

In this work, we consider a limited version of the aspect identification and mention extraction task, which we call *multi-aspect sentence labeling*. In our limited setting, we assume that aspects are fixed—e.g., food, service, and ambiance for restaurant reviews—and that it is sufficient to identify a single aspect for each sentence in a document.

In particular, we evaluate 4 topic models, weakly supervised with aspect-specific seed words (see Section III-A1), and label each sentence according to its latent topic distribution. Formally, for each sentence s and topic k, we calculate the probability, p_k^s, of words in s assigned to k, averaged over n samples, and use $\arg\max_k p_k^s$ as the label for s.

1) Weak Supervision with Minimal Prior Knowledge: To encourage topic models to learn latent topics that correlate directly with aspects, we augment them with a weak supervised signal in the form of aspect-specific seed words. Rather than directly using the seed words to do bootstrapping, as in [19] and [20], we use them to define an asymmetric prior on the word-topic distributions. This approach guides the latent topic learning towards more coherent aspect-specific topics, while also allowing us to utilize large-scale unlabeled data.

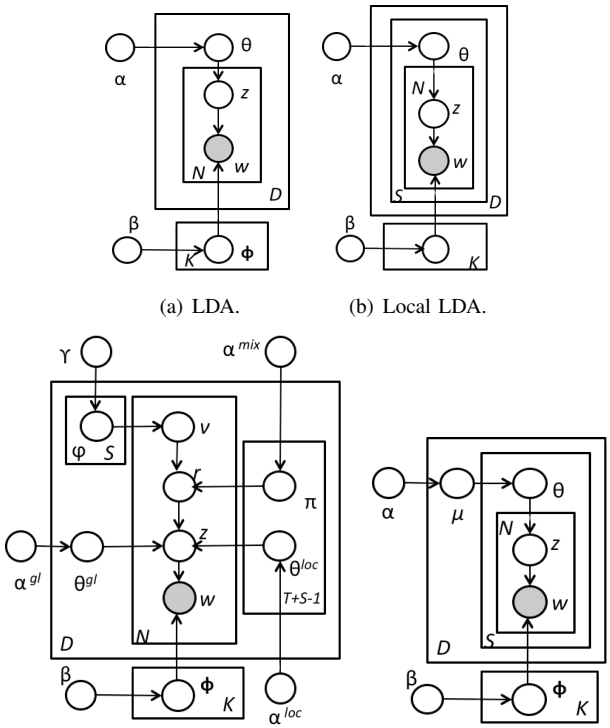
For example, we define our prior knowledge (seed words) for the original LDA model as a conjugate Dirichlet prior to the multinomial word-topic distributions ϕ. By integrating with the symmetric smoothing prior β, we define a combined conjugate prior for each seed word w in $\phi \sim Dir\left(\{\beta + C_w\}_{w \in V}\right)$, where C_w can be interpreted as an equivalent sample size—i.e., the impact of our asymmetric prior is equivalent to adding C_w pseudo counts to the sufficient statistics of the topic to which w belongs. When we do not have prior knowledge for a word w, we set $C_w = 0$.

B. Multi-aspect Rating Prediction

The second phase of multi-aspect sentiment analysis is *multi-aspect rating prediction* [7], [17], [20], [21]—in which each aspect of a document is assigned polar (i.e., positive, negative, neutral), numeric, or "star" (i.e., 1-5) ratings.

Specifically, we consider two settings: (1) multi-aspect rating prediction with indirect supervision, and (2) supervised multi-aspect rating prediction. In (1), aspect ratings are predicted based only on the text and overall rating of each review. Specifically, we train a regression model on the given overall ratings and, for each aspect, apply the model to the corresponding aspect-labeled sentences (see Section III-A).

In (2), the supervised multi-aspect rating prediction setting, we augment and compare standard supervised regression learners with features derived from unsupervised topic



(a) LDA. (b) Local LDA.

(c) MG-LDA. (d) STM.

Figure 1. Plate notations for topic models described in Section IV.

models (without seed words). Following [7], we create features based on the output of each topic model by concatenating standard n-gram features with their associated sentence-level topic assignments, and then evaluate supervised classifiers trained on those features.

IV. Topic Models

In their most basic form, topic models exploit word co-occurrence information to capture latent topics in a corpus. Approaches to both tasks described in Section III use these latent topics to model multiple aspects within a document, however the quality of these topics varies depending on the topic model used. In this work we consider 4 topic models, described here. Graphical representations for each of these models appear in Figure 1, in plate notation.

1) LDA and Local LDA: The first two topic models that we consider are based on Latent Dirichlet Allocation (LDA) [5]. LDA is a probabilistic generative model in which documents are represented as mixtures over latent topics. Formally, LDA assumes that a corpus is generated according to the following generative story line:

- For each topic k:
  - Choose word-topic mixture: $\phi_k \sim Dir(\beta)$
- For each document d:
  - Choose document topic proportions: $\theta_d \sim Dir(\alpha)$
  - For each word w in document d:
    - Choose topic: $z_{d,w} \sim \theta_d$
    - Choose word: $w \sim \phi_{z_{d,w}}$

While LDA can effectively model word co-occurrence at the document level, [6] argue that review aspects are more likely to be discovered from sentence-level word co-occurrence information. They propose Local LDA, in which sentences are modeled as documents are in standard LDA.

2) Multi-grain LDA: In response to limitations of standard LDA for multi-aspect work, [7] propose Multi-Grain LDA (MG-LDA). MG-LDA jointly models document-specific themes (global topics), and themes that are common throughout the corpus intended to correspond to ratable aspects, called local topics. Additionally, while the distribution over global topics is fixed for a given document (review), local topic proportions are varied across the document according to sentence-level sliding windows. Formally, each document d is generated as follows:

- Choose global topic proportions: $\theta^{gl} \sim Dir(\alpha^{gl})$
- For each sliding window v of size T:
  - Choose local topic proportions: $\theta^{loc}_{d,v} \sim Dir(\alpha^{loc})$
  - Choose granularity mixture: $\pi_{d,v} \sim Beta(\alpha^{mix})$
- For each sentence s:
  - Choose window proportions: $\psi_{d,s} \sim Dir(\gamma)$
- For each word w in sentence s of document d:
  - Choose sliding window: $v_{d,w} \sim \psi_{d,s}$
  - Choose granularity: $r_{d,w} \sim \pi_{d,v_{d,w}}$
  - Choose topic: $z_{d,w} \sim \{\theta^{gl}, \theta^{loc}_{d,v}\}_{r_{d,w}}$
  - Choose word: $w \sim \phi^{r_{d,w}}_{z_{d,w}}$

When $T = 1$, MG-LDA generalizes to a combination of standard and Local LDA, where α^{mix} regulates the tradeoff between document- and sentence-level topic proportions.

3) Segmented Topic Model: Lastly, we introduce the Segmented Topic Model (STM) [9], which jointly models document- and sentence-level topic proportions using a two-parameter Poisson Dirichlet Process (PDP). Documents d are generated as follows:

- Choose document topic proportions: $\theta_d \sim Dir(\alpha)$
- For each sentence s:
  - Choose topic proportions: $\theta_s \sim PDP(\theta_d, a, b)$
- For each word w in sentence s:
  - Choose topic: $z_{d,w} \sim \theta_s$
  - Choose word: $w \sim \phi_{z_{d,w}}$

STM can be considered an extension of Local LDA that additionally considers document-level topic distributions induced from the individual sentence-level topic distributions.

4) Inference: While exact inference for the models just presented is largely intractable [5], approximate techniques such as variational inference or Gibbs sampling can be used instead. Following [26], we use a collapsed Gibbs sampling

approach for inference.[1] The exact sampling algorithms are excluded for brevity. We instead refer the reader to [26] for the LDA and Local LDA sampler, [7] for the MG-LDA sampler, and [9] for the STM sampler.

V. Experimental Setup

A. Dataset and Preprocessing

Tasks and models discussed in Section III and Section IV are evaluated on three datasets. The first dataset contains 73,495 reviews and their associated overall, food, service, and ambiance aspect ratings for all restaurants in the New York/Tri-State area appearing on OpenTable.com, and is used for our multi-aspect rating prediction task. After excluding reviews that were too short (< 50 words) or too long (> 300 words), we were left with 29,596 reviews.

Since the OpenTable dataset does not contain gold-standard labeled sentences, we evaluate our multi-aspect sentence labeling performance on a second, annotated dataset, of 652 restaurant reviews from CitySearch.com, introduced by [27]. Each sentence in this corpus has been manually labeled with one or more of the following six aspects: food, service, ambiance, price, anecdotes, or miscellaneous.

Finally, we evaluate multi-aspect rating prediction on [20]'s TripAdvisor hotel review corpus. For each review, this corpus contains an associated overall rating, as well as ratings for 7 aspects: value, room, location, cleanliness, check-in/front desk, service, and business services. After removing reviews missing any of the first 6 aspect-ratings, and (as before) excluded reviews that were too short or too long, we were left with 66,512 reviews.

Datasets were tokenized and sentence split using the Stanford POS Tagger [28]. For topic models, we removed singleton words, and stop words not appearing in the sentiment lexicon introduced by [29].

B. Supervised Classifiers for Multi-aspect Rating Prediction

We consider two supervised machine-learning approaches to multi-aspect rating prediction. The first is linear ϵ-Support Vector Regression (SVR) [30]. We use the LIBSVM toolkit [31] with default parameters.[2] The second is Perceptron Ranking (PRank) [18], an online ordinal regression classifier that has been used in related work [7], [17], [19], [32]. We use the implementation by [17].[3]

Classifiers are trained on unit-normalized binary unigram[4] presence features. We also experimented with raw and normalized frequency counts and raw binary features, but found that normalized binary features work best.

[1]In this work we sample all models for 1,000 iterations, with a 500-iteration *burn-in* and a *sampling-lag* of 10.

[2]Pilot experiments suggest that these values give near-optimal performance compared to parameters fully tuned by grid-search.

[3]http://people.csail.mit.edu/bsnyder/naacl07/

[4]While bigram and trigram features can be considered, unigram features better highlight differences between competing topic models.

Table I
Seed words for restaurant reviews.

Aspect	Seed Words
food	food, chicken, beef, steak
service	service, staff, waiter, reservation
ambiance	ambiance, atmosphere, room, experience
price	price, value, quality, worth

Finally, pilot experiments suggest that the optimal number of iterations for PRank is data-dependent, and can heavily influence performance. Consequently, except where specified, the number of iterations for PRank is always tuned via nested cross validation on the training set.

C. Topic Model Hyperparameters

Unless otherwise stated, topic model hyperparameters are assigned the following values: α: 0.5 for STM, 0.1 for LDA, Local LDA and MG-LDA (including α^{gl}, α^{loc} and α^{mix}); β: 0.1 for all models; the window size v for MG-LDA is 3; and a and b for STM are 0.1 and 1, respectively. The values for LDA and MG-LDA follow [7], [25], and those for Local LDA follow [6]. Some experimentation was performed with different hyperparameter choices, but downstream performance was not significantly affected.

VI. Results and Discussion

A. Multi-aspect Sentence Labeling

Topic models were weakly supervised using seed words in Table I. The pseudo count C_w for seed words was heuristically set to be 3000 ($\sim$1% of the number of reviews), although we show in Section VI-A1 that performance is robust to variations of this parameter. Assuming that the majority of sentences are aspect-related, we set the number of topics K to 5, thereby allowing a single "background" topic.[5] We also tried other topic numbers in the range of [5-30] with a step of 5, with performance decreasing with increasing K, in most cases.[6]

For evaluation, we used all 1,490 singly-labeled sentences from the annotated portion of the CitySearch corpus for the three main aspects (food, service, and ambiance), following [6] and [8]. Because LDA, MG-LDA and STM are document-level models, inference is performed on all 652 documents, and then performance is evaluated on the 1,490-sentence subset. Note that none of the OpenTable data is labeled with respect to sentence-level aspects.

Results are given in term of precision (P), recall (R), and F-1 score in Table II. The majority baseline labels all sentences according to the most common aspect label, food.

[5]Note that the number of global topics for MG-LDA was set to 10.

[6]While we restrict ourselves to only one set of seed words for each aspect, it is also possible to enlarge the topic number K by providing more than one set of seed words for the major aspects, such as food for the restaurant domain, to reflect the reality that there could be many subtopics of major aspects, such as the subtopics drink, bakery and main dishes shown in [6]. However, that strategy would involve more fine-tuning of seed words for each subtopic, and is therefore left to future work.

Table II
MULTI-ASPECT SENTENCE LABELING RESULTS.

	Accuracy	**Food**			**Service**			**Ambiance**		
		P	R	$F1$	P	R	$F1$	P	R	$F1$
Majority	0.595	0.595	**1**	0.746	0	0	0	0	0	0
LDA	0.477	0.646	0.554	0.597	0.469	0.494	0.481	0.126	0.179	0.148
MG-LDA	0.760	0.888	0.772	0.826	0.637	0.648	0.642	0.609	0.876	0.719
STM	0.794	0.954	**0.776**	0.856	0.674	0.759	0.714	**0.611**	**0.908**	**0.731**
Local LDA	**0.803**	**0.969**	0.775	**0.861**	**0.731**	**0.810**	**0.768**	0.573	0.892	0.698
SVM	**0.830**	0.814	**0.975**	**0.887**	**0.874**	0.670	0.759	**0.860**	0.538	0.662

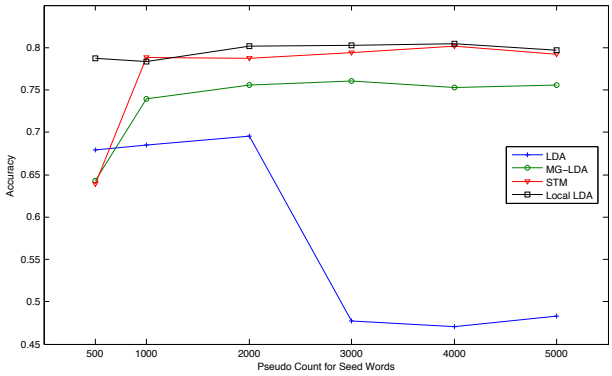


Figure 2. Influence of pseudo counts.

As an upper bound, we also test a fully supervised SVM classifier on the labeled data with 5-fold cross-validation.

We can see that weakly supervised topic models achieve good performance on this task, and at best are comparable to the supervised SVM classifier, confirming that adding prior knowledge can encourage latent topics to correlate directly with aspects. Among the topic models themselves, Local LDA gives the highest accuracy and is also the best at labeling food and service aspects; STM achieves similar results and is the best performing topic model for the ambiance aspect, followed by MG-LDA and LDA.

These results can be explained as follows. Since most sentences usually focus on just one or two aspects, sentence-level word co-occurrence information is more appropriate than document-level co-occurrences for studying aspects. Indeed, while a review may talk about several aspects simultaneously, the document-level word co-occurrence may not be able to well distinguish the individual aspects from each other. Through directly modeling the word co-occurrences within sentences, Local LDA better captures aspect information, while standard LDA fails to differentiate between words in different aspects, even given seed words.

While both STM and MG-LDA simultaneously model document- and sentence-level word co-occurrences, the former indirectly models document-level co-occurrences via sentence-level co-occurrences and a PDP prior. The latter, MG-LDA, models both document- and sentence-level co-occurrences directly, which may therefore consider some aspects to be global topics, when they are in fact specic to a type of restaurants, as mentioned in [7].

Table III
ENTITY-LEVEL MULTI-ASPECT RATING PREDICTION RESULTS FOR TRIPADVISOR DATA.

	L_1 **error**	$\rho_{\mathbf{aspect}}$	$\rho_{\mathbf{review}}$	$MAP@10$
SVR_Ovr	0.311	0	0.800	**0.429**
LDA	0.645	−0.149	0.454	0.143
MG-LDA	**0.400**	**0.407**	0.622	0.129
STM	0.517	0.218	0.694	**0.286**
Local LDA	0.433	0.335	**0.729**	0.229
SVR	**0.238**	**0.715**	**0.846**	0.400

1) Influence of Pseudo Counts: We also examine the influence of the seed-word pseudo-count parameter, C_w, with results shown in Figure 2. We observe that performance is reasonable across a variety of values of C_w, and the relative ordering between models is stable. Notably, there is a dramatic drop in performance for LDA at $C_w = 3,000$. By looking at the corresponding LDA topics, we found that with large C_w, LDA separates the food aspect into two topics, one focusing on main dishes (due to the seed words for food) and the other focusing on dessert. This dramatically decreases overall performance, since only a single label is assigned to each sentence.

B. Multi-aspect Rating Prediction with Indirect Supervision

For multi-aspect rating prediction with indirect supervision, we assume that we only have access to overall ratings in the training data, and no gold-standard aspect ratings. We label sentences with aspects using weakly supervised topic models on both the OpenTable and TripAdvisor datasets (see Section III-B). Seed words for TripAdvisor come from [20]. For TripAdvisor, we also set $C_w = 6,000$, and use $K = 8$ topics (with 15 global topics for MG-LDA).

Because not all aspects are discussed in every review, we chose to combine all reviews for each entity (hotel or restaurant) into a single "super"-review. Ground-truth aspect ratings are obtained by averaging the overall/aspect ratings for each "super"-review. After excluding "super"-reviews containing fewer than 10 reviews, we were left with 913 restaurants and 1,604 hotels.

We then predict aspect ratings based on the aspect-labeled sentences by using a support vector regression (SVR) model trained on all combined vectors for each kind of entity (hotel or restaurant) and their overall ratings. The baseline approach always uses the predicted overall rating as aspect ratings for each entity, called SVR_Ovr. As an upper bound,

Table IV
MULTI-ASPECT RATING PREDICTION RESULTS FOR RESTAURANT DATA.

Restaurant	Food	Service	Ambiance
Greek Taverna - Glen Rock	4.19 (3.9)	3.31 (3.2)	3.9 (3.6)
Milonga Wine and Tapas	4.0 (4.1)	3.54 (3.1)	3.97 (3.7)
Equus Tavern	3.87 (3.8)	3.97 (4.1)	3.83 (3.6)

we also test a fully supervised SVR model (SVR) trained with ground-truth aspect ratings. For both SVR_Ovr and SVR, we use 5-fold cross validation.

In addition to L_1 error (absolute difference) [21], we use three other metrics from [20]. The first metric is MAP@10, which measures how well the predicted ratings keep the top entities on the top. The other two metrics are ρ_{aspect} and $\rho_{reviews}$, which are two averaged Pearson correlations between the predicted and the ground-truth ratings for all aspects within each review, and for each aspect across all entities. The former assesses whether the predicted ratings give the correct preference order over the different aspects within each review, e.g., the reviewer likes food more than service. The latter measures how well the predicted aspect ratings rank entities for each aspect, in order to answer questions such as "which restaurant has the best food."

Due to space constraints, we only show averaged results over all aspects for the hotel dataset in Table III. We observe that, with the exception of LDA for ρ_{aspect}, topic models provide positive correlations. MG-LDA and Local LDA show a medium correlation (larger than 0.3) with the gold standard on ρ_{aspect}, which means that even without access to the ground-truth aspect ratings, we can still reasonably predict the relative preference order over aspects by using the aspect-labeled sentences given by the weakly supervised topic models. Although STM and MG-LDA perform well on some metrics, Local LDA is always among the top two in terms of all metrics among topic models. LDA performs the worst. Not surprisingly, SVR performs the best in terms of three metrics with access to the ground-truth aspect ratings, while SVR-Ovr does quite well in three metrics, but cannot provide information on ρ_{aspect}.

For qualitative evaluation, we select 3 restaurants with the same overall rating of 3.7 (on average) but different aspect ratings, and compare the predicted ratings given by Local LDA. The prediction results are shown in Table IV with ground-truth ratings in parentheses. We observe that although all three restaurants have the same overall rating, the aspect ratings are quite different: Greek Taverna - Glen Rock and Milonga Wine and Tapas has higher ratings for food, and Equus Tavern has better service. This kind of detailed aspect information is important for users who have different aspect preferences.

C. Supervised Multi-aspect Rating Prediction

We also evaluate multi-aspect rating prediction for each classifier introduced in Section V-B, trained with and without features derived from topic models introduced in Section IV, in addition to baseline n-gram features (unigrams). Topic models trained in this section do not make use seed words.

Topic model features are created following [7]. For each sentence s and topic k, we calculate the proportion, p_k^s, of words in s assigned to k, averaged over 50 samples. We then bucket the corpus-wide proportions as evenly as possible into five buckets, such that $b_k^s \in \{1, 2, 3, 4, 5\}$ corresponds to the bucket containing p_k^s. Then, since a sentence will typically contain small proportions of many topics, we limit our consideration to only the top-3 topics per sentence, ordered by p_k^s, which we denote k_1^{s*}, k_2^{s*} and k_3^{s*}. Finally, for each word w in sentence s, we construct three binary features of the form: $\left(w, k_i^{s*}, b_{k_i^{s*}}^s\right)$ for $i \in \{1, 2, 3\}$.

We report 5-fold cross-validated performance for each method on subsets of the OpenTable and TripAdvisor data introduced in Section V-A. For each dataset we select a balanced (according to overall rating) random subset of 5,000 reviews. The remaining reviews are used to train the unsupervised topic models.

Results appear in Table V. Interestingly, we find that the PRank baseline performs worse than the SVR baseline across all aspects and datasets. This is perhaps unsurprising, since PRank was originally proposed for online learning, and is very sensitive to both its parameterization and data ordering. While more experiments are necessary, these results suggest that despite PRank's recent popularity, it is perhaps an ineffective baseline for aspect-rating prediction.

We also observe that adding features derived from topic models can increase performance (albeit slightly) over even a strong (SVR) baseline. However, in contrast to previous work by [7], we find that the choice of topic model makes little difference in this case. Indeed, LDA often outperforms other more complicated models on this supervised task.

D. Further Discussion on Aspect-based Summarization

In addition to the concise aspect-based opinion summary shown in Section VI-B, we can choose sentences from reviews based on their aspects and rating scores to provide aspect-based review summaries for a given entity. Since the aspect label for each sentence has a probability, as mentioned in Section III-A, we set a threshold to filter out unconfident sentences for each aspect (e.g., 0.75). We predict the rating of each sentence by using an SVR model trained on the overall ratings of the 5,000 balanced restaurant reviews mentioned in Section VI-C, and then we select the sentences with the highest and lowest scores for each aspect.

In Table VI, we show a sample aspect-based summary (with ground-truth ratings in parenthesis) generated in this way for Mesa Grill, one of the most popular restaurants on OpenTable.com. We can see that reviewers had different experience or preference. For example, in terms of ambiance, one user thinks that the restaurant is *full of energy*, while another considers it *too cramped*. Such detailed summaries could be helpful to both consumers and service providers.

Table V
SUPERVISED MULTI-ASPECT RATING PREDICTION RESULTS, WITH MODELS RUN TO GENERATE 15 TOPICS (45 GLOBAL TOPICS FOR MG-LDA). RESULTS WERE SIMILAR ACROSS A VARIETY OF TOPIC NUMBER CHOICES.

Learner	Model	OpenTable (L_1 error)				TripAdvisor (L_1 error)						
		Over.	*Food*	*Serv.*	*Amb.*	*Over.*	*Check.*	*Serv.*	*Value*	*Loc.*	*Rooms*	*Clean.*
PRank	*Baseline*	0.798	0.821	1.052	1.071	0.687	0.818	0.856	0.946	0.828	0.932	0.900
	LDA	0.638	0.683	0.806	0.817	0.563	0.640	0.682	0.770	0.668	0.737	0.721
	Local LDA	0.650	0.703	0.815	0.841	0.569	0.657	0.685	0.761	0.680	0.757	0.716
	MG-LDA	0.650	0.707	0.812	0.841	**0.554**	0.656	0.685	0.767	0.672	0.764	0.722
	STM	0.642	0.686	0.812	0.838	0.574	0.647	0.689	0.750	0.679	0.754	0.723
SVR	*Baseline*	0.654	0.700	0.810	0.811	0.585	0.651	0.708	0.737	0.695	0.747	0.725
	LDA	**0.637**	**0.679**	0.790	**0.781**	0.560	**0.628**	**0.667**	0.738	**0.663**	**0.732**	**0.709**
	Local LDA	0.651	0.686	**0.786**	0.804	0.576	0.654	0.688	**0.731**	0.688	0.742	0.729
	MG-LDA	0.656	0.693	0.787	0.804	0.576	0.648	0.681	0.743	0.676	0.744	0.725
	STM	0.650	0.682	0.794	0.794	0.571	0.643	0.686	0.741	0.675	0.741	0.718

Table VI
ASPECT-BASED COMPARATIVE SUMMARY FOR MESA GRILL RESTAURANT.

Aspect	Summary	Rating
Food 3.90 (3.69)	[+] The [food] is delicious, the grits are phenomenal and I love the breads they bring before the meal with the pepper jelly.	4.62
	[–] One entree was not even edible it was overcooked and dry.	0.85
Service 3.53 (3.87)	[+] The staff is professional and friendly.	5.09
	[–] Our server hovered over us until we got our appetizers, trying to push more booze, but then disappeared, and we had to wait for about half an hour between apps and main meal, with no one coming over to check in with us about what was going on.	1.08
Ambiance 3.66 (3.71)	[+] Atmosphere was great; full of energy and a great open bar area.	4.30
	[–] The place was too cramped as you feel like the restaurant management has squeezed too many tables in the seating area.	1.85

VII. CONCLUSION

We investigate the role of unsupervised and weakly supervised topic modeling approaches to multi-aspect sentiment analysis. We show that weakly supervised topic models perform quite well on multi-aspect sentence labeling tasks, and can also be used to aid multi-aspect rating prediction with only indirect supervision. In combination, they can also support interesting applications for aspect-based review summarization. Finally, we find that incorporating features derived from unsupervised topic models provides substantial increases in performance, but only for weak prediction models like PRank. With a stronger model, like SVR, this improvement is diminished.

ACKNOWLEDGMENT

We thank Shuo Chen, Long Jiang, Lillian Lee, Chenhao Tan, Ainur Yessenalina, and Jingbo Zhu, as well as members of the Cornell NLP seminar group and the anonymous reviewers for useful comments and discussion. This work was supported in part by National Science Foundation Grants BCS-0904822, IIS-0968450, IIS-1111176; a gift from Google; the Jack Kent Cooke Foundation; and a Research Grant under No. 149607 from the HKSAR Research Grant Council.

REFERENCES

[1] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*. Association for Computational Linguistics, 2002, pp. 79–86.

[2] B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," in *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 2005, pp. 115–124.

[3] S. Baccianella, A. Esuli, and F. Sebastiani, "Multi-facet rating of product reviews," in *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*. Springer-Verlag, 2009, pp. 461–472.

[4] L. Qu, G. Ifrim, and G. Weikum, "The bag-of-opinions method for review rating prediction from sparse text patterns," in *Proceedings of the 23rd International Conference on Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 913–921.

[5] D. Blei, A. Ng, and M. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993–1022, 2003.

[6] S. Brody and N. Elhadad, "An unsupervised aspect-sentiment model for online reviews," in *Proceedings of ACL:HLT*, 2010, pp. 804–812.

[7] I. Titov and R. McDonald, "Modeling online reviews with multi-grain topic models," in *Proceeding of the 17th international conference on World Wide Web*. ACM, 2008, pp. 111–120.

[8] W. Zhao, J. Jiang, H. Yan, and X. Li, "Jointly modeling aspects and opinions with a maxent-lda hybrid," in *Proceedings*

of the 2010 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, 2010, pp. 56–65.

[9] L. Du, W. Buntine, and H. Jin, "A segmented topic model based on the two-parameter poisson-dirichlet process," *Machine learning*, vol. 81, no. 1, pp. 5–19, 2010.

[10] B. Pang and L. Lee, "Opinion mining and sentiment analysis," *Foundations and Trends in Information Retrieval*, vol. 2, no. 1-2, pp. 1–135, 2008.

[11] M. Hu and B. Liu, "Mining opinion features in customer reviews," in *Proceedings of the National Conference on Artificial Intelligence*. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2004, pp. 755–760.

[12] ——, "Mining and summarizing customer reviews," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 168–177.

[13] A. Popescu and O. Etzioni, "Extracting product features and opinions from reviews," in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2005, pp. 339–346.

[14] L. Zhuang, F. Jing, and X. Zhu, "Movie review mining and summarization," in *Proceedings of the 15th ACM international conference on Information and knowledge management*. ACM, 2006, pp. 43–50.

[15] K. Lerman, S. Blair-Goldensohn, and R. McDonald, "Sentiment summarization: Evaluating and learning user preferences," in *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2009, pp. 514–522.

[16] S. Blair-Goldensohn, K. Hannan, R. McDonald, T. Neylon, G. Reis, and J. Reynar, "Building a sentiment summarizer for local service reviews," in *WWW Workshop on NLP in the Information Explosion Era*, 2008.

[17] B. Snyder and R. Barzilay, "Multiple aspect ranking using the good grief algorithm," in *Proceedings of NAACL HLT*, 2007, pp. 300–307.

[18] K. Crammer and Y. Singer, "Pranking with ranking," in *Proceedings of NIPS*, 2001, pp. 641–647.

[19] J. Zhu, H. Wang, B. Tsou, and M. Zhu, "Multi-aspect opinion polling from textual reviews," in *Proceeding of the 18th ACM Conference on information and Knowledge Management*. ACM, 2009, pp. 1799–1802.

[20] H. Wang, Y. Lu, and C. Zhai, "Latent aspect rating analysis on review text data: a rating regression approach," in *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2010, pp. 783–792.

[21] C. Sauper, A. Haghighi, and R. Barzilay, "Incorporating content structure into text analysis applications," in *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2010, pp. 377–387.

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

[23] C. Lin and Y. He, "Joint sentiment/topic model for sentiment analysis," in *Proceeding of the 18th ACM conference on Information and knowledge management*. ACM, 2009, pp. 375–384.

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

[25] I. Titov and R. McDonald, "A joint model of text and aspect ratings for sentiment summarization," *Urbana*, vol. 51, pp. 308–316, 2008.

[26] T. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. Suppl 1, p. 5228, 2004.

[27] G. Ganu, N. Elhadad, and A. Marian, "Beyond the stars: Improving rating predictions using review text content," in *Proceedings of the 12th International Workshop on the Web and Databases*. Citeseer, 2009.

[28] K. Toutanova, D. Klein, C. Manning, and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," in *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*. Association for Computational Linguistics, 2003, pp. 173–180.

[29] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2005, pp. 347–354.

[30] A. Smola and B. Sch
"olkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

[31] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27:1–27:27, 2011, software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[32] N. Gupta, G. Di Fabbrizio, and P. Haffner, "Capturing the stars: Predicting ratings for service and product reviews," in *Proceedings of the NAACL HLT 2010 Workshop on Semantic Search*. Association for Computational Linguistics, 2010, pp. 36–43.