Limbic: Author-Based Sentiment Aspect Modeling Regularized with Word Embeddings and Discourse Relations

Zhe Zhang
Watson Group
IBM Corporation
Research Triangle Park, NC 27703-9141
zhangzhe@us.ibm.com

Munindar P. Singh
Department of Computer Science
North Carolina State University
Raleigh, NC 27695-8206
singh@ncsu.edu

Abstract

We propose Limbic, an unsupervised probabilistic model that addresses the problem of discovering aspects and sentiments and associating them with authors of opinionated texts. Limbic combines three ideas, incorporating authors, discourse relations, and word embeddings. For discourse relations, Limbic adopts a generative process regularized by a Markov Random Field. To promote words with high semantic similarity into the same topic, Limbic captures semantic regularities from word embeddings via a generalized Pólya Urn process. We demonstrate that Limbic (1) discovers aspects associated with sentiments with high lexical diversity; (2) outperforms state-of-the-art models by a substantial margin in topic cohesion and sentiment classification.

1 Introduction

How can we understand opinionated texts, e.g., social media postings, expressing sentiments about various entities? Three phenomena are key. First, even for similar entities, authors may differ both on aspects and sentiments about those aspects. For example, when reviewing a hotel, Alice may consider aspects such as *Concierge* and *Room*, whereas Bob may consider aspects such as *Nearby* and *Room*. Capturing similarities and differences among authors can help produce recommendations for services that are better aligned with a user's expectations (Wang et al., 2013). Second, reviews exhibit discourse structure, i.e., relations between propositions, which carries valuable information about sentiment. Third, crucial relationships between rare words are lost because each review may be short and use distinct rare words.

Probabilistic topic models (Hofmann, 1999; Blei et al., 2003) provide an unsupervised means to learn latent constructs from texts. Author-specific topic discovery associates texts with their authors (Rosen-Zvi et al., 2004; Kim et al., 2012; Diao and Jiang, 2013) but ignores sentiments. Sentiment analysis methods jointly model aspects and sentiments but exclude either authors (Lazaridou et al., 2013), discourse relations (Mukherjee et al., 2014; Poddar et al., 2017), or both (Jo and Oh, 2011; Lin et al., 2012; Kim et al., 2013).

Word co-occurrence sparsity plagues existing approaches, which model documents as distributions over latent topics and estimate them from word co-occurrence. Since word frequency follows a power law, most words are rare and representative words of a topic rarely co-occur, especially in short opinionated texts, despite semantic proximity. For example, a reviewer would not use both *spotless* and *immaculate* to express a positive sentiment toward the cleanliness of a hotel room. Losing information about word relatedness impedes learning effectiveness, producing topics that are not semantically cohesive.

We contribute Limbic, an unsupervised probabilistic model for discovering author-based aspects and sentiments from opinionated texts that incorporates discourse-level topic modeling and semantic cohesion. (1) It associates authors and sentiment-aspect pairs by generating a mixture over sentiments and aspects for each author. (2) It captures discourse relations by applying a Markov Random Field over Sentiment Expression Units (SEUs), i.e., text elements describing sentiment-aspect pairs. (3) It promotes words with high semantic similarity into the same topic by incorporating semantic regularities from word embeddings using a generalized Pólya Urn process.

We empirically compare Limbic with state-of-the-art models using datasets from two domains. Qualitatively, Limbic discovers aspect-sentiment pairs with higher lexical diversity. Quantitatively, Limbic obtains substantial improvements in topic cohesion and sentiment classification.

Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3412–3422
Brussels, Belgium, October 31 - November 4, 2018. ©2018 Association for Computational Linguistics

2 Model and Inference in Limbic

We now introduce our proposed model.

2.1 Sentiment Expression Unit (SEU)

Existing topic models represent documents as bags of words or as sentences. Bag-of-words models, e.g., LDA (Blei et al., 2003), AT (Rosen-Zvi et al., 2004), JST (Lin et al., 2012), JAST (Mukherjee et al., 2014), and AATS (Poddar et al., 2017), rely on word co-occurrence at the document level, which is problematic when applied to opinionated texts. Sentence-based models, e.g., ASUM (Jo and Oh, 2011), assume that words appearing in a sentence belong to the same aspect and sentiment, which often fails to hold in real text. For instance, the TripAdvisor review sentence *Service was good and friendly, location is good and my room was spacious but oldish*, exhibits three aspects, *Service*, *Location*, and *Room*, and two sentiments. Zhang and Singh's (2014) segmentation algorithm leverages transition cues to convert sentences into segments. Although transition cues are good indicators for capturing sentiment change, their algorithm disregards syntactic information in sentences, which also helps reveal changes of aspects and sentiments.

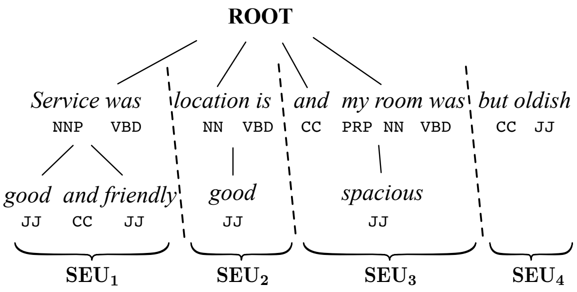

Figure 1: Generate SEUs from a sentence.

We propose a concept of sentiment expression unit (SEU). Each SEU contains either a sentiment, or an aspect, or both. We extract SEUs by incorporating both discourse and syntactic information. We first split sentences in reviews into snippets based on contradiction transition cues, such as *but*. Then we apply a grammar parser on each snippet. We extract phrases from snippets by using two syntactic patterns commonly observed in opinionated texts including (1) existential (EX) with verb (VB) and adjective (JJ) and (2) noun (NN) with verb (VB) and adjective (JJ). If a phrase matches a pattern, we identify it as an SEU. Otherwise, the phrase joins its following phrases iteratively until the combination matches a pattern. Figure 1 demonstrates the process of generating SEUs from the above hotel review sentence.

2.2 Discourse Relation

Markov Random Field (MRF) is a probabilistic framework to model statistical dependencies between variables. Limbic applies an MRF to capture the discourse relations between SEUs. Given a document containing N SEUs, let a_i and s_i be the aspect and sentiment assignments of SEU_i, respectively. Limbic creates an undirected edge $\langle s_i, s_j \rangle$ between the sentiment assignments of this SEU and its preceding SEU. Let r be the discourse relation between SEUs, Limbic imposes a binary potential on the edge.

Limbic focuses on two discourse relations frequently observed in opinionated texts: *Comparison* and *Expansion*. Comparison highlights prominent differences between two SEUs and often signals a change of sentiment regardless of the change of aspect. For example, in SEU_1: {The location was great} and SEU_2: {but it was just too noisy}, we see that *but* indicates a sentiment difference. Other transition cues for Comparison include *however*, *in contrast*, and such.

Expansion extends the discourse and indicates a continuation of sentiment across SEUs. For example, in SEU_3: {There are no safes here which is unfortunate} and SEU_4: {And speaking of unfortunate, the breakfast is hardly impressive}, we see that *and* and *unfortunate* indicate the negative sentiment in SEU_3 continues toward aspect *Breakfast* in SEU_4. Other transition cues for Expansion include *also*, *moreover*, and such.

Formally, $\mathcal{R}_{r,i,j}$ asserts discourse relation r between SEU_i and SEU_j. For Comparison, $\mathcal{R}_{c,i,j}$ holds if $s_i \neq s_j$, SEU_j contains Comparison cues, and (1) SEU_j contains syntactic patterns described in Section 2.1 and $a_i \neq a_j$ or (2) SEU_j contains incomplete syntactic patterns and $a_i = a_j$.

For Expansion, $\mathcal{R}_{e,i,j}$ holds if $s_i = s_j$, SEU_j contains Expansion cues, and (1) SEU_j contains syntactic patterns and $a_i = a_j$ or (2) SEU_j contains incomplete syntactic patterns and $a_i \neq a_j$.

Given document d, the joint probability of its sentiment assignments is:

$$p(\boldsymbol{s}|\theta_d) = \prod_i p(s_i|\theta_d) \exp\{\lambda \sum_{r=1}^{R} (I(\mathcal{R}_{r,i-1,i}))\}, \tag{1}$$

where R is the number of discourse relation types; θ_d is the sentiment distribution of d; I is an identity function that returns 1 if its argument is true; λ controls reinforcing the effects of discourse relations.

Take the expansion relation, for example. During the sampling process, Equation 1 generates a large value if two SEUs share an expansion relation and have the same sentiments and yields a small value if the two SEUs have different sentiments. Therefore, SEUs in an expansion relation have a high probability to be associated with the same sentiment.

2.3 Generative Process

Figure 2 shows Limbic's model. With $Dir(\cdot)$ and $Mul(\cdot)$ as Dirichlet and multinomial distributions, hyperparameter α is the Dirichlet prior of the word distribution ϕ, β is the Dirichlet prior of the sentiment distribution θ, and γ is the Dirichlet prior of the aspect distribution ψ. Given a set of reviews D written by a set of authors U with regards to a set of aspects T and a set of sentiments S, the generative process in Limbic is as follows.

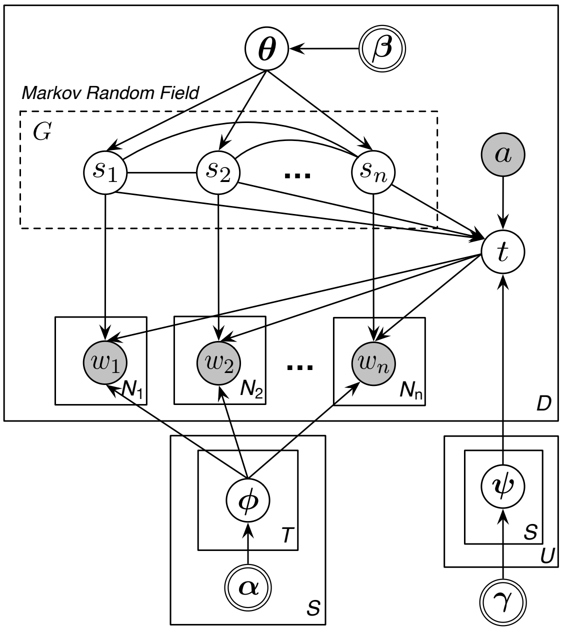

Figure 2: Generative process of Limbic.

First, for each pair of aspect t and sentiment s, draw a word distribution $\phi_{t,s} \sim Dir(\alpha)$. Second, for each author a and each sentiment s, draw an aspect distribution $\psi_{s,a} \sim Dir(\gamma)$. Third, given a review d written by a, draw a sentiment distribution $\theta_d \sim Dir(\beta)$, and for each SEU in d, (a) choose a sentiment s using Equation 1; (b) given s, choose an aspect $t \sim Mul(\psi_{s,a})$; (c) given t and s, sample word $w \sim Mul(\phi_{t,s})$.

2.4 Model Inference

Limbic estimates $p(\boldsymbol{s}, \boldsymbol{t}|\boldsymbol{w}, a)$, the posterior distribution of latent variables, sentiments $\boldsymbol{s}$ and aspects $\boldsymbol{t}$, given all words used in reviews written by author a. We factor the joint probability of the assignments of sentiments, aspects, and words for a:

$$p(\boldsymbol{s}, \boldsymbol{t}, \boldsymbol{w}|a, \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma}) = p(\boldsymbol{w}|\boldsymbol{s}, \boldsymbol{t}, \boldsymbol{\alpha}) p(\boldsymbol{t}|\boldsymbol{s}, a, \boldsymbol{\gamma}) p(\boldsymbol{s}|\boldsymbol{\beta}). \tag{2}$$

By integrating over $\Phi = \{\boldsymbol{\phi_i}\}_{i=1}^{S\times T}$, we calculate the first term of Equation 2 as follows.

$$p(\boldsymbol{w}|\boldsymbol{s},\boldsymbol{t},\boldsymbol{\alpha}) = \int p(\boldsymbol{w}|\boldsymbol{s},\boldsymbol{t},\Phi) p(\Phi|\boldsymbol{\alpha}) \mathrm{d}\Phi$$
$$= \left(\frac{\Gamma(\sum_{w=1}^{W}\alpha_w)}{\prod_{w=1}^{W}\Gamma(\alpha_w)}\right)^{S\times T} \times \prod_{s=1}^{S}\prod_{t=1}^{T} \frac{\prod_{w=1}^{W}\Gamma(n_{s,t}^{w}+\alpha_w)}{\Gamma[\sum_{w=1}^{W}(n_{s,t}^{w}+\alpha_w)]}, \tag{3}$$

where W is the size of the vocabulary; $n_{s,t}^{w}$ equals the number of occurrences of the word w that are assigned to sentiment s and aspect t; and $\Gamma(\cdot)$ is the Gamma function.

Next, by integrating over $\Psi_a = \{\boldsymbol{\psi_i}\}_{i=1}^{S}$, we calculate the second term in Equation 2 as follows.

$$p(\boldsymbol{t}|\boldsymbol{s},\boldsymbol{\gamma},a) = \int p(\boldsymbol{t}|\boldsymbol{s},\Psi_a,a) p(\Psi_a|\boldsymbol{\gamma}) \mathrm{d}\Psi_a$$
$$= \left(\frac{\Gamma(\sum_{t=1}^{T}\gamma_t)}{\prod_{t=1}^{T}\Gamma(\gamma_t)}\right)^{S} \times \prod_{s=1}^{S} \frac{\prod_{t=1}^{T}\Gamma(n_{s,a}^{t}+\gamma_t)}{\Gamma[\sum_{t=1}^{T}(n_{s,a}^{t}+\gamma_t)]}, \tag{4}$$

where $n_{s,a}^{t}$ equals the number of SEUs in author a's reviews associated with sentiment s and aspect t.

Similarly, for the third term in Equation 2, by integrating over $\Theta = \{\boldsymbol{\theta_i}\}_{i=1}^{D}$, we obtain

$$p(\boldsymbol{s}|\boldsymbol{\beta}) = \int p(\boldsymbol{s}|\Theta) p(\Theta|\boldsymbol{\beta}) \mathrm{d}\Theta$$
$$= \left(\frac{\Gamma(\sum_{s=1}^{S}\beta_s)}{\prod_{s=1}^{S}\Gamma(\beta_s)}\right)^{D} \times \prod_{d=1}^{D} \frac{\prod_{s=1}^{S}\Gamma(n_{d}^{s}+\beta_s)}{\Gamma[\sum_{s=1}^{S}(n_{d}^{s}+\beta_s)]} \tag{5}$$
$$\times \prod_{l=1}^{L} \exp\{\lambda \sum_{r=1}^{R} (I(\mathcal{R}_{r,i-1,i}))\},$$

where D is the number of reviews; n_d^s is the number of times that an SEU from review d is associated with sentiment s; and n_d is the number of SEUs in review d; L is the number of SEUs.

We obtain the conditional probability for a via Gibbs sampling (Liu, 1994)

$$
\begin{aligned}
&p(s_i = s, t_i = t | \boldsymbol{s_{-i}}, \boldsymbol{t_{-i}}, \boldsymbol{w}, a) \\
&\quad \propto \frac{n^t_{s,a,-i} + \gamma_t}{\sum_{t=1}^{T} (n^t_{s,a,-i} + \gamma_t)} \\
&\quad \times \frac{n^s_{d,-i} + \beta_s}{\sum_{s=1}^{S} (n^s_{d,-i} + \beta_s)} \\
&\quad \times \frac{\prod_v \prod_{c=0}^{C_v^i - 1} (n^v_{t,s,-i} + \alpha_v + c)}{\prod_{c=0}^{C_i - 1} (n_{t,s,-i} + \sum_{w=1}^{W} \alpha_w + c)} \\
&\quad \times \exp\{\lambda \sum_{r=1}^{R} (I(\mathcal{R}_{r,i-1,i}))\}
\end{aligned} \tag{6}
$$

where $n^t_{s,a}$, as in Equation 4, is the number of SEUs associated with sentiment s and aspect t from reviews written by author a; n^s_d is the number of SEUs from review d associated with sentiment s; C_i is the number of words in SEU$_i$; C^i_v is the number of words v in SEU$_i$; $n^v_{t,s}$ is the number of words v assigned sentiment s and aspect t; $n_{t,s}$ is the number of words assigned sentiment s and aspect t in all reviews; an index of $-i$ means we exclude SEU$_i$ from the count; R, I, and $\mathcal{R}$ are as in Equation 1.

Equations 7, 8, and 9, respectively, approximate the probabilities of word w occurring given sentiment s and aspect t; of aspect t of an SEU occurring given sentiment s and author a; of sentiment s occurring given document d.

$$
\phi_{s,t,w} = \frac{n^w_{s,t} + \alpha_w}{\sum_{w=1}^{W} (n^w_{s,t} + \alpha_w)}. \tag{7}
$$

$$
\psi_{s,t,a} = \frac{n^t_{s,a} + \gamma_t}{\sum_{t=1}^{T} (n^t_{s,a} + \gamma_t)}. \tag{8}
$$

$$
\theta_{d,s} = \frac{n^s_d + \beta_s}{\sum_{s=1}^{S} (n^s_d + \beta_s)}. \tag{9}
$$

Incorporating Word Embeddings. Word embedding approaches (Mikolov et al., 2013; Pennington et al., 2014) leverage local contextual information surrounding words to map the words into continuous vector representations. Word embeddings are known to effectively capture semantic and syntactic regularities among words. Based on word embeddings trained using Word2Vec on a hotel review dataset, we observe that the generated word embeddings correctly link opinionated words that are semantically correlated even though they do not co-occur frequently. For example, the three closest words of *spotless* are *immaculate*, *clean*, and *well appointed*.

A Generalized Pólya Urn Process. To promote words with high semantic similarity into the same topic, Limbic incorporates semantic regularities from word embeddings using a generalized Pólya Urn process (Mimno et al., 2011). Start with an urn containing colored balls. At each time step, randomly choose a ball from the urn, observe its color, and return it to the urn with one replicated ball of the same color. A Pólya Urn model describes a random sampling process with reinforcement. In a generalized Pólya Urn process, given a sampled ball with a color, we put back that ball along with a certain number of balls of similar colors. When applied to document generation, balls of different colors represent distinct words. The similarity of colors represents semantic similarity of the words.

Given words v and w in vocabulary W, we compute their semantic similarity $sim(v, w)$ based on the cosine similarity between their word embeddings. For word v, we create its similarity word set $\mathbb{S}_v$ by adding all words $w \in W$ for which $sim(v, w)$ is higher than a predefined threshold ϵ. During sampling, if word v is drawn, we reinforce $w \in \mathbb{S}_v$ via a predefined weight ρ which controls the reinforcement of semantically similar words.

Sentiment Alignment. Widely used Word embedding approaches, such as Word2Vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), and fastText (Bojanowski et al., 2017), are semantically oriented and do not explicitly encode sentiment information in the generated word-vector representations. Hence, semantically related words with opposite polarity may have close vectors. For example, *smell* and *aroma* are synonyms but *smell* often expresses a negative sentiment toward aspect *Cleanliness* whereas *aroma* is often positive. Simply promoting all words may adversely affect the generated topics. Therefore, we calculate the sentiment alignment of each word in a vocabulary based on its average cosine similarity to the words in a general sentiment word list. In the sampling process, we promote words only if their sentiments align with sampled sentiments.

3 Evaluation

To assess Limbic's effectiveness, we prepare online review datasets from two domains. TripUser is a collection of hotel reviews from TripAdvisor. It contains 28,165 reviews posted by 202 randomly selected reviewers, each of whom contributes at least 100 hotel reviews. YelpUser is a set of restaurant reviews from Yelp Dataset Challenge (2017). It contains 23,873 restaurant reviews posted by 144 users, each of whom contributes at least 100 reviews. Table 1 reports statistics on the datasets.

Table 1: Summary of the evaluation datasets.

Statistic	TripUser	YelpUser
Number of reviews	28,165	23,873
Number of SEUs	484,805	359,191
Average SEUs / review	17	15
Average words / SEU	7	6

We remove stop words and HTML tags, expand typical abbreviations, and mark special named entities using a rule-based algorithm (e.g., replace a monetary amount by #MONEY#) and the Stanford Named Entity Recognizer (Finkel et al., 2005). To handle negation, we employ the Stanford Dependency Parser to detect negations. For any word in a negation relation, we add the negated term as a prefix of the word, e.g., *not_work*. Finally, we split each review into SEUs. Datasets and source code are publicly available for research purposes (Limbic, 2018).

3.1 Parameter Settings

Limbic includes three hand-tuned hyperparameters that influence its sampling via a smoothing effect on the associated multinomial distribution. It uses a short list of sentiment words shown in Table 2 as prior knowledge to set asymmetric priors.

Table 2: Sentiment words used as prior knowledge.

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

Consider hyperparameter α, the Dirichlet prior of the word distribution. For any word in the positive list, $\alpha = 0$ if the word appears in an SEU assigned a negative sentiment, and $\alpha = 5$ if the word appears in an SEU assigned a positive sentiment, and conversely for words in the negative list. For all remaining words, we set $\alpha = 0.05$. And, hyperparameter $\beta = 5$ for both sentiments is the Dirichlet prior of the sentiment distribution. Using T as the number of aspects, hyperparameter $\gamma = \frac{50}{T}$ is the Dirichlet prior of the aspect distribution. We set the number of sentiments, S, to two (positive and negative), although our approach generalizes to additional sentiment categories.

For each fold in cross validation, we pretrain two sets of Word2Vec (Mikolov et al., 2013) word embeddings with 300 dimensions and a window size of five using the training split in TripUser (hotels) and YelpUser (restaurants). We exclude words with frequency lower than three. We set the reinforcement weight ρ to 0.3 and 0.1 for hotel and restaurant reviews, respectively, and set the similarity threshold ϵ to 0.6. For all models, we perform 1,000 Gibbs iterations with a burn-in phase of 200 and a sampling gap of 50 iterations.

3.2 Sentiment Aspect Discovery

Our first experiment shows how Limbic discovers sentiment-aspect pairs. We apply Limbic and all baseline models (AT, JST, ASUM, and AATS) to TripUser and YelpUser with the number of aspects set to 30. We manually assign an aspect for each cluster of words. ASUM generates the best results among baseline models. For brevity, we show only some aspects identified by Limbic and ASUM.

Table 3 (top) shows the results on hotel reviews. We see that Limbic discovers word clusters with higher lexical diversity than ASUM. For example, for aspect *Decoration*, in addition to words, *decor*, *modern* and *design*, Limbic discovers words *contemporary*, *minimalist*, *chic*, and so on. For aspect *Service*, comparing with ASUM, Limbic extracts an expanded list of sentiment words including *competent*, *knowledgeable*, and so on.

Limbic discovers finer and more distinctive word clusters than ASUM. For example, for aspect *Cleanliness*, ASUM generates a word cluster that includes negative sentiment words toward multiple entities, such as *carpet* and *hallway*. Limbic generates two distinctive word clusters for aspect

Table 3: Top words discovered from hotel (top section) and restaurant (bottom section) reviews.

Decoration		*Service*		*Cleanliness*			*Environment*
ASUM	Limbic	ASUM	Limbic	ASUM	Limbic	Limbic	Limbic
room	**contemporary**	staff	staff	carpet	smell	carpet	beautiful
modern	decor	friendly	friendly	smell	room	wallpaper	setting
furniture	colour	helpful	helpful	stain	cigarette	old	peaceful
decor	**tasteful**	desk	attentive	bathroom	smoke	furniture	relaxing
wood	room	front	courteous	room	floor	paint	golf
lobby	**marble**	english	hostess	dirty	odor	need	lush
design	**chic**	professional	professional	furniture	elevator	stain	gorgeous
wall	lobby	efficient	**gracious**	wall	reek	worn	environment
color	modern	service	**accommodating**	hallway	odour	carpeting	countryside
look	**elegant**	reception	**welcoming**	paint	smelt	bedspread	atmosphere
style	**artwork**	attentive	desk	smoke	smoking	remodel	ground
bathroom	**stone**	polite	**knowledgeable**	worn	hallway	update	course
dark	stylish	speak	**competent**	clean	stair	room	place
ceiling	**minimalist**	courteous	front	old	non-smoking	look	quiet
decorate	design	pleasant	service	tile	lift	bathroom	surroundings
Service		*Decoration*		*Portion*		*Mexican*	*Seafood*
ASUM	Limbic	ASUM	Limbic	ASUM	Limbic	Limbic	Limbic
friendly	friendly	decor	decor	portion	portion	asada	crab
server	service	atmosphere	atmosphere	size	size	carne	lobster
staff	attentive	place	**sleek**	small	half	taco	scallop
service	staff	modern	interior	large	large	tostada	shrimp
attentive	helpful	feel	vibe	plate	huge	burrito	roll
helpful	**efficient**	restaurant	**ambiance**	huge	serving	bean	risotto
nice	server	cool	clean	price	big	refried	salmon
waitress	**prompt**	inside	modern	big	could	corn	mussel
owner	**consistently**	clean	**ambience**	share	salad	guacamole	calamari
bartender	nice	vibe	**cozy**	#MONEY#	eat	cabbage	bisque
waiter	**professional**	nice	**contemporary**	enough	plate	torta	tuna
greet	host	like	place	half	share	pinto	jumbo
manager	great	interior	**comfortable**	generous	bowl	guac	clam
hostess	**quick**	space	**stylish**	bowl	sandwich	jalapeno	ahi
table	**knowledgeable**	look	inside	inch	generous	flour	tempura

Cleanliness. One cluster contains words, such as *smoke* and *reek*, which describe bad odor in room and hallway. The other cluster contains words such as, *worn* and *stain*, describes negative sentiments toward carpets. By capturing word semantic relatedness, Limbic discovers highly diverse aspects, including those that arise rarely in reviews, such as *peaceful*, *relaxing*, and *lush*, as positive words describing aspect *Environment*.

Limbic yields promising results for restaurant reviews. In Table 3 (bottom), we see that Limbic yields more specific sentiment words than ASUM. Aspect *Service* in Limbic contains additional positive words, *efficient*, *prompt*, *knowledgeable*, and so on. For aspect *Decoration*, Limbic produces *sleek*, *ambiance*, and so on. By incorporating constraints from discourse relations, Limbic yields aspects that are more sentiment coherent. For example, we see that positive aspect *Portion* in ASUM contains the negative word *small* whereas words in aspect *Portion* in Limbic are all positive.

We observe that restaurants associate more complex aspects than hotels—presumably, because of the large variety of cuisines and thus, on average, smaller data relevant to a cuisine. Titov and McDonald's (2008b) Multi-Grain LDA (MG-LDA) model performs well for hotel reviews but discovers only few ratable aspects from restaurant reviews, which they ascribe to the relatively small occurrences of words describing aspects for specific cuisines (e.g., *Italian*) and general categories (e.g., *Meat*), compared with the words describing major aspects, such as *Service*. In contrast, Limbic discovers words describing specific cuisines, such as *Mexican* and *Seafood*.

3.3 Quantitative Evaluation

Whether topics (word clusters) are semantically cohesive is an important factor in assessing topic modeling approaches. Normalized Pointwise Mutual Information (NPMI) (Lau et al., 2014) has strong correlation with human-judged topic coherence ratings and is widely used for accessing topic modeling approaches (Nguyen et al., 2015a,b;

Table 4: Topic coherence: Hotel reviews.

NPMI	T=10	T=20	T=30	T=40	T=50	T=60
AT	3.64	4.04	4.37	4.49	4.86	5.14
AATS	5.63	9.08	10.41	10.78	11.05	11.00
JST	8.99	10.78	11.45	11.54	11.56	11.46
ASUM	9.48	10.64	11.02	11.33	11.39	11.56
Limbic	**16.04**†	**17.16**†	**17.75**†	**17.65**†	**17.16**†	**16.60**†
W2V	T=10	T=20	T=30	T=40	T=50	T=60
AT	0.10	0.10	0.10	0.09	0.09	0.09
AATS	0.13	0.16	0.18	0.18	0.18	0.18
JST	0.15	0.18	0.18	0.19	0.19	0.19
ASUM	0.17	0.18	0.18	0.18	0.18	0.18
Limbic	**0.35**†	**0.37**†	**0.38**†	**0.37**†	**0.35**†	**0.34**†

Table 5: Topic coherence: Restaurant reviews.

NPMI	T=10	T=20	T=30	T=40	T=50	T=60
AT	5.64	5.21	5.30	5.65	6.54	7.94
AATS	6.05	8.02	9.03	9.35	9.90	9.95
JST	9.46	11.13	11.73	11.92	**12.14**	**12.31**
ASUM	8.81	9.7	9.92	10.09	10.07	10.04
Limbic	**11.72**†	**13.41**†	**13.77**†	**13.06**†	12.08	11.30
W2V	T=10	T=20	T=30	T=40	T=50	T=60
AT	0.16	0.15	0.14	0.13	0.13	0.14
AATS	0.11	0.14	0.16	0.17	0.18	0.18
JST	0.21	0.21	0.20	0.20	0.20	0.19
ASUM	0.20	0.19	0.18	0.18	0.17	0.17
Limbic	**0.28**†	**0.29**†	**0.27**†	**0.25**†	**0.25**†	**0.25**†

Yang et al., 2017). More recently, O'Callaghan et al. (2015) propose a topic coherence measure, W2V, based on word embeddings. For completeness, we adopt both metrics. Topics with higher scores of NPMI and W2V are semantically more coherent. We compare Limbic with four baselines: AT, JST, ASUM, and AATS, using both TripUser and YelpUser based on the top 15 words in each sentiment-aspect pair. For each number of aspects, we perform five-fold cross-validation. We perform the two-tailed exact permutation test (Good, 2005) on the improvement of Limbic over the best performing baseline. (Throughout, $*$ and $\dagger$ indicate significance at 0.05 and 0.001, respectively.)

Table 4 shows average NPMI and W2V scores of each model on hotel reviews for different numbers of aspects. We observe that Limbic statistically outperforms the other models for both metrics in all settings. Limbic yields substantial improvements, with average gains over the second best models of 6.00 and 0.18 in NPMI and W2V, respectively, which validates that the incorporation of semantic regularities helps Limbic promote semantically equivalent and related words into the same aspect-sentiment pair. Of the baseline models, AT yields the lowest topic coherence. AATS outperforms AT but does not perform well when the number of aspects is small, possibly due to the undesirable mixture of words with different aspects, topics, and sentiments in individual sentences. ASUM, and JST yield comparable results that are consistently better than AATS. Table 5 shows similar conclusions for restaurant reviews.

3.4 Sentiment Classification

We now evaluate Limbic for document-level sentiment classification vis à vis JST, ASUM, and AATS. For comparison purposes, we add a supervised baseline, BiLSTM, using the bidirectional LSTM model (Schuster and Paliwal, 1997). BiLSTM uses 100 as hidden state size and 0.2 as both the recurrent dropout rate and the dropout rate in the last layer. For training, we run 20 epochs with a minibatch size of 1,000. We use two datasets, TripUser and YelpUser. To collect ground-truth labels, we use integer ratings (three and above as positive and rest as negative). Note that our review datasets are imbalanced. Our results are based on five-fold cross-validation (80% of each author's reviews for training and 20% for testing) with the two-tailed exact permutation test. As our principal evaluation metrics, we adopt accuracy; the receiver operating characteristic (ROC) curve; and area under the curve (AUC). ROC and AUC are standard metrics used for evaluating classifiers on data with class imbalance (Bradley, 1997; Hoens and Chawla, 2013).

Tables 6 and 7 report accuracy and AUC on hotel and restaurant reviews. AATS yields high accuracy but low AUC due to a strong bias toward the majority class. Compared with AATS, JST yields higher AUC for both datasets but lower accuracy for TripUser. ASUM outperforms JST, indicating that sentences are more effective as units of sentiment analysis than bags of words. Limbic significantly outperforms ASUM in all settings. For hotel reviews, Limbic attains average gains of 4.0% and 2.3% in accuracy and AUC, respectively. For restaurant reviews, Limbic yields average gains of 5.1% and 3.0% in accuracy and AUC, respectively. In Figure 3 and 4, we compare the ROC curves of Limbic with baselines. The ROC curves show how the true positive rate (TPR) (vertical axis) varies with the false positive rate (FPR) (horizontal axis) by moving the decision boundary. We see that for all FPRs, Limbic yields the highest TPRs. Its ROC curves dominate other models' curves. The results demonstrate that, among all models, Lim-

Table 6: Accuracy and AUC of sentiment classification on hotel reviews.

	T=10		T=20		T=30		T=40		T=50		T=60	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
BiLSTM	0.907	0.820	0.907	0.820	0.907	0.820	0.907	0.820	0.907	0.820	0.907	0.820
AATS	0.729	0.485	0.794	0.454	0.809	0.443	0.824	0.475	0.835	0.468	0.839	0.482
JST	0.601	0.813	0.609	0.818	0.634	0.826	0.641	0.830	0.654	0.835	0.665	0.836
ASUM	0.793	0.828	0.804	0.832	0.819	0.829	0.835	0.838	0.850	0.835	0.872	0.829
Limbic	**0.838**†	**0.849***	**0.859**†	**0.853***	**0.868**†	**0.857***	**0.870**†	**0.859***	**0.885**†	**0.858***	**0.890**	**0.858***

Table 7: Accuracy and AUC of sentiment classification on restaurant reviews.

	T=10		T=20		T=30		T=40		T=50		T=60	
	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC
BiLSTM	0.876	0.841	0.876	0.841	0.876	0.841	0.876	0.841	0.876	0.841	0.876	0.841
AATS	0.763	0.500	0.789	0.472	0.785	0.474	0.804	0.477	0.813	0.499	0.816	0.492
JST	0.579	0.708	0.589	0.713	0.595	0.724	0.606	0.731	0.629	0.727	0.642	0.733
ASUM	0.773	0.775	0.799	0.778	0.811	0.778	0.836	0.775	0.859	0.752	0.867	0.744
Limbic	**0.872**†	**0.797***	**0.873**†	**0.803**†	**0.874**†	**0.795**†	**0.876**†	**0.795**†	**0.877**†	**0.798**‡	**0.876**†	**0.794**†

bic achieves the best tradeoff between positive and negative sentiment classes.

3.5 Model Analysis

To understand the contributions of incorporating authors, discourse relations, and word embeddings, we evaluate variants of Limbic for SEU-level sentiment classification on two datasets: tSEU and tSEU(D). We create tSEU by randomly selecting 200 hotel reviews by seven authors. We manually annotate the sentiments of each SEU, obtaining 2,692 SEUs. We create tSEU(D) by selecting reviews in tSEU containing at least one Comparison or Expansion. We define three variants of Limbic (L): L_A with just authors, no discourse relations or word embeddings; L_{AD} with authors and discourse relations but no word embeddings; L_{AW} with authors and word embeddings but no discourse relations. Table 8 compares Limbic with L_A L_{AD}, and L_{AW}. We observe that for both datasets, incorporating discourse relations improves accuracy. By incorporating word embeddings, L_{AW} yields better accuracy than L_{AD}, showing that word embeddings add more value to Limbic than discourse relations do.

Table 8 (lower part) reports p-values (two-tailed test using a χ^2 distribution) from McNemar's Test on pairwise comparisons (Alpaydin, 2010, p. 501). We see that Limbic is significantly different from each variant, including L_A (omitted for space).

Table 8: SEU sentiment classification accuracy.

Accuracy	L_A	L_{AD}	L_{AW}	L
tSEU	0.702	0.723	0.733	**0.750**
tSEU(D)	0.692	0.715	0.716	**0.735**

p value	L_{AD}:L_A	L_{AW}:L_A	L_{AD}:L_{AW}	L:L_{AD}	L:L_{AW}
tSEU	0.015	0.003	0.331	0.005	0.002
tSEU(D)	0.017	0.042	0.934	0.055	0.003

3.6 Related Work

Sentiment and aspect discovery are often based on Latent Dirichlet Allocation (LDA) (Blei et al., 2003). LDA represents a document (for us, a review) as a mixture of topics, each topic being a multinomial distribution over words. The learning process approximates the topic and word distributions based on their co-occurrence in documents. Titov and McDonald's (2008b) model handles global and local topics involved in documents, and their (2008a) framework discovers topics using aspect ratings provided by reviewers. JST (Lin et al., 2012) and ASUM (Jo and Oh, 2011) model a review via multinomial distributions of topics and sentiments and use them to condition the probability of generating words. Kim et al. (2013) extend ASUM by allowing its probabilistic model to discover a hierarchical structure of aspect-based sentiments. Lazaridou et al. (2013) introduce discourse transitions into the document generating process as aspect and sentiment shifters. Although the above models produce good results, they omit author information, which is an intrinsic attribute of opinionated texts.

Rosen-Zvi et al.'s Author Topic model (AT) (2004) captures authorship by building a topic distribution for each author. When generating a word

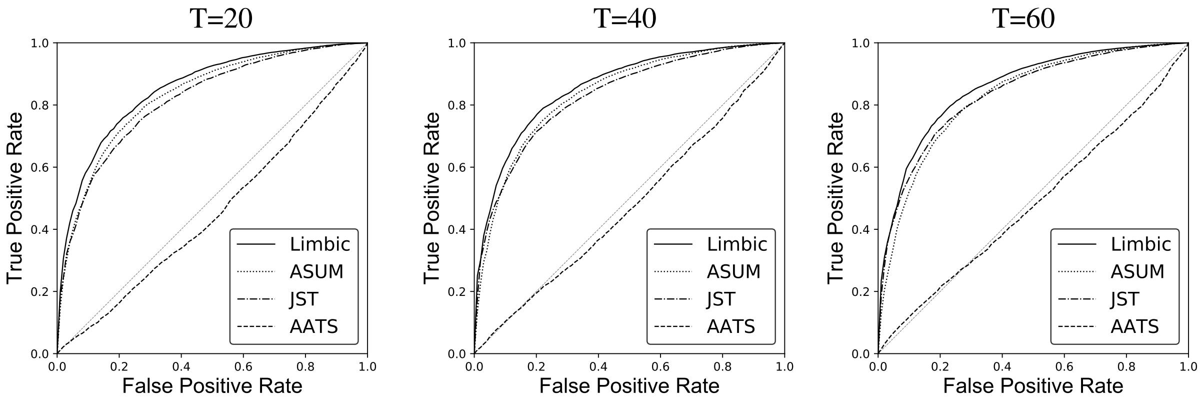

Figure 3: ROC curves comparing the performance of three models on hotel reviews.

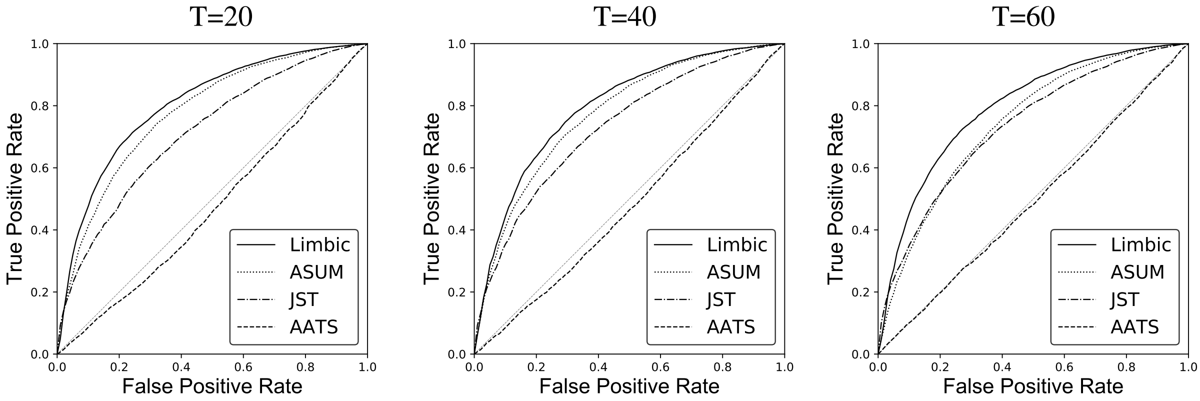

Figure 4: ROC curves comparing the performance of three models on restaurant reviews.

in a document, AT conditions the probability of the topic assignment on the author of the document. Kim et al.'s (2012) topic model captures entities mentioned in documents and models the probability of generating a word as conditioned on both entity and topic. Diao and Jiang's (2013) jointly model topics, events, and users on Twitter. Although these models capture the author associated with a text, they do not handle sentiments.

Mukherjee et al.'s (2014) JAST model jointly considers authors, sentiments, topics, and ratings. JAST does not consider discourse relations and word semantic similarity in its generative process. Poddar et al. (2017) propose a model that jointly considers author, aspect, sentiment, and the non-repetitive generation of aspect sequences. The model uses a Bernoulli process to capture the non-repetitive nature of aspect sequences. This mechanism does not consider discourse relations or syntactic information.

4 Conclusion and Discussion

Limbic provides an unsupervised method to discover aspects and sentiments from opinionated texts. By incorporating authors as a factor, Limbic allows for reviews written by the same or similar authors to exhibit an idiosyncratic preference toward certain aspects and sentiments. It assigns aspects of SEUs by sampling author-specific aspect distributions. This makes the model more suitable for opinionated texts in which aspects and sentiments are tightly bound to authors who follow their specific criteria and preferences when writing reviews. By incorporating a Markov Random Field and word embeddings into its sampling process, Limbic imposes constraints associated with discourse relations, effectively captures word semantic relatedness, and generates word clusters with high topic cohesion and lexical diversity. In future work, we plan to extend Limbic to capture long-distance discourse relations and the influence decay of discourse relations between SEUs as their distance increases.

5 Acknowledgments

Thanks to Chung-Wei Hang and the anonymous reviewers for helpful comments, corrections and inspiration.

References

Ethem Alpaydin. 2010. *Introduction to Machine Learning*, 2nd edition. MIT Press, Cambridge, Massachusetts.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Andrew P. Bradley. 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, 30(7):1145–1159.

Qiming Diao and Jing Jiang. 2013. A unified model for topics, events and users on Twitter. In *Proceedings of the* 18^{th} *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1869–1879, Seattle.

Jenny Rose Finkel, Trond Grenager, and Christopher D. Manning. 2005. Incorporating non-local information into information extraction systems by Gibbs sampling. In *Proceedings of the* 43^{rd} *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 363–370, Ann Arbor.

Phillip I. Good, editor. 2005. *Permutation, Parametric, and Bootstrap Tests of Hypotheses*, 3rd edition. Springer Series in Statistics. Springer-Verlag, New York.

T. Ryan Hoens and Nitesh V. Chawla. 2013. Imbalanced datasets: From sampling to classifiers. In Haibo He and Yunqian Ma, editors, *Imbalanced Learning: Foundations, Algorithms, and Applications*, pages 43–59. John Wiley & Sons, Hoboken, New Jersey.

Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In *Proceedings of* 15^{th} *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 289–296, Stockholm.

Yohan Jo and Alice Haeyun Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of the* 4^{th} *ACM International Conference on Web Search and Data Mining (WSDM)*, pages 815–824, Hong Kong.

Hyungsul Kim, Yizhou Sun, Julia Hockenmaier, and Jiawei Han. 2012. ETM: Entity topic models for mining documents associated with entities. In *Proceedings of the* 12^{th} *IEEE International Conference on Data Mining (ICDM)*, pages 349–358, Brussels.

Suin Kim, Jianwen Zhang, Zheng Chen, Alice H. Oh, and Shixia Liu. 2013. A hierarchical aspect-sentiment model for online reviews. In *Proceedings of the* 27^{th} *AAAI Conference on Artificial Intelligence (AAAI)*, pages 804–812, Bellevue.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the* 14^{th} *Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 530–539, Gothenburg.

Angeliki Lazaridou, Ivan Titov, and Caroline Sporleder. 2013. A Bayesian model for joint unsupervised induction of sentiment, aspect and discourse representations. In *Proceedings of the* 51^{st} *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1630–1639, Sofia, Bulgaria.

Limbic. 2018. https://research.csc.ncsu.edu/mas/code/limbic/. Accessed: 08/23/2018.

Chenghua Lin, Yulan He, Richard Everson, and Stefan M. Rüger. 2012. Weakly supervised joint sentiment-topic detection from text. *IEEE Transactions on Knowledge and Data Engineering*, 24(6):1134–1145.

Jun S. Liu. 1994. The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 89(427):958–966.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the* 27^{th} *Annual Conference on Neural Information Processing Systems (NIPS)*, pages 3111–3119, Lake Tahoe.

David M. Mimno, Hanna M. Wallach, Edmund M. Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the* 16^{th} *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 262–272, Edinburgh.

Subhabrata Mukherjee, Gaurab Basu, and Sachindra Joshi. 2014. Joint author sentiment topic model. In *Proceedings of the* 14^{th} *International Conference on Data Mining (SDM)*, pages 370–378, Philadelphia.

Dat Quoc Nguyen, Richard Billingsley, Lan Du, and Mark Johnson. 2015a. Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics (TACL)*, 3:299–313.

Thang Nguyen, Jordan L. Boyd-Graber, Jeffrey Lund, Kevin D. Seppi, and Eric K. Ringger. 2015b. Is your anchor going up or down? Fast and accurate supervised topic models. In *Proceedings of the* 16^{th} *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 746–755, Denver.

Derek O'Callaghan, Derek Greene, Joe Carthy, and Pádraig Cunningham. 2015. An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, 42(13):5645–5657.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the* 19^{th} *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha.

Lahari Poddar, Wynne Hsu, and Mong-Li Lee. 2017. Author-aware aspect topic sentiment model to retrieve supporting opinions from reviews. In *Proceedings of the* 22^{nd} *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 472–481, Copenhagen.

Michal Rosen-Zvi, Thomas L. Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *Proceedings of the* 20^{th} *Conference in Uncertainty in Artificial Intelligence (UAI)*, pages 487–494, Banff, Canada.

Mike Schuster and Kuldip K. Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Ivan Titov and Ryan T. McDonald. 2008a. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of the* 46^{th} *Annual Meeting on Association for Computational Linguistics (ACL)*, pages 308–316, Columbus, Ohio.

Ivan Titov and Ryan T. McDonald. 2008b. Modeling online reviews with multi-grain topic models. In *Proceedings of the* 17^{th} *International Conference on World Wide Web (WWW)*, pages 308–316, Beijing.

Feng Wang, Weike Pan, and Li Chen. 2013. Recommendation for new users with partial preferences by integrating product reviews with static specifications. In *Proceedings of the* 21^{st} *International Conference on User Modeling, Adaptation, and Personalization (UMAP)*, pages 281–288, Rome.

Weiwei Yang, Jordan L. Boyd-Graber, and Philip Resnik. 2017. Adapting topic models using lexical associations with tree priors. In *Proceedings of the* 22^{nd} *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1901–1906, Copenhagen.

Yelp. 2017. Yelp dataset challenge. https://www.yelp.com/dataset_challenge/. Accessed: 05/20/2018.

Zhe Zhang and Munindar P. Singh. 2014. ReNew: A semi-supervised framework for generating domain-specific lexicons and sentiment analysis. In *Proceedings of the* 52^{nd} *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 542–551, Baltimore.