Author-aware Aspect Topic Sentiment Model to Retrieve Supporting Opinions from Reviews

Lahari Poddar, Wynne Hsu, Mong Li Lee
School of Computing
National University of Singapore
{lahari, whsu, leeml}@comp.nus.edu.sg

Abstract

User generated content about products and services in the form of reviews are often diverse and even contradictory. This makes it difficult for users to know if an opinion in a review is prevalent or biased. We study the problem of searching for supporting opinions in the context of reviews. We propose a framework called SURF, that first identifies opinions expressed in a review, and then finds similar opinions from other reviews. We design a novel probabilistic graphical model that captures opinions as a combination of aspect, topic and sentiment dimensions, takes into account the preferences of individual authors, as well as the quality of the entity under review, and encodes the flow of thoughts in a review by constraining the aspect distribution dynamically among successive review segments. We derive a similarity measure that considers both lexical and semantic similarity to find supporting opinions. Experiments on TripAdvisor hotel reviews and Yelp restaurant reviews show that our model outperforms existing methods for modeling opinions, and the proposed framework is effective in finding supporting opinions.

1 Introduction

In order to make an informed decision when booking a hotel online, a user will often read through its reviews looking for specific feedbacks. For example, if he or she plans to do an early check-in and comes across a review that mentions a hassle-free early check-in as shown in Figure 1, it will be helpful to know whether other guests had similar experiences. If a review complains about bed bugs or noise from construction nearby, then it is important to know if that was an occasional problem based on a single user's experience or happens frequently. However, it is impossible for an individual to go through the large volume of reviews to verify whether an opinion is prevalent.


"Weekend Stay"
Reviewed 2 weeks ago
I stayed in this hotel for three nights with my family. ***I made a reservation online, and even though we arrived early in the morning the front-desk staff was very accommodating and helped us in doing an early check-in.*** The hotel is located just a few blocks away from all the major attractions and restaurants. We ate most of our meals outside since the food at the hotel had very limited choices and wasn't much good either. We had a spacious room with a clean bathroom and a comfy bed but no wifi.

Figure 1: A sample hotel review

In this work, we study the problem of finding supporting sentences from reviews that corroborate the opinions expressed in a target review sentence. This is useful as it enables users to easily look for appropriate comments on the specific issues they are interested in.

A review is a collection of sentences where each sentence may have multiple segments separated by punctuations or conjunctions. Each segment expresses an opinion that can be represented as a combination of aspect, topic and sentiment. An aspect refers to the overall theme of a segment, a topic is the specific subject or issue discussed and the sentiment for each topic can be neutral, positive or negative. Table 1 shows the segments and the possible latent aspect, topics and sentiment for a sentence of the review in Figure 1.

Review Sentence	Segments	Aspect	Topic	Sentiment
We had a big room with clean bathroom and a comfy bed, but no wifi	We had a big room with clean bathroom	room	room	positive
			bathroom	positive
	a comfy bed	room	bed	positive
	no wifi	amenities	wifi	negative

Table 1: Opinion structure for a review sentence

Given an opinion (in a target segment), we say that a review supports the opinion, if it contains some segment whose aspect, topic and sentiment are similar to those in the target segment. Finding such supporting reviews is a challenge since

reviews are typically short unstructured text and discuss a wide range of topics on various aspects with differing sentiments and vocabulary used.

Topic modeling have been widely used to reduce the effect of huge vocabulary by grouping words in topics. However, the fundamental assumption of topic models is the independence of topics even in the same document. This fails to capture the natural coherence present in reviews, which rarely consist of isolated, unrelated sentences, but are composed of collocated, structured and coherent groups of sentences (Hovy, 1993). We observe that an author's train of thoughts when writing a review is often linear, i.e., he or she will finish discussing one aspect before moving on to the next. In Figure 1, we see that the user first commented on *Service* (*"front-desk staff was very accommodating"*), then the *Location* aspect, followed by the comment on *Food*, and finally moved on to *Room*. This shows that aspects discussed in a review are not chosen from a simple *independent mixture*, but rather, words in close proximity tend to discuss the same aspect and within a review the aspects discussed in the current segment will affect the possible aspects for the successive segments.

We explicitly model this by constraining aspect transition between segments using a review specific Markov chain. Each segment is assumed to discuss a single aspect and possible aspects for a segment are made dependent on the aspects of the previous segments. By tracking aspects of previous segments we are able to ensure constrained aspect sampling for accurate modeling of a review structure. This non-iterative nature of discourse has not been considered by existing works.

For opinion modeling, capturing the sentiment expressed for an aspect is important. Recent works (Kim et al., 2013; Jo and Oh, 2011; Moghaddam and Ester, 2011; Wang et al., 2010; Titov and McDonald, 2008a,b) have developed models to capture aspect and sentiment. However, they do not consider the preferences of authors, or the inherent quality of the entity for the aspect. In a hotel review, the sentiment expressed for *service* depends on both the service standard of the hotel (evident from the sentiment distribution of *service* of all reviews for the hotel) and the expectation of the author for *service* (evident from the sentiment distribution of the author on *service* across all hotels) (Poddar et al., 2017). We take this into account by making the sentiment distribution of a review dependent on both entity and author.

We propose an Author-aware Aspect Topic Sentiment model (Author-ATS) to capture the diverse opinions, taking into account user preferences and thought patterns. The model considers a word to be generated from a hierarchy of aspect, topic and sentiment and encodes the coherent structural property of a review by dynamically constraining aspect distributions. We also develop a non-parametric version of Author-ATS based on Dirichlet Process called Author-ATS (DP).

We develop a SUpporting Review Framework (SURF) that utilizes the Author-ATS model to compute the lexical and semantic similarity of an opinion in a target segment to those in the review corpus, and returns the top-k supporting reviews. Experiments on real world review datasets show the effectiveness of Author-ATS in modeling opinions compared to existing topic models. Furthermore, SURF outperforms keyword-based approaches and word embedding based similarity measures in finding supporting opinions. To the best of our knowledge, this is the first work to find supporting reviews for an opinion expressed in user generated contents.

2 Related Work

There has been substantial research to mine online reviews using topic models (Paul and Girju, 2010; Trabelsi and Zaiane, 2014; Lin and He, 2009; Jo and Oh, 2011; Mukherjee and Liu, 2012; Chen et al., 2013). The Topic Aspect Model (TAM) (Paul and Girju, 2010) jointly discovers aspects and topics from documents. The aspect and topic are independent and each aspect affects all topics in similar manner. However, in reviews, the topics discussed are often closely related to an aspect. JTV (Trabelsi and Zaiane, 2014) encodes topic-viewpoint dependency, but assumes that a document contains only one aspect. JST (Lin and He, 2009) assumes that there is a single sentiment polarity for a review and the topics are chosen conditioned on that, while ASUM in (Jo and Oh, 2011) assumes that all words in a sentence are associated with the same topic and sentiment. In contrast, our proposed model handles the more realistic scenario where sentiments may vary depending on the topics discussed in a review.

For incorporating author information, the User-Sentiment topic model (Zhao et al., 2012) considers the topic-sentiment distribution only from

the author perspective and ignores the characteristics of the entity. Supervised topic model (Li et al., 2014) uses explicit ratings to infer sentiments. PDA-LDA (Zhang and Wang, 2015) associates its Dirichlet prior distribution with user and item topic factors. The work in (Yang et al., 2015) models aspects and sentiments based on the demography of authors. However, such demographic information are not always available and it cannot model the bias or preference of an individual.

Additionally, most topic models are concerned about the discourse at word level, and ignore the document structure. HTMM (Gruber et al., 2007) models topic coherence by considering topic transition between sentences. HTSM (Rahman and Wang, 2016) extends HTMM by capturing sentiment shifts along with topic coherence. Both models do not capture the non-repetitive discourse of reviews. Progressive topical dependency model (Du et al., 2010, 2015) captures the sequential nature of ideas among segments, especially in movies or books. However, unlike books, the sequence of topics in reviews is not significant. Rather, once a topic has been discussed in a review, it is unlikely to be mentioned again in a later segment. From this perspective, it is similar to labeled LDA (Ramage et al., 2009) where topic distribution of a document is constrained. However, unlike labeled LDA, the possible aspects of a segment are dynamically constrained depending on previously sampled aspects.

3 Author-ATS Model

Author-ATS models an opinion as hierarchical dependent mixtures, where words are generated from a three-level hierarchical structure of aspects, topics and sentiments. We assume there are A distinct aspects for a domain, for each aspect there are Z topics and for each aspect-topic pair S possible sentiments. We treat a segment as the basic semantic unit, discussing a particular aspect. A review r is a collection of D_r segments where each segment is a document d, consisting of N_d words. We now describe the assumptions and detailed construction of the proposed model.

3.1 Constrained Aspect Generation

We explicitly model the behavior that after an author has finished discussing an aspect and has moved on to the next, he or she is unlikely to return to it again. We assume that each document d discusses a single aspect a_d. The aspect distribution σ_r is drawn from a Dirichlet with parameter α. In order to model the linear writing style of authors, we constrain the possible aspects that can be sampled from σ_r. Whenever an author starts writing a segment, he or she can choose to either (a) talk about an aspect not yet discussed, or (b) continue with the aspect of the previous segment. This is captured by imposing the constraint that the aspect of the j^{th} document is dependent on the aspects of the $(j-1)^{th}, (j-2)^{th}, \cdots, 1^{st}$ documents of the same review.

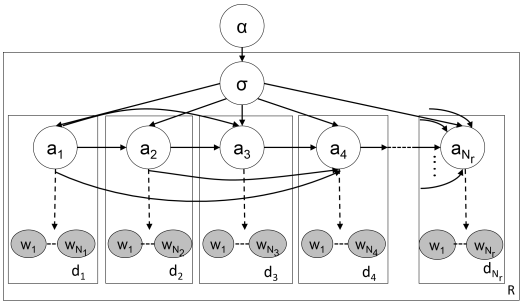


Figure 2: Constrained aspect generation in Author-ATS. Aspects in review form a Markov chain.

With this we relax the *independent mixture* assumption of the standard LDA model for aspects and form a review-specific Markov chain (see Figure 2). Such a higher order Markov chain would normally incur intractable computational complexity due to the exponential size of transition probability matrix. However, in our case, the transition probability can be determined by overall aspect distribution of the review, σ_r and a list of possible aspects for the segment. Since we assume a non-repetitive nature of discourse, the number of possible aspects for a segment is monotonically decreasing for successive segments. This special property enables us to devise a dynamic programming strategy to solve the problem with linear complexity.

Each document is associated with a binary aspect vector Λ. We restrict the sampled aspect of a document to be drawn from only the aspects that are turned on, in Λ of that document. For a document d, $\Lambda_d =< l_1, \cdots, l_A >$ where each $l_a \in \{0, 1\}$ and A is the total number of aspects. Traditionally, for a document d, an aspect a_d is sampled from a multinomial distribution σ_r. Here, we restrict the possible sampled aspects to the list Λ_d. A value of 1 for the entry l_a indicates that the aspect a can be sampled, while 0 indicates that the aspect should not be sampled.

We generate Λ_d by tossing a Bernoulli coin for each aspect a with prior probability Φ_a for value 0. We set Φ_a as the sampling probability for aspects which have been sampled for a previous document. This ensures that an aspect which has been discussed before has lesser probability of coming up again. We set $\Phi_a = 0$ for aspects not sampled in the past, and for the aspect of (immediately) preceding segment. This models aspect coherency in a review document where an author either chooses to discuss a new aspect or continues to talk about the current one.

We define the list of possible aspects for the document d to be $\lambda_d = \{a \mid \Lambda_d[a] = 1\}$. We sample an aspect a_d from σ_r with the constraint that $a_d \in \lambda_d$ i.e. an aspect can be sampled for a document only if it is turned on in the binary aspect vector for the document and thereby exists in the list of possible aspects for the document. Thus, the aspect transition probability among documents becomes dependent on σ_r and the vector λ_d. Unlike regular topic models, Author-ATS is no longer invariant to reshuffling of words and is able to model linear aspect coherency in a review.

3.2 Author-Entity dependent Sentiment Distribution

We account for the dual role of entity and author in a review, by observing that the sentiments expressed are influenced by both the quality of the entity being reviewed and the preferences of the author. We use two Dirichlet distributions to derive sentiment, namely, entity-dependent distribution (ξ) and author-dependent distribution (χ). For each aspect-topic combination, ξ is drawn from a Dirichlet distribution with prior γ^1 and χ is drawn from a Dirichlet distribution with prior γ^0.

Since online reviews describe experiences of people, some words tend to appear frequently (e.g.: 'hotel','trip' or 'mobile', 'phone' for hotel and mobile reviews respectively). We call them *domain stopwords* as they are not specific to any aspect. We use a binary switching variable y_i to determine the type for the i^{th} word. If $y_i = 0$, then the word is aspect neutral (domain stopword); and if $y_i = 1$, it is aspect dependent.

The generative process of the model is as follows:

- Draw a multinomial word distribution ϕ_0 for domain stopwords and ϕ_1 for each aspect, topic and sentiment words from Dir (ω).
- For each author u, draw a multinomial sentiment mixture χ for each aspect and topic from Dir(γ^0)

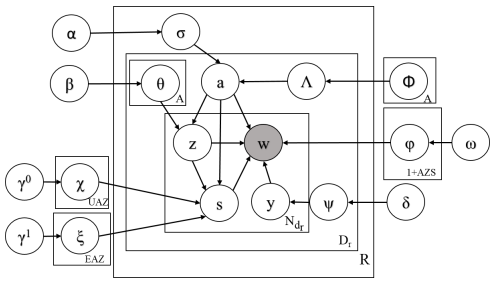


Figure 3: Graphical representation of Author-ATS

- For each entity e, draw a multinomial sentiment mixture ξ for each aspect and topic from Dir (γ^1)
- For each review r:
  1. Draw multinomial aspect mixture σ from Dir(α)
  2. For each document $d \in r$:
     (a) Draw Λ_d from Bernoulli (Φ)
     (b) Draw a type mixture ψ from Beta (δ_0, δ_1)
     (c) Sample an aspect a_d from σ s.t. $a_d \in \lambda_d$
     (d) For sampled aspect a_d, draw a topic mixture θ from Dir (β)
     (e) For each word position i where $0 \leq i \leq N_d$
        i. Sample a type y_i from ψ
        ii. Sample a topic z_i from θ
        iii. Sample a sentiment s_i from χ and ξ
        iv. Sample a word w_i from $\begin{cases} \phi_0 & \text{if } y_i = 0, \\ \phi_1 & \text{if } y_i = 1 \end{cases}$

Note that for the first document of a review, we set λ_0 to the set of all possible aspects, such that there is no constraint when sampling for the first segment of a review. Figure 3 shows the plate notation for Author-ATS model.

3.3 Bayesian Inference

We employ collapsed Gibbs sampling for inference. Markov chain introduced for aspect coherency makes the aspects non-exchangeable, hence sampling an aspect for a segment will also affect all subsequent segments. Since the exact sampling for this would be computationally expensive, we propose the following approximate posterior considering only the previous segments, which has been shown to work well in similar cases previously (Mimno et al., 2011).

We sample an aspect (a_d) for each document based on the posterior probability of the type, topic and sentiment assignment of each word in the document and the aspects sampled for preceding documents in the review.

$$P(a_d|\vec{a_{-d}},\vec{y}_{-d},\vec{z}_{-d},\vec{s}_{-d},\vec{w}) \propto P(a_d|a_{1:\vec{d}-1}) \prod_{z=1}^{Z}\prod_{s=1}^{S} \frac{\sum_{w=1}^{W} B(n_w^{a_d,z,s}+\omega)}{\sum_{w=1}^{W} B(n_w^{a_d,z,s,-d}+\omega)} \quad (1)$$

$$P(a_d|a_{1:\vec{d}-1}) \propto \begin{cases} \frac{n_{a_d}^{r,-d}+\alpha}{\sum_{a\in\lambda_d} n_a^{r,-d}+|\lambda_d|*\alpha} & \text{if } a_d \in \lambda_d \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where $B(\vec{x})$ is the multidimensional extension of the Beta function. The notation $n_a^{b,-c}$ refers to the number of times a has been assigned to b excluding current occurrence c, e.g. $n_{a_d}^{r,-d}$ denotes the number of documents in review r that has been assigned aspect a_d excluding current document d.

The target aspect a_d is dependent on the aspects sampled for the 1^{st} to $(d-1)^{th}$ documents of the review, denoted by $a_{1:\vec{d-1}}$. We restrict the target aspect a_d to belong to the set defined by Λ_d of the document d to achieve coherence among aspects respecting the nature of discourse observed in review writing styles. This constrained aspect sampling differentiates Author-ATS from existing topic modeling works on review text by explicitly modeling the topic coherence of opinionated text.

After sampling the aspect for the document, we jointly sample the latent type, topic and sentiment for each word within the document. The posterior for the i^{th} word of document d (written by author u for entity e) is given as:

$$P(y_i, z_i, s_i | a_d, \vec{w}, \vec{y}_{-i}, \vec{z}_{-i}, \vec{s}_{-i}) \propto P(y_i|d) * P(z_i|a_d, d) * P(s_i|a_d, z_i, u, e, d) * P(w_i|y_i, a_d, z_i, s_i,) \quad (3)$$

$$\propto \frac{n_{y_i}^{d,-i} + \delta_{y_i}}{\sum_{y=0}^{1}(n_y^{d,-i} + \delta_y)} * \frac{n_{z_i}^{d,a_d,-i} + \beta}{\sum_{z=1}^{Z} n_z^{d,a_d,-i} + Z\beta} * \Big(q_1 \frac{n_{s_i}^{u,a_d,z_i,-i} + \gamma^0}{\sum_{s=1}^{S} n_s^{u,a_d,z_i,-i} + S\gamma^0} + q_2 \frac{n_{s_i}^{e,a_d,z_i,-i} + \gamma^1}{\sum_{s=1}^{S} n_s^{e,a_d,z_i,-i} + S\gamma^1}\Big) * \frac{n_{w_i}^{\zeta,-i} + \omega}{\sum_{w=1}^{W} n_w^{\zeta,-i} + W\omega} \quad (4)$$

$$y_i = 0 \Rightarrow \zeta = y_i$$
$$y_i = 1 \Rightarrow \zeta = a_d, z_i, s_i$$

For sampling sentiment, instead of using a single Dirichlet density we use a Dirichlet mixture as the prior (Sjölander et al., 1996; Smucker et al., 2005). It is a weighted combination of two individual Dirichlet densities χ and ξ. Mixture coefficients q_1, q_2 are set to 0.5, giving equal weights to both author and entity. This ensures that the chosen sentiment reflects both the entity's quality for that topic as well as the author's preferences.

3.4 Non-parametric Author-ATS (DP) Model

While the number of aspects for a domain are limited, the number of topics for each aspect may vary significantly and can be difficult to estimate. For restaurants, the topics for *ambiance* are fewer (e.g. music, crowd etc.) compared to *food*. This motivates us to propose a non-parametric version of the Author-ATS model where the number of topics can be automatically discovered.

In this non-parametric version, topic inference is done through Chinese Restaurant Process (CRP), a popular variant of Dirichlet Process (DP). In a Chinese restaurant with infinite number of tables, each with infinite capacity, CRP determines if a customer chooses to sit at an occupied table (with a probability proportional to the number of customers already sitting at the table), or an unoccupied one. Following the idea of CRP, each observed aspect dependent word can either be assigned to an existing topic or to a new topic. The conditional distributions for the Gibbs sampler are omitted due to space constraints.

4 SURF Framework

Given a target sentence in a review SURF computes its similarity with other review sentences using the distributions learned by Author-ATS and returns a list of supporting reviews.

A sentence supports another sentence if they are either lexically or semantically similar. Two sentences are **lexically similar** if they share keywords that are important for an aspect. Whereas two sentences can be **semantically similar** if they share the same sentiment for an aspect and topic even though they use different words. For example, *"The hotel was quite close to space needle"* and *"Major attractions are just walking distance from the hotel"* have high semantic similarity as they both talk about the same aspect 'location' on the topic 'attractions' with a positive sentiment.

We treat each review sentence as a vector and lexical similarity ($lexical\_sim$) is computed as cosine-similarity between the two vectors. The i^{th} entry of a vector signifies importance of the corresponding word to its assigned aspect computed using the *tf-idf* weighting scheme. We define the tf-idf of a word w w.r.t. an aspect a as:

$$tf(w, a) = \sum_{d=1}^{D} P(w|d, a)$$

$$P(w|d, a) = \begin{cases} P(w) & \text{if } w \text{ assigned to } a \text{ in } d \\ 0 & \text{otherwise} \end{cases}$$

$$idf(w, \mathbf{A}) = log \frac{A}{1 + |a \in \mathbf{A} : \exists d \in \mathbf{D}, P(w|d, a) > 0|}$$

$P(w)$ is the generation probability obtained from Author-ATS model. Since words are important with respect to an aspect, unlike traditional tf-idf, these values are computed across reviews on the whole corpus. Words frequently used for describing an aspect often tend to converge across reviews, even though written by different users.

Two sentences are considered semantically similar if they share the same sentiment for an aspect. Let C be the set of words in a sentence. Aspect-topic probability of a sentence is defined as the ratio of generation probability of words generated from the aspect-topic pair a, z to the summation of generation probabilities of all the words.

$$P(C|a,z) = \frac{\sum_{w \in C} P(w|w \text{ has aspect } a \text{ and topic } z)}{\sum_{w \in C} P(w)}$$

We define sim_0 to measure the similarities between two sentences (C_1 and C_2) having the same aspect, topic and sentiment, and sim_1 to measure the similarities of two sentences with the same aspect and sentiment but discussing different topics.

$$sim_0(C_1, C_2, a) = \sum_{z=1}^{Z} P(C_1|a,z)P(C_2|a,z)$$

$$sim_1(C_1, C_2, a) = \sum_{z_1, z_2 \in [1 \cdots Z] z_1 \neq z_2} P(C_1|a,z_1)P(C_2|a,z_2)$$

The semantic similarity between two sentences is:

$$semantic\_sim(C_1, C_2, a) = sim_0(C_1, C_2, a) + \delta sim_1(C_1, C_2, a)$$

where δ is a damping factor with value less than 1.

Lexical-semantic similarity (LSS) of two sentences with same sentiment for an aspect is measured as a weighted combination of their $lexical\_sim$ and $semantic\_sim$ as defined above.

Ranking of Reviews. Given a review sentence, we employ kNN search to find the k most similar sentences for each of its aspects according to LSS measure. Since a target sentence C may contain multiple aspects, we determine the importance of an aspect a to C as follows:

$$Imp(C,a) = \frac{\sum_{w \in C} P(w|w \text{ has aspect } a)}{\sum_{w \in C} P(w)}$$

For each aspect a with $Imp(C,a) > 0$, we return the top $k * Imp(C,a)$ sentences from the review corpus. Proportionately allocating supporting sentences from each aspect in the top-k results diversifies the result set and ensures that a user is able to find information about whichever aspect of the target sentence she wished to verify.

5 Experiments

We perform two sets of experiments to evaluate our proposed framework. We first compare Author-ATS with state-of-the-art topic models using perplexity on test data. Then we evaluate the performance of SURF, for the task of retrieving supporting opinions using human annotation, against keyword based search engine Lucene and a competent word embedding model Word2Vec. We use two real world datasets: (a) hotel reviews from TripAdvisor (Wang et al., 2010), and (b) restaurant reviews from yelp.com. Table 2 shows the statistics of the two datasets.

Dataset	# entity	# author	# review	# sentence	# vocab
TripAdvisor	12,773	781,403	1,621,956	20,244,293	980,323
Yelp	578	16,981	25,459	232,107	56,200

Table 2: Statistics of datasets used

We pre-process both datasets by removing domain independent stopwords[1]. We retain some negation stopwords (e.g.: *not, can't, didn't*) and join them with the next word (so that 'not good' is treated as a single unit) to help discover sentiment properly. We use common punctuations like '.', '?', '!' to split into sentences. To further split a sentence into segments we use punctuations used to separate clauses like ',', ';' and conjunctions like 'and', 'however', 'but' as separators. We use a domain independent subjectivity lexicon[2] to initialize sentiment distributions. Since aspect words may consist of highly co-occurring words (e.g. 'front-desk', 'walking distance') we use Pointwise Mutual Information (PMI) (Manning and Schütze, 1999) to find such collocations. Bigrams with PMI greater than a threshold (we use 0.05 in our experiments) are treated as a single word.

To make the discovered aspects understandable and intuitive, we provide a few seed words to the models. The seeds are only used during initialization and subsequent iterations of Gibbs sampling are not dependent on them. Table 3 lists the aspect seed words used for both domains.

Aspects	Seed Words
Value for Money	value, rate, price
Room	room, bed, bathroom, clean
Location	location, walk, minute
Service	staff, reservation, front-desk
Food	restaurant, breakfast, buffet
Amenities	pool, parking, internet, wifi

(a) TripAdvisor Dataset

Aspects	Seed Words
Value for Money	value, rate, portions, price
Service	ambience, wifi, music, service
Food	steak, rice, burger, cocktail

(b) Yelp Dataset

Table 3: Sample Aspect Seed Words

[1] http://www.ranks.nl/stopwords

[2] http://mpqa.cs.pitt.edu/lexicons/subj_lexicon

5.1 Evaluation of Author-ATS Model

In this set of experiments, we examine the ability of Author-ATS to capture the opinions in reviews.

Perplexity is derived from the likelihood of unseen test data and is a standard measure for evaluating topic models. The lower the perplexity, the less confused the model is on seeing new data, implying a better generalization power. We compare with the following state-of-the-art opinion models:

LDA (Blei et al., 2003) : A topic model where words are generated from a latent topic dimension.

TAM (Paul and Girju, 2010): A topic model for opinion mining where words are generated from a two-level hierarchy of aspect and topic.

JTV (Trabelsi and Zaiane, 2014): A topic model especially for contentious documents where each word has a topic and a viewpoint.

We also implement a baseline model **ATS** based on three-level Aspect-Topic-Sentiment hierarchy. We use this model to show the performance gain by just considering a hierarchical dependency between these dimensions while capturing an opinion. For Author-ATS and ATS, we use 6 aspects, 5 topics for each aspect and 3 sentiments. For fair comparison, we keep the total number of dimensions as close as possible across models. We partition our dataset into train (80%) and test (20%) sets and report five fold cross validation results.

Table 4 shows that ATS outperforms other models in both datasets due to its hierarchical modeling of words. Author-ATS further improves the performance by considering author and entity characteristics as well as the thought patterns of the authors. We note that the performance of the non-parametric model is comparable with Author-ATS, making it easier to use the model for any new domain without having much prior knowledge.

Model	**TripAdvisor**	**Yelp**
LDA	5070	5737
TAM	2980	3468
JTV	3430	4370
ATS	2385	3337
Author-ATS	**2212**	**2784**
Author-ATS(DP)	2300	2829

Table 4: Perplexity values for different models.

Table 5 shows the top words extracted by Author-ATS as domain stopwords. Although these words do not convey any aspect information, they are domain dependent and are not found in a general stopword dictionary.

From Table 6, we observe that the majority of the words are correctly clustered in aspects, and further into specific topics. For example, the first topic for aspect *Room* is about in-room experience ('bed','king-size','view'), whereas the second topic seems to be about bathroom ('shower', 'towels', 'tub'). We also observe that the model is able to obtain contextual sentiment terms which are aspect-topic coherent. For example, words such as 'noise', 'night', 'hear' could be assigned negative sentiment labels for topic 0 of *Room* due to the context in which they are used, e.g., when describing a room, these words probably indicate a noisy room bothering their sleep at night.

Dataset	Domain Stopwords
TripAdvisor	hotel, nice, stay, trip, times, day, place, back
Yelp	good, place, food, time, order, bit, make

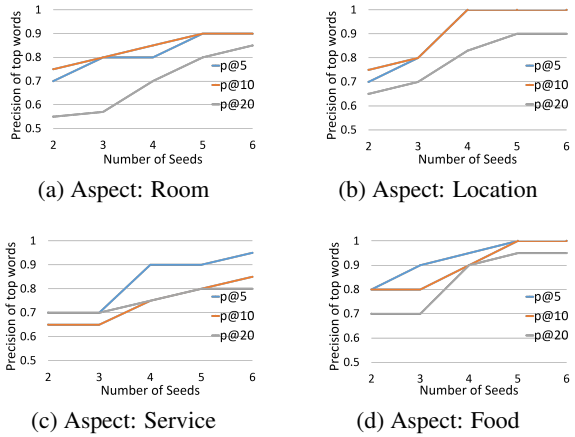
Table 5: Domain stopwords from Author-ATS.

Aspect: **Room**		
Topic 0		
Positive	Negative	Neutral
bed	noise	room
comfortable	night	floor
spacious	sleep	view
king-size	window	size
clean	hear	modern
Topic 1		
Positive	Negative	Neutral
bathroom	small	room
large	door	bathroom
tub	barely	shower
shower	tiny	water
shampoo	kitchen	towels

Aspect: **Service**		
Topic 0		
Positive	Negative	Neutral
staff	night	staff
extremely	greet	call
welcoming	problem	front-desk
care	asked	service
friendly	manager	shuttle
Topic 1		
Positive	Negative	Neutral
card	called	check-in
reservation	upgrade	day
airport	manager	arrived
polite	rude	directions
excellent	questions	time

Table 6: Top words for aspect-topic-sentiments found by Author-ATS for TripAdvisor dataset.

Impact of Seed Words We vary the number of seed words for an aspect and examine its effect on the aspect discovery. We use $p@n$, the fraction of correctly discovered aspect words among the top n words, to evaluate the quality of the results.



(a) Aspect: Room (b) Aspect: Location

(c) Aspect: Service (d) Aspect: Food

Figure 4: Impact of varying number of seeds.

The average precision of top-n words for different aspects is obtained by taking the average over all combinations $\binom{6}{m}$ of seed words where m is the number of selected seed words, $2 \leq m \leq 6$. Figure 4 shows the results. We observe that the

average precision increases with the number of seeds, and stabilizes when $m \geq 4$. This demonstrates that providing a handful of seed words can go a long way for discovering intended, explainable domain specific aspects.

5.2 Evaluation of SURF

We now evaluate Author-ATS model and LSS measure on retrieving sentences that are *relevant* to a target sentence. A sentence is considered relevant if it expresses similar opinions as the target sentence. A sentence with multiple aspects is relevant if it expresses at least one of the opinions in the target sentence. Precision of the top-k answers are manually determined by three annotators and conflicts are resolved by majority voting.

Recall that LSS considers both lexical and semantic similarity. The computation of semantic similarity requires the aspect-topic-sentiment distribution which is only available in the baseline ATS and Author-ATS models. We define a similarity measure called CJSD that can be used by the various topic models to facilitate comparison. CJSD measures the lexical similarity of two sentences as the cosine similarity of their tf-idf vectors, while the semantic similarity is measured by the similarity of their topic distributions using Jensen-Shannon Divergence(JSD) as follows:

$$CJSD(s_1, s_2) = \lambda\, cosine\_sim(s_1, s_2) + (1-\lambda)\, JSD(s_1, s_2)$$

We randomly select 5 hotels from TripAdvisor and 5 restaurants from Yelp datasets. For each hotel/restaurant, we randomly pick 10 target sentences and retrieve their supporting sentences.The topic distributions of these sentences are obtained using LDA, TAM, JTV, and the proposed models ATS and Author-ATS.

Table 7 shows the average precision for top 5, 10 and 20 results retrieved using various topic models with similarity measure CJSD. We see that Author-ATS model always outperforms other topic models for the task of retrieving supporting sentences. This is consistent with the perplexity results of the models obtained previously.

Table 8 shows the average precision using variants of the proposed model with LSS. Clearly, using LSS always yields a better precision compared to using CJSD, with the best performer being the Author-ATS with LSS combination. SURF framework utilizes this combination for retrieving top-k supporting reviews.

	TripAdvisor			Yelp		
	p@5	p@10	p@20	p@5	p@10	p@20
LDA	0.56	0.48	0.45	0.43	0.42	0.42
TAM	0.58	0.53	0.52	0.49	0.47	0.47
JTV	0.51	0.47	0.53	0.41	0.41	0.43
ATS	0.62	0.60	0.55	0.60	0.57	0.44
Author-ATS	**0.68**	**0.62**	**0.61**	**0.60**	**0.58**	**0.56**

Table 7: Average precision using CJSD

	TripAdvisor			Yelp		
	p@5	p@10	p@20	p@5	p@10	p@20
ATS	0.69	0.62	0.58	0.62	0.59	0.58
Author-ATS	**0.74**	**0.66**	**0.60**	**0.68**	**0.64**	**0.62**
Author-ATS (DP)	0.64	0.63	0.57	0.62	0.56	0.54

Table 8: Average precision using LSS

Next, we compare SURF with the following:

Lucene: A popular keyword based ranking method. We used its default combination of vector space model and boolean model for retrieval.

Word2Vec: (Mikolov et al., 2013) A state-of-the-art algorithm for word embeddings using neural network. Supporting sentences are ranked with Word Mover's distance using the word embeddings. We train on TripAdvisor dataset using CBOW algorithm with context window set to 5 as recommended by the authors. We do not train Word2Vec on the Yelp dataset as it is too small. We set the vector dimension to 500 based on grid search. We also compare with Word2Vec model pre-trained on the large GoogleNews dataset[3].

Table 9 shows the average precision for the top 5, 10 and 20 results retrieved using Lucene, Word2Vec and SURF. Word2Vec performs better when trained on review data, compared to the model trained on general news data. This confirms that domain knowledge is important. It is evident from the results that SURF significantly outperforms existing approaches for opinion search.

	p@5	p@10	p@20
Lucene	0.67	0.58	0.52
Word2Vec (GoggleNews)	0.62	0.48	0.39
Word2Vec (TripAdvisor)	0.70	0.61	0.51
SURF	**0.74**	**0.66**	**0.60**

(a) TripAdvisor

	p@5	p@10	p@20
Lucene	0.61	0.54	0.49
Word2Vec (GoogleNews)	0.52	0.47	0.37
SURF	**0.68**	**0.64**	**0.62**

(b) Yelp

Table 9: Comparison with Lucene and Word2Vec

For evaluating the coherence of retrieved set of supporting reviews for an aspect, we look at their corresponding user given aspect ratings. For each aspect of each review sentence, we retrieve its top-k supporting sentences. Then we compute the

[3]https://code.google.com/archive/p/word2vec/

Target Sentence : bedroom had the most comfortable mattress, feather soft pillows as well as a set of firmer ones, so they thought about keeping every guest comfortable		
Supporting Sentences by **SURF**	Supporting Sentences by **Lucene**	Supporting Sentences by **Word2Vec**
Aspect : **Room** **Statement**: bill clinton suite was huge with two baths, a wonderful jacuzzi and a comfortable bed	bed was very comfortable, as were the large pillows	The room had a microwave, coffemaker, hairdryer, bottled water replenished each day **(x)**
Aspect : **Room** **Statement**: the beds are the most comfortable of any hotel I have stayed in	we were recommending it for our out of town wedding guests, and wanted to make sure they were comfortable **(x)**	It really is a shame because the bed and pillows were super comfortable and we could have had a great night sleep on both nights
Aspect : **Room** **Statement**: the beds were comfortable and they had good selection of towels	who would have imagined that somebody actually thought about where a guest would watch tv **(x)**	I think they took regular sized hotel rooms and divided them into a sitting room with a bedroom with a door, keeping the bathroom to divide the two areas **(x)**

(a) Target Sentence with Single Aspect

Target Sentence : the check in was quick, with friendly polite service, and the room was very big with a very comfortable king size bed		
Supporting Sentences by **SURF**	Supporting Sentences by **Lucene**	Supporting Sentences by **Word2Vec**
Aspect : **Room** **Statement**: bed was extremely comfortable, I'm hard to please in the department because I sleep on a sleep number bed at home	the room was a great size; bed was very comfortable	The first room assigned was very small and dingy with one king sized bed that just fit **(x)**
Aspect : **Room** **Statement**: room size was large and bed was comfortable	king size bed was comfy	bathroom was well furnished with soap, shampoo/ conditioner, very large, soft towels - perfect **(x)**
Aspect : **Service** **Statement**: service is very friendly	our room faced denny park **(x)**	the room was large and the bed very comfortable and our room faced the street and it was very quiet

(b) Target Sentence with Multiple Aspects

Table 10: Sample Supporting Sentences Retrieved by SURF, Lucene and Word2Vec. Aspects shown for SURF are discovered by Author-ATS model.

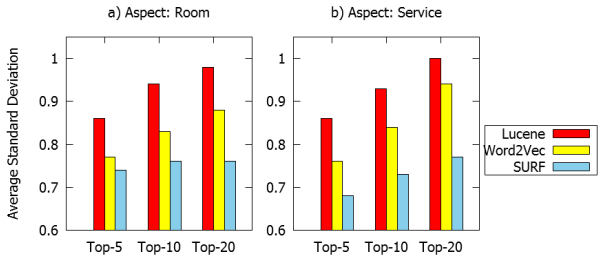


Figure 5: Average standard deviations of aspect ratings for supporting reviews. Smaller deviation implies greater coherence.

standard deviation of the ratings for that aspect in the retrieved supporting reviews. We aggregate the standard deviation values for each aspect over all the reviews and look at the average value. Figure 5 shows results for two aspects from the TripAdvisor dataset. Other aspects also had similar trends.

We rank the retrieved results based on their similarity to the target sentence. Naturally, the longer the retrieved list, the larger is the average standard deviation. We see that SURF has a smaller average standard deviation compared to Word2Vec and Lucene. The gap between the performance of SURF and the other methods also widens as the size of the retrieved results increases. This demonstrates SURF's superiority in retrieving reviews with similar opinions.

Table 10 shows samples of supporting sentences extracted by the different methods. We observe that the sentences retrieved by SURF are semantically similar although the words may be quite different from the target sentence. In contrast, Lucene may retrieve irrelevant sentences matching a keyword used in a totally different context. Word2Vec considers words used in proximity of one another (e.g. *bed, pillow* with *microwave, coffemaker* etc.) to be similar which clearly does not always imply conformity of opinions.

Furthermore, the retrieved results of SURF are categorized according to their aspects making them easy to interpret. Particularly if a target sentence has multiple aspects, then SURF will retrieve results for each aspect. For example, for the second target sentence shown in Table 10, the results contain supporting statements for both *room* and *service*. If a user then wishes to view more results for one of those aspects it will be possible for SURF to fetch more results only for that aspect.

6 Conclusion

We studied the problem of finding supporting sentences to help a user get an idea of consensus about an entity. To this end, we developed a hierarchical topic model to jointly infer aspect-topic-sentiment, and a fine-grained similarity measure. Author-ATS model encodes the coherent writing style of a review by constraining the aspect distributions dynamically. It considers the sentiment distribution of a review to have influence of both the author and the entity. Experimental results on two datasets indicate that the proposed approach is promising compared to existing techniques. With growing amount of user generated content on the web, and more people relying on them to make decisions, we believe that the ability to verify opinions will become increasingly important.

References

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*.

Zhiyuan Chen, Arjun Mukherjee, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. 2013. Exploiting domain knowledge in aspect extraction. In *Proc. of EMNLP*.

Lan Du, Wray Buntine, and Huidong Jin. 2010. Sequential latent dirichlet allocation: Discover underlying topic structures within a document. In *Proc. of IEEE ICDM*.

Lan Du, John K Pate, and Mark Johnson. 2015. Topic segmentation with an ordering-based topic model. In *Proc. of AAAI*.

Amit Gruber, Yair Weiss, and Michal Rosen-Zvi. 2007. Hidden topic markov models. In *International Conference on Artificial Intelligence and Statistics*.

Eduard H Hovy. 1993. Automated discourse generation using discourse structure relations. *Artificial intelligence*, 63(1):341–385.

Yohan Jo and Alice H Oh. 2011. Aspect and sentiment unification model for online review analysis. In *ACM International Conference on Web Search and Data Mining*.

Suin Kim, Jianwen Zhang, Zheng Chen, Alice H Oh, and Shixia Liu. 2013. A hierarchical aspect-sentiment model for online reviews. In *Proc. of AAAI*.

Fangtao Li, Sheng Wang, Shenghua Liu, and Ming Zhang. 2014. Suit: A supervised user-item topic model for sentiment analysis. In *Proc. of AAAI*.

Chenghua Lin and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proc. of ACM CIKM*.

Christopher D Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT press.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proc. of NIPS*.

David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proc. of EMNLP*.

Samaneh Moghaddam and Martin Ester. 2011. Ilda: interdependent lda model for learning latent aspects and their ratings from online product reviews. In *Proc. of SIGIR*.

Arjun Mukherjee and Bing Liu. 2012. Aspect extraction through semi-supervised modeling. In *Proc. of ACL*.

Michael Paul and Roxana Girju. 2010. A two-dimensional topic-aspect model for discovering multi-faceted topics. In *Proc. of AAAI*.

Lahari Poddar, Wynne Hsu, and Mong Li Lee. 2017. Quantifying aspect bias in ordinal ratings using a bayesian approach. In *Proc. of IJCAI*.

Md Mustafizur Rahman and Hongning Wang. 2016. Hidden topic sentiment model. In *Proc. of World Wide Web*.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. 2009. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proc. of EMNLP*.

Kimmen Sjölander, Kevin Karplus, Michael Brown, Richard Hughey, Anders Krogh, I Saira Mian, and David Haussler. 1996. Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Computer applications in the biosciences: CABIOS*.

Mark D Smucker, David Kulp, and James Allan. 2005. Dirichlet mixtures for query estimation in information retrieval. *Center for Intelligent Information Retrieval*.

Ivan Titov and Ryan McDonald. 2008a. Modeling online reviews with multi-grain topic models. In *Proc. of WWW*.

Ivan Titov and Ryan T McDonald. 2008b. A joint model of text and aspect ratings for sentiment summarization. In *Proc. of ACL*.

Amine Trabelsi and Osmar R Zaiane. 2014. Mining contentious documents using an unsupervised topic model based approach. In *Proc. of IEEE ICDM*.

Hongning Wang, Yue Lu, and Chengxiang Zhai. 2010. Latent aspect rating analysis on review text data: a rating regression approach. In *Proc. of SIGKDD*.

Zaihan Yang, Alexander Kotov, Aravind Mohan, and Shiyong Lu. 2015. Parametric and non-parametric user-aware sentiment topic models. In *Proc. of SIGIR*.

Wei Zhang and Jianyong Wang. 2015. Prior-based dual additive latent dirichlet allocation for user-item connected documents. In *Proc. of IJCAI*.

Tong Zhao, Chunping Li, Qiang Ding, and Li Li. 2012. User-sentiment topic model: refining user's topics with sentiment information. In *ACM SIGKDD Workshop on Mining Data Semantics*.