Parametric and Non-parametric User-aware Sentiment Topic Models

Zaihan Yang Alexander Kotov Aravind Mohan Shiyong Lu
Department of Computer Science
Wayne State University
Detroit, MI 48202 USA
{zaihan.yang|kotov|aravind.mohan|shiyong}@wayne.edu

ABSTRACT

The popularity of Web 2.0 has resulted in a large number of publicly available online consumer reviews created by a demographically diverse user base. Information about the authors of these reviews, such as age, gender and location, provided by many online consumer review platforms may allow companies to better understand the preferences of different market segments and improve their product design, manufacturing processes and marketing campaigns accordingly. However, previous work in sentiment analysis has largely ignored these additional user meta-data. To address this deficiency, in this paper, we propose parametric and non-parametric User-aware Sentiment Topic Models (USTM) that incorporate demographic information of review authors into topic modeling process in order to discover associations between market segments, topical aspects and sentiments. Qualitative examination of the topics discovered using USTM framework in the two datasets collected from popular online consumer review platforms as well as quantitative evaluation of the methods utilizing those topics for the tasks of review sentiment classification and user attribute prediction both indicate the utility of accounting for demographic information of review authors in opinion mining.

Categories and Subject Descriptors

H.2.8 [**Database Applications**]: Data mining; I.2.7 [**Natural Language Processing**]: Text analysis

Keywords

Opinion Mining; Topic Models; Dirichlet Process

1. INTRODUCTION

The emergence of online consumer review platforms, such as Amazon[1], Tripadvisor[2], and MSN Autos[3], allowed consumers to publicly express their opinions about a wide variety of products and services. The popularity of such platforms has resulted in large amounts of online review content created by a demographically diverse user base. Due to popularity and public availability, online consumer reviews have become an increasingly important and valuable source of information not only for consumers, who often base their decisions about purchasing a product or using a service on opinions of other people, but also for companies and manufacturers, who are trying to understand consumer preferences in different market segments and adjust their product design, manufacturing processes and marketing campaigns accordingly. However, the large volume of on-line reviews has made manual analysis and summarization of reviews, even for a single market segment, a very labor-intensive and time-consuming task. The need to automate such analysis gives rise to a novel problem of *summarization of contrasting opinions about aspects of products or services by different demographic groups of consumers*, which we introduce and address in this work.

Previous studies in opinion mining have largely focused on three major tasks: identification and extraction of opinion aspects, when the reviews are segmented into fine-grained aspects (topics); detection of sentiment polarity (positive, negative or neutral) towards these aspects; and summarization of aspects by sentiment polarity. Although recently proposed unsupervised topic models, such as Joint Sentiment Topic Model (JST) [17], Aspect and Sentiment Unification Model (ASUM) [8] as well as its hierarchical extension (HASM) [10], allow to summarize both the major aspects as well as the sentiments towards them in collections of on-line reviews, they ignore demographic information about review authors, such as their age, gender and location. However, such information may play an important role in opinion mining and sentiment analysis, since different demographic groups of consumers may have different opinions about the same product aspect.

Meta-data about review authors (e.g. location, gender and age), often provided by on-line consumer review platforms in user profiles, can be viewed as a discrete set of textual labels (or tags) associated with individual reviews, which introduce an element of supervision into sentiment summarization using topic models. Therefore, demographic groups of consumers (or market segments) in this setting can be defined as groups of on-line review authors sharing one or several user meta-data tags. For example, if such tags are organized along the dimensions of age (e.g. "18-25", "26-35", "35-50"), gender ("male" or "female") and location (e.g "san francisco", "new york", ...), then some of the possible market segments are "females", "people aged 18-25", "26-35 year-old males", "females living in new york", etc. In the simplest case, market segments correspond to a set of all distinct user meta-data tags in a given collection of reviews, however arbitrary market segments can be dynamically created by combining two or more tags depending on the required resolution of topic summaries. The textual con-

[1]http://www.amazon.com/
[2]http://www.tripadvisor.com/
[3]http://www.msn.com/en-us/autos/

SIGIR'15, August 09 - 13, 2015, Santiago, Chile.

ACM 978-1-4503-3621-5/15/08 ...$15.00.
DOI: http://dx.doi.org/10.1145/2766462.2767758 .

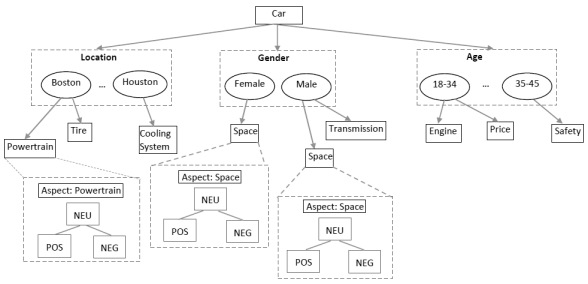


Figure 1: **Topical structure of automotive reviews with respect to market segments defined by distinct user meta-data tags.**

tent of reviews can be analyzed to identify specific product aspects, while the user meta-data can be used to jointly determine collective preferences of different market segments. We hypothesize that establishing associations between the content of reviews and demographic properties of review authors can facilitate fine-grained understanding of product adoption by different customers, from which the companies can benefit by reshaping their product development, marketing and consumer relationship strategies. Since each review may have multiple associated user meta-data tags and there can be a large number of such tags (and corresponding market segments), summarization of reviews across all market segments can be challenging.

In this work, we propose User-Aware Sentiment Topic Models (USTM for short), a framework for modeling user meta-data, topical aspects and sentiments in a unified way. Each of the topic models in the USTM framework identifies several topical aspects frequently discussed in reviews by each market segment. Each aspect is in turn a two-level topical hierarchy, the first level of which corresponds to the summary of objective comments related to this aspect and its positive and negative subtopics constitute the second level. Figure 1 provides an example of the topical structure identified by USTM in automotive reviews with respect to market segments defined by distinct user meta-data tags grouped into the demographic dimensions of location, gender and age.

The proposed USTM framework includes the following 4 topic models:

- User-aware Sentiment Topic Model with Fixed number of Topics and Sentence-based sentiment assignment (**USTM-FT(S)**) is a parametric topic model that extends the Partially Labeled Topic Model [24] to jointly model market segments defined by user meta-data tags, topical aspects and sentiment-based subtopics in reviews based on the assumption that each market segment is associated with the same pre-defined number of topical aspects. Similar to ASUM [8] and HASUM [10], USTM-FT(S) also assumes that all words within the same review sentence can be associated with different topics, but can either be neutral or assigned to the *sentiment selected for the entire sentence*;

- The User-aware Sentiment Topic Model with Fixed number of Topics and Word-based sentiment assignment (**USTM-FT(W)**) is a parametric topic model that is different from USTM-FT(S) in that it assumes that *sentiment is associated with each individual word* rather than an entire sentence;

- **USTM-DP(S)** is a non-parametric alternative to USTM-FT(S), which is based on the Dirichlet Process and allows *different number of topical aspects per each market segment*;

- **USTM-DP(W)** is a non-parametric alternative to USTM-FT(W), which is based on the Dirichlet Process and assumes that *sentiment is associated with each individual word rather than an entire sentence*.

Overall, the key contributions of this work are two-fold:

1. we propose novel parametric and non-parametric topic models for market segment-based opinion summarization, which incorporate demographic information about the authors of reviews in the form of textual user meta-data tags and linguistic information in the form of asymmetric sentiment priors;
2. we experimentally demonstrate the effectiveness of the proposed topic models for opinion mining as well as for the tasks of sentiment classification and user attribute prediction on two real world data sets.

The remainder of this paper is organized as follows. Section 2 provides a brief overview of previous relevant work. All topic models within USTM framework are presented in Section 3. Experimental results are reported in Section 4 and Section 5 concludes the paper.

2. RELATED WORK

The topic models proposed in this paper build upon the previous work along the following two research directions.

Aspect-based sentiment analysis. In the past decade, several machine learning methods [6, 23, 14] have been proposed for opinion mining and sentiment analysis at the word/phase, sentence and document levels. In recent years, there is a surging interest in aspect-based sentiment analysis, which aims at extracting aspects of entities commented on in reviews and opinions towards them. A majority of the previously proposed approaches for review aspect discovery rely on natural language processing techniques such as dependency relations [6], supervised sequence labeling [7] and centering theories [19]. However, some of these approaches are either supervised and require training data in the form of manually identified aspects or aspect terms, or they cannot group extracted terms from multiple reviews into high-level cross-user latent topics (aspects).

More recently, probabilistic topic models have become the main tool for aspect-based opinion mining due to their ability to identify and concisely represent latent topics in collections of reviews. Several topic models for opinion analysis extending the basic topic models have been proposed. The initial work in this direction is the Topic-Sentiment Model (TSM) [20] which considers each document as a mixture of topics and sentiments. However, TSM is based upon PLSA [5] and, thus, is prone to overfitting. Titov and McDonald proposed the MG-LDA [28] model and a further extended MAS [27] model both of which aim to determine the local and global topics from reviews with structured aspects and numeric rating associated with each aspects. Both MG-LDA and MAS assume that at least one aspect is rated in one review, which is impractical; Lin et al. proposed the LDA-based Joint-Sentiment/Topic model (JST) [17], which assumes that each sentiment has a multinomial distribution over topics, and that each sentiment-topic pair has a multinomial distribution over words. This model, however, cannot accurately distinguish the different sentiments of each topic. Lin et al. later proposed a Reverse JST model (R-JST) [18] by reversing the association between sentiments and topics in JST, which, however, performed poorly. Jo et al. [8] recently developed the Aspect-Sentiment Unification Model (ASUM), which is based on the assumption that all words in one sentence are associated with the same topic and sentiment. Mukherjee et al. [22] proposed a Seeded

Aspect and Sentiment Model, which discovers aspect-based sentiments given sets of seed words for aspect categories. Sauper et al. [25] developed a topic model to jointly identify properties and attributes of review snippets rather than complete reviews. Wang et al. [30] introduced Latent Aspect Rating Analysis, a new aspect-level sentiment analysis task aiming to discover both topical aspects and each individual reviewer's latent rating for each aspect. Moghaddam [21] et al. addressed the same latent rating prediction problem and proposed three different versions of topic models to solve it. Nevertheless, most of the previously proposed topic modeling approaches for aspect-level sentiment analysis largely ignore other valuable supportive information, such as the meta-data of review authors.

Several previous works [26, 4, 31, 16, 11] utilized user information in opinion mining and information retrieval, although in a different way from this work. In particular, [26] and [4] are not topic modeling-based approaches, while [31] and [16] model users as random variables and not the attributes of users in the topic modeling process. However, all these approaches are parametric and require to specify a pre-defined number of topics (aspects) per review. To overcome this limitation, USTM framework includes non-parametric topic models based on Dirichlet Process. The work that is the closest to ours is [10] Hierarchical Aspect Sentiment Unification Model (HASUM), which extends the ASUM model by integrating it with the recursive Chinese Restaurant Process [9], a modified version of the nested Chinese Restaurant Process [1], thus allowing to identify hierarchical aspect-sentiment structure. However, HASUM does not consider user meta-data. Li [15] et al. also explored the structure in on-line reviews.

Supervised/partially supervised topic models. Many researchers have recently directed their attention to incorporating supervision in the form of additional meta-data associated with textual content into topic models. Supervised LDA [2] extended the traditional Latent Dirichlet Allocation [3] by adding a response variable associated with each document. Partially Labeled Topic Model [24] is based on the assumption that if a document is associated with a set of labels (tags), then those tags play a direct role in generating its content. In particular, PLDA and PLDP introduce an additional layer of latent variables that determine associations of each word with a document tag and topic. Topic models incorporating the locations of Twitter users extracted from their profiles have been shown to improve microblog retrieval in [12] and [13]. Although these models shed some light on how to incorporate useful meta-data into topic modeling process, none of them has been applied to aspect-based sentiment analysis. Therefore, the USTM framework proposed in this paper can be viewed as an extension and unification of the previous work on supervised/partially supervised topic modeling and aspect-based sentiment analysis.

3. USTM FRAMEWORK

In this section, we introduce the four different topic models that constitute the framework for User-Aware Sentiment Topic Modeling. Table 1 provides a summary of notations used in our discussion of the topic models in the USTM framework.

Given a collection $C = \{d_1, \ldots, d_m\}$ of M reviews, in which each review consists of $\{w_1, \ldots, w_{N_{dw}}\}$ words and is associated with $\{t_1, \ldots, t_{N_{dt}}\}$ user meta-data attributes (tags), the primary goal of topic models in USTM framework is to discover associations between user meta-data attributes, topical aspects and sentiments associated with those aspects across different demographic groups of users. In the following sections, we discuss how parametric and non-parametric topic models in USTM framework achieve this goal.

Table 1: **Summary of notations used in this paper**

Symbol	Description
	Counts
M	number of reviews
W	vocabulary size
T	number of distinct user meta-data tags (attributes)
K	number of topics associated with each tag (parametric models)
S	number of sentiment polarities
K_t	number of topics associated with tag t (non-parametric models)
N_{dw}	the number of words in review d
N_{dt}	the number of tags in review d
N_{ds}	the number of sentences in review d
N^w_{di}	the number of words in the sentence i of review d
	Random Variables
ψ	proportion of review words assigned to different tags
θ	proportion of review words assigned to different tag-specific topics
ϕ	sentiment-specific topics for each user meta-data attribute (tag)
ξ	trinomial/binomial distribution of sentiments for tag-specific topics
ϑ	binomial distribution of (review, tag, topic) triples over sentiments
t_{di}	tag assigned to word i in review d
z_{di}	topic assigned to word i in review d
p_{di}	subjectivity assigned to word i in review d
s^s_{di}	sentiment assigned to sentence i of review d
s^w_{dij}	sentiment assigned to word j in sentence i of review d
	Hyper-Parameters
α	Dirichlet prior for θ
β^w_s	weight of word w in the Dirichlet prior for the topics with sentiment s
$\hat{\beta}_s$	sum of the weights of all words in the Dirichlet prior for the topics with sentiment s
β	hyper-parameter of the uniform Dirichlet prior
η	hyper-parameter of the uniform Dirichlet prior for ψ
γ	hyper-parameter of the uniform Dirichlet prior for ξ
δ	hyper-parameter of the uniform Dirichlet prior for ϑ

3.1 Parametric models

3.1.1 USTM-FT(W)

USTM-FT(W) extends the PLDA model [24] by jointly modeling user meta-data attributes and sentiments in a generative process. In particular, USTM-FT(W) associates a multinomial distribution over topics with each tag (or a combination of tags) from a set of tags $\mathbf{\Lambda_d}$ for d, a trinomial distribution over sentiments (neutral, positive or negative) with each tag-specific topic and a multinomial distribution over words with sentiment-specific topics for each tag. It generates the reviews according to the following generative process:

1. for each tag t, topic z and sentiment s, draw a distribution over words $\phi_{tzs} \sim Dir(\beta_s)$
2. for each review d:
   (a) draw a distribution over tags $\psi_d \sim Dir(\eta)$
   (b) for each tag $t \in \mathbf{\Lambda_d}$, draw a distribution over topics $\theta^t_d \sim Dir(\alpha)$
   (c) for each pair (t, z), draw a distribution over sentiments $\xi^{tz}_d \sim Dir(\gamma)$
   (d) for each word position i in d:
      i. draw a tag $t \sim \psi_d$
      ii. for the sampled tag t, draw a topic $z \sim \theta^t_d$
      iii. for the sampled tag t and topic z, draw a sentiment $s \sim \xi^{tz}_d$

iv. draw a word $w \sim \phi_{tzs}$ from the topic corresponding to the sampled tag t, topic z and sentiment s

The graphical model for USTM-FT(W) in plate notation is presented in Figure 2. Since this model is parametric, each user metadata tag is associated with a fixed number of topics, which is a parameter that needs to be specified a priori. Sentiment information is incorporated into USTM-FT(W) using asymmetric Dirichlet priors for positive and negative sentiment-specific topics, in which each word w in the corpus vocabulary is assigned a weight β_s^w ($s \in 0, 1, 2$), which can be pre-defined or learned off-line by bootstrapping from a set of seed words for the corresponding sentiment polarity s. In general, positive sentiment words will be assigned larger weight in the prior for positive topics than in the prior for negative topics and vice versa. Posterior inference of USTM-FT(W) model parameters is done using Gibbs sampling. At each state of the Markov chain for the Gibbs sampler, the latent tag $t_{d,i}$, topic $z_{d,i}$ and sentiment $s_{d,i}$ are sampled for each word i in review d according to the following formula:

$$P(t_{d,i}=j, z_{d,i}=k, s_{d,i}=s | t_{\neg d,i}, z_{\neg d,i}, s_{\neg d,i}, \alpha, \gamma, \beta_s^w, \hat{\beta}_s)$$
$$\propto \frac{n_{d,j,k,\cdot,\cdot}^{(\neg d,i)} + \alpha}{n_{d,\cdot,\cdot,\cdot,\cdot}^{(\neg d,i)} + N_{dt} * K * \alpha} \cdot \frac{n_{\cdot,j,k,\cdot,w}^{(\neg d,i)} + \beta}{n_{\cdot,j,k,\cdot,\cdot}^{(\neg d,i)} + V * \beta} \cdot \frac{n_{d,j,k,s,\cdot}^{(\neg d,i)} + \gamma}{n_{d,j,k,\cdot,\cdot}^{(\neg d,i)} + 3 * \gamma} \times$$
$$\times \frac{n_{\cdot,j,k,s,w}^{(\neg d,i)} + \beta_s^w}{n_{\cdot,j,k,s,\cdot}^{(\neg d,i)} + \hat{\beta}_s}$$

where $n_{d,j,k,\cdot,\cdot}^{(\neg d,i)}$ is the total number of words in review d that have been assigned to tag j and topic k; $n_{d,\cdot,\cdot,\cdot,\cdot}^{(\neg d,i)}$ is the total number of words in review d; $n_{\cdot,j,k,\cdot,w}^{(\neg d,i)}$ is the total number of times the word w has been assigned to tag j and topic k in the entire corpus; $n_{\cdot,j,k,\cdot,\cdot}^{(\neg d,i)}$ is the total number of words in the entire corpus that have been assigned to tag j and topic k; $n_{d,j,k,s,\cdot}^{(\neg d,i)}$ is the total number of words in review d that have been assigned to tag j, topic k and sentiment s; $n_{\cdot,j,k,s,w}^{(\neg d,i)}$ is the total number of times the word w has been assigned to tag j, topic k, and sentiment s in the entire corpus; $n_{\cdot,j,k,s,\cdot}^{(\neg d,i)}$ is the total number of words in the entire corpus that have been assigned to tag j, topic k, and sentiment s. All these counts exclude the word i for which the associated tag, topic and sentiment are being sampled.

After sampling is complete, the distributions for latent variables θ, ϕ, ξ are calculated as follows:

$$\theta_{djk} = \frac{n_{d,j,k,\cdot,\cdot} + \alpha}{n_{d,j,\cdot,\cdot,\cdot} + K * \alpha} \tag{1}$$

$$\phi_{jksw} = \frac{n_{\cdot,j,k,s,w} + \beta_s^w}{\sum_{w=1}^{V} n_{\cdot,j,k,s,w} + \hat{\beta}_s} \tag{2}$$

$$\xi_{djks} = \frac{n_{d,j,k,s,\cdot} + \gamma}{n_{d,j,k,\cdot,\cdot} + 3 * \gamma} \tag{3}$$

3.1.2 USTM-FT(S)

Previous studies [8, 10] indicate that assigning sentiment to the entire sentence rather than each individual word might be a better strategy for opinion mining. Following this idea, we propose USTM-FT(S), which is based on the assumption that all the words in a given sentence have the same sentiment, but can be associated with different tags and topics. We also distinguish the subjectivity of each word, which indicates whether a word is a topic word (e.g. "car", "hotel", "engine", "breakfast", etc.) or a sentiment word (e.g. "great", "wonderful", "awful", etc.). Therefore, given the sentiment $s_{d,m} \in \{1, 2\}$ (1 = positive and 2 = negative) assigned to the sentence m in document d and the subjectivity $p_{d,m,i} \in \{0, 1\}$ assigned to the word i in sentence m, the sentiment of this word $s_{d,m,i} \in \{0, 1, 2\}$ (0 = neutral, 1 = positive and 2 = negative) is determined as $s_{d,m} * p_{d,m,i}$. USTM-FT(S) generates each review according to the following generative process:

1. for each tag t, topic z and sentiment s, draw a distribution over words $\phi_{tzs} \sim Dir(\beta_s)$
2. for each review d:
   (a) for each tag $t \in \mathbf{\Lambda_d}$, draw a distribution over topics $\theta_d^t \sim Dir(\alpha)$
   (b) for each pair (t, z), draw subjectivity distribution over words ϑ_d^{tz}
   (c) for sentence m in review d:
      i. draw a sentiment $s \sim \xi_d^{tz}$
      ii. for each word i in sentence m of review d:
         A. draw a tag $t \sim \psi_d$
         B. for the sampled tag t, draw a topic $z \sim \theta_d^t$
         C. for the sampled tag t and topic z, draw a subjectivity $p \sim \vartheta_d^{tz}$
         D. draw a word $w \sim \phi_{tz\{s*p\}}$ from the topic corresponding to the sampled tag t, topic z and sentiment $s * p$

The graphical model for USTM-FT(S) in plate notation is presented in Figure 3. Each state of the Markov chain for the Gibbs sampler used for posterior inference of parameters of USTM-FT(S) consists of two steps. In the first step, we sample the sentiment for each sentence in a review based upon the tag, topic and subjectivity assignments to each word in the sentence in the previous iteration of the Gibbs sampler. In particular, the probability of choosing sentiment s for sentence m in review d can be computed by multiplying the probabilities of assigning every word i in sentence m to the sentiment s:

$$P(s_{d,m}^s = s) = \frac{n_{d,s,\cdot}^{(\neg d,m)} + \gamma}{n_{d,\cdot,\cdot}^{(\neg d,m)} + 2 * \gamma} \cdot \prod_{w \in s_{d,m}} P(w_{d,m,i} | t_{d,m,i}, z_{d,m,i}, \tag{4}$$

$$w_{d,m,i}^s) = \frac{n_{d,s,\cdot}^{(\neg d,m)} + \gamma}{n_{d,\cdot,\cdot}^{(\neg d,m)} + 2 * \gamma} \cdot \prod_{w \in s_{d,m}} \frac{n_{\cdot,j,k,s*p,w} + \beta_{s*p}^w}{n_{\cdot,j,k,s*p,\cdot} + \hat{\beta}_{s*p}}$$

where $n_{d,s,\cdot}^{(\neg d,m)}$ indicates the number of sentences, which have been assigned sentiment s in review d.

In the second step, based on the chosen sentiment for the entire sentence, we further sample the latent tag $t_{d,m,i}$, topic $z_{d,m,i}$ and sentiment $s_{d,m,i}$ for each word in that sentence, according to the following formula:

$$P(t_{d,m,i}=j, z_{d,m,i}=k, s_{d,m,i}=(s*p) | t_{\neg d,i}, z_{\neg d,i}, s_{\neg d,i}, \alpha, \gamma, \beta_{s*p}^w,$$
$$\hat{\beta}_{s*p}) \propto \frac{n_{d,j,k,\cdot,\cdot}^{(\neg d,i)} + \alpha}{n_{d,\cdot,\cdot,\cdot,\cdot}^{(\neg d,i)} + N_{dt} * K * \alpha} \cdot \frac{n_{d,j,k,s*p,\cdot}^{(\neg d,i)} + \delta}{n_{d,j,k,\cdot,\cdot}^{(\neg d,i)} + 2 * \delta} \cdot \frac{n_{\cdot,j,k,\cdot,w}^{(\neg d,i)} + \beta}{n_{\cdot,j,k,\cdot,\cdot}^{(\neg d,i)} + V * \beta} \times$$
$$\times \frac{n_{\cdot,j,k,s*p,w}^{(\neg d,i)} + \beta_{s*p}^w}{n_{\cdot,j,k,s*p,\cdot}^{(\neg d,i)} + \hat{\beta}_{s*p}}$$

After sampling is complete, the distributions for latent variables θ, ϕ, ξ are calculated similar to Equations 1, 2 and 3.

3.2 Non-parametric models

3.2.1 Dirichlet Process

A major limitation of parametric USTM models is that they require to specify a fixed number of topics per each market segment

a priori, while such number cannot be easily estimated. Moreover, although it is possible to empirically determine the optimal setting for the number of topics by optimizing an evaluation metric (e.g., perplexity), such an approach is generally impractical. To overcome this deficiency, we propose two non-parametric topic models, USTM-DP(W) and USTM-DP(S), which build upon the Dirichlet Process (DP for short) and allow to automatically discover the latent topical structure in collections of reviews annotated with user meta-data.

In non-parametric Bayesian statistics, a Dirichlet process is a method of assigning a probability distribution over other probability distributions. Given a Dirichlet Process $DP(H, \alpha)$ which is characterized by a base distribution (or a base measure) H and a concentration parameter α, a draw $G \sim DP(H, \alpha)$ will return a random distribution over some values that can be drawn from H. If we further draw a parameter $\theta_i \sim G$ and use it as a prior for a mixture model, we get the Dirichlet Process mixture model (DPM), from which we can draw observed data points.

The Chinese Restaurant Process (CRP) metaphor is one typical representation of the Dirichlet Process, which generates partitions of variables that exhibit the same clustering structure as the one created by the Dirichlet Process. The CRP process can be described as assignment of dining tables to new customers, who enter a restaurant with an infinite number of tables. In the initial state, all the tables are empty, and the probability of the ith customer, z_i, who enters the restaurant, to choose the tth table is:

$$p(z_i = t) = \begin{cases} \frac{n_t}{\sum_t n_t + \alpha}, & \text{for an existing table} \\ \frac{\alpha}{\sum_t n_t + \alpha}, & \text{for a new table} \end{cases} \quad (5)$$

where n_t indicates the number of customers who are sitting at the table t. Following the same idea, we can assign each observed word either to a new or to an existing topic (the number of topics can be infinite).

3.2.2 USTM-DP(W) and USTM-DP(S)

USTM-DP(W) and USTM-DP(S) are the non-parametric counterparts of USTM-FT(W) and USTM-FT(S), in which the LDA-based topic inference is replaced by the Dirichlet Process. In both of these models, a word w is assigned to a user meta-data tag and one of its sentiment-specific topics in proportion to how often other words have been assigned to the given tag, topic and sentiment, or to a new sentiment-specific topic created for some tag in proportion to the concentration parameter α.

Similar to USTM-FT(W), USTM-DP(W) assigns sentiment to each individual word. The following Gibbs Sampler updating formula shows the probability of choosing the tag j, topic k and sentiment s for the i^{th} word in review d:

$$P(t_{d,i} = j, z_{d,i} = k, s_{d,i} = s | t_{\neg d,i}, z_{\neg d,i}, s_{\neg d,i}, \alpha, \gamma, \beta_s^w, \hat{\beta}_s)$$
$$\propto \begin{cases} \frac{n_{d,j,k,\cdot,\cdot}^{(\neg d,i)}}{n_{d,\cdot,\cdot,\cdot,\cdot}^{(\neg d,i)} + \alpha} \cdot \frac{n_{d,j,k,s,\cdot}^{(\neg d,i)} + \gamma}{n_{d,j,k,\cdot,\cdot}^{(\neg d,i)} + 3*\gamma} \cdot \frac{n_{\cdot,j,k,\cdot,w}^{(\neg d,i)} + \beta}{n_{\cdot,j,k,\cdot,\cdot}^{(\neg d,i)} + V*\beta} \cdot \frac{n_{\cdot,j,k,s,w}^{(\neg d,i)} + \beta_s^w}{n_{\cdot,j,k,s,\cdot}^{(\neg d,i)} + \hat{\beta}_s}, \\ \text{for an existing topic} \\ \frac{\alpha}{n_{d,\cdot,\cdot,\cdot,\cdot}^{(\neg d,i)} + \alpha} \cdot \frac{n_{d,j,\cdot,s,\cdot}^{(\neg d,i)} + \gamma}{n_{d,j,\cdot,\cdot,\cdot}^{(\neg d,i)} + 3*\gamma} \cdot \frac{\beta}{V*\beta} \cdot \frac{\beta_s^w}{\hat{\beta}_s}, \text{ for a new topic} \end{cases}$$

USTM-DP(S) assigns sentiment on a sentence-based level. Similar to USTM-FT(S), it first samples a sentiment for the ith sentence in review d based on Equation 4 and then samples the tag $t_{d,m,i}$, topic $z_{d,m,i}$ and sentiment $s_{d,m,i}$ for each word in that sentence, according to the following formula:

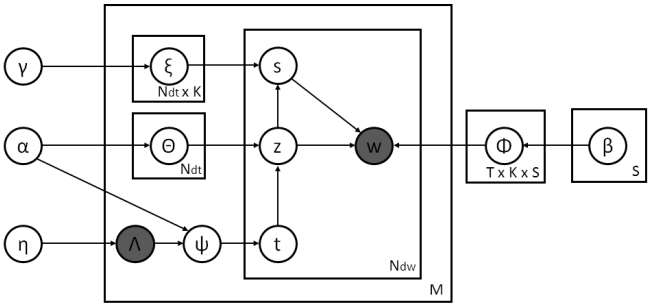


Figure 2: **Graphical model of USTM-FT(W).**

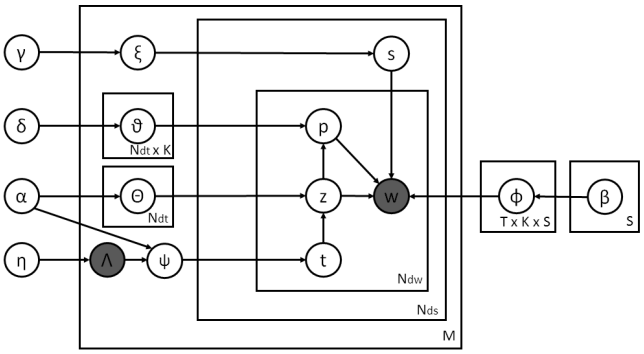


Figure 3: **Graphical model of USTM-FT(S).**

$$P(t_{d,m,i} = j, z_{d,m,i} = k, s_{d,m,i} = (s*p) | t_{\neg d,i}, z_{\neg d,i}, s_{\neg d,i}, \alpha, \gamma, \beta_s^w, \hat{\beta}_s)$$
$$\propto \begin{cases} \frac{n_{d,j,k,\cdot,\cdot}^{(\neg d,i)}}{n_{d,\cdot,\cdot,\cdot,\cdot}^{(\neg d,i)} + \alpha} \cdot \frac{n_{d,j,k,s*p,\cdot}^{(\neg d,i)} + \delta}{n_{d,j,k,\cdot,\cdot}^{(\neg d,i)} + 2*\delta} \cdot \frac{n_{\cdot,j,k,\cdot,w}^{(\neg d,i)} + \beta}{n_{\cdot,j,k,\cdot,\cdot}^{(\neg d,i)} + V*\beta} \cdot \frac{n_{\cdot,j,k,s*p,w}^{(\neg d,i)} + \beta_{s*p}^w}{n_{\cdot,j,k,s*p,\cdot}^{(\neg d,i)} + \hat{\beta}_{s*p}}, \\ \text{for an existing topic} \\ \frac{\alpha}{n_{d,\cdot,\cdot,\cdot,\cdot}^{(\neg d,i)} + \alpha} \cdot \frac{n_{d,j,\cdot,s*p,\cdot}^{(\neg d,i)} + \delta}{n_{d,j,\cdot,\cdot,\cdot}^{(\neg d,i)} + 2*\delta} \cdot \frac{\beta}{V*\beta} \cdot \frac{\beta_{s*p}^w}{\hat{\beta}_{s*p}}, \text{ for a new topic} \end{cases}$$

4. EXPERIMENTAL EVALUATION

Experimental evaluation of the proposed topic models was performed according to the following four aspects: perplexity on the test data, qualitative analysis of the discovered topics, review sentiment classification and prediction of demographic attributes of review authors. The results of an experimental evaluation for each of these aspects are provided below.

4.1 Experimental setup

4.1.1 Data sets

Experimental evaluation of the proposed models was conducted on two real world data sets. This first data set (further referred to as **Auto**) consists of reviews crawled from FordForum.com[4], a public automobile on-line review website, which provides the meta-data of review authors, such as location, gender and occupation (in this work, we are only interested in location and gender). The second data set (further referred to as **Hotel**) consists of reviews of hotels crawled from TripAdvisor[5] along with the meta-data of review authors, such as location, gender and age[6].

[4]http://www.fordforum.com/forum/

[5]http://www.tripadvisor.com

[6]both datasets are available at https://github.com/teanalab/USTM

Table 2: **Summary of statistics of experimental datasets.**

Dataset	# reviews	# tags	voc. size	# tokens	avg. len
Auto	11254	401	4952	362,225	32.19
Hotel	7266	324	6430	441,489	60.75

Table 3: **Most popular demographic attributes in experimental datasets. The number in parenthesis indicates the number of reviews associated with the label. Only these labels were considered in experiments.**

Dataset	Locations	Gender	Age groups
Hotel	london, uk (288) new york, new york (184) toronto, canada (119) boston, massachusetts (116) sydney, australia (101)	female (2285) male (1853)	35-49 (1788) 25-34 (1076) 50-64 (1006) 65 (145) 18-24 (112)
Auto	de pere, wi (1535) yorkshire, ny (1263) denver, colorado (577) iowa (442) fresno, ca (134)	male (2335) female (183)	

For both of these datasets, we performed pre-processing by lowercasing all words, removing stop-words and retaining only those words, which appear in the English dictionary of the spell-checking program *aspell*. Since we intend to discover sentiment topics, we kept certain stopwords like *not*, *don't*, *doesn't* and *won't*. We consider only those reviews in each data set, which are associated with at least one user label. From those reviews, we further filtered out the words that are either too rare (i.e. appear less than 5 times in the entire corpus) or too common (i.e. appear in more than 30% of the reviews). Sentences were segmented based on punctuation ('.', '?', '!' and new line). Various statistics of experimental datasets are summarized in Table 2.

All experimental results reported in this paper were obtained by considering user meta-data tags belonging to three dimensions (location, gender and age). To avoid the sparsity issue, in addition to the gender tags, we considered the top 5 location tags in both data sets as well as the top 5 age tags in the **Hotel** dataset in terms of the number of reviews associated with them (there is no age information for the review authors in the **Auto** data set). Statistics of the user meta-data tags considered in the experiments reported in this work are presented in Table 3 .

4.1.2 Sentiment seed words and asymmetric priors

Sentiment information is incorporated into our models via asymmetric Dirichlet priors (β) for sentiment-specific topics. For example, the words with strongly positive sentiment polarity like "good", "better" or "great" will have the higher weights in the Dirichlet prior for positive topics than for negative topics, so that these words will have a higher probability to be sampled for positive aspects of reviews.

In order to calculate the weights for words in the asymmetric priors for sentiment-specific topics, we utilized the sentiment lexicons from the previous studies as seed words with strong sentiment polarity. In particular, we used the **PARADIGMhasm** sentiment lexicon [10] consisting of 31 positive and 33 negative words, and the **MPQA** [29] sentiment lexicon consisting of 2718 positive words and 4911 negative words. After filtering out the words from the MPQA lexicon, which are not in the vocabulary of our datasets, we obtained sentiment lexicons consisting of 341 positive and 356 negative words for the **Auto** dataset, and of 671 positive and 572 negative words for the **Hotel** dataset.

Table 4: **Summary of statistics of considered bigrams.**

Datasets	# bi-grams	# pos. HASM	# neg. HASM	# pos. MPQA	# neg. MPQA
Auto	27791	785	1114	3511	3099
Hotel	7751	924	88	1596	477

We first set the neutral, positive and negative priors for all words in the corpus to 0.01. Then for each word found in the known sentiment lexicon with polarity s, we set its weight in the Dirichlet prior for the topics with the same sentiment polarity β_s^w to 2.0, and its weight in the Dirichlet prior for the topics with the opposite sentiment polarity to 0.001. For example, using this approach, the weights of a word that belongs to a positive sentiment lexicon in the Dirichlet priors for neutral, positive and negative topics will be 0.01, 2.0, 0.001, respectively. All results reported in this work were obtained by setting the hyper-parameters γ, η, and δ to 0.1, β to 0.01, α to $50/K$ and 10^{-8} for parametric and non-parametric models, respectively.

4.1.3 Unigrams vs. Bigrams

Bigrams are generally more informative than unigrams in expressing sentiments for topics. For example, if the fragment of a review "this hotel has a great location" is represented as a unigram bag of words {"hotel", "great", "location"}, it is hard to tell whether the hotel is great or the location is great. The bigram "great location", however, clearly indicates positive sentiment about the aspect of "location". In order to compare the different representations of topics, we replaced individual words in reviews with bigrams and ran the topic models in the USTM framework on the resulting collection. To create bigrams, we combined all pairs of consecutive words in reviews and retained only those bigrams, which appear in more than five and less than 30% of all reviews in each collection. Table 4 shows the statistics of considered bigrams in both collections. To determine the weight of each bigram in the Dirichlet priors for topics of different sentiment polarity, we assumed that if only one word in the bigram belongs to the known sentiment lexicon and it is not preceded by a negation like "not", then the weight of the entire bigram is determined based on this lexicon. If a bigram is preceded by a negation, then the weight is set based on the sentiment lexicon of the opposite sentiment polarity. If both terms in the bigram belong to the same subjectivity lexicon, then the weight is set based on the sentiment lexicon of this subjectivity and, if words in the bigram belong to different subjectivity lexicons, we considered such bigram as neutral.

4.2 Results

In this section, we evaluate the different aspects of performance of the proposed models with respect to the state-of-the-art baselines Aspect-Sentiment Unification Model (ASUM) [8] and Joint Sentiment-Topic Model (JST) [17], which jointly model aspects and sentiments but do not consider demographic attributes of review authors. ASUM associates aspect and sentiment with an entire sentence and uses asymmetric priors to incorporate sentiment information for individual words. JST associates sentiment and aspect with each word rather than an entire sentence and uses symmetric priors for sentiment-specific topics. Both ASUM and JST are parametric models and require to specify the number of topics a priori.

In all experiments, we use the following settings of parameters for ASUM and JST. For ASUM we set α to $50/K$ and γ to 1. The weights of the seed words from the sentiment lexicons were set to 1.0 in the Dirichlet priors for the topics with the same polarity and to 0 in the priors for the topics with the opposite polarity. The

weights of all other words in the priors were set to 0.001. For JST, we set α to $50/K$, β to 0.01, γ for positive document-sentiment associations to 0.01 and negative document-sentiment associations to 5.0.

4.2.1 *Influence of sentiment lexicons*

In the first set of experiments, we evaluate the performance of different models in the USTM framework depending on unigram or bigram representation of reviews (Uni for Unigram and Bi for Bigram) as well as the sentiment lexicon (P for PARADIGMhasm and M for MPQA) used to create the Dirichlet priors for topics with different sentiment polarity. Performance of the model is evaluated in terms of perplexity, which is a measure derived from the likelihood of the data in the testing subset (10% of all reviews) under the model estimated on the training subset (90% of all reviews) of each dataset. The results for this set of experiments are summarized in Table 5.

Table 5: **Perplexity of parametric and non-parametric models using different lexicons for unigram and bigram review representations. Lower perplexity is better.**

Model	Auto	Hotel
Parametric models		
USTM-FT(W)+Uni+P	849.37	2146.48
USTM-FT(W)+Uni+M	959.80	2512.26
USTM-FT(S)+Uni+P	744.10	2099.34
USTM-FT(S)+Uni+M	782.93	2257.53
USTM-FT(W)+Bi+P	1324.27	3188.09
USTM-FT(W)+Bi+M	1435.38	3221.79
USTM-FT(S)+Bi+P	1236.58	2749.09
USTM-FT(S)+Bi+M	1314.75	3066.19
Non-parametric models		
USTM-DP(S)+Uni+P	704.85	1792.55
USTM-DP(W)+Uni+P	835.45	2007.32
Baselines		
ASUM+Uni+P	1190.92	2317.13
ASUM+Uni+M	1247.09	2563.96
ASUM+Bi+P	1358.87	2883.01
ASUM+Bi+M	1581.50	3160.39
JST+Uni+P	1091.01	2093.46
JST+Uni+M	1127.07	2272.31
JST+Bi+P	1297.85	2709.09
JST+Bi+M	1457.08	2982.82

As follows from Table 5, the lowest perplexity on both datasets is achieved by the non-parametric topic models using the PARADIGMhasm lexicon as a source of sentiment seed words. Second, sentence-based sentiment association consistently results in lower perplexity than word-based sentiment association across all lexicons and priors for both parametric and non-parametric topic models. Furthermore, USTM models perform consistently better on unigrams than on bigrams and when using PARADIGMhasm lexicon instead of MPQA. The proposed topic models also outperform ASUM on most combinations, except when using bigram representation of the **Hotel** dataset and have lower perplexity than JST on the **Auto** data set.

4.2.2 *Qualitative Topic Models results*

The primary goal of incorporating user meta-data into topic modeling process is to summarize the opinions of different market segments about various aspects of products and services. Therefore, we select several examples of sentiment-specific topics from the list of all topics discovered for different market segments that are designated by one or several user demographic attributes. Sample topics presented in Tables 6 and 7 were discovered by the USTM-FT(S) model in the Auto dataset using unigram-based representation of reviews. 10 terms with the highest weight are reported for each topic.

Table 6: **Sample topics discovered in the Auto dataset for the tag "male".**

#	sent.	topic words
1	NEU	new car battery engine thanks start help oil pump know
	POS	good new recommend great car thank engine oil help battery
	NEG	problem bad new car battery thanks junk help start oil
2	NEU	truck engine drive new thanks time transmission car know check
	POS	good drive recommend truck new engine driving thank vehicle right
	NEG	problem truck new bad engine issue transmission driving vehicle help
3	NEU	fuel engine help start light new pump car running truck
	POS	good fuel engine thank great start recommend help light thanks
	NEG	problem bad fuel engine start truck help light got running
4	NEU	report check cost vehicle using service free try government link
	POS	report recommend vehicle check using free service good history provides
	NEG	problem check vehicle link try time report alternative history service
5	NEU	report fuel help vehicle check thanks know new try rear
	POS	good report recommend help check engine new thank know service
	NEG	problem check help vehicle thanks sorry cost try bad report

Table 7: **Sample topics discovered in the Auto dataset for the tag "female".**

#	sent.	topic words
1	NEU	engine said light oil air driving check explorer probably knows
	POS	good engine light mechanic new power runs said oil wanted
	NEG	engine problem light coils run said know check explorer notice
2	NEU	truck coding ads heard ranger opinion work day repair second
	POS	truck having happy work great dealership heard radio funny best
	NEG	wrong problem told getting truck took cars saturn driver days
3	NEU	new don battery bought used plug right truck player wiring
	POS	new good love enjoy plug bought control code don positive
	NEG	new bad sorry free shift plug bought months mess negative
4	NEU	view allow account clicking try easiest set access email fuel
	POS	thank view says sharing fuel account controls set runs stay
	NEG	view problem account fuel pump info files says allow set
5	NEU	car drive got check went transmission know lot use focus
	POS	car thanks got miles good offer replaced mph help thought
	NEG	car went wrong focus escape auto help drive sensors miles

Several observations can be made based on examples in Tables 6 and 7. First, the proposed topic models can obtain coherent sentiment-specific topics for different aspects discussed in reviews by the corresponding market segments. For example, the first, second and third topics in Table 6 are related to battery, transmission and lights, respectively. Second, the proposed topic models can assign subjective terms to the correct sentiment-specific topics corresponding to the same aspect of reviews. Third, aspects are different between males and females. In particular, males tend to focus on mechanical aspects of the car, while females, although paying attention to some mechanical topics like Topic 1 (engine light) in Table 7, also talk about more general topics like dealerships (Topic 2) and account information (Topic 4).

Table 8 illustrates the difference between the sentiment-specific topics discovered by different age groups in the **Hotel** dataset, and Table 10 shows the example sentiment-specific topics for different combinations of gender and location.

As follows from Table 8, one major difference between the users in the age group "65" and "18-24" is that older people care more about tips as well as how quiet and comfortable their rooms are, while younger people pay more attention to the location of the hotel, breakfast and friendliness of staff. Tables 9 and 10 show sam-

Table 8: **Sample topics discovered in the Hotel dataset for the age tags "65" and "18-24".**

65	
NEU	breakfast times subway square place street tip close tips quiet
POS	new tips comfortable york floor desk excellent amazing square suite
NEG	service hotels lobby walk time little terrible bathroom breakfast place
18-24	
NEU	new night nice lobby central city floor free price time
POS	york times square check night time friendly lovely fantastic minutes
NEG	bed floor service bathroom breakfast trip disappointing beds problem night

Table 9: **Sample topics discovered in the Auto dataset for the location tag "fort worth, texas" and gender "female" and "male".**

female	
NEU	conditioner heat heater explorer turn car truck air help engine amp
POS	hot good conditioner vents heater dash air correct small blow
NEG	amp left turn signal bad explorer suggestions bulbs negative
male	
NEU	defrost heater hot vents time trying works fix fine waste
POS	battery water engine flow check truck mods thermostat deg correct
NEG	original heated happening part food bad truck works circuit care

ple topics discovered for females and males in "fort worth, texas" and "minnesota", respectively. Several interesting observations can be made based on these examples. First, both males and females in Texas are talking about the climate control systems in the car ("heater" and "conditioner"), while review authors in Minnesota are more focused on the "battery" and "starter" issues. Therefore, location appears to have a noticeable influence on the aspects people care about, since the summer temperatures in Fort Worth, TX are typically very high and therefore reviews often mention climate control systems (e.g. "conditioner"), while extremely cold winters in Minnesota seem to cause battery and starter issues. Second, gender influences preferences towards particular makes and models of cars, since "truck" and "mustang" are more frequently mentioned in reviews written by males than by females.

Table 11 presents an example of the topics discovered from the bigram-based representation of the **Auto** dataset. As follows from Table 11, bigram-based representation of reviews results in much more direct and closer associations between the topical aspects and sentiments than unigram-based representation. In this example, females in "dover, ohio" seem to be more interested in aesthetic aspects of vehicles, such as interior design (e.g. "seats"), while males tend to care more about the functional components of a car ("injector", "engine", etc.).

4.2.3 Review sentiment classification

Automatic detection of the overall sentiment of a review is one of the fundamental problems in opinion analysis. In this section, we report the results of using the associations between the words, user meta-data tags and sentiment aspects discovered by the topic models in USTM framework for the task of review sentiment prediction. We adopt a probabilistic approach and estimate $P(s|d)$, a distribution of predicted sentiments for a given review d. In particular, we

Table 10: **Sample topics discovered in the Auto dataset for the location tag "minnesota" and gender "female" and "male".**

female	
NEU	cold new know looking work fine headers control start thanks
POS	starter remote cold lights computer battery positive snow good focus
NEG	problem engages battery truck gear half hour negative bad
male	
NEU	fuel thanks caliper gas valve drier tank start gt wiring
POS	battery wires freeze positive power engine relay large tank
NEG	fuel bad car check mustang air cowl valve defrost

Table 11: **Examples of bigram topics discovered in the Auto dataset.**

location "dover ohio" and gender "female"	
NEU	ford bronco, make sure, lincoln ls, back seat, air control, fuse box, cherry bombs, seat belt, power loss, plugs wires
POS	car club, seat belt, back seat, key goes, wheel base, bench seat, good deal, local ford, ford dealer, heard good
NEG	isn't bad, bad wire, bad wheel, gone bad, go wrong, find bad, wrong time, connection bad, negative battery, bad water
location "dover ohio" and gender "male"	
NEU	throttle body, plugs wires, position sensor, ignition parts, power steering, fuel system, timing belt, system cleaner, rough running, rear wheels
POS	dash light, nice car, back seat, battery side, fuel injector, drive truck, trans fluid, right side, high end, engine light
NEG	engine runs, automatic transmission, bad wheel, ford thunderbird, go wrong, fluid changed, correct level, started acting, bad wire, bad wheel

compare the probabilities of assigning positive $P(s = pos|d)$ and negative $P(s = neg|d)$ sentiments to a review and classify the review as positive if $P(s = pos|d) > P(s = neg|d)$. Since the proposed topic models do not directly provide $P(s|d)$ as a result of posterior inference, we derive this distribution from the sentiment-based topics for each market segment, ϕ, by marginalizing out the topics, z, and user attributes (meta-data tags), t, as follows:

$$P(s|d) \propto P(d|s) = \prod_{w \in \boldsymbol{w}_d} P(w|s) = \prod_{w \in \boldsymbol{w}_d} \sum_{t=1}^{N_{dt}} \sum_{z=1}^{K_t} P(w|s,z,t) \quad (6)$$

Since the **Auto** dataset does not provide any information based on which the sentiment polarity of each review could be automatically derived, we only used the **Hotel** dataset for this experiment. In this dataset, each review is associated with graded ratings for six different aspects: service, value, sleep quality, room, location and cleanliness. Each of these ratings is a numeric score between 0 (lowest) and 5 (highest). The overall sentiment polarity of each review in the golden standard created for this dataset was determined automatically based on the average score for these six aspects as follows. If the average score for a given review is equal or greater then 3, then the review was considered as overall positive. If the average score of a review is equal or less than 2, then the review was considered as overall negative. Reviews with the average score between 2 and 3 were considered as neutral. Out of 9411 reviews with known ratings, 8553 were labeled as positive, 351 were labeled as negative and 507 were labeled as neutral, using the above method.

For this experiment, we used unigrams as lexical units and both PARADIGMhasm and MPQA lexicons as the sets of seed words for deriving sentiment-specific priors. Since ASUM and JST only consider positive or negative sentiments, we evaluate the performance of all models based only on those reviews, for which the ground truth labels are either positive or negative and are associated with at least one user attribute label in Table 3. The reported results are macro-averaged based on 5-fold cross validation. Figure 4 illustrates the change in accuracy by varying the number of topics per tag for the USTM topic models as well as the ASUM and JST baselines.

Several interesting observations can be made based on Figure 4. First, USTM-FT(W) has a comparatively stable prediction performance as there is a small change in accuracy when the number of topics per tag changes. The accuracy of USTM-FT(S) model, however, significantly improves as the number of topics per tag increases, while the performance of both ASUM and JST gradually drops. Table 12 compares the best accuracy, precision, recall, F1 score of the topic models in the USTM framework with ASUM and JST baselines for the review sentiment classification task.

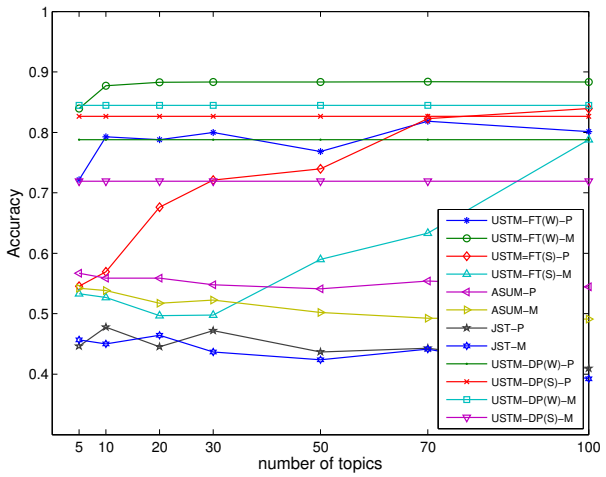


Figure 4: **Accuracy of sentiment prediction by varying the number of topics per tag (Hotel dataset)**

Table 12: **Performance the proposed models and the baselines for the task of predicting review sentiment. Best values for each performance metric on each dataset is highlighted in bold.**

Model	Precision	Recall	F1	Accuracy
USTM-FT(W)+P	**0.9651**	**0.8282**	**0.8914**	0.8194
USTM-DP(W)+P	0.9613	0.7789	0.8605	0.7878
USTM-FT(S)+P	0.9555	0.8279	0.8871	**0.8396**
USTM-DP(S)+P	0.9630	0.8193	0.8854	0.8264
ASUM + P	0.9626	0.5725	0.7180	0.5668
JST + P	0.9563	0.4812	0.6402	0.4778
USTM-FT(W)+M	**0.9685**	**0.8915**	**0.9284**	**0.8836**
USTM-DP(W)+M	0.9644	0.8294	0.8918	0.8447
USTM-FT(S)+M	0.9534	0.7658	0.8494	0.7880
USTM-DP(S)+M	0.9648	0.7217	0.8257	0.7190
ASUM + M	0.9654	0.5579	0.7071	0.5421
JST + M	0.9506	0.4588	0.6189	0.4643

Several important conclusions can be derived based on the results in Table 12. First, topic models in the proposed USTM framework significantly outperform the ASUM and JST baselines in terms of both accuracy and F1 score. Second, topic models assigning sentiment to each word individually outperform the models assigning sentiment on a per-sentence basis. Third, tuning the number of topics for parametric topic models allows to significantly improve their classification performance, ultimately outperforming the non-parametric topic models. Fourth, in most cases (except USTM-FT(W)-P and USTM-DP(W)-P), using PARADIGMhasm lexicon to derive sentiment-specific priors results in better accuracy than using MPQA lexicon. Since the experimental datasets are domain specific, larger generic lexicons may not translate into better performance. Finally, although the high precision of all models can be attributed to the small number of negative reviews in experimental dataset, USTM models still show better performance in predicting positive reviews than the baseline algorithms.

4.2.4 User attribute prediction

Predicting the attributes of the author of a review based on its lexical content is another interesting opinion mining task, for which the topic models in the proposed USTM framework is a natural choice. Similar to the sentiment classification task, the distribution over user attributes (or meta-data tags), $P(t|d)$, for each review, d, can also be estimated using sentiment-based topics for each market segment, ϕ, by marginalizing out the topics, z, and sentiments, s, as follows:

$$P(t|d) \propto P(d|t) = \prod_{w \in \boldsymbol{w}_d} P(w|t) = \prod_{w \in \boldsymbol{w}_d} \sum_{z=1}^{K_t} \sum_{s=1}^{S} P(w|s,z,t) \quad (7)$$

Since each review can be associated with several user attributes, we use Mean Average Precision (MAP), which takes into account the positions of the actual user attributes in the list of predicted ones, as a measure of performance of the topic models for this task. For this experiment, we used both the **Auto** and **Hotel** datasets and considered only the reviews that are associated with at least one of 100 most frequent tags in each dataset. We also used only unigrams as lexical units and PARADIGMhasm as the seed set for deriving the sentiment priors. The reported MAP is obtained using 5-fold cross validation and macro-averaged over the folds.

First, to optimize the configuration and evaluate the impact of different number of topics on the attribute prediction performance of parametric topic models (USTM-FT(W) and USTM-FT(S)), we varied the number of topics from 5 to 100 and recorded MAP for each setting. Figure 5 presents the results of this experiment along with the performance of non-parametric models (UMTM-DP(W) and UMTM-DP(S)) for comparison.

As follows from Figures 5a and 5b, the user attribute prediction accuracy reaches the maximum value when the number of topics is set to 20 for the **Auto** data set. For the **Hotel** data set, it reaches the maximum value at 70 and 100 for USTM-FT(S) and USTM-FT(W) respectively. Table 13 summarizes and compares the best results achieved by the proposed parametric and non-parametric topic models for the task of predicting the attributes of review authors on both experimental datasets.

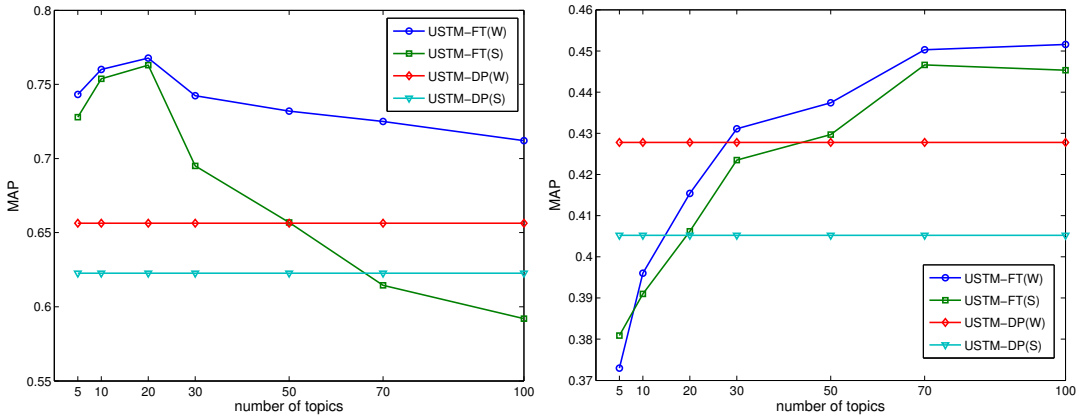
Table 13: **Performance of the topic models in the USTM framework for the task of predicting the attributes of review authors.**

Dataset	Model	MAP
Auto	USTM-FT(W)+P	**0.7678**
Auto	USTM-DP(W)+P	0.6563
Auto	USTM-FT(S)+P	0.7630
Auto	USTM-DP(S)+P	0.6226
Hotel	USTM-FT(W)+P	**0.4516**
Hotel	USTM-DP(W)+P	0.4278
Hotel	USTM-FT(S)+P	0.4466
Hotel	USTM-DP(S)+P	0.4052

As follows from Table 13, the proposed models can be used to predict the attributes of review authors with reasonable accuracy. Furthermore, analysis of the results presented in Table 13 leads to three important conclusions. First, all proposed models consistently perform better on the **Auto** dataset than on the **Hotel** data set. Second, similar to the sentiment prediction task, topic models assigning sentiment to each word individually are more accurate than the models assigning sentiment on a per-sentence level. Third, the optimized parametric topic models achieved better accuracy than non-parametric ones for this task.

5. CONCLUSION AND FUTURE WORK

In this paper, we introduced a novel research problem of market segment-based summarization of contrasting opinions about different aspects of products or services in on-line consumer reviews, which has extensive practical applications. We also proposed two parametric extensions of LDA and two non-parametric extensions of the Dirichlet Process to address this problem. The proposed models incorporate asymmetric sentiment priors and jointly model



(a) **Auto dataset** (b) **Hotel dataset**

Figure 5: **Accuracy of user attribute prediction for user-sentiment topic models by varying the number of topics**

demographic information of review authors and topical aspects of reviews. Qualitative analysis of sentiment-based topics discovered by the proposed models in two real-world collections of on-line consumer reviews using both unigrams and bigrams as lexical units indicates that incorporating user information into opinion analysis of on-line consumer reviews allows to better understand the preferences of different demographic groups of customers. We also demonstrated through quantitative evaluation that the proposed models can be used to accurately predict the overall sentiment of reviews as well as the demographic attributes of their authors.

We envision future work to proceed along the following two directions. First, machine learning techniques can be leveraged to learn the sentiment-specific priors for each individual word rather than a group of words. The second direction can focus on leveraging natural processing techniques, such as chunkers and part-of-speech taggers, to improve aspect-sentiment summaries by partitioning the text of reviews into more descriptive lexical units and better accounting for negations.

6. REFERENCES

[1] D. M. Blei, T. L. Griffiths, and M. I. Jordan. The Nested Chinese Restaurant Process and Bayesian Nonparametric Inference of Topic Hierarchies. *JACM*, 57(7), 2010.

[2] D. M. Blei and J. D. McAuliffe. Supervised Topic Models. In *Proceedings of the 20th NIPS*, pages 121–128, 2007.

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] W. Gao, N. Yoshinaga, N. Kaji, and M. Kitsuregawa. Modeling User Leniency and Product Popularity for Sentiment Classification. In *Proceedings of the 6th IJCNLP*, pages 1107–1111, 2013.

[5] T. Hofmann. Probabilistic Latent Semantic Indexing. In *Proceedings of the 22nd SIGIR*, pages 50–57, 1999.

[6] M. Hu and B. Liu. Mining and Summarizing Customer Reviews. In *Proceedings of the 10th ACM SIGKDD*, pages 168–177, 2004.

[7] N. Jakob and I. Gurevych. Extracting Opinion Targets in a Single- and Cross-domain Setting with Conditional Random Fields. In *Proceedings of the 2010 EMNLP*, pages 1035–1045, 2010.

[8] Y. Jo and A. H. Oh. Aspect and Sentiment Unification Model for Online Review Analysis. In *Proceedings of the 4th WSDM*, pages 815–824, 2011.

[9] J. H. Kim, D. Kim, S. Kim, and A. Oh. Modeling Topic Hierarchies with the Recursive Chinese Restaurant Process. In *Proceedings of the 21st ACM CIKM*, pages 783–792, 2012.

[10] S. Kim, J. Zhang, Z. Chen, A. H. Oh, and S. Liu. A Hierarchical Aspect-Sentiment Model for Online Reviews. In *Proceedings of the 27th AAAI*, pages 526–533, 2013.

[11] A. Kotov and E. Agichtein. The Importance of Being Socially-Savvy: Quantifying the Influence of Social Networks on Microblog Retrieval. In *Proceedings of the 22nd CIKM*, pages 1905–1908, 2013.

[12] A. Kotov, V. Rakesh, E. Agichtein, and C. K. Reddy. Geographical Latent Variable Models for Microblog Retrieval. In *Proceedings of the 37th ECIR*, pages 635–647, 2015.

[13] A. Kotov, Y. Wang, and E. Agichtein. Leveraging Geographical Metadata to Improve Search over Social Media. In *Proceedings of the 22nd WWW*, pages 151–152, 2013.

[14] H. Lakkaraju, C. Bhattacharyya, I. Bhattacharya, and S. Merugu. Exploiting Coherence for the Simultaneous Discovery of Latent Facets and associated Sentiments. In *Proceedings of the 11th SDM*, pages 498–509, 2011.

[15] F. Li, C. Han, M. Huang, X. Zhu, Y.-J. Xia, S. Zhang, and H. Yu. Structure-aware Review Mining and Summarization. In *Proceedings of the 23rd COLING*, pages 653–661, 2010.

[16] F. Li, S. Wang, S. Liu, and M. Zhang. SUIT: A Supervised User-Item Based Topic Model for Sentiment Analysis. In *Proceedings of the 28th AAAI*, pages 1636–1642, 2014.

[17] C. Lin and Y. He. Joint Sentiment/Topic Model for Sentiment Analysis. In *Proceedings of the 18th CIKM*, pages 375–384, 2009.

[18] C. Lin, Y. He, R. Everson, and S. Ruger. Weakly Supervised Joint Sentiment-Topic Detection from Text. *IEEE Transactions on Knowledge and Data Engineering*, 24(6):1134–1145, 2012.

[19] T. Ma and X. Wan. Opinion Target Extraction in Chinese News Comments. In *Proceedings of the 23rd COLING*, pages 782–790, 2010.

[20] 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 WWW*, pages 171–180, 2007.

[21] S. Moghaddam and M. Ester. ILDA: Interdependent LDA Model for Learning Latent Aspects and Their Ratings from Online Product Reviews. In *Proceedings of the 34th SIGIR*, pages 665–674, 2011.

[22] A. Mukherjee and B. Liu. Aspect Extraction Through Semi-supervised Modeling. In *Proceedings of the 50th ACL*, pages 339–348, 2012.

[23] A.-M. Popescu and O. Etzioni. Extracting Product Features and Opinions from Reviews. In *Proceedings of the 2005 EMNLP-HLT*, pages 339–346, 2005.

[24] D. Ramage, C. D. Manning, and S. Dumais. Partially Labeled Topic Models for Interpretable Text Mining. In *Proceedings of the 17th SIGKDD*, pages 457–465, 2011.

[25] C. Sauper, A. Haghighi, and R. Barzilay. Content Models with Attitude. In *Proceedings of the 49th ACL-HLT*, pages 350–358, 2011.

[26] C. Tan, L. Lee, J. Tang, L. Jiang, M. Zhou, and P. Li. User-level Sentiment Analysis Incorporating Social Networks. In *Proceedings of the 17th ACM SIGKDD*, pages 1397–1405, 2011.

[27] I. Titov and R. Mcdonald. A Joint Model of Text and Aspect Ratings for Sentiment Summarization. In *Proceedings of 46th ACL-HLT*, pages 308–316, 2008.

[28] I. Titov and R. McDonald. Modeling Online Reviews with Multi-grain Topic Models. In *Proceedings of the 17th WWW*, pages 111–120, 2008.

[29] P. D. Turney and M. L. Littman. Measuring Praise and Criticism: Inference of Semantic Orientation from Association. *ACM TOIS*, 21(4):315–346, 2003.

[30] H. Wang, Y. Lu, and C. Zhai. Latent Aspect Rating Analysis on Review Text Data: A Rating Regression Approach. In *Proceedings of the 16th SIGKDD*, pages 783–792, 2010.

[31] T. Zhao, C. Li, Q. Ding, and L. Li. User-sentiment Topic Model: Refining User's Topics with Sentiment Information. In *Proceedings of the ACM SIGKDD Workshop on Mining Data Semantics*, 2012.