Structured Neural Topic Models for Reviews

Babak Esmaeili
Northeastern University
esmaeili.b@husky.neu.edu

Hongyi Huang
Viewpoint School
f.huang19@viewpoint.org

Byron C. Wallace
Northeastern University
b.wallace@northeastern.edu

Jan-Willem van de Meent
Northeastern University
j.vandemeent@northeastern.edu

Abstract

We present Variational Aspect-based Latent Topic Allocation (VALTA), a family of autoencoding topic models that learn aspect-based representations of reviews. VALTA defines a user-item encoder that maps bag-of-words vectors for combined reviews associated with each paired user and item onto structured embeddings, which in turn define per-aspect topic weights. We model individual reviews in a structured manner by inferring an aspect assignment for each sentence in a given review, where the per-aspect topic weights obtained by the user-item encoder serve to define a mixture over topics, conditioned on the aspect. The result is an autoencoding neural topic model for reviews, which can be trained in a fully unsupervised manner to learn topics that are structured into aspects. Experimental evaluation on large number of datasets demonstrates that aspects are interpretable, yield higher coherence scores than non-structured autoencoding topic model variants, and can be utilized to perform aspect-based comparison and genre discovery.

1 Introduction

In recent years, the field of natural language processing (NLP) has decisively shifted away from bag-of-words representations towards neural models. These models represent text using embeddings that are learned from data in an end-to end manner. A potential drawback to such embeddings is that learned representations tend to be *entangled*, in the sense that an embedding is a monolithic vector that encodes some unknown set of characteristics of the input data. When one is interested in training a model solely for a particular task, entanglement is not necessarily a problem, so long as the trained model achieves sufficiently robust predictive performance. However, there are cases where it is desirable to learn a representation that factors into distinct, complementary sets of features, i.e., is *disentangled*.

One reason we may want a disentangled representation is interpretability. Separating representations into distinct factors that correspond to identifiable subsets of features, such as the topic and political leaning of an opinion piece, allows one to more easily reason about which features informed a prediction. A second reason to induce disentangled representations is data efficiency. Suppose that were to train a model on images that contain K categories of shapes which assume L categories of colors. If a model can separate shape from color, then it should generalize to shape and color combinations not observed in the training data. This means that we can hope to train such a model on $O(K+L)$ examples, rather than a dataset in which all combinations of features are present, which would require $O(KL)$ examples. Learning disentangled representations thus provides a strategy for factorizing a problem in a high-dimensional feature space into problems in lower-dimensional feature spaces.

In computer vision, there has been considerable effort to develop methods for inducing disentangled representations in semi- and un-supervised settings [1–9]. Many of these approaches define deep generative models such as variational autoencoders (VAE) [10, 11], or Generative Adversarial Networks (GANs) [12, 13]. In NLP, work on learning disentangled representations has been more limited [14–17]. A large body of pre-neural work exists on aspect-based topic models that derive from Latent Dirichlet Allocation (LDA) [18]. This includes approaches for sentiment analysis [19–24], and models in the factorial LDA family [25–27].

There has been relatively little work in NLP on learning disentangled representations with neural architectures. One reason for this is that work on deep generative models for

Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89.

text is not as well-established as work for images. Early approaches in this space, such as the Neural Variational Document Model (NVDM) [28] and autoencoding LDA [29] developed neural topic models in which the generative model is either an LDA-style mixture, or a SAGE-style [30] log-linear combination over topics. More recently there have been some efforts to develop deep generative models with interpretable aspects, chiefly the work by Hu et al. [31], which combines a recurrent VAE architecture with a set of aspect discriminators to induce a structured representation.

In this paper we explore the effectiveness of neural topic models for learning disentangled representations. We treat review datasets as a particular case study, where we consider the task of learning structured representations for both the reviewer and the reviewed item. Reviews comprise several variables of interest, such as the aspect(s) of the item being discussed, user sentiment regarding each aspect, and characteristics of the item for each aspect (i.e., sub-aspects). More concretely, in any review corpus, items will very likely share certain *aspects*, each affecting the rating separately. For example, in the case of restaurant reviews, all establishments will serve food and have a location. Similarly, every beer will have an aroma and appearance. More generally, each aspect may contain nested *sub-aspects*: A restaurant can serve Italian, Chinese, or fast food; and a beer can be dark or light in appearance.

In this paper we develop autoencoding models that induce representations of review texts that capture this structure. Such representations can perform aspect-based item comparisons, and also provide one sort of interpretability. To realize these goals, VALTA combines topic modelling and recommender systems into a structured VAE framework. We model reviews in a structured manner by associating an aspect with each sentence in a review, and use aspect-specific topics to define a log-linear likelihood, similar to the one used in SAGE [30], the NVDM [28], and ProdLDA [29]. Topic and aspect weights are predicted based on a user and item embedding. The result is a highly structured model, in which both aspects and sub-aspects are interpretable, and topics have a high predictive power in terms of perplexity and coherence scores. These learned representations can be used for downstream tasks such as genre discovery, representation quality, and aspect-retrieval.

Symbol	Description
V	vocabulary size
A	number of aspects
K	number of sub-aspects
H	number of hidden units
u	user index
i	item index
θ	parameters of generative model
ϕ	parameters of inference model
$\boldsymbol{x}_{i,u}$	review written by user u about item i
$\boldsymbol{x}_{i,u,s}$	sentence s of $\boldsymbol{x}_{i,u}$
$\omega_{i,u,s}$	aspect log probabilities of $\boldsymbol{x}_{i,u,s}$
$\boldsymbol{z}_{i,u,s}$	aspect assignment of $\boldsymbol{x}_{i,u,s}$
ρ_i	hidden representation of item i
ρ_u	hidden representation of user u
$\rho_{i,u}$	hidden representation of $\boldsymbol{x}_{i,u}$
$\boldsymbol{\psi}_{i,u}$	aspect-specific topic distributions of $\boldsymbol{x}_{i,u}$
$\boldsymbol{\lambda}_i$	aspect-importance of item i
$\boldsymbol{\lambda}_u$	aspect-preference of user u
β_0	global rating bias
β_i	item rating bias
β_u	user rating bias
$r_{u,i}$	true rating by user u for item i
$\hat{r}_{u,i}$	prediction of rating by user u for item i

Table 1: Summary of notation used throughout.

2 Background and Preliminaries

Review datasets have been widely studied in the context of recommender systems [32]. Matrix factorization techniques [33–35] are widely used to predict ratings by representing each user and item by a K dimensional vector, which we sometimes refer to as an embedding. Since these approaches consider the ratings alone, they ignore the text of the review, which is a key source of information. McAuley and Leskovec [36] proposed combining topic models and matrix factorization techniques for learning ratings to learn topics and ratings simultaneously. Subsequent approaches aimed to exploit review text in addition to rating [36–40]. These efforts have shown that topic models indeed can act as a good regularizer for rating prediction, particularly for users or items with few reviews [36, 40]. In the last few years, both recommender systems and topic modelling approaches have shifted towards deep learning methods [28, 29, 37], many of which also exploit text to predict ratings.

While neural recommender systems can achieve good predictive performance, it is unclear how they do so, because learned feature vectors are optimized only to code indiscriminately for (unknown) predictive combinations of attributes. Such entangled representations thus do not reveal any information about the structure of the data, which in turn hinders model interpretability and generalizability. By imbuing representations with probabilistic semantics, we can design models to explicitly tease out structured embeddings, components of which may then be re-used.

In prior work, deep generative models have been proposed to learn representations of text via variational autoencoders [10, 11]. VAEs jointly optimize a generative network and an inference network. The former, $p_\theta(\boldsymbol{x}, \boldsymbol{z})$, specifies a distribution over set of hidden variables $\boldsymbol{z}$ and observed variables $\boldsymbol{x}$. The latter is a conditional distribution $q_\phi(\boldsymbol{z}|\boldsymbol{x})$. Defining $q(x)$ as the empirical distribution, these two models are

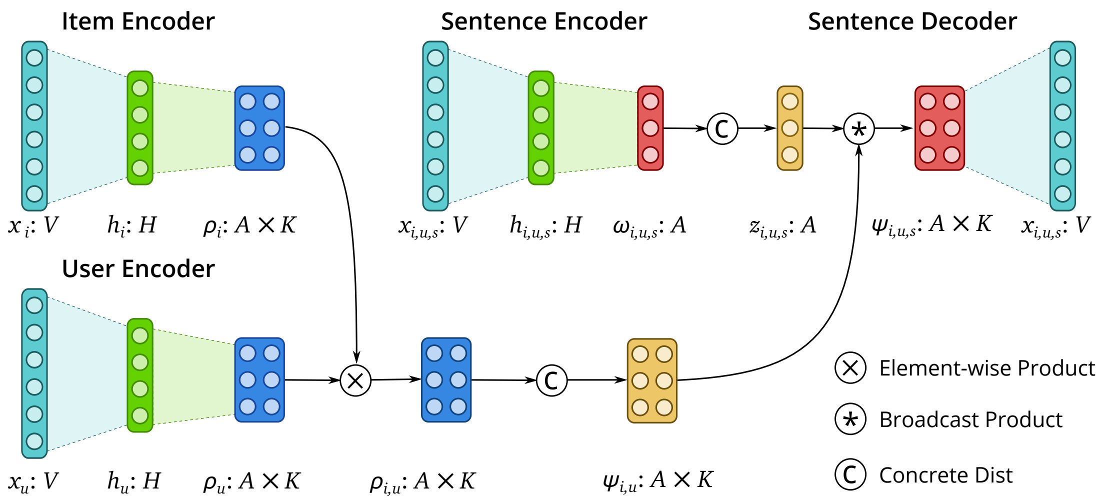

Figure 1: VALTA learns representations for reviews using a structured autoencoder that consists of a sentence-level encoder $q_\phi(\boldsymbol{z}_{i,u,s} \mid \boldsymbol{x}_{i,u,s})$, which infers assignments to aspects, and a document-level user-item encoder $q_\phi(\boldsymbol{\psi}_{i,u} \mid \boldsymbol{x}_i, \boldsymbol{x}_u)$, which infers per-aspect topic weights, and a sentence-level decoder $p_\theta(\boldsymbol{x}_{i,u,s} \mid \boldsymbol{z}_{i,u,s}, \boldsymbol{\psi}_{i,u})$.

trained by optimizing the evidence lower bound (ELBO),

$$\mathcal{L}(\theta, \phi) := \mathbb{E}_{q(\boldsymbol{x})}\left[\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p_\theta(\boldsymbol{x}, \boldsymbol{z})}{q_\phi(\boldsymbol{z} \mid \boldsymbol{x})}\right]\right]. \qquad (1)$$

Variants of VAEs have been used to develop autoencoding topic models [28, 29]. These models achieve predictive performance (in terms of the perplexity score) that is competitive with other bag-of-word models, but lack the explicit structure that we aim to capture here. More specifically: the prior in existing VAE-based approaches takes the form of a Gaussian with diagonal covariance matrix, where each dimension of the Gaussian corresponds to a topic. By contrast, we here aim to characterize groups of topics that correspond to specific aspects of interest. One means of realizing this would be to posit A Gaussians, one per aspect, and each of these might comprise its own set of K topics. Motivated by this idea, we propose a structured prior for VAEs to capture this information from a collection of reviews.

Table 1 presents the notation we use throughout this paper.

3 Methodology

To learn structured representations of reviews, we begin by identifying key axes of variation in review datasets. We define three variable categories: items, users, and review texts. We assume A aspects of interest for all items. For example, in the case of beer reviews, these aspects correspond to properties such as appearance and aroma. We further decompose these aspects into K topics. A topic within appearance might be, e.g., dark versus pale beer. Reflecting these structural assumptions, our model defines aspect-specific embeddings that in turn yield distributions over topics. Thus, a representation of a sweet, dark beer should place a relatively large mass on the dark topic of appearance, and a large mass on the sweet topic of taste.

The relative importance of aspects may vary for both items and users. A restaurant, for example, may be located on the water or on a famous city street, in which case the location is likely to be its most salient aspect. Similarly, lagers are not typically renowned for their smell. Users will also have their own weightings of aspect importance. A particular user may be concerned primarily with food quality over price, and might prefer Chinese food. Others, meanwhile, may emphasize location or ambience.

The words contained in a review are a function of aspects and topics, and their relative importance for particular user-item *pairs*. To accurately learn topics and predict ratings, we now introduce variables that are defined at the review level. A naive approach would be to encode the review text $\boldsymbol{x}_{i,u}$ and then train the generative model to learn both topics and ratings for item i and user u. We propose an approach that is directly motivated by the observation that the aspects and topics discussed in review $\boldsymbol{x}_{i,u}$ will depend on a combination of the aspect preferences of u, and the relative salience of the respective aspects for i. Therefore, rather than encoding $\boldsymbol{x}_{i,u}$ directly, we encode information about i and u separately, and then combine these representations to yield a joint embedding for $\boldsymbol{x}_{i,u}$ and predict the rating $r_{i,u}$.

It is likely that most reviews will contain at least some words about all aspects (although the prevalence of individual aspects will vary across reviews). Thus it is intuitive to attempt to infer which parts of a review talk about which aspect. In our model we make the simplifying assumption that every sentence within a review discusses only a single aspect. One could alternatively assign aspects at the word- or paragraph-level. However, sentence-level assignment constitutes an intuitive compromise, and is also consistent with prior work [41, 42]. Note that while the aspect assignment varies between sentences within a review, the topic distributions should be fixed for that particular review.

Following prior work [36, 37, 40], we assume the input representation for item i is a bag-of-words vector encoding the words used across all reviews written about this item. Similarly, we define the input vector for user u as a bag-of-words induced over all reviews that they have written. This representation has been shown to perform well in terms of capturing characteristics of items and users [36, 38, 39], but it does not take into account the relative importance of different aspects with respect to both i and u. Nor have such models explicitly accounted for the intuitive observation that different parts of reviews (probably) discuss different aspects, which we achieve via sentence-wise aspect assignments $\{\boldsymbol{z}_{i,u,s}\}$ based on encoded sentences $\{\boldsymbol{x}_{i,u,s}\}$ (one per each sentence in the review written by u for i).

We provide a schematic of our model in Figure 1. The inference and generative models are defined to codify the structure discussed above. Specifically, given the topic distribution $\boldsymbol{\psi}_{i,u}$ for i and u, sentence aspect assignments $\{\boldsymbol{z}_{i,u,s}\}$, and the review $\boldsymbol{x}_{i,u}$, we define the inference model

$$\begin{aligned} &q_\phi(\boldsymbol{\psi}_{i,u}, \boldsymbol{z}_{i,u} \mid \boldsymbol{x}_i, \boldsymbol{x}_u, \boldsymbol{x}_{i,u}) \\ &\quad = q_\phi(\boldsymbol{\psi}_{i,u} \mid \boldsymbol{x}_i, \boldsymbol{x}_u) \prod_s q_\phi(\boldsymbol{z}_{i,u,s} | \boldsymbol{x}_{i,u,s}). \end{aligned} \tag{2}$$

An obvious choice for the likelihood model is to define a decoder for review $\boldsymbol{x}_{i,u}$. However, this will entangle the different aspects discussed in said review. To ensure that the generative model associates different dimensions with specific aspects, we define our generative network at a sentence level

$$\begin{aligned} &p_\theta(\boldsymbol{x}_{i,u}, \boldsymbol{z}_{i,u}, \boldsymbol{\psi}_{i,u}) \\ &= \prod_s p_\theta(\boldsymbol{x}_{i,u,s} | \boldsymbol{z}_{i,u,s}, \boldsymbol{\psi}_{i,u}) p(\boldsymbol{z}_{i,u,s}) \prod_a p(\boldsymbol{\psi}_{i,u,a}). \end{aligned} \tag{3}$$

We note that the K-dimensional topic distribution $\boldsymbol{\psi}_{i,u}$ is fixed at the review level. This reflects the assumption that given the item and user, the specific topics of interest will not change, as the opinion of the user and the characteristics of the item are fixed. The only axis of variation is the user's decision regarding which aspect to write about in any given sentence. However, we must ensure that the generative model focuses only on the assigned aspect, rather than topics of all aspects. We enforce this by multiplying the columns of $\boldsymbol{\psi}_{i,u}$ with the (nearly) one-hot topic assignment vector $\boldsymbol{z}_{i,u,s}$: $\boldsymbol{\psi}^s_{i,u} = \boldsymbol{z}_{i,u,s} \circledast \boldsymbol{\psi}_{i,u}$ where $\circledast$ is the broadcasting operation. Because $\boldsymbol{z}_{i,u,s}$ resembles a one-hot vector, this effectively masks the topic distributions pertaining to other (unassigned) topics. Thus, only the topic distribution corresponding to a single (selected) aspect is responsible for reconstructing $\boldsymbol{x}_{i,u,s}$.

We use the generative model for text introduced in prior work [29]. This model induces log probabilities for each word via feeding the $\boldsymbol{\psi}^s_{i,u}$ through a single layer neural network followed by applying a log softmax

$$\log p_\theta(\boldsymbol{x}_{i,u,s} | \boldsymbol{\psi}^s_{i,u}) = \boldsymbol{x}_{i,u,s} \log \left(\frac{\exp(\hat{w}_v)}{\sum_{v=1}^{V} \exp(\hat{w}_v)} \right).$$

Where the log word probabilities $\hat{w} = g_\theta(\boldsymbol{\psi}^s_{i,u})$ are computed using a single-layer perceptron g_θ with weights θ.

3.1 Concrete Distribution

An important factor in our model is the choice of $p(\boldsymbol{z})$ and $p(\boldsymbol{\psi})$. The appropriate choice for $p(\boldsymbol{z})$ and $p(\boldsymbol{\psi})$ are discrete and Dirichlet distributions respectively, as they represent aspect *assignment* and topic *proportions*. This is problematic in practice because discrete variables are not amenable to the reparameterization trick, thus precluding use of estimation via standard backpropagation algorithms. In the case of Dirichlet distributions, several methods have been proposed to allow for sampling via reparameterization [43, 44]. However, in practice these methods dramatically increase training time in our implementation because the base system, PyTorch, does not provide GPU implementations for these distributions at the time of writing. In this work, we choose to model both variables $\boldsymbol{z}_{i,u,s}$ and $\boldsymbol{\psi}_{i,u,a}$ using the Concrete distribution, a relaxation of discrete distributions implemented via a Gumbel softmax [45, 46]. The Gumbel distribution can be sampled in a reparameterized way by drawing $u \sim \text{Uniform}(0, 1)$ and then computing $g = \log(\log(u))$. If $\boldsymbol{z}$ has aspect log-probabilities $\omega_1, \omega_2, \ldots, \omega_A$, then we can sample from a continuous approximation of the discrete distribution by sampling a set of $g_a \sim \text{Gumbel}(0, 1)$ i.i.d. and applying the transformation

$$\boldsymbol{z}_a = \frac{\exp((\omega_a + g_a)/\tau)}{\sum_a \exp((\omega_a + g_a)/\tau)}.$$

where τ is a temperature parameter controlling relaxation. The sample $\boldsymbol{z}$ is a continuous approximation of the desired one-hot representation. The role of τ is critical in our model, as it dictates the *peakiness* of the samples. In the case of $\boldsymbol{z}_{i,u,s}$, we keep the temperature low to enforce the assumption that each sentence is only talking about a single aspect. However, we do not wish for the topic proportions to be close to a one-hot vector, as this would restrict items to contain a single topic within each aspect. To encourage $q_\phi(\boldsymbol{\psi})$ to mimic a Dirichlet distribution, we set the temperature to higher values (greater than 1.0), thereby encouraging all K dimensions within each aspect to contribute. In our experiments, we have observed that sampling $\boldsymbol{\psi}_{i,u}$ with a low temperature results in only a few dimensions within each aspect learning something meaningful (having non-zero mutual information) about the review which is not desired.

Dataset	Aspects	#users	#items	#reviews	#sentences
Beer (Beeradvocate)	Aroma, Taste, Mouthfeel, Look	4,923	2,017	127,346	1,515,517
Restaurant (Yelp)	Price, Ambiance, Food, Service	13,847	6,588	140,139	1,416,317
Clothing (Amazon)	Formality, Appearance, Type	12,203	73,903	80,285	447,920
Movie (Amazon)	Genre, Awards, Screen Play	7,590	2,288	100,489	1,446,690

Table 2: Review dataset statistics.

3.2 Rating Prediction

A good representation of a review should not contain only informative topics, but should also assist in accurately predicting the rating linked to the review. In this subsection, we extend VALTA to predict ratings in combination with learning aspects and topics. We take several factors into account when predicting the rating. As discussed above, we assume that users have different aspect preferences and that items exhibit different aspect-importance. To extract aspect-importance vectors $\boldsymbol{\lambda}_i$ and $\boldsymbol{\lambda}_u$ for item i and user u respectively, we use the weights of the sentence encoder $f_\phi^{aspect}(h)$ that is responsible for predicting the aspect

$$\boldsymbol{\lambda}_i = f_\phi^{aspect}(h_i), \qquad \boldsymbol{\lambda}_u = f_\phi^{aspect}(h_u).$$

As $f_\phi^{aspect}(\cdot)$ is trained at the sentence level (and so compelled to extract words associated with aspects), we can re-use its weights to extract an aspect-importance vector from a collection of reviews. We then average these two embeddings to obtain aspect-importance for a particular pair

$$\boldsymbol{\lambda}_{i,u} = \frac{1}{2}(\boldsymbol{\lambda}_i + \boldsymbol{\lambda}_u)$$

One could consider learning different embeddings for items and users that are different from the one in topic models. However, as discussed in [36], *coupling* the embeddings for items and users with their topic models representation helps to learn topics that explain the diversities in ratings. Thus, in our model we re-use the input to the concrete distribution $\rho_{i,u}$ to predict the rating associated with each aspect as

$$\hat{r}_{i,u,a} = \sum_{k=1}^{K} \rho_{i,u,a,k}.$$

Based on this structure we predict the overall ratings as

$$\hat{r}_{i,u} = \beta_0 + \beta_i + \beta_u + \frac{1}{A}\sum_{a=1}^{A} \underbrace{\boldsymbol{\lambda}_{i,u,a}}_{\textit{aspect importance}} \overbrace{\hat{r}_{i,u,a}}^{\textit{aspect rating}} \tag{4}$$

where β_0 is the global rating bias, and β_i and β_u are item i and user u bias respectively. This approach is similar to the family of aspect-aware latent factor models (ALFM) proposed in [40]. Following prior work, we use the mean squared error (MSE) loss for the recommender model component. This may also be interpreted in a probabilistic way where we model $\hat{r}_{u,i}$ as a Gaussian distribution: $p(\hat{r}_{u,i}; r_{u,i}) := \mathcal{N}(\hat{r}_{i,u}; r_{i,u}, 1)$.

3.3 VALTA

In this subsection, we put everything together to define a unified objective for VALTA. For clarity, we decompose our objective into the four terms, which together define a lower bound on the log marginal likelihood, analogous to the VAE objective defined in equation 1

$$\mathcal{L}^{\textit{VALTA}}(\theta, \phi) = \mathcal{L}^{\boldsymbol{x}}_{gen} + \mathcal{L}^{r}_{mse} + \mathcal{L}^{\boldsymbol{\psi}}_{KL} + \mathcal{L}^{\boldsymbol{z}}_{KL}. \tag{5}$$

where the terms are responsible for reconstructing the review text, predicting the rating, matching the aspect distribution in the encoder to prior, and matching the topic distributions in the encoder to the prior respectively.

The first term is the expected log likelihood of the review

$$\mathcal{L}^{\boldsymbol{x}}_{gen}(\theta, \phi) = \mathbb{E}\left[\log \prod_s p_\theta(\boldsymbol{x}_{i,u,s} | \boldsymbol{\psi}_{i,u}, \boldsymbol{z}_{i,u,s})\right].$$

Note that this expectation is are defined w.r.t to the inference model $q_\phi(\cdot)$, which we omit for simplicity.

The second term is the likelihood of the rating $r_{i,u}$

$$\mathcal{L}^{r}_{mse}(\theta) = \log p_\theta(r_{i,u} | \boldsymbol{x}_i, \boldsymbol{x}_u).$$

Finally, as with a normal VAE we incorporate two regularization terms in the form of KL divergences between the encoder distribution and the prior

$$\mathcal{L}^{\boldsymbol{z}}_{KL}(\theta, \phi) = -\mathbb{E}\left[\log \prod_s \frac{q_\phi(\boldsymbol{z}_{i,u,s} | \boldsymbol{x}_{i,u,s})}{p_\theta(\boldsymbol{z}_{i,u,s})}\right],$$
$$\mathcal{L}^{\boldsymbol{\psi}}_{KL}(\theta, \phi) = -\mathbb{E}\left[\log \frac{q_\phi(\boldsymbol{\psi}_{i,u} | \boldsymbol{x}_i, \boldsymbol{x}_u)}{p_\theta(\boldsymbol{\psi}_{i,u})}\right].$$

4 Related Work

A comprehensive literature review of recommender systems is beyond the scope of this work. Here, we discuss models that exploit both rating and reviews to jointly learn topics and predict ratings. We divide these models to three classes: 1) probabilistic topic models; 2) deep learning-based approaches; 3) VAEs. VALTA belongs to the last category.

In the first class, the most closely related approach to our work is the aspect-aware topic model (ATM) [40], which considers a similar decomposition of reviews to aspects and sub-aspects. In the same paper, the authors also propose an

	CitySearch		BeerAdvocate	
	ACC	F1-Score	ACC	F1-Score
LDA	0.477	0.597	0.447	0.468
Local-LDA	0.803	0.861	0.758	0.761
SVM	0.830	0.887	0.647	0.604
VALTA	**0.885**	**0.931**	**0.769**	**0.794**

Table 3: Multi-aspect sentence labelling results.

aspect-aware latent factor model (ALFM) which exploits the parameters learned from the ATM to predict ratings. While VALTA shares the idea of further decomposition of aspects with ATM, it is trained to learn topics and predict ratings jointly rather than sequentially.

Another related model is factorial-LDA [47] which learns a facorized topic structure. Their approach to learn structured topics is different to VALTA in that factorial-LDA learns topics as *tuples* while VALTA learns topics as hierarchies. Other approaches similar to ours are [36, 37, 48, 49]. However, they are all purely probabilistic models and furthermore, they do not consider hierarchical topics.

VAEs have also been used for collaborative filtering [50, 51]. However, as pointed out by [52], these approaches tend to under-fit the data. In the vision domain, the idea of capitalizing on more complex priors in VAEs has become a popular idea and has been strongly associated with disentanglement [4, 53, 54]. However, this idea has not been emphasized as much in natural language processing. Two recent, closely related deep learning-based methods are [38, 39]. Both exploit the review texts for a pairs (i, u) to predict ratings. While these models perform well in terms of predicted ratings, they are not designed to learn topics.

Aspect classification has also been separately studied in the context of semantic analysis [41, 42, 55, 56]. To our knowledge, VALTA is the only VAE-based approach that considers hierarchical topics. Furthermore, our encoder architecture is unique in that it couples a sentence-level decoder with an item and user encoder.

5 Experiments

To assess the structured representation learned by VALTA, we evaluate a number of tasks and datasets. The experiments are designed to evaluate the quality of aspects and topics, structure of the representations, and rating-prediction. We implement all VAE-based models in Probabilistic Torch [3], a library for deep probabilistic models. In all experiments, we set the temperature for $\boldsymbol{z}_{i,u,s}$ and $\boldsymbol{\psi}_{i,u,a}$ to 0.66 and 5.0 respectively, which were we choose based on the quality of the learned topics. The default temperature for Gumbel softmax is 0.66, and we found it to be sufficiently low to capture sentence's aspect-assignments. For $\boldsymbol{\psi}_{i,u,a}$, we had to increase the temperature to encourage learning a diverse set of topics for each aspect. The number of hidden units H for all models is 256, followed by a $\tanh$ function. We ran all experiments for 200 epochs with batch-size 100 (at the review level). The optimizer we used was ADAM with default hyperparameters. We use $(A = 5, K = 2)$ for the beer review and $(A = 10, K = 3)$ for all other datasets. In order to be able to visualize all topics, we choose K here to be relatively small (e.g. 2 or 3). A on the other hand was chosen based on the dataset. In the beer example, we know that there are 5 aspects, so we set A to 5. For all other datasets we choose $A = 10$ and let the KL regularizer prune out the unnecessary aspects.

The main findings in our experiments are as follows.

1. VALTA can disentangle aspects of a review in a fully unsupervised manner. We demonstrate this on the CitySearch and BeerAdvocate datasets, which have been annotated with aspect-specific ratings [41, 57].
2. VALTA learns word distributions for every aspect and topic that have a higher coherence score [58] than baseline methods at both the sentence and review level. This indicates interpretability.
3. VALTA learns a representation that can be used to make aspect-based comparisons of items and users.
4. In all but one dataset, VALTA produces the most accurate rating predictions of all models considered.

5.1 Datasets and Preprocessing

A summary of the data we use in this paper is shown in Table 2. We focus on datasets that exhibit clear structure. We chose the *BeerAdvocate* dataset[1] and restaurant data from *Yelp Dataset Challenge*[2] from as they both contain explicit aspects. For the other two datasets, we selected Clothing and Movie reviews from Amazon [59, 60].

We preprocessed the datasets as follows. We used the Spacy library [61] (*version 2.0.12*) to remove all English standard stop words, and we removed all words that occurred fewer than five times overall. We also used Spacy for sentence

[1] http://snap.stanford.edu/data/
[2] https://www.yelp.com/dataset_challenge/

	BeerAdvocate	
	Sentence	Review
LDA	-	0.5756
Local-LDA	1.5196	-
NVDM	-	0.2338
ProdLDA	-	0.2208
VALTA	**1.817**	**0.655**

Table 4: Average Topic Coherence (NPMI). We compare either the sentence-level or review-level coherence score, depending on the input-level employed in the baseline.

Appearance		Aroma-Taste		Palate		Semantic	
golden	black	roasted	citrus	mouthfeel	mouthfeel	lagers	try
yellow	tans	coffee	grapefruit	bodied	watery	heineken	hype
white	glass	vanilla	pine	smooth	rjt	macro	recommend
orange	pour	chocolate	hops	carbonation	bodied	import	overall
hazy	head	bourbon	lemon	medium	refreshing	euro	founders
color	pitch	oak	floral	drinkability	carbonation	lager	favorite
gold	lacing	malts	clove	drinkable	crisp	bmc	stouts
copper	color	sweet	malt	alcohol	dry	worse	stout
straw	brown	aroma	aroma	finish	finish	bad	ipa
amber	ginger	malt	grass	mouth	thirst	skunky	cheers

Food			Service		Ambiance	Payment	
rice	pepperoni	bagel	service	friendly	walls	card	minutes
chicken	provolone	eggs	friendly	service	love	debit	card
sauce	mozzarella	hash	staff	staff	restaurant	stamp	seated
shrimp	bagel	scrambled	attentive	baristas	located	minutes	table
pork	onions	brown	food	helpful	wall	receipt	debit
fried	mushrooms	ham	helpful	employees	decor	cash	waited
beef	arugula	lox	server	customer	hidden	register	asked
noodles	knots	benedict	atmosphere	barista	ceiling	cards	wait
spicy	artichoke	poached	great	clean	gem	order	gratuity
salmon	cheese	capers	knowledgeable	rude	lighting	cashier	told

Table 5: Top 10 words learned by VALTA: beer (top) and restaurants (bottom).

segmentation with default configuration. The resultant vocabulary size for these datasets varies from 20000 to 30000. We also filtered reviews such that we include only items and users for which we have at least five reviews.

5.2 Baselines

We compare our model to a diverse set of baseline models, including probabilistic and VAE-based models. We also include a simple version of our own model that we call Variational Sentence Latent Topic Allocation (VSLTA) which follows the implementation of VALTA except for modelling the aspects. In other words, the representation learned by VSLTA is only a flat K dimensional vector rather than matrix of size $(A \times K)$. The full list of baselines are: LDA [18], Local-LDA [19], HFT [36], MF [33], NVDM [28], ProdLDA [29], and VSLTA.

5.3 Interpretablity

In Table 5, we show the top 10 words in topics for the BeerAdvocate and the Yelp data. Top words are decided based on the decoder weights with the strongest connection with each topic. Words associated in sub-aspects are clearly related to each other. For example, in the beer dataset, the topic "dark" contains words such as "black", "tans", and "brown". Furthermore, we can see that the topics within every aspect are also correlated with one another. In the beer example, if we look at the topic neighbours of "dark", we can see the topic "yellow". Note that the "dark" and "yellow" topics are learned within the *same* aspect in our model. The same pattern can be observed in the Yelp data where we can recover topics corresponding to food types, such as "Chinese", "Pizza", and "Breakfast".

5.4 Quantitative Assessment

We perform several quantitative evaluations of our model. We first demonstrate that we can successfully disentangle different aspects at the sentence level. We evaluate this on the two available sentence-annotated datasets: CitySearch [57] and BeerAdvocate [41], each containing 652 and 450 reviews respectively, in which sentences are labeled with an aspect by an annotator. We then compare the labels to inferred aspect assignments. Prior work on sentence aspect classification shows that Local-LDA is one of the most successful at capturing aspects in an unsupervised way [42]. Therefore, we compare against both LDA and Local-LDA. We also train fully supervised SVM classifier[3] on the labeled data. As presented in Table 3, VALTA outperforms other approaches in terms of both accuracy and F1-score.

Next, we quantitatively evaluate the top-words learned in topics. According to Lau et al. [58], NPMI is a good metric for qualitative evaluation of topics in terms of matching human judgment. Details of how we compute NPMI scores can be found in Appendix Section A. We measure NPMI

[3] We use the SVM implementation in scikit-learn 0.19.2 [62].

Model	ACC
LDA	0.13
NVDM	0.48
ProdLDA	0.45
VALTA	**0.53**

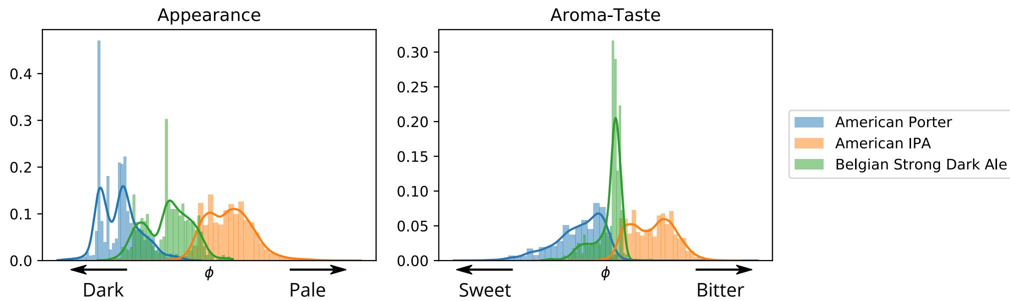

Figure 2: Left: table of category classification results on beer data. Right: normalized density of latent values $\psi_{i,u,a}$ of the encoded reviews of three different style of beer.

both at the sentence and the review level to take both aspects and topics into account. For every baseline, we only compared at the input-level that it was trained on. For example, LDA is trained at the review level while Local-LDA is trained at the sentence level. We report results in Table 4. We can see that VALTA performs significantly better than baselines, both at the sentence and review levels. Note that we keep the overall number topics for all other baselines to be the same as VALTA ($A \times K$).

5.5 Genre Discovery and Aspect-Based Analysis

In Table 5, we show that we can successfully learn a structured representation of aspects and topics. An interesting question is whether based on this representation, we have managed to cluster the items in a reasonable way. Furthermore, can we now perform an aspect-based comparison of different items? In this section, we investigate this question both qualitatively and quantitatively. We hypothesize that if the learned representation of the item is sufficiently rich in capturing the structure of the data, then even a simple classifier should be able to accurately distinguish between the categories. After training VALTA and other baselines, we fit a multi-class SVM to classify the items.[4] Results are shown in Figure 2 (left). VALTA outperforms other approaches with respect to clustering items in an unsupervised manner due to is structured nature.

To inspect VALTA's ability to enable aspect-based comparison, we perform the following experiments: for both the BeerAdvocate and Yelp restaurant data, we manually select three categories of items that are different with respect to every aspect. We encode all the reviews associated with these items and we plot the histogram of parameters $\rho_{i,u}$, as well as their kernel density estimation in different aspects.

Figure 2 shows that item representations cluster appropriately within topics. For example, American Porter is a sweet dark beer, thus the histogram of $\rho_{i,u}$ for American Porter is on the side of "dark" in the appearance aspect and on the side of "sweet" in the taste aspect. American IPA on the other hand is bitter pale beer, thus it has very little overlap with American Porter in either aspects. Note that in the beer example, we trained with $K = 2$. Since $\rho_{i,u}$ are parameters of the Concrete distribution, we only need to see value for 1 as the other one provides no addition information.

[4] We use the SVC implementation in sklearn 0.19.2.

Dataset	MF	LDA	HFT	VSLTA	VALTA
Beer	1.32	0.714	0.552	0.611	**0.437**
Yelp	2.32	1.894	**1.225**	1.540	1.236
Clothing	0.568	0.444	**0.316**	0.400	**0.315**
Movies	0.242	0.217	0.197	0.201	**0.154**

Table 6: MSE comparison with baselines on the test data.

5.6 Recommendation Performance

As noted in the methodology section, VALTA's generative model is also trained to predict rating for pairs of users and items, based on their aspect and topic representation. Results on MSE are shown in Table 6. It can be observed that VALTA outperforms baselines in two of the datasets by taking both the rating and our structured review representation into account, and perform reasonably close to state-of-art (HFT) in other cases.

6 Conclusion

We have proposed VALTA, a novel VAE-based model that instantiates structured probabilistic topic models in combination with an inference neural network to learn aspect-based representations of reviews. VALTA uncovers interpretable aspects, and additional structure (sub-aspects) beneath these. These representations enable one to measure similarity with respect to individual aspects, and thus perform aspect-wise clustering. Furthermore, we demonstrated the these representations afford improved generalization, as assessed in zero-shot settings.

Our hope is that structured (disentangled) representations will see increased development and use in natural language processing (NLP) applications, as these may allow greater generalizability and transparency.

Acknowledgements

We would like to thank our reviewers and the area chair for their thoughtful comments. This work was supported by NSF award 1835309. JWM would additionally like to acknowledge generous support from the Intel Corporation, the 3M Corporation, and startup funds from Northeastern University. Wallace's contribution was supported by the Army Research Office (ARO), award W911NF1810328.

References

[1] Diederik P. Kingma and Max Welling. "Auto-Encoding Variational Bayes". In: *International Conference on Learning Representations*. 2013.

[2] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. "beta-VAE: Learning basic visual concepts with a constrained variational framework". In: *International Conference on Learning Representations (ICLR)*. 2016.

[3] N. Siddharth, Brooks Paige, Jan-Willem van de Meent, Alban Desmaison, Noah D. Goodman, Pushmeet Kohli, Frank Wood, and Philip Torr. "Learning Disentangled Representations with Semi-Supervised Deep Generative Models". In: *Neural Information Processing Systems (NIPS)*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett. Curran Associates, Inc., 2017.

[4] Babak Esmaeili, Hao Wu, Sarthak Jain, Alican Bozkurt, N Siddharth, Brooks Paige, Dana H Brooks, Jennifer Dy, and Jan-Willem van de Meent. "Structured disentangled representations". In: *stat* 1050 (2018), p. 12.

[5] Shengjia Zhao, Jiaming Song, and Stefano Ermon. "InfoVAE: Information maximizing variational autoencoders". In: *arXiv preprint arXiv:1706.02262* (2017).

[6] Shuyang Gao, Rob Brekelmans, Greg Ver Steeg, and Aram Galstyan. "Auto-Encoding Total Correlation Explanation". In: *arXiv preprint arXiv:1802.05822* (2018).

[7] A. Achille and S. Soatto. "Information Dropout: Learning Optimal Representations Through Noisy Computation". In: *Transactions on Pattern Analysis and Machine Intelligence* (2018).

[8] Hyunjik Kim and Andriy Mnih. "Disentangling by factorising". In: *arXiv preprint arXiv:1802.05983* (2018).

[9] Tian Qi Chen, Xuechen Li, Roger Grosse, and David Duvenaud. "Isolating Sources of Disentanglement in Variational Autoencoders". In: *arXiv preprint arXiv:1802.04942* (2018).

[10] Diederik P Kingma and Max Welling. "Auto-encoding variational bayes". In: *arXiv preprint arXiv:1312.6114* (2013).

[11] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. "Stochastic backpropagation and approximate inference in deep generative models". In: *arXiv preprint arXiv:1401.4082* (2014).

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. "Generative Adversarial Nets". In: *Advances in Neural Information Processing Systems 27*. Ed. by Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger. Curran Associates, Inc., 2014, pp. 2672–2680.

[13] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. "Infogan: Interpretable representation learning by information maximizing generative adversarial nets". In: *Advances in Neural Information Processing Systems*. 2016, pp. 2172–2180.

[14] Sebastian Ruder, Parsa Ghaffari, and John G Breslin. "A Hierarchical Model of Reviews for Aspect-based Sentiment Analysis". In: *Empirical Methods in Natural Language Processing (EMNLP)*. 2016.

[15] Ruidan He, Wee Sun Lee, Hwee Tou Ng, and Daniel Dahlmeier. "An Unsupervised Neural Attention Model for Aspect Extraction". In: *Association for Computational Linguistics (ACL)*. 2017.

[16] Yuan Zhang, Regina Barzilay, and Tommi Jaakkola. "Aspect-augmented Adversarial Networks for Domain Adaptation". In: *arXiv preprint arXiv:1701.00188* (2017).

[17] Sarthak Jain, Edward Banner, Marshall Iain J. van de Meent Jan-Willem, and Byron C. Wallace. "Learning Disentangled Representations of Texts with Application to Biomedical Abstracts". In: *Empirical Methods in Natural Language Processing (EMNLP)*. 2018.

[18] David M Blei, Andrew Y Ng, and Michael I Jordan. "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan (2003), pp. 993–1022.

[19] Samuel Brody and Noemie Elhadad. "An unsupervised aspect-sentiment model for online reviews". In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics. 2010, pp. 804–812.

[20] Christina Sauper, Aria Haghighi, and Regina Barzilay. "Incorporating content structure into text analysis applications". In: *Empirical Methods in Natural Language Processing (EMNLP)*. 2010.

[21] Christina Sauper, Aria Haghighi, and Regina Barzilay. “Content models with attitude”. In: *Association for Computational Linguistics (ACL)*. 2011.

[22] Arjun Mukherjee and Bing Liu. “Aspect extraction through semi-supervised modeling”. In: *Association for Computational Linguistics (ACL)*. 2012.

[23] Christina Sauper and Regina Barzilay. “Automatic aggregation by joint modeling of aspects and values”. In: *Journal of Artificial Intelligence Research* (2013).

[24] Suin Kim, Jianwen Zhang, Zheng Chen, Alice H Oh, and Shixia Liu. “A Hierarchical Aspect-Sentiment Model for Online Reviews”. In: *AAAI Conference on Artificial Intelligence (AAAI)*. 2013.

[25] Michael Paul and Roxana Girju. “A two-dimensional topic-aspect model for discovering multi-faceted topics”. In: *AAAI Conference on Artificial Intelligence (AAAI)*. 2010.

[26] Michael Paul and Mark Dredze. “Factorial LDA: Sparse multi-dimensional text models”. In: *Advances in Neural Information Processing Systems (NIPS)*. 2012.

[27] Byron C Wallace, Michael J Paul, Urmimala Sarkar, Thomas A Trikalinos, and Mark Dredze. “A large-scale quantitative analysis of latent factors and sentiment in online doctor reviews”. In: *Journal of the American Medical Informatics Association* 21.6 (2014), pp. 1098–1103.

[28] Yishu Miao, Lei Yu, and Phil Blunsom. “Neural variational inference for text processing”. In: *International Conference on Machine Learning*. 2016, pp. 1727–1736.

[29] Akash Srivastava and Charles Sutton. “Autoencoding variational inference for topic models”. In: *arXiv preprint arXiv:1703.01488* (2017).

[30] Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. “Sparse Additive Generative Models of Text”. In: *Proceedings of the International Conference on Machine Learning (ICML)*. 2011.

[31] Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. “Toward Controlled Generation of Text”. en. In: *International Conference on Machine Learning*. July 2017, pp. 1587–1596.

[32] James Bennett, Stan Lanning, et al. “The netflix prize”. In: *Proceedings of KDD cup and workshop*. Vol. 2007. New York, NY, USA. 2007, p. 35.

[33] Yehuda Koren, Robert Bell, and Chris Volinsky. “Matrix factorization techniques for recommender systems”. In: *Computer* (2009), pp. 30–37.

[34] Andriy Mnih and Ruslan R Salakhutdinov. “Probabilistic matrix factorization”. In: *Advances in neural information processing systems*. 2008, pp. 1257–1264.

[35] Yang Bao, Hui Fang, and Jie Zhang. “TopicMF: Simultaneously Exploiting Ratings and Reviews for Recommendation.” In: *AAAI*. Vol. 14. 2014, pp. 2–8.

[36] Julian McAuley and Jure Leskovec. “Hidden factors and hidden topics: understanding rating dimensions with review text”. In: *Proceedings of the 7th ACM conference on Recommender systems*. ACM. 2013, pp. 165–172.

[37] Qiming Diao, Minghui Qiu, Chao-Yuan Wu, Alexander J Smola, Jing Jiang, and Chong Wang. “Jointly modeling aspects, ratings and sentiments for movie recommendation (JMARS)”. In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM. 2014, pp. 193–202.

[38] Lei Zheng, Vahid Noroozi, and Philip S Yu. “Joint deep modeling of users and items using reviews for recommendation”. In: *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. ACM. 2017, pp. 425–434.

[39] Rose Catherine and William Cohen. “Transnets: Learning to transform for recommendation”. In: *Proceedings of the Eleventh ACM Conference on Recommender Systems*. ACM. 2017, pp. 288–296.

[40] Zhiyong Cheng, Ying Ding, Lei Zhu, and Mohan Kankanhalli. “Aspect-Aware Latent Factor Model: Rating Prediction with Ratings and Reviews”. In: *Proceedings of the 2018 World Wide Web Conference on World Wide Web*. International World Wide Web Conferences Steering Committee. 2018, pp. 639–648.

[41] Julian McAuley, Jure Leskovec, and Dan Jurafsky. “Learning attitudes and attributes from multi-aspect reviews”. In: *Data Mining (ICDM), 2012 IEEE 12th International Conference on*. IEEE. 2012, pp. 1020–1025.

[42] Bin Lu, Myle Ott, Claire Cardie, and Benjamin K Tsou. “Multi-aspect sentiment analysis with topic models”. In: *2011 11th IEEE International Conference on Data Mining Workshops*. IEEE. 2011, pp. 81–88.

[43] Francisco R Ruiz, Michalis Titsias RC AUEB, and David Blei. “The generalized reparameterization gradient”. In: *Advances in neural information processing systems*. 2016, pp. 460–468.

[44] Michael Figurnov, Shakir Mohamed, and Andriy Mnih. “Implicit Reparameterization Gradients”. In: *arXiv preprint arXiv:1805.08498* (2018).

[45] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. “The concrete distribution: A continuous relaxation of discrete random variables”. In: *arXiv preprint arXiv:1611.00712* (2016).

[46] Eric Jang, Shixiang Gu, and Ben Poole. “Categorical reparameterization with gumbel-softmax”. In: *arXiv preprint arXiv:1611.01144* (2016).

[47] Michael Paul and Mark Dredze. "Factorial LDA: Sparse Multi-Dimensional Text Models". In: *Advances in Neural Information Processing Systems 25*. Ed. by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. Curran Associates, Inc., 2012, pp. 2582–2590. URL: http://papers.nips.cc/paper/4784-factorial-lda-sparse-multi-dimensional-text-models.pdf.

[48] Tong Zhao, Julian McAuley, and Irwin King. "Improving latent factor models via personalized feature projection for one class recommendation". In: *Proceedings of the 24th ACM international on conference on information and knowledge management*. ACM. 2015, pp. 821–830.

[49] Yongfeng Zhang, Guokun Lai, Min Zhang, Yi Zhang, Yiqun Liu, and Shaoping Ma. "Explicit factor models for explainable recommendation based on phrase-level sentiment analysis". In: *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM. 2014, pp. 83–92.

[50] Xiaopeng Li and James She. "Collaborative variational autoencoder for recommender systems". In: *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM. 2017, pp. 305–314.

[51] Giannis Karamanolakis, Kevin Raji Cherian, Ananth Ravi Narayan, Jie Yuan, Da Tang, and Tony Jebara. "Item Recommendation with Variational Autoencoders and Heterogenous Priors". In: *arXiv preprint arXiv:1807.06651* (2018).

[52] Dawen Liang, Rahul G. Krishnan, Matthew D. Hoffman, and Tony Jebara. "Variational Autoencoders for Collaborative Filtering". In: *arXiv:1802.05814 [cs, stat]* (Feb. 2018). arXiv: 1802.05814 [cs, stat].

[53] Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. "Semi-supervised learning with deep generative models". In: *Advances in Neural Information Processing Systems*. 2014, pp. 3581–3589.

[54] Siddharth Narayanaswamy, T Brooks Paige, Jan-Willem Van de Meent, Alban Desmaison, Noah Goodman, Pushmeet Kohli, Frank Wood, and Philip Torr. "Learning disentangled representations with semi-supervised deep generative models". In: *Advances in Neural Information Processing Systems*. 2017, pp. 5925–5935.

[55] Soujanya Poria, Erik Cambria, and Alexander Gelbukh. "Aspect extraction for opinion mining with a deep convolutional neural network". In: *Knowledge-Based Systems* 108 (2016), pp. 42–49.

[56] Kim Schouten and Flavius Frasincar. "Survey on aspect-level sentiment analysis". In: *IEEE Transactions on Knowledge & Data Engineering* (2016), pp. 1–1.

[57] Gayatree Ganu, Noemie Elhadad, and Amélie Marian. "Beyond the stars: improving rating predictions using review text content." In: *WebDB*. Vol. 9. Citeseer. 2009, pp. 1–6.

[58] Jey Han Lau, David Newman, and Timothy Baldwin. "Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality". In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*. 2014, pp. 530–539.

[59] Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton Van Den Hengel. "Image-based recommendations on styles and substitutes". In: *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM. 2015, pp. 43–52.

[60] Ruining He and Julian McAuley. "Ups and downs: Modeling the visual evolution of fashion trends with one-class collaborative filtering". In: *proceedings of the 25th international conference on world wide web*. International World Wide Web Conferences Steering Committee. 2016, pp. 507–517.

[61] Matthew Honnibal and Ines Montani. "spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing". In: *To appear* (2017).

[62] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. "Scikit-learn: Machine Learning in Python". In: *Journal of Machine Learning Research* 12 (2011), pp. 2825–2830.