Exploring Latent Semantic Factors to Find Useful Product Reviews

Subhabrata Mukherjee* Kashyap Popat* Gerhard Weikum*

Abstract
Online reviews provided by consumers are a valuable asset for e-Commerce platforms, influencing potential consumers in making purchasing decisions. However, these reviews are of varying quality, with the useful ones buried deep within a heap of non-informative reviews. In this work, we attempt to automatically identify review quality in terms of its helpfulness to the end consumers. In contrast to previous works in this domain exploiting a variety of syntactic and community-level features, we delve deep into the semantics of reviews as to what makes them useful, providing interpretable explanation for the same. We identify a set of consistency and semantic factors, all from the text, ratings, and timestamps of user-generated reviews, making our approach generalizable across all communities and domains. We explore review semantics in terms of several latent factors like the expertise of its author, his judgment about the fine-grained facets of the underlying product, and his writing style. These are cast into a Hidden Markov Model – Latent Dirichlet Allocation (HMM-LDA) based model to jointly infer: (i) reviewer expertise, (ii) item facets, and (iii) review helpfulness. Large-scale experiments on five real-world datasets from Amazon show significant improvement over state-of-the-art baselines in predicting and ranking useful reviews.

1 Introduction

Motivation: With the rapid growth in e-Commerce, product reviews have become a crucial component for the business. As consumers cannot test the functionality of a product prior to purchase, these reviews help them make an informed decision to buy the product or not. As per a survey conducted by Nielsen Corporations, 40% of online consumers indicated that they would not buy electronics without consulting online reviews first [2].

Due to the increasing dependency on user-generated reviews, it is crucial to understand their quality — that can widely vary from being an excellent-detailed opinion to superficial criticizing or praising, to spams in the worst case. Without any indication of the review quality, it is overwhelming for consumers to browse through a multitude of reviews. In order to help consumers in finding useful reviews, most of the e-Commerce platforms nowadays allow users to vote whether a product review is helpful or not. For instance, any Amazon product review is accompanied with information like x out of y users found the review helpful. This helpfulness score ($x/y$) can be considered as a proxy for the review quality, and its usefulness to the end customers. In this work, we aim to automatically find the helpfulness score of a review based on certain consistency, and semantic aspects of the review like: whether the review is written by an expert, what are the important facets of the product outlined in his review, what do other experts have to say about the given product, timeliness of the review etc. — that are automatically mined as latent factors from review texts.

2 Related Research and their Limitations

Predicting Review Helpfulness and Spams: Prior works on predicting review usefulness mostly operate on shallow syntactic textual features like bag-of-words, part-of-speech tags, and tf-idf (term, and inverse document frequency) statistics [5,10]. These works, and other related works on finding review spams [4,13] classify extremely opinionated reviews as not helpful. Similarly, other works exploiting rating & activity features like frequency of user posts, average ratings of users and items [8,10,17] consider extreme ratings and deviations as indicative of unhelpful reviews. Some recent works incorporate additional information like community-specific characteristics (who-voted-whom) with explicit user network [10,19], and item-specific meta-data like explicit item facets and product brands [5,9].

Apart from the requirement of a large number of meta-features that restrict the generalizability of many of these models to any arbitrary domain, these shallow features do not analyze what the review is about, and, therefore, cannot explain why it should be helpful for a given product. Some of these works [8,17] identify expertise of a review’s author as an important feature. However, in absence of suitable modeling techniques, they consider prior reputation features like user activity, and low rating deviation as proxy for user expertise.

Latent Factors for Review Analysis: Prior approaches for analyzing review texts aim to learn latent topics [7], latent aspects and their ratings [6,11,20], and user-user interactions [21]. The author writing style is also used in [4]. However, these prior approaches do not factor in the temporal dynamics and user expertise.

Modeling Expertise: Our model for capturing user expertise draws motivation from [12,15,16] with signifi-
3 Overview of our Approach and Contributions

Our work aims to overcome the limitations of prior works by exploring the semantics and consistency of a review to predict its helpfulness for a given product. The first step towards understanding the semantics of a review is to uncover the facet descriptions of the target product outlined in the review. We treat these facets as latent and use Latent Dirichlet Allocation (LDA) to discover them as topic clusters. The second step is to find the expertise of the users who wrote the review, and their description of the different (latent) facets of the product.

In this work, we model expertise as a latent variable that evolves over time using Hidden Markov Model (HMM).

We make use of distributional hypotheses like: expert users agree on what are the important facets of a product, and their description (or, writing style) of those facets influences the helpfulness of a review. We also derive several consistency features — all from the given quintuple ⟨userId, itemId, rating, reviewText, timepoint⟩ — like prior user reputation, item prominence, and timeliness of a review. Finally, we leverage the interplay between all of the above factors in a joint setting to predict the review helpfulness.

For interpretable explanation, we derive interesting insights from the latent word clusters used by experts — for instance, reviews describing the underlying “theme and storytelling” of movies and books, the “style” of music, and “hygiene” of food are considered most helpful for the respective domains.

In summary, we make the following novel contributions:

a) Model: We propose an approach to leverage the semantics and consistency of reviews to predict their helpfulness. We propose a Hidden Markov Model – Latent Dirichlet Allocation (HMM-LDA) based model that jointly learns the (latent) product facets, (latent) user expertise, and his writing style from observed words in reviews at explicit timepoints.

b) Algorithm: We introduce an effective learning algorithm based on an iterative stochastic optimization process that reduces the mean squared error of the predicted helpfulness scores with the ground scores, as well as maximizes the log-likelihood of the data.

c) Experiments: We perform large-scale experiments with real-world datasets from five different domains in Amazon, together comprising of 29 million reviews from 5.7 million users on 1.9 million items, and demonstrate substantial improvement over state-of-the art baselines for prediction and ranking tasks.

4 Review Helpfulness Factors

In this section, we outline the components of our model that analyze the semantics and consistency of reviews.

**Item Facets:** It is essential to understand the different facets of an item in a review. For instance, a camera review can focus on facets like “resolution”, “zoom”, “price”, “size”, or a movie review can focus on “narration”, “cinematography”, “acting”, “direction” etc. However, not all facets are equally important for an item. For example, a review downrating a camera for “late delivery” by the seller is not as helpful to end consumers as opposed to downrating it due to grainy resolution or shaky zoom. Therefore, a helpful review should focus on the important facets of an item. We model facets as latent variables, where the item’s latent facet distribution in the review is indicative of how detailed and diverse the review is.

**Review Writing Style:** Words used to describe the facets play a crucial role in making the review useful to the consumers. An important aspect of an expert writing style is to use precise, domain-specific vocabulary to describe a facet in details, rather than using generic words. For instance, contrast this expert camera review: “60D focus screen is ‘grainy’; It is the ‘precision matte’ surface that helps to increase contrast and minimize depth of field for manual focusing. The Ef-s screen is even more so for use with fast primes…” with this amateur one: “This camera is pure garbage. It is the worst one I have ever owned. I bought it last xmas on a deal and have thrown it away and replaced it with a decent camera.” We learn a language model from the latent facets and user expertise that helps to distinguish the writing style of an experienced user from an amateur one.

**Reviewer Expertise:** Prior works used proxy features like user activity and reputation (e.g. number of reviews written, feedback from community etc.) to harness users’ expertise under the hypothesis that expert reviews are positively correlated to review helpfulness. We explicitly model user expertise, drawing motivation from recent works with substantial modifications for tractable inference (refer Section 5.4). Expertise is not static, but evolves over time. A user who was amateur at the time of entering a community, may have become an expert now. We model expertise as a latent variable that evolves over time, exploiting the hypothesis that users at similar levels of expertise have similar rating behavior, facet preferences, and writing style.
5 Joint Model for Review Helpfulness

5.1 Incorporating Consistency Factors

Let $u \in U$ be a user writing a review at time $t \in T$ on an item $i \in I$. Let $d = \{w_1, w_2, ..., w_{|W_d|}\}$ be the corresponding review text with a sequence of words $\langle w \rangle$, and rating $r \in R$. Each such review is associated with a helpfulness score $h \in [0 - 1]$. Let $b_i$ be the corresponding timeliness of the review computed as $exp^{-\left(t-t_{i,0}\right)}$, where $t_{i,0}$ is the first review on the item $i$.

Let $\beta_u$ be the average helpfulness score of user $u$ over all the reviews written by her (capturing user reputation), and $\beta_1$ be the average helpfulness score of all reviews for item $i$ (capturing item prominence). Let $\tau_u$ be the average rating assigned by the user over all items, $\tau_i$ be the average rating assigned to the item by all users, and $\tau_g$ be the average global rating over all items and users. Consistency features include prior item and user reputation, deviation features, and burst.

Let $\xi$ be a tensor of dimension $E \times Z$, where $E$ is the number of expertise levels of the users, and $Z$ is the number of latent facets of the items. $\xi_{e,z}$ depicts the opinion of users at (latent) expertise level $e \in E$ about the (latent) facet $z \in Z$. Therefore, the distributional hypotheses (outlined in Section 4) are intrinsically integrated in $\xi$ that is estimated from the reviews’ text, conditioned on reviews’ helpfulness scores.

The estimated helpfulness score $\hat{h}(u, i)$ of a review by user $u$ on item $i$ is a function $f$ of the following consistency and latent factors, parametrized by $\Psi$:

$$\hat{h}(u, i) = f(\beta_u, \beta_i, |r - \tau_u|, |r - \tau_i|, |r - \tau_g|, b_i, \xi; \Psi)$$

Here, $f$ can be a polynomial, radial basis, or a simple linear function for combining the features. The objective is to estimate the parameters $\Psi$ (of dimension: $6 + E \times Z$) that reduces the mean squared error of the predicted helpfulness scores with the ground scores:

$$\Psi^* = \arg\min_{\Psi} \frac{1}{|U|} \sum_{u,i \in U \cap I} (h(u, i) - \hat{h}(u, i))^2 + \mu \|\Psi\|_2^2$$

where, we use $L_2$ regularization for the parameters to penalize complex models. There are several ways to estimate the parameters like alternate least squares, gradient-descent, and Newton based approaches.

5.2 Incorporating Latent Facets

We use principles of Latent Dirichlet Allocation (LDA) [1] to learn the latent facets associated to an item. Each review $d$ on an item is assumed to have a Multinomial distribution $\theta$ over facets $Z$ with a symmetric Dirichlet prior $\alpha$. Each facet $z$ has a Multinomial distribution $\phi_z$ over words drawn from a vocabulary $W$ with a symmetric Dirichlet prior $\delta$. Exact inference is not possible due to the intractable coupling between $\Theta$ and $\Phi$. Two popular ways for approximate inference are MCMC techniques like Collapsed Gibbs Sampling and Variational Inference.

5.3 Incorporating Latent Expertise

Expertise influences both the facet distribution $\Theta$, as users at different levels of expertise have different facet preferences,
and the language model $\Phi$ as the writing style is also different for users at different levels of expertise. Therefore, we parametrize both of these distributions with user expertise similar to the prior work in [16], with some major modifications (discussed in the next section).

Consider $\Theta$ to be a tensor of dimension $E \times Z$, and $\Phi$ to be a tensor of dimension $E \times Z \times W$, where $\theta_{e,z}$ denotes the preference for facet $z \in Z$ for users at expertise level $e \in E$, and $\phi_{e,z,w}$ denotes the probability of the word $w \in W$ being used to describe the facet $z$ by users at expertise level $e$.

Now, expertise changes as users evolve over time. However, the transition should be smooth. Users cannot abruptly jump from expertise level 1 to 4 without passing through expertise levels 2 and 3. Therefore, at each timepoint $t+1$ (of posting a review), we assume a user at expertise level $e_t \in E$ to stay at $e_t$, or move to $e_t+1$ (i.e., expertise level is monotonically non-decreasing). This progression depends on how the writing style (captured by $\Phi$), and facet preferences (captured by $\Theta$) of the user is evolving with respect to other expert users in the community; as well as the rate of activity of the user — that we use as a hyper-parameter for controlling the rate of progression. Let $\gamma_u$, the activity rate of user $u$, be defined as: $\gamma_u = \frac{D_u}{D_u + D_{avg}}$, where $D_u$ and $D_{avg}$ denote the number of posts written by $u$, and the average number of posts written by any user in the community, respectively.

Let $\Pi$ be a tensor of dimension $E \times E$ with hyper-parameters $\langle \gamma_u \rangle$ of dimension $U$, where $\pi_{e_i,e_j}$ denotes the probability of moving to expertise level $e_j$ from $e_i$ with the constraint $e_j \in \{e_i, e_i+1\}$. However, not all users start at the same level of expertise, when they enter the community; some may enter already as experts. The algorithm figures this out during the inference process. We assume all users to start at expertise level 1 during parameter initialization.

During inference, we want to learn the parameters $\Psi, \xi, \Theta, \Phi, \Pi$ jointly for predicting review helpfulness.

5.4 Difference with Prior Works for Modeling Expertise Our generative process of user expertise is motivated by [12][16] with the following differences: (i) The prior works learn user-specific preferences for personalized recommendation. However, we assume users at the same level of expertise to have similar facet preferences. Therefore, the facet distribution $\Theta$ is conditioned only on the user expertise, and not the user explicitly, unlike the prior works. This helps us to reduce the dimensionality of $\Theta$, and exploit the correspondence between $\Theta$ and $\xi$ to tie the parameters of the consistency and latent factor models together for tractable inference. (ii) The prior work [16] incorporates supervision, for predicting ratings, only indirectly via optimizing the Dirichlet hyper-parameters $\alpha$ of the Multinomial facet distribution $\Theta$ — and cannot guarantee an increase in data log-likelihood over iterations. In contrast, we exploit (i) to learn the expertise-facet distribution $\Theta$ directly from the review helpfulness scores by minimizing the mean squared error during inference. This is also tricky as parameters of the distribution $\Theta$, for an unconstrained optimization, are not guaranteed to lie on the simplex — for which we do certain transformations, discussed during inference. Therefore, parameters are strongly coupled in our model, not only reducing mean squared error, but also leading to a near smooth increase in data log-likelihood over iterations (refer Figure[2]).

5.5 Generative Process Consider a corpus $D = \{d_1, \ldots, d_D\}$ of reviews written by a set of users $U$ at timestamps $T$. For each review $d \in D$, we denote $u_d$ as its user, $t_d$ as the timestamp of the review. Reviews are assumed to be ordered by timestamps, i.e., $t_d < t_{d'j}$ for $i < j$. Each review $d \in D$ consists of a sequence of $N_d$ words denoted by $d = \{w_1, \ldots, w_{N_d}\}$, where each word is drawn from a vocabulary $W$ with unique words indexed by $\{1 \ldots W\}$, and $Z$ is the number of facets.

Let $e_d \in \{1, 2, \ldots, E\}$ denote the expertise value of review $d$. Since each review $d$ is associated with a unique timestamp $t_d$ and unique user $u_d$, the expertise value of a review refers to the expertise of the user at the time of writing it. Following Markovian assumption, the user’s expertise level transitions follow a distribution $\Pi$ with the Markovian assumption $e_{u_d} \sim \pi_{e_{u_d-1}}$, i.e. the expertise level of user $u_d$ at time $t_d$ depends on her expertise level when writing the previous review at time $t_{d-1}$.

Once expertise level $e_d$ of user $u_d$ for review $d$ is known, her facet preferences are given by $\theta_{e_d}$. Thereafter, the facet $z_{d,u}$ of each word $w$ in $d$ is drawn from a Multinomial $\{\theta_{e_d}\}$. Now that the expertise level of a user, and her facets of interest are known, we can generate the language model $\Phi$ and individual words in the review — where the user draws a word from the Multinomial distribution $\phi_{e_d,z_{d,u}}$ with a symmetric Dirichlet prior $\delta$.

Refer Figure[1] for the generative process.

The joint probability distribution is given by:

\[
P(E, Z, W, \Theta, \Phi; U; \langle \gamma_u \rangle, \delta) \propto \prod_{u \in U} \prod_{d \in D_u} P(\pi_{e_d}; \gamma_u) \cdot P(e_d|\pi_{e_d}) \cdot \left( \prod_{j=1}^{N_d} P(z_{d,j}|\theta_{e_d}) \cdot P(\phi_{e_d,z_{d,j}}|\phi_{e_d,z_{d,j}}) \cdot P(w_{d,j}|\phi_{e_d,z_{d,j}}) \right)
\]

5.6 Inference Given a corpus of reviews indexed by $\langle\text{userId}, \text{itemId}, \text{rating}, \text{reviewText}, \text{timepoint}\rangle$, with corresponding helpfulness scores, our objective is to learn the parameters $\Psi$ that minimizes the mean squared error given by Equation 5.2.
In case $\xi$ was known, we could have directly plugged in its values (other features being observed) in Equation (5.1) to learn a model (e.g., using regression) with parameters $\Psi$. However, the dimensions of $\xi$, corresponding to both facets and user expertise, are latent that need to be inferred from text. Now, the parameter $\psi_{e,z}$ corresponding to $\xi_{e,z}$ learned from Equation (5.2) depicts the importance of the facet $z$ for users at expertise level $e$ for predicting review helpfulness. We want to exploit this observation to infer the latent dimensions from text.

During the generative process of a review document, for a user at expertise level $e$, we want to draw her facet of interest $z$ with probability $\theta_{e,z} \propto \psi_{e,z}$. However, we cannot directly replace $\Theta$ with $\Psi$ due to the following reason. The traditional parametrization of a Multinomial distribution ($\Theta$ in this case) is via its mean parameters. Any unconstrained optimization will take the parameters out of the feasible set, i.e. they may not lie on the simplex. Hence, it is easier to work with the natural parameters instead. If we consider the unconstrained parameters $\langle \psi_{e,z} \rangle$ (learned from Equation (5.2)) to be the natural parameters of the Multinomial distribution $\Theta$, we need to transform the natural parameters to the mean parameters that lie on the simplex (i.e. $\sum_z \theta_{e,z} = 1$). In this work, we follow the principle similar to [15] to do this transformation:

\begin{align}
\theta_{e,z} = \frac{\exp(\psi_{e,z})}{\sum_z \exp(\psi_{e,z})}
\end{align}

where, $\psi_{e,z}$ corresponds to the learned parameter for $\xi_{e,z}$.

Exploiting conjugacy of the Multinomial and Dirichlet distributions, we can integrate out $\Phi$ from the joint distribution in Equation (5.3) to obtain the posterior distribution $P(W|Z,E;\delta)$ given by:

\begin{align}
\prod_{e=1}^{E} \prod_{z=1}^{Z} \Gamma(\sum_{w} \delta) \prod_{w} \Gamma(n(e, z, w) + \delta) \prod_{w} \Gamma(\delta) \Gamma(\sum_{w} n(e, z, w) + \sum_{w} \delta)
\end{align}

where, $\Gamma$ denotes the Gamma function, and $n(e, z, w)$ is the number of times the word $w$ is used for facet $z$ by users at expertise level $e$.

We use Collapsed Gibbs Sampling [3], as in standard LDA, to estimate the conditional distribution for each of the latent facets $z_{d,j}$, which is computed over the current assignment for all other hidden variables, after integrating out $\Phi$. In the following equation, $n(e, z, .)$ indicates the summation of the counts over all possible $w \in W$. The subscript $-j$ denotes the value of a variable excluding the data at the $j$th position.

The posterior distribution $P(Z|\Phi, W, E)$ of the latent variable $Z$ is given by:

\begin{align}
P(z_{d,j} = k|z_{d,-j}, \Phi, w_{d,j} = w, e_d = e, d)
\propto \theta_{e,k} \cdot n(e, k, w) + \delta
\end{align}

\begin{align}
= \frac{\exp(\psi_{e,k}) \cdot n(e, k, w) + \delta}{\sum_{z} \exp(\psi_{e,z}) \cdot n(e, k, w) + \delta}
\end{align}

Similar to the above process, we use Collapsed Gibbs Sampling [3] also to sample the expertise levels, keeping all facet assignments $Z$ fixed. Let $n(e_{i-1}, e_i)$ denote the number of transitions from expertice level $e_{i-1}$ to $e_i$ over all users in the community, with the Markovian constraint $e_i \in \{e_{i-1}, e_i + 1\}$.

\begin{align}
P(e_i|e_{i-1}, e_{-i}, u; \gamma_u) = \frac{n(e_{i-1}, e_i) + I(e_{i-1} = e_i) + \gamma_u}{n(e_{i-1}, .) + I(e_{i-1} = e_i) + \gamma_u}
\end{align}

where $I(.)$ is an indicator function taking the value 1 when the argument is true (a self-transition, in this case, where the user has the same expertise level over subsequent reviews), and 0 otherwise. The subscript $-i$ denotes the value of a variable excluding the data at the $i$th position.

The conditional distribution for the expertise level transition is given by:

\begin{align}
P(E|U, Z, W; \langle \gamma_u \rangle) \propto P(E|U; \langle \gamma_u \rangle) \cdot P(Z|E) \cdot P(W|Z, E)
\end{align}

Using Equations (5.5), (5.6), and (5.7) we obtain the conditional distribution for updating latent variables $E$ as:

\begin{align}
P(e_{ud} = e_i|e_{ud-1} = e_{i-1}, u_d = u, \{z_{i,j} = z_j\}, \{w_{i,j} = w_j\}, e_{-i})
\propto \frac{n(e_{i-1}, e_i) + I(e_{i-1} = e_i) + \gamma_u}{n(e_{i-1}, .) + I(e_{i-1} = e_i) + \gamma_u}
\cdot \left( \prod_{j} \frac{\exp(\psi_{e_{i,z}})}{\sum_{z} \exp(\psi_{e_{i,z}})} \cdot \frac{n(e_i, z_j, w_j) + \delta}{n(e_i, z_j, .) + W \cdot \delta} \right)
\end{align}
Consider a document $d$ containing a sequence of words $\{w_j\}$ with corresponding facets $\{z_j\}$. The first factor models the probability of the user $u_d$ reaching expertise level $e_{u_d}$ for document $d$; whereas the second and third factor models the probability of the facets $\{z_j\}$ being chosen at the expertise level $e_{u_d}$, and the probability of observing the words $\{w_j\}$ with the facets $\{z_j\}$ and expertise level $e_{u_d}$, respectively. Following the Markovian assumption, we only consider the expertise levels $e_{u_d}$ and $e_{u_d} + 1$ for sampling, and select the one with the highest conditional probability.

Samples obtained from Gibbs sampling are used to approximate the expertise-facet-word distribution $\Phi$:

$$
\phi_{e,z,w} = \frac{n(e, z, w) + \delta}{n(e, z, w) + \mathbf{W} \cdot \delta}
$$

Once the generative process for a review $d$ with words $\{w_j\}$ is over, we can estimate $\xi$ from $\Phi$ as the proportion of the $z^*$th facet in the document written at expertise level $e$ as:

$$
\xi_{e,z} \propto \sum_{j=1}^{N_d} \phi_{e,z,w_j}
$$

In summary, $\xi$, $\Phi$, and $\Theta$ are linked via $\Psi$:

i) $\Psi$ generates $\Theta$ via Equation 5.4

ii) $\Theta$ and $\Phi$ are coupled in Equations 5.3, 5.5

iii) $\Phi$ generates $\xi$ using Equation 5.10

iv) $\Psi$ is learned via regression (with $\xi$ as latent features) using Equations 5.1, 5.2 so as to minimize the mean squared error for predicting review helpfulness.

Overall Processing Scheme: Exploiting results from the above discussions, the overall inference is an iterative stochastic optimization process consisting of the following steps:

i) Sort all reviews by timestamps, and estimate $E$ using Equation 5.8 by Gibbs sampling. During this process, consider all facet assignments $Z$ and $\Psi$, from the earlier iteration fixed.

ii) Estimate facets $Z$ using Equation 5.5 by Gibbs sampling, keeping the expertise levels $E$ and $\Psi$, from the earlier iteration fixed.

iii) Estimate $\xi$ using Equations 5.9 and 5.10

iv) Learn $\Psi$ from $\xi$ and other consistency factors using Equations 5.1, 5.2 by regression.

v) Estimate $\Theta$ from $\Psi$ using Equation 5.4

Regression: For regression, we use the fast and scalable Support Vector Regression implementation from LibLinear (www.csie.ntu.edu.tw/ cjlin/liblinear) that uses Trust Region Newton method for learning the parameters $\Psi$.

| Factors | Books | Music | Movie | Electronics | Food |
|---------|-------|-------|-------|-------------|------|
| #Users  | 2,588,991 | 1,134,684 | 889,176 | 811,034 | 256,059 |
| #Reviews| 929,264 | 556,814 | 251,059 | 82,067 | 74,258 |
| #Items  | 12,886,488 | 6,396,350 | 7,911,684 | 1,241,778 | 568,454 |
| #Reviews| 4.98 | 5.64 | 8.89 | 1.53 | 2.22 |
| #Items  | 13.86 | 11.48 | 31.26 | 15.13 | 7.65 |
| #Reviews| 9.71 | 5.95 | 7.90 | 8.91 | 4.24 |

Table 1: Dataset statistics. Votes indicate the total number of helpfulness votes (both, for and against) cast for a review. Total number of users = 5, 679, 944, items = 1, 895, 462, and reviews = 29, 004, 754.

Test: Given a test review with $\langle user=u, item=i, words=\{w_j\}, rating=r, timestamp=t\rangle$, we find its helpfulness score by plugging in the consistency features, and latent factors in Equation 5.1 with the parameters $\langle \Psi, \beta_u, \beta_i, r_u, r_i, r_g \rangle$ having been learned from the training data. $\xi$ is computed over the words $\{w_j\}$ using Equation 5.10 — with the distribution $\Phi$ having been learned during training.

6 Experiments

Setup: We perform experiments with data from Amazon in five different domains: (i) movies, (ii) music, (iii) food, (iv) books, and (v) electronics. The statistics of the dataset (snap.stanford.edu/data/) is given in Table 1. In total, we have 29 million reviews from 5.6 million users on 1.8 million items from all of the five domains combined. We extract the following quintuple for our model $\langle user=\text{id}, \text{itemId}, \text{timestamp}, \text{rating}, \text{helpfulnessVotes} \rangle$ for each domain. During training, for movies, books, music, and electronics, we consider only those reviews for which at least $y \geq 20$ users have voted about their helpfulness (including for, and against) to have a robust dataset (similar to the setting in [9,17]) for learning. Since the food dataset has less number of reviews, we lowered this threshold to five. For test, we used the 3 most recent reviews of each user as withheld test data (similar to the setting in [12,16]), that received at least five votes (including for, and against). The same data is used for all the models for comparison. We group long-tail users with less than 10 reviews in training data into a background model, treated as a single user, to avoid modeling from sparse observations. We do not ignore any user. During the test phase for a “long-tail” user, we take her parameters from the background model. We set the number of facets as $Z = 50$, and number expertise levels as $E = 5$, for all the datasets.

Tasks and Evaluation Measures: We use all the models for the following tasks:

1) Prediction: Here the objective is to predict the helpfulness score of a review as $x/y$, where $x$ is the number of users who voted the review as helpful out of $y$ number of users. As evaluation measures, we report:
(i) **Mean squared error** of the predicted scores with the ground helpfulness scores (using Equation 5.2), and (ii) **Squared correlation coefficient** ($R^2$) that gives an indication of the goodness of fit of a model, i.e., how well the regression function approximates the real data points, with $R^2 = 1$ indicating a perfect fit.

2) **Ranking**: A more suitable way of evaluation is to compare the ranking of reviews from different models sorted on their (predicted) helpfulness scores — where the reviews at the top of the rank list should be more helpful than the ones below — and compute rank correlation with the gold/reference rank list (sorted by ground-truth helpfulness scores $x/y$) using the following measures: (i) **Spearman correlation** ($\rho$) that assesses how well the relationship between two variables can be described using a monotonic function, and (ii) **Kendall-Tau correlation** ($\tau$) that measures the number of concordant and discordant pairs, to find whether the ranks of two elements agree or not based on their scores, out of the total number of combinations possible.

### 6.1 Quantitative Comparison Baselines:

We consider the following baselines to compare our work:

(a) P.O’ Mahony et al. (RecSys, 2009) \[17\] use several rating based features as proxy for reviewer reputation and sentiment; review length and letter cases for content; and review count statistics for social features.

(b) Lu et al. (WWW, 2010) \[10\] use syntactic features (part-of-speech tags of words), sentiment (using a lexicon to find word polarities), review length and reviewer rating statistics to predict the quality of a review.

(c) Kim et al. (EMNLP, 2006) \[5\] use structural (review length statistics), lexical (tf-idf), syntactic (part-of-speech tags), semantic (explicit product features, and sentiment of words), and meta-data related features to rank the reviews based on their helpfulness.

(d) Liu et al. (ICDM 2008) \[9\] predict the helpfulness of reviews on IMDB based on: reviewer expertise, syntactic features, and timeliness of a review. The authors use reviewer preferences for explicit facets (pre-defined genres of movies in IMDB) as proxy for their expertise, part-of-speech tags of words for syntactic features, and review publication dates for “timeliness” of reviews. This baseline is the closest to our work as we attempt to model similar factors. However, we model reviewer expertise explicitly, and the facets as latent — therefore not relying on any additional item meta-data (like, genres) or proxies for user authority.

For all of the above baselines, we use all the features from their works that are supported by our dataset — for instance, we could not use the social network, and explicit product meta-data absent in our dataset — for a fair comparison. Table 2 shows the comparison of the **Mean Squared Error** (MSE) and **Squared Correlation Coefficient** ($R^2$) for review helpfulness predictions, as generated by our model with the four baselines. Our model consistently outperforms all baselines in reducing the MSE. Table 9 shows the comparison of the Spearman ($\rho$) and Kendall-Tau ($\tau$) correlation between the rank list of helpful user reviews, as generated by all the models, and the gold rank list.

The most competitive baseline for our model is \[9\]. Due to a high overlap in consistency features of our model with this baseline, our performance improvement can be attributed to the incorporation of latent factors in our model. We perform paired sample t-tests, and find that our performance improvement over all the baselines is statistically significant at $p$-value $< 2e-16$. We perform the best for the domains movies, music, and books with large number of reviews, and relatively worse in the domains food, electronics due to data sparsity for which user maturity could not be captured well.

### 6.2 Qualitative Comparison

**Log-likelihood of data and convergence:** The inference of our model is quite involved with the coupling between several variables, and the alternate stochastic optimization process. Figure 2 shows the increase in the data log-likelihood of our model per-iteration for different domains. We observe that the model is stable, and achieves a near smooth increase in the data log-likelihood per-iteration. It also converges quite fast between 20 – 30 iterations depending on the complexity of the dataset. For electronics the convergence is quite rapid as the data is quite sparse, and the model does not find sufficient evidence for categorizing users to different expertise levels; this behavior is reflected in all the experiments involving the electronics dataset.

**Language model and facet preference divergence:** Figures 3a and 3b show the heatmaps of the Kullback-Leibler (KL) divergence for facet preferences and language models of users at different expertise levels, as computed by our model conditioned on review helpfulness — given by $D_{KL}(\theta_e || \theta_i)$ and $D_{KL}(\phi_e || \phi_i)$ respectively, where $\Theta$ and $\Phi$ are given by Equations 5.4 and 5.9 respectively.

The main observation is that the KL divergence is higher — the larger the difference is between the expertise levels of two users. This confirms our hypothesis that expert users have a distinctive writing style and facet preferences different than that of amateurs — captured by the joint interactions between review helpfulness, reviewer expertise, facet preferences, and writing style. We also note that the increase in divergence with the increase in gap between expertise levels is not smooth for food and electronics due to sparsity of per-user data.
Table 2: Prediction Task: Performance comparison of our model versus baselines. Our improvements over the baselines are statistically significant at p-value < 2.2e−16 using paired sample t-test.

| Models                      | Mean Squared Error (MSE) | Squared Correlation Coefficient ($R^2$) |
|-----------------------------|--------------------------|---------------------------------------|
|                            | Movies | Music | Books | Food | Electr. | Movies | Music | Books | Food | Electr. |
| Our model                   | 0.058  | 0.059 | 0.055 | 0.053| 0.050   | 0.438  | 0.405 | 0.397| 0.345| 0.197  |
| a) P.O’Mahony et al. [17]   | 0.067  | 0.069 | 0.069 | 0.060| 0.064   | 0.325  | 0.295 | 0.249| 0.312| 0.134  |
| b) Lu et al. [10]           | 0.093  | 0.087 | 0.077 | 0.072| 0.071   | 0.111  | 0.128 | 0.139| 0.134| 0.056  |
| c) Kim et al. [15]          | 0.107  | 0.125 | 0.094 | 0.075| 0.161   | 0.211  | 0.025 | 0.211| 0.065| 0.065  |
| d) Liu et al. [14]          | 0.091  | 0.091 | 0.082 | 0.075| 0.063   | 0.076  | 0.053 | 0.076| 0.039| 0.043  |

Table 3: Ranking Task: Correlation comparison between the ranking of reviews and gold rank list — our model versus baselines. Our improvements over the baselines are statistically significant at p-value < 2.2e−16 using paired sample t-test.

| Models                      | Spearman ($\rho$) | Kendall-Tau ($\tau$) |
|-----------------------------|-------------------|----------------------|
|                            | Movies | Music | Books | Food | Electr. | Movies | Music | Books | Food | Electr. |
| Our model                   | 0.657  | 0.610 | 0.603 | 0.533| 0.394   | 0.475  | 0.440 | 0.435| 0.387| 0.280  |
| a) P.O’Mahony et al. [17]   | 0.591  | 0.554 | 0.496 | 0.541| 0.340   | 0.414  | 0.390 | 0.347| 0.398| 0.237  |
| b) Lu et al. [10]           | 0.330  | 0.349 | 0.334 | 0.367| 0.205   | 0.224  | 0.242 | 0.230| 0.259| 0.144  |
| c) Kim et al. [15]          | 0.489  | 0.166 | 0.474 | 0.551| 0.261   | 0.342  | 0.114 | 0.334| 0.414| 0.184  |
| d) Liu et al. [14]          | 0.268  | 0.292 | 0.258 | 0.199| 0.159   | 0.183  | 0.161 | 0.178| 0.143| 0.112  |

Top words used by experts in most helpful reviews.

Music: album, lyrics, recommend, soundtrack, touch, songwriting, features, rare, musical, ears, lyrical, enjoy, absolutely, musically, individual, bland, soothing, released, inspiration, share, mainstream, deeper, flawesome, wonderfully, eclectic, heavily, critics, presence

Books: serious, complex, claims, content, illustrations, picture, genre, beautifully, literary, witty, critics, complicated, argument, premise, scholarship, talented, divine, twists, exceptional, obsession, commentary, landscape, exposes, influenced, accomplished, oriented

Movies: scene, recommend, screenplay, business, depth, justice, humanity, packaging, perfection, ficks, sequels, propaganda, anamorphic, cliché, cinematic, pretentious, goofy, ancient, marvelous, perspective, outrageous, intensity, mildly, immensely, bland, subplot, anticipation

Electronics: adapter, wireless, computer, sounds, camera, range, drives, mounted, photos, shots, packaging, antenna, ease, careful, broken, cards, distortion, stick, media, application, worthless, clarity, technical, memory, steady, dock, items, cord, systems, amps, skin, watt

Food: expensive, machine, months, clean, chips, texture, spicy, odor, inside, processed, robust, packs, weather, sticking, alot, press, poured, swallow, reasonably, portions, beware, fragrance, basket, volume, sweetness, terribly, caused, scratching, serves, sensation, sipping, smelled

Top words used by amateurs in least helpful reviews.

Music: will, good, favorite, cool, great, genius, earlier, notes, attention, place, putting, superb, style, room, beauty, realize, brought, passionate, difference, god, fresh, musical, grooves, consists, tapes, depressing, interview, short, rock, appeared, learn, brothers

Books: will, book, time, religious, liberal, material, interest, utterly, moves, movie, consistent, false, committed, question, turn, coverage, decade, novel, understood, worst, leader, history, kind, energy, fit, dropped, current, doubt, fan, books, building, travel, sudden, fails

Movies: movie, hour, gay, dont, close, previous, feature, type, months, meaning, wait, boring, absolutely, truth, generation, going, fighting, runs, fantastic, kids, quiet, kill, lost, angles, previews, crafted, teens, help, believes, brilliance, touches, sea, hardcore, continue, album

Electronics: order, attach, replaced, write, impressed, install, learn, tool, offered, details, turns, snap, price, digital, well, buds, fit, runs, fantastic, kids, quiet, kill, lost, angles, previews, crafted, teens, help, believes, brilliance, touches, sea, hardcore, continue, album

Food: night, going, haven, sour, fat, avoid, sugar, coffee, store, bodied, graham, variety, salsa, reasons, favorite, delicate, purpose, brands, worst, litter, funny, partially, sesame, handle, excited, close, awful, happily, fully, fits, effects, virgin, salt, returned, powdery, meals, great

Interpretable explanation by salient words used by experts for helpful reviews: Table 4 shows a snapshot of the latent word clusters, as used by experts and amateurs, for helpful reviews and otherwise, as generated by our model.

We observe that the most helpful reviews pertaining to music talk about its essence and style; for books they describe the theme and writing style; for movies they write about screenplay and storytelling; for food reviews these are mostly concerned about hygiene and allergens; for electronics they discuss about specific product features. Note that prior works [5,9] used explicit product features, that we were able to automatically discover as latent features from reviews. The least helpful reviews mostly describe some generic concepts, praise or criticize an item without going in depth about the facets, and are generally quite superficial in nature.

7 Conclusion

We proposed an approach to predict useful product reviews by exploiting the joint interaction between user expertise, writing style, timeliness, and review consistency using Hidden Markov Model – Latent Dirichlet Allocation. Unlike prior works exploiting a variety of syntactic and domain-specific features, our model uses only the information of a user reviewing an item at an explicit timepoint for this task — making our approach generalizable across all communities and domains. Additionally, we provide interpretable explanation as to why a review is helpful, in terms of salient words from latent word clusters — that are used by experts to describe important facets of the item in the review. Our experiments on real-world datasets from Amazon like books, movies, music, food, and electronics demonstrate effectiveness of our approach over state-of-the-art baselines.
Figure 2: Increase in log-likelihood (scaled by $10^e + 0^7$) of the data per-iteration in the five domains.

(a) Our model: Facet preference divergence with expertise learned from review helpfulness.

(b) Our model: Language model divergence with expertise learned from review helpfulness.

Figure 3: Facet preference and language model KL divergence with expertise.

Acknowledgment
This research was partly supported by ERC Synergy Grant 610150 (imPACT).

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3, 2003.
[2] N. Corporations. Global online shopping report. [http://www.nielsen.com/us/en/insights/news/2010/global-online-shopping-report.html](http://www.nielsen.com/us/en/insights/news/2010/global-online-shopping-report.html). [Online; accessed 10-Jun-2016].
[3] T. Griffiths. Gibbs sampling in the generative model of latent dirichlet allocation. Technical report, 2002.
[4] N. Jindal and B. Liu. Opinion spam and analysis. In *WSDM*, 2008.
[5] S.-M. Kim, P. Pantel, T. Chklovski, and M. Pennacchiotti. Automatically assessing review helpfulness. In *EMNLP*, 2006.
[6] H. Lakkaraju, C. Bhattacharyya, I. Bhattacharya, and S. Merugu. Exploiting coherence for the simultaneous discovery of latent facets and associated sentiments. *SDM*, 2011.
[7] C. Lin and Y. He. Joint sentiment/topic model for sentiment analysis. *CIKM*, 2009.
[8] J. Liu, Y. Cao, C.-Y. Lin, Y. Huang, and M. Zhou. Low-quality product review detection in opinion summarization. In *EMNLP-CoNLL*, 2007.
[9] Y. Liu, X. Huang, A. An, and X. Yu. Modeling and predicting the helpfulness of online reviews. In *ICDM*, 2008.
[10] Y. Lu, P. Tsaparas, A. Ntoulas, and L. Polanyi. Exploiting social context for review quality prediction. In *WWW*, 2010.
[11] J. McAuley and J. Leskovec. Hidden factors and hidden topics: Understanding rating dimensions with review text. *RecSys*, 2013.
[12] J. J. McAuley and J. Leskovec. From amateurs to connoisseurs: modeling the evolution of user expertise through online reviews. *WWW*, 2013.
[13] A. Mukherjee, A. Kumar, B. Liu, J. Wang, M. Hsu, M. Castellanos, and R. Ghosh. Spotting opinion spammers using behavioral footprints. In *KDD*, 2013.
[14] S. Mukherjee, G. Basu, and S. Joshi. Joint author sentiment topic model. *SDM*, 2014.
[15] S. Mukherjee, S. Günemann, and G. Weikum. Continuous experience-aware language model. *KDD*, 2016.
[16] S. Mukherjee, H. Lamba, and G. Weikum. Experience-aware item recommendation in evolving review communities. In *ICDM*, 2015.
[17] M. P. O’Mahony and B. Smyth. Learning to recommend helpful hotel reviews. In *RecSys*, 2009.
[18] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. *UAI*, 2004.
[19] J. Tang, H. Gao, X. Hu, and H. Liu. Context-aware review helpfulness rating prediction. In *RecSys*, 2013.
[20] H. Wang et al. Latent aspect rating analysis without aspect keyword supervision. *KDD*, 2011.
[21] R. West, H. S. Paskov, J. Leskovec, and C. Potts. Exploiting social network structure for person-to-person sentiment analysis. *Transactions of the Association for Computational Linguistics*, 2(2), 2014.