Supervised Topic Modeling Using Hierarchical Dirichlet Process-Based Inverse Regression: Experiments on E-Commerce Applications

Weifeng Li, Junming Yin, and Hsinchsun Chen, *Fellow, IEEE*

Abstract—The proliferation of e-commerce calls for mining consumer preferences and opinions from user-generated text. To this end, topic models have been widely adopted to discover the underlying semantic themes (i.e., topics). Supervised topic models have emerged to leverage discovered topics for predicting the response of interest (e.g., product quality and sales). However, supervised topic modeling remains a challenging problem because of the need to prespecify the number of topics, the lack of predictive information in topics, and limited scalability. In this paper, we propose a novel supervised topic model, *Hierarchical Dirichlet Process-based Inverse Regression* (HDP-IR). HDP-IR characterizes the corpus with a flexible number of topics, which prove to retain as much predictive information as the original corpus. Moreover, we develop an efficient inference algorithm capable of examining large-scale corpora (millions of documents or more). Three experiments were conducted to evaluate the predictive performance over major e-commerce benchmark testbeds of online reviews. Overall, HDP-IR outperformed existing state-of-the-art supervised topic models. Particularly, retaining sufficient predictive information improved predictive R-squared by over 17.6 percent; having topic structure flexibility contributed to predictive R-squared by at least 4.1 percent. HDP-IR provides an important step for future study on user-generated texts from a topic perspective.

Index Terms—Bayesian nonparametrics, hierarchical dirichlet process, topic modeling, sufficient dimension reduction, variational inference

1 Introduction

The proliferation of e-commerce has given rise to a significant amount of user-generated text, which contains salient information about consumer preferences and opinions [1],[2],[3]. Topic models are a major family of text analysis techniques for exploring the underlying semantic themes (i.e., topics) within textual data [4],[5],[6],[7]. However, prior research necessitates not only understanding the semantic themes but also integrating predictive analytics on variables of interest, such as customer sentiment [3], product quality [8], affect [2], and more. Standard topics models (e.g., LDA) are unsupervised and therefore incapable of making such predictions. To this end, the supervised topic modeling techniques have emerged, which can simultaneously discover the underlying semantic themes and leverage these themes for prediction [1]. Both the discovered themes and the predicted response variables provide valuable insights about consumer preferences and opinions. Supervised topic models have a number of important e-commerce applications, including customer feedback assessment [2],[8], online review evaluation [3], consumer sentiment analysis [3], product attributes mining [8], and customer preferences identification [9].

However, supervised topic modeling remains a challenging problem. First, most supervised topic models require prespecifying the number of topics *a priori* [10]. Such specification may result in model misspecification when the specified number of topics misrepresent the true underlying topic structure. For example, customer reviews for new products may contain unseen topics about new features. Prespecifying the number of topics inhibits the incorporation of such unseen topics, leading to unreliable topics and inaccurate predictions. Second, existing supervised topic models treat the proportion of topic mixtures as a reduced dimension representation of the original document and make predictions based on such representations. It is unclear whether these representations contain sufficient predictive information about the response [11]. Statistically speaking, sufficiency entails that the reduced dimension representation preserves all the information from original documents for making predictions. The missing information in the supervised topic modeling process may diminish the prediction accuracy. Third, large text corpora often span several million documents, leaving many supervised topic models unscalable [7]. Most supervised topic models adopt sampling-based inference algorithms, which require hundreds of iterations over each variable across all documents before convergence [12]. Therefore, the scalability of these models is limited.

In this paper, we propose a novel supervised topic model called *Hierarchical Dirichlet Process-based Inverse Regression* (HDP-IR). Specifically, the *Hierarchical Dirichlet Process*

- W. Li and H. Chen are with the Aritificial Intelligence Lab, Department of Management Information Systems, University of Arizona, Tucson, AZ 85721-0108. E-mail: weifeng.li@uga.edu, hsinchun@email.arizona.edu.
- J. Yin is with the Department of Management Information Systems, Eller College of Management, and Statistics Graduate Interdisciplinary Program (GIDP), University of Arizona, 430 McClelland Hall, 1130 E. Helen St., PO Box 210108, Tucson, AZ 85721-0108. E-mail: junmingy@email.arizona.edu.

Manuscript received 2 Feb. 2017; revised 4 Oct. 2017; accepted 11 Dec. 2017. Date of publication 25 Dec. 2017; date of current version 27 Apr. 2018.
(Corresponding author: Weifeng Li.)
Recommended for acceptance by D. Barbosa.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TKDE.2017.2786727

TABLE 1
A Taxonomy of Major Supervised Topic Models

Model	Topic Structure	SDR	Inference	Testbed & Performance
Supervised LDA (sLDA) [15]	Fixed	No	Variational	Movie reviews (5,006): 0.5 pR^2; Webpages (4,078): 0.095 pR^2
Dirichlet-Multinomial Regression (DMR) [16]	Fixed	No	Sampling	Academic papers: ~ 65% recall
DiscLDA [11]	Fixed	Yes	Sampling	News (19,997): 17% error rates
Labeled LDA (L-LDA) [17]	Fixed	No	Sampling	Webpages (4,000): 52.12% MicroF1
Dependency LDA (D-LDA) [18]	Fixed	No	Sampling	News (30,658): 54.1% MicroF1; Legal docs (19,800): 46.7% MicroF1
MedLDA [19]	Fixed	No	Sampling	News (19,997): ~ 83% Accuracy
Inverse Regression Topic Model (IRTM) [20]	Fixed	Yes	Variational	Amazon (13,528):0.996 MAE; Yelp(152,280): 0.704 MAE; Press release (72,224): 0.826 MAE
Supervised Hierarchical Dirichlet Process (sHDP) [21]	Non-fixed	No	Sampling	News (1,518): ~ 60% Accuracy; Movie reviews (10,662): ~ 0.3 pR^2; Webpages (3,880): ~ 0.08 pR^2

(HDP) is a nonparametric topic modeling technique that allows for a flexible number of topics. *Inverse Regression* (IR) is a *sufficient dimension reduction* (SDR) technique that makes predictions with provably sufficient information. HDP-IR characterizes the corpus with a flexible number of topics, which prove to retain statistically sufficient information for improved predictive performance. Moreover, we develop an efficient inference algorithm for model estimation that is capable of examining large-scale corpora with millions of documents. Evaluation of HDP-IR in comparison with the state-of-the-art baseline techniques reveals that both increasing the topic structure flexibility and using sufficient dimension reduction could improve the predictive performance on user-generated review text in e-commerce applications, and the proposed inference algorithm is highly effective in terms of its scalability.

This paper is organized as follows. Section 2 provides a review of related work on major supervised topic models and identifies research gaps. Section 3 briefly introduces the background of HDP and IR. Section 4 details our proposed model and the algorithms for estimating the model and making predictions. Section 5 includes the experimental evaluation of the proposed model in comparison with the state-of-the-art baseline techniques. Section 6 provides the conclusion and future directions.

2 Related Work

Topic modeling aims to analyze large text corpora by discovering the underlying semantic themes (i.e., topics) that are consistent across documents [13], [14]. A topic model is a probabilistic model explaining how observed documents relate to underlying topics. In topic models, a topic $\boldsymbol{\beta}_k$ is often represented as a multinomial distribution over words in the vocabulary: $[\beta_k^1, \ldots, \beta_k^W]$, where β_k^w is the probability of word w. Then, the collection of words $\boldsymbol{w}_d$ in each document d is generated from a mixture of K topics: $[\theta_1, \ldots, \theta_K]$, where θ_k is the proportion of the topic $\boldsymbol{\beta}_k$ within the document. Topic models have been widely studied and applied in various research contexts [4], [5], [6], [7]. Nonetheless, these topic models are unsupervised, incapable of making predictions [11]. Supervised topic models (STMs) are capable of predicting the response y_d of each document d based on the underlying topics. Formally, given D document-response pairs $\{(\boldsymbol{w}_1, y_1), \ldots, (\boldsymbol{w}_D, y_D)\}$, STMs estimate K topics $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$ that are predictive of the response. When given a new, unlabeled document $\boldsymbol{w}_{\text{new}}$, STMs can predict y_{new} based on examining its underlying topic structure.

We summarize major STMs from four aspects through a taxonomy: Topic Structure, Sufficient Dimension Reduction, Inference, and Testbed & Performance (Table 1). *Topic structure* concerns how STMs organize the topics. The major topic structure parameter is whether the number of topics need to be prespecified and fixed *a priori*. The *SDR* aspect examines whether the topics of the STM contain statistically sufficient information for making predictions. The *Inference* aspect concerns what technique each STM uses for model estimation. The two major inference techniques are the variational algorithm and the sampling-based algorithm. We will explain the differences between these two algorithms later. The *Testbed & Performance* aspect reports the predictive performance of each model as measured by various performance metrics on different testbeds.

As shown in Table 1, most of the STMs have fixed topic structure, requiring the number of topics to be prespecified. This is because these STMs are derivatives of the well-known latent Dirichlet allocation (LDA). LDA is a parametric topic model, requiring the number of topics to be specified *a priori*. In many research contexts, it is often difficult to determine the correct number of topics [10]. Hence, model selection is included to determine the optimal number of topics under certain evaluation metrics. However, the optimal numbers of topics under different metrics are not often consistent. For example, Blei and Mcauliffe [15] used the per-word held-out log-likelihood as a benchmark for selecting the optimal number of topics. In their testbeds, the selected number of topics failed to yield the best predictive performance as measured with predictive R-squared (pR^2). In STMs based on parametric topic models, the fixed topic structure does not have the flexibility to accommodate the potential topics in new documents, therefore causing model misspecification. Lately, the nonparametric topic modeling approach has drawn great attention for its capability to provide topic structure flexibility. Particularly, the hierarchical Dirichlet Process (HDP) is a leading nonparametric topic model that can accommodate an unlimited number of topics [22]. Compared with the Dirichlet distribution used in LDA, which assigns proportions to a fixed number of topics, HDP uses the Dirichlet process that can generate a countably infinite number of topics to be shared across different documents. HDP-based STMs have the potential to overcome the aforementioned limitation of most STMs. For example, Dai and Storkey [21] extended HDP to the supervised context by modeling the joint distribution of documents and the response, which has achieved promising performance improvement over Supervised LDA (sLDA) as measured by pR^2.

The sufficiency of dimension reduction [23] is not usually guaranteed in many previous STMs. In STMs, dimension reductions of documents are the document-specific topic mixtures, which are further used to inform the prediction [11]. For example, sLDA used the empirical topic vector z_d as the dimension reduction projection of each document for subsequent Generalized Linear Model regression [15]. DiscLDA achieved dimension reduction by using the average transformed topic mixtures [11]. The sHDP model is built on a Generalized Linear Model to regress the response variable on the average of the dimension reduction generated from HDP [21]. Yet, the dimension reductions provided by most STMs are not sufficient [11]. Sufficient dimension reduction entails a comprehensive, succinct representation, which retains as much predictive information as the original document. Formally, given a document $\boldsymbol{w} \in \mathbb{R}^W$ and its response y, SDR is the low-dimensional representation $R(\boldsymbol{w}) \in \mathbb{R}^k$, where $k < W$, such that conditioning on SDR $R(\boldsymbol{w})$, the response y is independent of the original document $\boldsymbol{w}$: $y \perp\!\!\!\perp \boldsymbol{w}|R(\boldsymbol{w})$. With sufficient predictive information, STMs with SDR have the potential to make more accurate predictions than STMs with non-SDR. In [20], the Inverse Regression Topic Model (IRTM) leveraged the inverse regression model to attain SDR, which led to more accurate predictions than STMs with non-SDR, such as sLDA and Dirichlet-Multinomial Regression (DMR). We will elaborate IRTM further in Section 3.

Inference concerns computing the model's posterior distributions. The posterior distributions are the conditional distributions of the model's variables given data, which are key to prediction. There are two major families of inference algorithms: *sampling-based inference* algorithms and *variational inference* algorithms [12]. *Sampling-based inference* algorithms, such as Gibbs sampling, approximate the posterior by empirically sampling from the conditional distribution of each variable within the model and further combining these conditional distributions into the posterior distribution [24]. *Sampling-based inference* algorithms usually require hundreds of iterations to "burn-in" each variable [12] and are therefore not scalable to large corpus containing hundreds of thousands of documents. In DMR, Mimno and McCallum sampled their model for 1,000 iterations, in addition to the "burn-in" period of 250 iterations [16]. Each iteration requires sampling for every variable from all documents. Therefore, STMs with *sampling-based inference* are often limited by the time complexity. As shown in Table 1, most testbeds in *sampling-based inference* studies contain fewer than 20,000 documents. The one exception was D-LDA, where Rubin et al. approximated the sampling of certain variable with direct assignment to expedite the sampling process [18]. On the other hand, *variational inference* algorithms seek to optimize an approximate distribution that is the closest to the posterior distribution as measured with *Kullback-Leibler divergence* [12]. By transforming a sampling problem into an optimization problem, *variational* algorithms require significantly fewer iterations than *sampling-based* algorithms and are therefore amenable to relatively large-scale corpora. Essentially, the variational algorithm is a coordinate ascent algorithm based on maximizing the marginal likelihood, iterating between optimizing document-level variables (e.g., the mixing proportion of topics) for each document and estimating corpus-level variables (e.g., topics). The variational algorithm equips STMs with the capability to analyze large corpora with significantly reduced complexity. For instance, Rabinovich and Blei developed a variational EM algorithm for inferring IRTM, which was able to process a relatively large research testbed containing 152,280 Yelp reviews [20]. The state-of-the-art variational framework is *Stochastic Variational Inference* (SVI) [12]. Instead of iterating through the document-level variables for the entire corpus, SVI performs stochastic optimization on the document-level variables for random subsamples of the corpus. SVI therefore has the potential to process large-scale corpora containing hundreds of thousands of documents to millions of documents.

In terms of testbeds, prior STMs have been applied to a variety of domains, including news reports [11], [18], [19], web pages [15],[21], and academic papers [16]. An emerging domain is e-commerce customer reviews (e.g., movie reviews [21], product reviews [20]). Customer reviews are increasingly available for a variety of products and services, significantly reshaping the e-commerce landscape. Analyzing customer reviews has important implications for a number of stakeholders [25]. For example, a manufacturer can identify the important features of a product or prospective buyers can assess product quality. Given the differences in the content, language usage, and communication structure, we are still far from a thorough understanding of how STMs can effectively examine customer reviews. Moreover, as mentioned previously, existing testbeds rarely exceeded a hundred thousand documents due to the model inference limitation. Collections of customer reviews may contain millions of documents. It is unclear how STMs can perform on such large-scale testbeds.

In terms of the performance metrics, we find that prior studies mostly adopted predictive R-squared (pR^2) and mean absolute error (MAE) for measuring the predictive performance on customer review testbeds. pR^2 captures the fit between the predicted response and the ground-truth by assessing the proportion of variation in the true response that can be explained by the predicted response. MAE captures the prediction error by measuring the average of the prediction errors between the predicted response and the true response. The two metrics evaluate different aspects of the model predictions and are therefore complementary to each other. Using both metrics provides a comprehensive assessment of the predictive performance. Nonetheless, little research has jointly used both metrics.

To summarize, we reorganize major STMs in Table 1 based on their model features. Table 2 presents a two-by-two matrix of existing STM modeling contexts and potential STM directions. The vertical dimension differentiates STMs based on whether the model is a parametric model or a non-parametric model. The horizontal dimension discriminates STMs providing SDRs from others.

TABLE 2
Major Supervised Topic Models Research Framework

	Non-SDR	SDR
Parametric	sLDA [15], DMR [16], L-LDA [17], D-LDA [18], MedLDA [19]	DiscLDA [11], IRTM [20]
Non-parametric	sHDP [21]	Our proposed model: HDP-IR

In the top left quadrant, most of the STMs are parametric models providing dimension reductions that are not sufficient. These models have two major limitations. First, parametric models have the restriction of specifying the number of topics *a priori*. Second, non-SDRs in these models are not capable of capturing as much predictive information from the document, thus reducing predictive performance. So far, prior studies have attempted to solve the two major limitations of STMs separately. The sHDP model (in the bottom left quadrant) leverages the nonparametric topic model to accommodate data with a flexible number of topics, achieving better accuracy than sLDA [21]. In the top right quadrant, DiscLDA [11] and IRTM [20] attempted to improve the prediction by leveraging SDRs. To the best of our knowledge, little has been done to provide a comprehensive model that addresses both limitations (the bottom right quadrant, our proposed HDP-IR model).

3 Background for Proposed Model

Our proposed model draws upon *nonparametric topic modeling* and *inverse regression*. This section gives a brief introduction about the backgrounds of these two streams of research. *Nonparametric topic modeling* provides an alternative to methods that use model selection procedures to choose a fixed number of topics. *Inverse regression* techniques support dimension reductions with provably sufficient predictive information about the original document for prediction.

3.1 Nonparametric Topic Modeling

As mentioned above, standard parametric topic models (e.g., LDA) face the challenge of prespecifying the number of topics. The nonparametric topic modeling approach overcomes this challenge by leveraging Bayesian nonparametrics, in which the number of topics does not need to be specified a priori and can be inferred from the data. The hierarchical Dirichlet Process (HDP) extension of the LDA model [22] can perform as well as the best LDA model in terms of held-out perplexity, while doing so without any model selection procedure.

The major building block of HDP is the Dirichlet process (DP), a probability distribution of discrete distributions over the topic space [26],[27]. Specifically, a Dirichlet Process, $\mathrm{DP}(\alpha, G_0)$, is specified by a positive *concentration parameter* α and a *base distribution* G_0. The *concentration parameter* specifies how concentrated the discrete distributions over topics drawn from the DP are. When the concentration parameter is small, the discrete distributions mostly concentrate on a few topics. As the concentration parameter increases, the discrete distributions gradually spread out probability weights to other topics. The *base distribution* determines the topic space and the expectation of the discrete distributions drawn from the DP. Modeling topics with the DP is advantageous because the discrete distribution drawn from the DP has a unique combination of properties. First, the topics drawn from this discrete distribution exhibit the *clustering* property; this property allows the words within a document to be clustered according to different topics. Second, the number of topics does not need to be specified a priori and can potentially grow with the size of corpus.

We further demonstrate these two properties using the *stick-breaking construction*, a statistically equivalent view of the DP from a constructive perspective [28]. Sethuraman [28] showed that a topic distribution $G \sim \mathrm{DP}(\alpha, G_0)$ can be formed through the following *stick-breaking construction* process: First, we generate two independent sequences of random variables: the topic sequence $\boldsymbol{\beta} = (\boldsymbol{\beta}_k)_{k=1}^{\infty}$, where $\boldsymbol{\beta}_k \sim G_0$ is topic k, and the "stick" probability sequence $\boldsymbol{\pi}' = (\pi'_k)_{k=1}^{\infty}$, where $\pi'_k \sim \mathrm{Beta}(1, \alpha)$ relates to the probability of topic k. Second, we define the probability sequence $\boldsymbol{\pi} = (\pi_k)_{k=1}^{\infty}$ as $\pi_k = \pi'_k \prod_{l=1}^{k-1}(1 - \pi'_l)$ (often denoted as $\boldsymbol{\pi} \sim \sigma(\boldsymbol{\pi}')$), where π_k is the probability of topic k. This resembles the breaking of a unit-length stick with the "stick" probability sequence $\boldsymbol{\pi}'$. Then, combining the probability sequence and the topic sequence, the topic distribution G is defined as $G = \sum_{k=1}^{\infty} \pi_k \delta_{\boldsymbol{\beta}_k}$, a *discrete* distribution over *a countably infinite* number of topics $(\boldsymbol{\beta}_k)_{k=1}^{\infty}$ with probability $(\pi_k)_{k=1}^{\infty}$. The same topic can appear in different draws from this *discrete* distribution [29]. Furthermore, this *stick-breaking construction* provides a guideline for developing variational inference algorithms for estimating nonparametric topic models [30].

So far, the topic distribution of each document d can be modeled with a draw from the DP: $G_d \sim \mathrm{DP}(\alpha, G_0)$. Particularly, each word w_{dn} in document d is generated from a topic $\boldsymbol{\beta}_{dn}$ that is drawn from the document topic distribution G_d. However, topic modeling requires the topics to be shared not only within each document but also across different documents. Following the intuition of topic sharing within each document, HDP extends DP to enable topic sharing across different documents by imposing a shared DP prior (called the corpus-level DP) onto the *base distributions* of the DPs for each document (called the document-level DP)

$$\begin{aligned} &\text{For the entire corpus}: \quad G_0 \sim \mathrm{DP}(\gamma, H), \\ &\text{For each document } d: \quad G_d \sim \mathrm{DP}(\alpha, G_0), \end{aligned} \tag{1}$$

where γ is the *concentration parameter* of the corpus-level DP, α is the *concentration parameter* of the document-level DP and H is the *baseline distribution* of the topics. As such, the *base distribution* of the document-level DP G_0 is discrete (with probability one) and therefore topics drawn for different documents are resamples from the same set of topics, thus achieving sharing of topics across documents.

Generally, the HDP topic model defines a set of document topic distributions G_d, one for each document, governed by the corpus-level topic distribution G_0, which includes a countably infinite number of topics. For each document, the HDP topic model assumes a sequence of topics $(\boldsymbol{\beta}_{d1}, \boldsymbol{\beta}_{d2}, \ldots)$ and each topic $\boldsymbol{\beta}_{dn}$ defines the probability distribution of word w_{dn}. Preliminary attempts to embed HDP in a STM were made in [21]. In this study, Dai and Storkey extended the application of HDP to supervised data by incorporating a Generalized Linear Model. The proposed supervised HDP (sHDP) model outperformed parametric STMs on multiple testbeds. However, the dimension reduction of each document in sHDP is not sufficient. While prior studies are encouraging, we are motivated to provide an STM based on HDP to make more accurate predictions by including SDR.

3.2 Inverse Regression

Inverse regression (IR) is a prominent SDR technique for textual data [31]. Classical regression analysis focuses on estimating the conditional distribution of the response given a document $\boldsymbol{w} \in \mathbb{R}^W$: $p(y|\boldsymbol{w})$. Due to the high dimensionality of textual data, classical regression analysis is not capable of efficiently estimating the conditional distribution, because an accurate estimation would require the sample size D to grow exponentially in the number of words W, which imposes both computational and statistical challenges [31]. To achieve dimension reduction, IR estimates the inverse conditional distribution of the document given the response $p(\boldsymbol{w}|y)$, because this inverse conditional distribution proves to lie on a lower dimensional subspace [23]. To prove this, we first assume a true model where the dimension reduction projection exists: $y = f(\boldsymbol{b}_1'\boldsymbol{w}, \ldots, \boldsymbol{b}_K'\boldsymbol{w}, \epsilon)$, where f determines the relationships between the document and the response, $\boldsymbol{b}_1, \ldots, \boldsymbol{b}_K$ are projection vectors, $\boldsymbol{b}_1'\boldsymbol{w}, \ldots, \boldsymbol{b}_K'\boldsymbol{w}$ are the true dimension reduction of $\boldsymbol{w}$, and ϵ is the mean-zero error. The centered inverse regression curve $\mathbb{E}(\boldsymbol{w}|y) - \mathbb{E}(\boldsymbol{w})$ lies on the K-dimensional subspace spanned by $\Sigma_{\boldsymbol{ww}}\boldsymbol{b}_k$'s, where $\Sigma_{\boldsymbol{ww}}$ is the covariance matrix of $\boldsymbol{w}$ [31]. As such, the inverse conditional distribution of document $p(\boldsymbol{w}|y)$ can be projected onto a K-dimensional ($K < W$) subspace without compromising any predictive information about the response.

Multinomial inverse regression (MNIR) is the state-of-the-art IR model that provides provable SDRs of documents [32]. MNIR extended the traditional IR by specifying the inverse conditional distribution for documents to be multinomial. MNIR performs multinomial logistic regression of word counts $\boldsymbol{w}$ onto the response y: $\boldsymbol{w}|y \sim \text{Multinomial}(\boldsymbol{q}|y)$ with the word frequency vector $\boldsymbol{q} = [q^1, \ldots, q^W]$. The frequency of word w relates to the response y through a logistic link: $q^w \propto \exp(\alpha_w + \phi_w y)$. The intercept α_w determines the "neutral" probability of word w when the response is zero. The coefficient ϕ_w uses an independent sparsity-inducing *Laplace* prior with mean-zero, whose *maximum a posteriori* (MAP) estimation is equivalent to the LASSO estimator [33] such that the coefficients of the words that are not correlated with the response are minimized. Alternatively, each coefficient ϕ_w can be viewed as the influence of the response on the frequency of word w within the document. With the sufficiency factorization of multinomial logistic regression, MNIR proves to yield an SDR projection $\boldsymbol{\Phi}'\boldsymbol{w}_d$, where $\boldsymbol{\Phi}' = [\phi_1, \ldots, \phi_W]$ are the vector of coefficients. Namely, the response y is independent of the original document $\boldsymbol{w}$ given the SDR projection $\boldsymbol{\Phi}'\boldsymbol{w}_d$: $y \perp\!\!\!\perp \boldsymbol{w}|\boldsymbol{\Phi}'w$. Consequently, with this SDR projection $\boldsymbol{\Phi}'\boldsymbol{w}_d$, we can ignore the original document $\boldsymbol{w}_d$ when making predictions. In other words, the SDR projection $\boldsymbol{\Phi}'\boldsymbol{w}_d$ retains as much predictive information about the response y as the original document $\boldsymbol{w}_d$. Further, the prediction can be readily implemented using classical regression analysis for estimating $p(y|\boldsymbol{\Phi}'\boldsymbol{w})$.

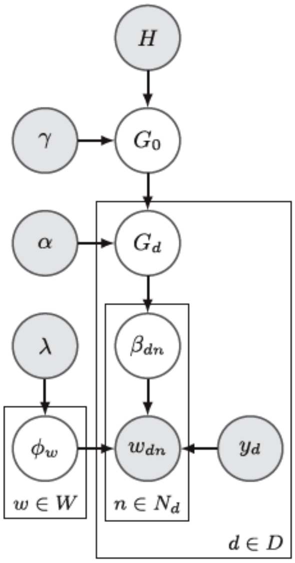

Fig. 1. A graphical model representation of hierarchical dirichlet process-based inverse regression model.

While effective in capturing the predictive information about the response, MNIR cannot capture the underlying semantic themes within the corpus. Drawing upon MNIR, the Inverse Regression Topic Model (IRTM) further accommodates for the topic heterogeneity within the corpus by replacing the word intercept α_w with the logarithm of word weight in each topic, $\ln \beta_k^w$, such that each word can have different intercepts under different topics [20]. As such, IRTM can be viewed as the combination of MNIR and LDA. The underlying intuition of IRTM is to attribute the relationship between documents and response to (1) the topic structure that is independent from the response and (2) the influence of the response on the document. From the topic modeling perspective, IRTM can be viewed as rescaling the LDA topics with the multinomial regression weights from MNIR, so that documents are jointly generated by the topics and the response. Nonetheless, as we mentioned previously, using standard parametric topic models requires the number of topics be specified *a priori*.

4 Proposed Model: Hierarchical Dirichlet Process-Based Inverse Regression (HDP-IR)

We propose a novel supervised topic model, the Hierarchical Dirichlet Process-based Inverse Regression model (HDP-IR). Our major methodological contribution to the literature is two-fold. First, HDP-IR combines the advantages of both nonparametric topic modeling and inverse regression: (1) HDP-IR avoids the model selection complications and can capture the uncertainty regarding the number of topics; (2) HDP-IR provides a SDR for each document, which can improve the predictive performance. Second, we design a scalable variational inference algorithm for fitting HDP-IR such that it can be applied to large-scale corpora (hundreds of thousands or millions of documents). Following prior STM literature, we design HDP-IR under a hierarchical Bayesian modeling framework. The structure of HDP-IR can be represented by a graphical model as shown in Fig. 1.

4.1 Model Representation

At a high level, HDP-IR contains three components: the *nonparametric topic modeling* component, the *inverse regression* component, and the *coupling* component. The *nonparametric topic modeling* component builds on HDP to capture the uncertainty regarding the number of topics. The *inverse regression* component leverages MNIR to model the response and the document. The *coupling* component combines the previous two components by integrating the topics into the logistic regression within the MNIR model. The design of each component is discussed in the remainder of this section.

4.1.1 The Nonparametric Topic Modeling Component (G_0, G_d, β_{dn} in Fig. 1)

Consistent with prior topic modeling literature [15], we define a topic $\boldsymbol{\beta}$ as a probability distribution over words in vocabulary. Specifically, $\boldsymbol{\beta}$ is a vector $[\beta^1, \dots, \beta^W]$, where β^w is the probability of word w in this topic. As mentioned in related work, parametric topic models assume a fixed finite number of K topics $(\boldsymbol{\beta}_1, \dots, \boldsymbol{\beta}_K)$ shared across the corpus. In contrast, the nonparametric topic model, HDP, relaxes this critical assumption and allows the number of topics to grow with the size of corpus. Specifically, at the corpus level, a random distribution of topics G_0 is generated from $\mathrm{DP}(\gamma, H)$, which provides a countably infinite number of topics to be shared across the corpus; at the document level, each document $d \in D$ generates a distribution of topics G_d from $\mathrm{DP}(\alpha, G_0)$ (Equation (1) in Section 3.1). We emphasize that G_0 is the base topic distribution here. As a result, each document has the sames set of topics but with different probabilities, thus sharing the topics within the corpus.

4.1.2 The Inverse Regression Component (ϕ_w, y_d in Fig. 1)

Drawing upon MNIR, the *inverse regression* component models the relationship between each word w_{dn} and the response variable y_d through a multinomial logistic regression. Specifically, each word w_{dn} is generated from a multinomial distribution: $w_{dn} \sim \mathrm{Multinomial}(q_d^1, \dots, q_d^W)|y_d$, where q_d^w is the frequency of word w in document d given the response variable y_d. Further, the relation between q_d^w and y_d is modeled through a logistic link: $q_d^w \propto \exp(\alpha_d^w + \phi_w y_d)$, where α_d^w is the intercept term and ϕ_w is the coefficient. The coefficient ϕ_w follows an independent fat-tailed and zero-mean *Laplace* distribution in order to induce sparsity such that some coefficients are shrunk to zero. This is because the *maximum a posteriori* (MAP) estimation of zero-mean Laplace-distributed coefficients is equivalent to the LASSO shrinkage [33]. Using the multinomial logistic regression setting has two notable strengths. First, as pointed out in the literature review, modeling using the multinomial logistic regression guarantees an SDR of the document through $R(\boldsymbol{w}_d) = \boldsymbol{\Phi}' \boldsymbol{w}_d$, where $\boldsymbol{\Phi}$ is the coefficient vector $[\phi_1, \dots, \phi_W]'$. Second, the coefficients $\{\phi_w\}$ can capture the influence of the response y_d on the observed words $\{w_{dn}\}$.

4.1.3 The Coupling Component ($\phi_w, \beta_{dn}, w_{dn}, y_d$ in Fig. 1)

In the logistic model within the inverse regression component (i.e., $q_d^w \propto \exp(\alpha_d^w + \phi_w y_d)$), α_d^w plays a vital role when the response is zero. That is, when $y_d = 0$, the frequency of word w is $q_d^w \propto \exp(\alpha_d^w)$. In other words, $\exp(\alpha_d^w)$ is proportional to the frequency of word w in the absence of the response. Recall that in topic models, the probability of word w under topic $\boldsymbol{\beta}_k$ is β_k^w. We therefore incorporate β_k^w into $q_d^w \propto \exp(\alpha_d^w + \phi_w y_d)$ by defining $\alpha_d^w = \ln \beta_k^w$. Modeling the intercept term using the logarithm of the probability of word w in topics has two strengths compared to other alternatives. First, instead of having a fixed intercept for each word as in MNIR, each word can have a different intercept (i.e., probability in the absence of the response) under different topics, which helps capture the topic structure. Second, the property of $\sum_{w=1}^{W} \beta_k^w = 1$ helps to identify the multinomial logistic regression in the MNIR such that there is no need to specify a null category for the model.

Drawing upon Sethuraman's *stick-breaking construction* [28], we reconstruct the original HDP-IR model to facilitate the development of inference algorithm because the DPs in HDP-IR cannot be readily represented in the posterior. Specifically, we apply Sethuraman's *stick-breaking construction* to both corpus-level DP (i.e., $G_0 \sim \mathrm{DP}(\gamma, H)$) and document level-DP (i.e., $G_d \sim \mathrm{DP}(\alpha, G_0)$). The advantage of this construction is that the conditionals within the resulting posterior are all in closed form [30]. The generative process of the reconstructed model is described as follows.

(1) Draw coefficients for each word,
    $\phi_w \sim \mathrm{Laplace}(\tau), w \in \{1, \dots, W\}$.
(2) Draw an infinite number of topics,
    $\boldsymbol{\beta}_k \sim \mathrm{Dirichlet}(\eta), k \in \{1, 2, 3, \dots\}$.
(3) Draw corpus-level breaking proportions,
    $v_k \sim \mathrm{Beta}(1, \gamma), k \in \{1, 2, 3, \dots\}$.
(4) For each document d,
  a) Draw document-level topic indices,
     $c_{di} \sim \mathrm{Multinomial}(\sigma(\boldsymbol{v})), i \in \{1, 2, 3, \dots\}$.
  b) Draw document-level breaking proportions,
     $\pi_{di} \sim \mathrm{Beta}(1, \alpha), i \in \{1, 2, 3, \dots\}$.
  c) For each word n,
    i) Draw topic assignment,
       $z_{dn} \sim \mathrm{Multinomial}(\sigma(\boldsymbol{\pi}_d))$.
    ii) Draw word,
       $w_{dn} \sim \mathrm{Multinomial}(\boldsymbol{q}_d)$, where

$$q_d^w = \frac{\beta_{c_{dz_{dn}}}^w \exp(\phi_w y_d)}{\sum_{u=1}^{W} \beta_{c_{dz_{dn}}}^u \exp(\phi_u y_d)}.$$

We further elaborate the generative process of this reconstruction of HDP-IR. Per-word coefficients ϕ_w are generated from the zero-mean, fat-tailed *Laplace* distribution to achieve LASSO shrinkage (Step 1). Topics $(\boldsymbol{\beta}_k)_{k=1}^{\infty}$ are generated from the Dirichlet distribution (Step 2). The corpus-level breaking proportion of each topic v_k defines the relative prevalence of each topic within the corpus (Step 3). For each document, we create a document-level distribution over topics: we first generate the document-level topics by drawing topic indices $\boldsymbol{c}_d$ from $\sigma(\boldsymbol{v})$ (Step 4.1); we then generate the document-level breaking proportion of each topic π_{di}, which defines the relative prevalence of each topic within the document (Step 4.2). For each word, we create a distribution over words: we first generate the topic by drawing topic index z_{dn} from $\sigma(\boldsymbol{\pi}_d)$ (Step 4.3.1); we then use this topic (i.e., $\boldsymbol{\beta}_{c_{dz_{dn}}}$) in the multinomial logistic model for generating the word (Step 4.3.2).

4.2 Inference and Prediction

We develop an efficient inference algorithm for fitting the HDP-IR model based on the *Stochastic Variational Inference* framework [12]. The traditional variational inference algorithm needs to perform coordinate ascent over both document-level variables for all documents (i.e., E-step) and corpus-level variables in each iteration (i.e., M-step). When the traditional variational algorithm examines large corpora containing hundreds of thousands of documents, the computational complexity associated with the E-step grows significantly. Based on stochastic optimization, SVI incorporates random subsampling into the E-step and then uses the resulting accumulated document-level sufficient statistics to optimize the corpus-level variables through natural gradient ascent. In the E-step, SVI randomly subsamples from the corpus and optimizes the document-level variables based on the subsampled documents. Then, SVI uses the optimized document-level variables from the subsample as noisy approximations of the collective document-level variables for optimizing the corpus-level variables. The rest of this section summarizes the algorithmic logic of our inference algorithm and prediction algorithm.

The objective of inference is to infer the following posterior distribution given data, which can then be further used in prediction

$$\begin{aligned} & p(\boldsymbol{v}, \boldsymbol{\beta}, \boldsymbol{\Phi}, \boldsymbol{\pi}, c, \boldsymbol{z} | \boldsymbol{w}, \boldsymbol{y}; \alpha, \gamma, \eta, \lambda) \\ & = \frac{p(\boldsymbol{v}, \boldsymbol{\beta}, \boldsymbol{\Phi}, \boldsymbol{\pi}, c, \boldsymbol{z}, w, \boldsymbol{y} | \alpha, \gamma, \eta, \lambda)}{p(\boldsymbol{w}, \boldsymbol{y} | \alpha, \gamma, \eta, \lambda)} \end{aligned} \tag{2}$$

The denominator $p(\boldsymbol{w}, \boldsymbol{y} | \alpha, \gamma, \eta, \lambda)$ is intractable to compute because it requires integrating over all other latent variables: $\boldsymbol{v}, \boldsymbol{\beta}, \boldsymbol{\Phi}, \boldsymbol{\pi}, \boldsymbol{c}, \boldsymbol{z}$. Hence, variational inference seeks to find an approximation distribution $q(\boldsymbol{v}, \boldsymbol{\beta}, \boldsymbol{\Phi}, \boldsymbol{\pi}, \boldsymbol{c}, \boldsymbol{z})$ that is the closest to the posterior distribution $p(\boldsymbol{v}, \boldsymbol{\beta}, \boldsymbol{\Phi}, \boldsymbol{\pi}, c, \boldsymbol{z} | \boldsymbol{w}, \boldsymbol{y}; \alpha, \gamma, \eta, \lambda)$ as measured by *Kullback-Leibler divergence*

$$\begin{aligned} & \mathrm{KL}(q(\boldsymbol{v}, \boldsymbol{\beta}, \boldsymbol{\Phi}, \boldsymbol{\pi}, \boldsymbol{c}, z) || p(\boldsymbol{v}, \boldsymbol{\beta}, \boldsymbol{\Phi}, \boldsymbol{\pi}, c, \boldsymbol{z} | \boldsymbol{w}, \boldsymbol{y}; \alpha, \gamma, \eta, \lambda)) \\ & = - \{\mathbb{E}_q[\ln p(\boldsymbol{v}, \boldsymbol{\beta}, \boldsymbol{\Phi}, \boldsymbol{\pi}, c, \boldsymbol{z}, w, \boldsymbol{y} | \alpha, \gamma, \eta, \lambda)] \\ & \quad - \mathbb{E}_q[\ln q(\boldsymbol{v}, \boldsymbol{\beta}, \boldsymbol{\Phi}, \boldsymbol{\pi}, \boldsymbol{c}, z)]\} + \log p(x) \\ & = - \mathcal{L}(q) + \log p(x) \end{aligned} \tag{3}$$

In other words, variational inference aims to maximize $\mathcal{L}(q)$, which is also known as *Evidence Lower BOund* (ELBO) [12]. Drawing upon the mean field theory [12], we define the approximation distribution q to be fully factorized

$$\begin{aligned} & q(v, \boldsymbol{\beta}, \boldsymbol{\Phi}, \boldsymbol{\pi}, \boldsymbol{c}, z) \\ & = q(\boldsymbol{v} | \boldsymbol{a}) q(\boldsymbol{\beta} | \boldsymbol{\rho}) q(\boldsymbol{\Phi}) q(\boldsymbol{\pi} | \boldsymbol{\psi}) q(\boldsymbol{c} | \boldsymbol{\zeta}) q(z | \boldsymbol{\xi}) \\ & = \prod_{k=1}^{K-1} q(v_k | a_k^{(1)}, a_k^{(2)}) \prod_{k=1}^{K} \prod_{w=1}^{W} q(\beta_k^w | \boldsymbol{\rho}_k) \prod_{w=1}^{W} q(\phi_w) \\ & \quad \times \prod_{d=1}^{D} \left[\prod_{i=1}^{I} q(\pi_{di} | \psi_{di}^{(1)}, \psi_{di}^{(2)}) \prod_{i=1}^{I} q(c_{di} | \boldsymbol{\zeta}_{di}) \prod_{n=1}^{N_d} q(z_{dn} | \boldsymbol{\xi}_{dn}) \right] \end{aligned} \tag{4}$$

where $\{v_k\}$, $\{\beta_{kw}\}$, and ϕ_w are corpus-level variables, $\{\pi_{di}\}$, $\{c_{di}\}$, and $\{z_{dn}\}$ are document-level variables, and $\boldsymbol{a}, \boldsymbol{\rho}, \boldsymbol{\Phi}, \boldsymbol{\psi}, \boldsymbol{\zeta}, \boldsymbol{\xi}$ are the corresponding variational parameters for optimizing the approximate posterior. We define the factor approximation distributions to be in the same exponential family of the corresponding conditional distributions as defined in our model (i.e., $q(v_k | a_k^{(1)}, a_k^{(2)}) \sim \mathrm{Beta}(a_k^{(1)}, a_k^{(2)})$, $q(\beta_{kw} | \boldsymbol{\rho}_k) \sim \mathrm{Dirichlet}(\boldsymbol{\rho}_k)$, etc.).

Having defined the form of $q(\boldsymbol{v}, \boldsymbol{\beta}, \boldsymbol{\Phi}, \boldsymbol{\pi}, \boldsymbol{c}, z)$, we further develop an SVI-based EM algorithm (Algorithm 1) to maximize $\mathcal{L}(q)$ by iteratively optimizing the variational parameters (i.e., $\boldsymbol{a}$, $\boldsymbol{\rho}$, $\boldsymbol{\Phi}$, $\boldsymbol{\psi}$, $\boldsymbol{\zeta}$, $\boldsymbol{\xi}$). In the E-step, we randomly subsample a set of documents from the corpus and update the variational parameters of the document-level variables for these documents by performing coordinate ascent on document-level variational parameters (i.e., $\boldsymbol{\psi}$, $\boldsymbol{\zeta}$, and $\boldsymbol{\xi}$). This step makes it possible to perform the M-step without iterating through all the documents within the corpus, enabling the scalability of our inference algorithm. In the M-step, we perform stochastic natural gradient ascent over the variational parameters of the corpus-level variables (i.e., $\boldsymbol{a}$, $\boldsymbol{\rho}$, and $\boldsymbol{\Phi}$) based on the subsampled document-level variables from the E-step. Prior studies have proved that such subsampled document-level variables are consistent approximations of the document-level variables of the entire corpus with respect to the objective ELBO function [34].

Algorithm 1. HDP-IR Inference Algorithm

```
Input: Corpus $D$ (with responses)
Output: Converged variational parameters: $\boldsymbol{a}, \boldsymbol{\rho}, \boldsymbol{\mu}, \boldsymbol{\tau}, \boldsymbol{\psi}, \boldsymbol{\zeta}, \boldsymbol{\xi}$
initialization
while ELBO has not yet converged do
    randomly sample documents $D_s$ from the corpus $D$
    {E-step:}
    for each document $d \in D_s$ do
        update the document-level breaking proportion parameter $\boldsymbol{\psi}_d^t$
        update the topic indices parameter $\boldsymbol{\zeta}_d^t$
        update the topic assignment parameter $\boldsymbol{\xi}_d^t$
    end for
    {M-step:}
    update the corpus-level breaking proportion parameter with document-level parameters from the subsampled documents: $\boldsymbol{a}^t = \boldsymbol{a}^{t-1} + \rho_t \partial \boldsymbol{a}(\{\boldsymbol{\psi}_d\}_{d \in D_s}, \{\boldsymbol{\zeta}_d\}_{d \in D_s}, \{\boldsymbol{\xi}_d\}_{d \in D_s})$
    update the topic parameter with document-level parameters from the subsampled documents: $\boldsymbol{\rho}^t = \boldsymbol{\rho}^{t-1} + \rho_t \partial \boldsymbol{\rho} (\{\boldsymbol{\psi}_d\}_{d \in D_s}, \{\boldsymbol{\zeta}_d\}_{d \in D_s}, \{\boldsymbol{\xi}_d\}_{d \in D_s})$
    update the coefficient parameters with document-level parameters from the subsampled documents: $\boldsymbol{\Phi}^t = \boldsymbol{\Phi}^{t-1} + \rho_t \partial \boldsymbol{\Phi}(\{\boldsymbol{\psi}_d\}_{d \in D_s}, \{\boldsymbol{\zeta}_d\}_{d \in D_s}, \{\boldsymbol{\xi}_d\}_{d \in D_s})$
    update the step size $\rho_t$ according to [12]
end while
```

4.2.1 E-Step

For each document within the subsample, we maximize $\mathcal{L}(q)$ by performing coordinate ascent over the document-level variational parameters: $\boldsymbol{\psi}$, $\boldsymbol{\zeta}$, and $\boldsymbol{\xi}$. This can be achieved by setting the ELBO derivatives with respect to these parameters to zero

$$\psi_{di}^{(1)} = 1 + \sum_{n=1}^{N_d} \xi_{dn}^i \tag{5}$$

$$\psi_{di}^{(2)} = \alpha + \sum_{n=1}^{N_d} \sum_{l=i+1}^{I} \xi_{dn}^{l} \tag{6}$$

$$\zeta_{di}^{k} \propto \exp\left(\mathbb{E}_q[\log \sigma_k(\boldsymbol{v})] + \sum_{n=1}^{N_d} \xi_{dn}^{i} \mathbb{E}_q[\log p(w_{dn}|\boldsymbol{\beta}_k, y_d)] \right) \tag{7}$$

$$\xi_{dn}^{i} \propto \exp\left(\mathbb{E}_q[\log \sigma_i(\boldsymbol{\pi}_d)] + \sum_{k=1}^{K} \zeta_{di}^{k} \mathbb{E}_q[\log p(w_{dn}|\boldsymbol{\beta}_k, y_d)] \right) \tag{8}$$

The $\mathbb{E}_q[\log p(w_{dn}|\boldsymbol{\beta}_k, y_d)]$ term in Equations (7) and (8) can be expanded as follows.

$$\begin{aligned} \mathbb{E}_q[\log p(w_{dn}|\boldsymbol{\beta}_k, y_d)] &= \mathbb{E}_q[\log \beta_k^{w_{dn}}] + \mathbb{E}_q[\phi_{w_{dn}} y_d] \\ &\quad - \mathbb{E}_q[\log \sum_{u=1}^{W} \beta_k^{u} \exp(\phi_u y_d)] \end{aligned} \tag{9}$$

The expectation of the denominator from the softmax function, $\mathbb{E}_q[\log \sum_{u=1}^{W} \beta_k^u \exp(\phi_u y_d)]$, does not have a closed form expression because of the non-conjugacy between the Dirichlet distribution and softmax function [35]. As a result, there are no tractable updates for variational parameters $\boldsymbol{\zeta}$ and $\boldsymbol{\xi}$. The *delta method* [36], [37] is therefore used to approximate this expectation term. For simplicity, we define this term as $\mathsf{E}(\boldsymbol{\rho}, \boldsymbol{\Phi})$, where $\boldsymbol{\rho}$ and $\boldsymbol{\Phi}$ are the topic and coefficient variational parameters to be optimized

$$\begin{aligned} \mathbb{E}_q[\log \sum_{u=1}^{W} \beta_k^u \exp(\phi_u y_d)] &\approx \log \sum_{u=1}^{W} \mathbb{E}_q[\beta_k^u] \mathbb{E}_q[\exp(\phi_u y_d)] \\ &\triangleq \mathsf{E}(\boldsymbol{\rho}, \boldsymbol{\Phi}) \end{aligned} \tag{10}$$

4.2.2 M-Step

We optimize the variational parameters for document-level variables. Since the optimized document-level variational parameters are from the random subsample of the corpus, we use the stochastic natural gradient ascent to optimize the corpus-level variational parameters (i.e., $\boldsymbol{a}$, $\boldsymbol{\rho}$, and $\boldsymbol{\Phi}$). The gradient for each corpus-level variational parameter is presented as follows:

$$\partial a_k^{(1)} = -a_k^{(1)} + 1 + \frac{D}{|D_s|} \sum_{d,i}^{D_s, I} \zeta_{di}^{k} \tag{11}$$

$$\partial a_k^{(2)} = -a_k^{(2)} + \gamma + \frac{D}{|D_s|} \sum_{d,i,l=k+1}^{D_s, I, K} \zeta_{di}^{l} \tag{12}$$

$$\partial \rho_k^w = \eta - \frac{D}{|D_s|} \sum_{d,n,i}^{D_s, N_d, I} \zeta_{di}^{k} \xi_{dn}^{i} \nabla_{\rho_{kw}} \mathsf{E}(\boldsymbol{\rho}, \boldsymbol{\Phi}) \tag{13}$$

$$\begin{aligned} \partial \phi_w =& - \lambda \partial |\phi_w| + \frac{D}{|D_s|} \sum_{d,n,k,i}^{D_s, N_d, K, K} \zeta_{di}^{k} \xi_{dn}^{i} \\ & \times \left(\mathbb{1}(w_d = w) y_d - \nabla_{\phi_w} \mathsf{E}(\boldsymbol{\rho}, \boldsymbol{\Phi}) \right) \end{aligned} \tag{14}$$

In these equations, $(\boldsymbol{\zeta}_d)_{d \in D_s}$ and $(\boldsymbol{\xi}_d)_{d \in D_s}$ are the optimized document-level and word-level topic proportion variational parameters for subsampled documents in D_s from the E-step. The $\frac{D}{|D_s|}$ in these equations helps approximate the optimized document-level variational parameters (with respect to ELBO) for all documents based on the subsample. For example, in Equation (11), the last term on the right-hand side, $\frac{D}{|D_s|} \sum_{d,i}^{D_s, I} \zeta_{di}^{k}$, approximates $\sum_{d,i}^{D,I} \zeta_{di}^{k}$, where D is the entire corpus. Consider that when the subsample is big enough to cover the entire corpus $D_s \to D$, Equation (11) becomes the exact gradient derived from the variational parameters for all documents

$$\partial a_k^{(1)} = -a_k^{(1)} + 1 + \sum_{d,i}^{D,I} \zeta_{di}^{k} \tag{15}$$

4.2.3 Prediction

The inference procedure computes the posterior distributions of the variables within the model, from which we can derive the sufficient dimension reduction. As suggested in our model, the SDR generated by HDP-IR contains two major components: (1) the topic-based dimension reduction from the nonparametric topic modeling component and (2) the non-topic-based dimension reduction from the inverse regression component. The former is the document-level topic proportion expectation (i.e., $\boldsymbol{\zeta}_d$), which has already been achieved in the inference. The latter is the inverse regression-based sufficient dimension reduction (i.e., $\boldsymbol{\Phi}' \boldsymbol{w_d}$), which can be calculated by taking the product of the optimized coefficients $\boldsymbol{\Phi}$ and the new document word count $\boldsymbol{w}_d$.

Following MNIR [32], we leverage the SDR for prediction using forward regression. Specifically, given a document $\boldsymbol{w}_d$, we use the SDR of the document $(\boldsymbol{\zeta}_d, \boldsymbol{\Phi}' \boldsymbol{w_d})$ to predict the response y_d. The predictive model is not restricted to a specific form. Rather, various forms of forward regression models can be applied, depending on the research context. In this study, we use the simplest regression model, linear regression, to predict the response y_d from the customer reviews (the document $\boldsymbol{w}_d$). Specifically, we use the following linear regression model

$$\mathbb{E}(y_d) = b_0 + \boldsymbol{b}_1 \boldsymbol{\zeta}_d + b_2 \boldsymbol{\Phi}' \boldsymbol{w_d}, \tag{16}$$

where b_0, $\boldsymbol{b}_1$, and b_2 are the regression coefficients. Algorithm 2 depicts the overall prediction procedure. We first achieve the SDRs for all the training documents (i.e., $\boldsymbol{\zeta}_d$, $\boldsymbol{\Phi}' \boldsymbol{w_d}$), then estimate the coefficients (i.e., b_0, $\boldsymbol{b}_1$, b_2) with the SDRs and the corresponding responses, and finally predict the expectation of the response based on the estimated coefficients.

Algorithm 2. HDP-IR Prediction Algorithm

```
Input: Training documents $D_{train}$, optimized coefficients $\boldsymbol{\Phi}$
Output: Prediction of the response expectation
initialization
\\ Training:
for each document $d \in D_{train}$ do
    calculate the topic-based dimension reduction $\boldsymbol{\zeta}_d$ through
    the E-step in Algorithm 1
    calculate the non-topic-based dimension reduction: $\boldsymbol{\Phi}' \boldsymbol{w_d}$
end for
estimate the coefficients in Equation 16 (i.e., $b_0$, $\boldsymbol{b}_1$, and $b_2$)
using the training documents: $\{(y_d, \boldsymbol{\zeta}_d, \boldsymbol{\Phi}' \boldsymbol{w_d})\}_{d \in D_{train}}$
\\ Predicting:
calculate the SDR for new document $d^*$: $(\boldsymbol{\zeta}_{d^*}, \boldsymbol{\Phi}' \boldsymbol{w_{d^*}})$
predict the response expectation: $\mathbb{E}(y_{d^*}) = b_0 + \boldsymbol{b}_1 \boldsymbol{\zeta}_d + b_2 \boldsymbol{\Phi}' \boldsymbol{w_{d^*}}$
```

5 Evaluation: E-Commerce Testbeds

Based on the design of HDP-IR, three experiments were conducted to evaluate the predictive performance of our proposed model. The first experiment was intended to evaluate how the topic modeling component in HDP-IR improved the predictive performance. We compared our proposed model to various state-of-the-art non-topic-based models. The second experiment compared the nonparametric STMs with parametric STMs to evaluate how the nonparametric technique helped improve the predictive performance. The third experiment sought to assess how the sufficient dimension reduction generated by the inverse regression component in HDP-IR contributed to the predictive performance. To this end, we compared the HDP-IR model to various state-of-the-art supervised topic models, such as sLDA and DMR. For all experiments, we evaluted the models through five-fold cross validation to prevent the evaluation bias induced by model misspecification. Conforming to the conventions in the STM literature [15], [20] and the SVI literature [37], we set the default parameters of priors as follows: $\gamma = 1.0$, $\alpha = 1.0$, $\lambda = 1.0$, and $\eta = 0.01$.

To evaluate the effectiveness of our proposed model, we conducted experiments on three e-commerce review testbeds used in prior studies. The first testbed is built on movie reviews from Rotten Tomatoes [38]. This testbed is composed of 5,006 movie reviews, each of which is associated with a numerical rating response ranging from 0 to 5. The second testbed is based on customer reviews from the Yelp Academic Dataset [39]. This testbed includes a total of 330,071 customer reviews. The reponses are stars rated on a scale of 0 to 5. The third testbed contains 1,422,518 Amazon reviews for various product categories, including books, electronics, housewares, and more. Each review corresponds to a numerical rating on a scale of 0 to 5 [40]. Since the responses of all three testbeds are numerical ratings, we standardized these responses by calculating their Z-scores: $y_z = \frac{y-\bar{y}}{\sigma(y)}$, where $\bar{y}$ is the mean of the responses for each testbed and $\sigma(y)$ is the standard deviation for each testbed.

As suggested in prior studies (see Table 1), the predictive performance on customer reviews is often measured by predictive R-squared (pR^2) and mean absolute error. In addition, we also report root mean square error (RMSE) to provide a comprehensive comparison between HDP-IR and baseline models. pR^2 assesses the fit between the prediction and the ground-truth with the proportion of variation in the true response that can be explained by the predicted response. Models with higher pR^2 have better predictive performance: $pR^2 \triangleq 1 - \frac{\sum_{d=1}^{D}(y_d-\hat{y}_d)^2}{\sum_{d=1}^{D}(y_d-\bar{y})^2}$. MAE is the average of the prediction errors between the prediction and the ground-truth. Lower MAE suggests lower prediction error: $\mathrm{MAE} \triangleq \frac{1}{D}\sum_{d=1}^{D}|\hat{y}_d - y_d|$. RMSE measures the standard deviation of the prediction errors between the prediction and the ground-truth. Lower RMSE suggests lower prediction error: $\mathrm{RMSE} \triangleq \sqrt{\frac{1}{D}\sum_{d=1}^{D}(\hat{y}_d - y_d)^2}$. Furthermore, we used one-sided Wilcoxon signed-rank test with 95 percent confidence for testing results significance because these performance metrics may not be normally distributed [41].

TABLE 3
pR^2 Results for HDP-IR and MNIR-Based Models

Model	Movie	Yelp	Yelp	Amazon	Amazon
	(N = 5 K)	(N = 5 K)	(N = 0.3 M)	(N = 5 K)	(N = 1.4 M)
HDP-IR	**0.681**	**0.661**	**0.722**	**0.595**	**0.598**
MNIR-LR	0.444 (p<0.001)	0.499 (p<0.001)	-	0.332 (p<0.001)	-
MNIR-PolyQ	0.463 (p<0.001)	0.575 (p<0.001)	-	0.481 (p<0.001)	-
MNIR-PolyC	0.476 (p<0.001)	0.577 (p<0.001)	-	0.554 (p<0.01)	-
Tree-LSTM	0.586 (p<0.001)	0.568 (p<0.001)	-	0.541 (p<0.001)	-

5.1 Experiment #1: HDP-IR versus Non-Topic-Based Models

To evaluate the improvement associated with the topic modeling component, we compared HDP-IR to the state-of-the-art non-topic-based algorithms, including Multinomial Inverse Regression (MNIR) models and Tree-structured Long Short-Term Memory model (Tree-LSTM). As suggested in [32], MNIR-based models provide an SDR of each document, which can be further used in predictive models. Based on [20] and [32], we extended MNIR with three predictive models: linear regression (MNIR-LR), polynomial regression with quadratic term (MNIR-PolyQ), and polynomial regression with cubic term (MNIR-PolyC). Tree-LSTM is the state-of-the-art recurrent neural network method for prediction. This method incorporates the tree-structured dependency relations into standard LSTM models and has demonstrated enhanced predictive performance over traditional approaches in sentiment classification [42]. The comparison against discriminative models as such can further demonstrate the benefits of modeling topics in HDP-IR. Our experiments confirmed that both MNIR models and Tree-LSTM were not able to converge on the two larger testbeds (i.e., Yelp Reviews and Amazon Reviews) because the scale of these testbeds exceeded the limits that these models can handle. Therefore, we randomly subsampled 5,000 reviews from each of the testbeds to evaluate MNIR and Tree-LSTM.

Table 3 shows the experimental results for predictive R-squared across all models applied to the three testbeds. Neither MNIR-based models nor Tree-LSTM were compared on the complete Yelp testbed (N = 330,071) or the complete Amazon testbed (N = 1,422,518). The inference algorithms of both MNIR and Tree-LSTM failed to converge on these testbeds due to their limited scalability. HDP-IR outperformed the baseline MNIR-based models on all three testbeds in terms of predictive R-squared. HDP-IR accounted for 1.7 percent (on the Amazon testbed) to 19.5 percent (on the Movie testbed) more of the variation in the true response than the baseline models. The improvements on all three testbeds were significant, as measured by the p-values from the Wilcoxon signed-rank test. For both the Yelp and Amazon testbeds, HDP-IR performed better on the complete testbeds than on the subsampled testbeds (N = 5,000). This suggests that the prediction bias caused by the noisy data points in

TABLE 4
MAE Results for HDP-IR and MNIR-Based Models

Model	Movie (N = 5 K)	Yelp (N = 5 K)	Yelp (N = 0.3 M)	Amazon (N = 5 K)	Amazon (N = 1.4 M)
HDP-IR	**0.457**	**0.477**	**0.474**	**0.347**	**0.376**
MNIR-LR	0.593 ($p<0.001$)	0.672 ($p<0.001$)	-	0.702 ($p<0.001$)	-
MNIR-PolyQ	0.579 ($p<0.001$)	0.613 ($p<0.001$)	-	0.627 ($p<0.001$)	-
MNIR-PolyC	0.574 ($p<0.001$)	0.616 ($p<0.001$)	-	0.548 ($p<0.001$)	-
Tree-LSTM	0.498 ($p<0.001$)	0.593 ($p<0.001$)	-	0.688 ($p<0.001$)	-

relatively small samples can be effectively corrected by increasing the sample size.

Tables 4 and 5 show the MAE and RMSE results for HDP-IR and all baseline models across the three testbeds. Consistent with the pR^2 results, the MAE results and the RMSE results show that HDP-IR had the best performance, with the lowest MAE prediction error and the lowest RMSE prediction error for all three testbeds. The Wilcoxon signed-rank test results supported that HDP-IR significantly outperformed the baseline MNIR-based models in terms of both MAE and RMSE across all three testbeds. Looking at the results of baseline models by testbeds, both MNIR-based models and Tree-LSTM had the lowest MAE prediction error and the lowest RMSE prediction error on the Movie testbed. This is not surprising since movie reviews have less topic heterogeneity than Yelp reviews and Amazon reviews, which span a variety of services and products.

As expected, including topic modeling component significantly improved the predictive performance of inverse regression models. This suggests that the learned topics were able to effectively capture the underlying semantic themes within the corpus, and these learned topics significantly contributed to prediction accuracy. Inverse regression models assume different documents to have the same distribution over words conditioned on the same response value. For documents with similar topics, this assumption is reasonable and effective; however, for documents covering a variety of topics, this assumption becomes less realistic and may affect the prediction accuracy. HDP-IR, on the other hand, relaxes this assumption with flexible topic structure. Specifically, HDP-IR learns the topics across all the documents with the nonparametric technique to capture topic heterogeneity and then incorporates these learned topics to differentiate the distribution of each document conditioned on the associated response. Furthermore, the experiment results emphasized the benefits of modeling topics using generative models. The state-of-the-art discriminative approach, Tree-LSTM, leveraged the word embedding language model, where each word was mapped to a unique embedding vector that capture its contextual semantics. While effective in many contexts, the word embedding language model cannot address homonyms adequately as each word is mapped to only one embedding vector. Therefore, Tree-LSTM tends to perform better on semantically coherent testbeds where each word tends to have fewer meanings within the entire corpus. This is evidenced by our experiment results: Tree-LSTM performed significantly better on the relatively coherent movie reviews. The topic modeling component in HDP-IR allows for each word to have multiple meanings as captured in its association with semantically different topics. Therefore, compared to MNIR-based models and Tree-LSTM, HDP-IR could have better predictive performance on datasets with topic heterogeneity.

TABLE 5
RMSE Results for HDP-IR and MNIR-Based Models

Model	Movie (N = 5 K)	Yelp (N = 5 K)	Yelp (N = 0.3 M)	Amazon (N = 5 K)	Amazon (N = 1.4 M)
HDP-IR	**0.581**	**0.553**	**0.512**	**0.615**	**0.644**
MNIR-LR	0.746 ($p<0.001$)	0.867 ($p<0.001$)	-	0.977 ($p<0.001$)	-
MNIR-PolyQ	0.733 ($p<0.001$)	0.799 ($p<0.001$)	-	0.861 ($p<0.001$)	-
MNIR-PolyC	0.724 ($p<0.001$)	0.797 ($p<0.001$)	-	0.798 ($p<0.001$)	-
Tree-LSTM	0.616 ($p<0.001$)	0.789 ($p<0.001$)	-	0.732 ($p<0.001$)	-

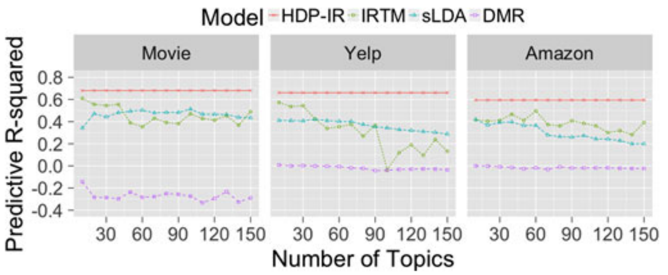

Fig. 2. pR^2 results for HDP-IR and baseline models with different number of topics.

5.2 Experiment #2: Nonparametric versus Parametric

To assess the improvement introduced by the nonparametric topic modeling technique, we compared HDP-IR to the baseline parametric STMs. The parametric STMs included open-sourced implementations of the state-of-the-art sLDA [15] and DMR [16], and our implementation of IRTM [20]. The specifications of these baseline models have already been discussed in *Related Work*. The comparison between HDP-IR and IRTM could directly reflect the predictive performance gain contributed by the nonparametric topic modeling component. We examined the predictive performance of the parametric STMs under different assumptions of the number of topics. We varied the number of topics in the parametric STMs from 10 to 150 in increments of 10 topics. This range covers the numbers of topics specified in most previous studies [15], [16], [11], [17], [18], [19], [20], [21].

Fig. 2 shows the predictive R-squared results for all models across the three testbeds, including the Movie testbed and the subsampled Yelp and Amazon testbeds. Overall, HDP-IR performed the best with the highest pR^2, followed by IRTM and sLDA. This pattern is consistent across all

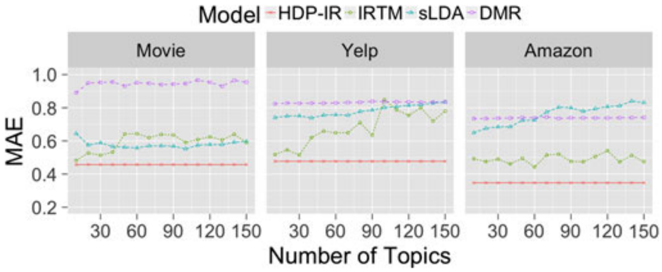

Fig. 3. MAE results for HDP-IR and baseline models with different number of topics.

three testbeds. Fig. 3 shows the MAE results for all models across the three testbeds. HDP-IR had the best predictive performance with the lowest MAE prediction error across all three testbeds. On the Amazon testbed, IRTM appeared to have better performance than the other two baseline models, while such advantage was not fully reflected on the other two testbeds. Fig. 4 shows the RMSE results for all models across the three testbeds. Similarly, HDP-IR had the best predictive performance with the lowest RMSE prediction error across all three testbeds. On the Amazon testbed, IRTM had better performance than the other two baseline models. As the major difference between IRTM and HDP-IR is the nonparametric topic modeling component, the experiment results clearly suggested the effectiveness of nonparametric topic modeling and quantified its contribution to overall predictive performance.

In comparison with parametric STMs, the nonparametric topic modeling technique improved the predictive performance of HDP-IR. The major underlying reason was that the parametric topic modeling approach requires specifying the number of topics, which might cause model overfitting or underfitting. Further, varying the number of topics could not significantly improve the predictive performance of the parametric STMs, as evidenced by the results of pR^2 (Fig. 2), MAE (Fig. 3), and RMSE (Fig. 4) on all three testbeds. This is because the nonparametric topic model is fundamentally different from the parametric topic model in terms of learning the topics. The nonparametric topic model considers the topics to be drawn from a topic space containing an infinite number of topics, whereas the parametric topic model assumes the topic space to have a limited number of topics. Therefore, the parametric topic model cannot attain as good performance as the nonparametric topic model can by merely changing the number of topics. For sLDA on the Yelp testbed and the Amazon testbed, the predictive performance became worse as the number of topics increased, which was consistent across all three performance metrics. This seems counterintuitive as the Yelp testbed and the Amazon testbed were supposed to have more topics due to the variety of services and products. A possible explanation is that sLDA assumes both the topic and the words to be generated based on the topic [16]. Such topics suffered from model overfitting on the held-out testing set because neither the response nor the words are generated from the topics in the same way as in the training set. Increasing the number of topics introduced more overfitted topics; therefore, the magnitude of the model overfitting became more significant.

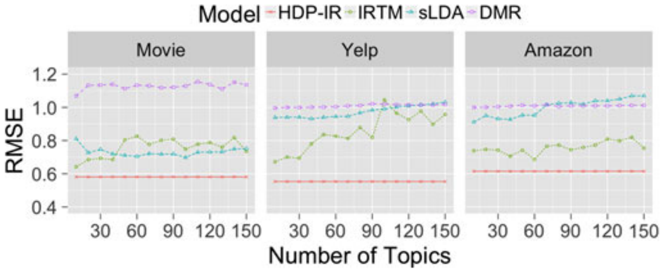

Fig. 4. RMSE results for HDP-IR and baseline models with different number of topics.

TABLE 6
pR^2 Results for HDP-IR and Baseline Topic-Based Models

Model	Movie (N = 5 K)	Yelp (N = 5 K)	Yelp (N = 0.3 M)	Amazon (N = 5 K)	Amazon (N = 1.4 M)
HDP-IR	**0.681**	**0.661**	**0.722**	**0.595**	**0.598**
sLDA	0.503 (p<0.001)	0.420 (p<0.001)	-	0.419 (p<0.001)	-
DMR	−0.142 (p<0.001)	0.008 (p<0.001)	-	0.003 (p<0.001)	-
HDP	-0.275 (p<0.001)	-0.214 (p<0.001)	0.013 (p<0.001)	-0.031 (p<0.001)	-0.035 (p<0.001)

5.3 Experiment #3: SDR versus Non-SDR

To evaluate the contribution of the inverse regression component to the predictive performance, we compared HDP-IR against baseline topic-based models without sufficient dimension reductions. Ideally, the best baseline model would include sHDP [21] because it is a nonparametric STM as well. However, the sHDP implementation was not available, so we identified a set of state-of-the-art open-source baseline topic-based models for comparison. Our first baseline model was a nonparametric STMs. This model performed linear regression on the posterior topic distribution from unsupervised HDP. Specifically, the nonparametric baseline method leveraged the posterior HDP topic distribution of each document as the covariates for predicting the response through a Generalized Linear Model. Inspired by [15], this baseline method could provide a meaningful comparison between SDR-based HDP and non-SDR-based HDP. Additionally, our baseline models also include parametric STMs, such as supervised LDA (sLDA) [15] and Dirichlet-Multinomial Regression [16]. Since most of these baseline models employed sampling-based inference algorithms (e.g., Gibbs sampling), which required hundreds of iterations over millions of variables, these models cannot scale well in the two larger testbeds (i.e., Yelp Reviews and Amazon Reviews). This was empirically validated by our preliminary experiment, where sLDA and DMR were not able to converge within a reasonable time frame (i.e., two weeks on a Windows system equipped with quad 3.0 Ghz processor and 32 GB memory). Therefore, we evaluated these models on the subsampled testbeds as described in Experiment #1.

Table 6 shows the experimental results for predictive R-squared across all the models for the three testbeds. HDP-IR outperformed the baseline topic-based models consistently across all three testbeds as measured by predictive R-squared. HDP-IR was able to predict 17.6 percent (on the

TABLE 7
MAE Results for HDP-IR and Baseline Topic-Based Models

Model	Movie (N = 5 K)	Yelp (N = 5 K)	Yelp (N = 0.3 M)	Amazon (N = 5 K)	Amazon (N = 1.4 M)
HDP-IR	**0.457**	**0.477**	**0.474**	**0.347**	**0.376**
sLDA	0.558 (p<0.001)	0.739 (p<0.001)	-	0.649 (p<0.001)	-
DMR	0.891 (p<0.001)	0.824 (p<0.001)	-	0.735 (p<0.001)	-
HDP	0.925 (p<0.001)	0.816 (p<0.001)	0.776 (p<0.001)	0.720 (p<0.001)	0.756 (p<0.001)

Amazon subsampled testbed) to 24.1 percent (on the Yelp testbed) more of the variation in the true response than the baseline topic-based models. Based on the Wilcoxon signed-rank test, HDP-IR significantly outperformed the baseline topic-based models, with p-values at the significance level of 0.001. Surprisingly however, the baseline nonparametric supervised topic model based on HDP, did not outperform the parametric supervised topic models. This may suggest that the topics learned by the nonparametric topic model under the unsupervised setting were not predictive of the document response. This confirmed our design of jointly considering the effects of the response (i.e., $\mathbf{\Phi}$) and the topic (i.e., $\boldsymbol{\beta}$) on document generation. Among the baseline models, sLDA consistently performed the best with the highest predictive R-squared across all three testbeds. HDP-IR achieved better performance with the SDR associated with the nonparametric topic modeling.

Tables 7 and 8 show the MAE results and the RMSE results for HDP-IR and the baseline topic-based models across the three testbeds. Again, HDP-IR outperformed the baseline topic-based models in terms of both the MAE prediction error and the RMSE prediction error across all three testbeds. The Wilcoxon signed-rank test also supported the significance of these results. The prediction errors were comparable with each other across different testbeds. Consistent with the pR^2 results, sLDA had the best performance among the baseline topic-based models across all three testbeds. Examining the baseline model results by testbeds, we found that sLDA had the lowest MAE prediction error and the lowest RMSE prediction error on the Movie testbed. In comparison, both DMR and HDP had relatively lower prediction errors in terms of both MAE and RMSE on the Yelp testbed and the Amazon testbed than on the Movie testbed.

Since the Yelp testbed and the Amazon testbed both had more topic heterogeneity than the Movie testbed, this suggests that DMR and HDP performed better at capturing the topics within the documents than sLDA. DMR is a generative model that considers the documents to be generated under the influence of both the response and the topics; therefore, the latent variables within DMR could differentiate the two types of influence and capture the actual topic heterogeneity within the corpus. Further, the HDP baseline method first learned the topics over the corpus under an unsupervised setting, which captured the actual topic heterogeneity as well. On the other hand, sLDA is a discriminative model that assumes both the documents and the response to be generated from the topics; hence, the learned topics were overfitted with the response. While predictive of the response, such topics cannot be readily generalized to new documents.

TABLE 8
RMSE Results for HDP-IR and Baseline Topic-Based Models

Model	Movie (N = 5 K)	Yelp (N = 5 K)	Yelp (N = 0.3 M)	Amazon (N = 5 K)	Amazon (N = 1.4 M)
HDP-IR	**0.581**	**0.553**	**0.512**	**0.615**	**0.644**
sLDA	0.705 (p<0.001)	0.931 (p<0.001)	-	0.911 (p<0.001)	-
DMR	1.069 (p<0.001)	0.996 (p<0.001)	-	0.999 (p<0.001)	-
HDP	1.129 (p<0.001)	1.047 (p<0.001)	0.964 (p<0.001)	0.982 (p<0.001)	1.034 (p<0.001)

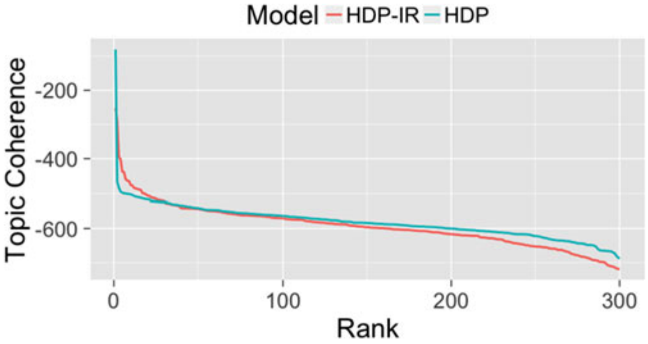

Fig. 5. Topic coherence results for HDP-IR and unsupervised HDP.

To further illustrate the effectiveness of HDP-IR in capturing the topics, we compare the quality of topics learned by our HDP-IR and the quality of unsupervised HDP topics. The quality of topics is measured using *topic coherence*, a commonly used topic quality measurement [43]. Topic coherence primarily measures the co-occurrences of the most probable keywords within each topic and has shown high correlation with the judgment of human annotator [43]. Higher topic coherence score suggests higher topic quality.

Fig. 5 shows the topic coherence result for HDP-IR topics and unsupervised HDP topics on the Movie testbed. We compared the rankings of topic coherence scores between the two models. Overall, the quality of HDP-IR topics is comparable to the quality of unsupervised HDP topics. For the higher quality topics (i.e., Top 100 coherent topics), HDP-IR learned more coherent topics than unsupervised HDP. This evidence shows that the major topics learned by HDP-IR could capture more semantic themes than unsupervised HDP as the HDP-IR model can differentiate the variation caused by neutral semantic themes and response variables of interest. On the other hand, the lower quality topics of unsupervised HDP learned are slightly more coherent than the lower quality topics of HDP-IR, as shown on the right hand side of the chart.

As expected, the SDR generated by the inverse regression component in HDP-IR contributed significantly to the predictive performance. The improvement induced by SDR can be measured by comparing HDP-IR and the HDP baseline model, which is a nonparametric STM based on non-sufficient dimension reduction of the documents. In this case, the SDR of HDP-IR was able to capture provably sufficient predictive information about the response, as reflected in the experimental results. Moreover, the SDR of HDP-IR outperformed the dimension reductions provided by other baseline topic-based models, such as sLDA and DMR. This is because under any true model of $y_d = f(\mathbf{w}_d)$, the statistical property of the SDR guarantees that the response is independent of the original document given the SDR: $y_d \perp\!\!\!\perp \mathbf{w}_d s | R(\mathbf{w}_d)$, where $R(\mathbf{w}_d) = (\boldsymbol{\zeta}_d, \boldsymbol{\Phi}'\mathbf{w}_d)$ is the SDR from HDP-IR. In the baseline STMs, the response is correlated with the original document even given the provided dimension reduction: $y_d \not\perp\!\!\!\perp \mathbf{w}_d | R^{\beta}(\mathbf{w}_d)$, where $R^{\beta}(\mathbf{w}_d) = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$ is the dimension reductions (i.e., topics) of the baseline STMs. As such, the predictions conditioned on the SDR $\mathbb{E}(y_d | \boldsymbol{\zeta}_d, \boldsymbol{\Phi}'\mathbf{w}_d)$ are theoretically more accurate than the predictions conditioned on the non-sufficient dimension reduction $\mathbb{E}(y_d | \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K)$.

6 Conclusions and Future Directions

In this study, we proposed a novel nonparametric supervised topic model, HDP-IR. HDP-IR leverages the nonparametric topic modeling approach to determine the topic structure from the data. Extending the inverse regression model, HDP-IR makes predictions with sufficient dimension reduction of the document to improve the predictive performance. Further, HDP-IR is able to examine large-scale corpora containing millions of documents with the help of a novel efficient inference algorithm based on the state-of-the-art *Stochastic Variational Inference*. Experimental results revealed that the proposed HDP-IR model significantly outperformed existing supervised topic models. The results also suggested that both the nonparametric topic modeling component and SDR could improve the predictive performance.

To the best of our knowledge, the proposed HDP-IR model is the first nonparametric topic model leveraging SDR to improve prediction accuracy. The proposed model provides an important step for future work seeking to study the user-generated text from a topic perspective. Based on our study, we have identified a few future research directions. In this study, we focused on predicting the univariate response. To inform multi-task learning, we would like to examine whether the multivariate response also fits in HDP-IR. We are also interested in exploring the role of additional factors such as the temporal factor in the extension of our proposed model. Moreover, the topic sharing idea is applicable to many other research domains, such as audio and image analysis. We therefore intend to examine the generalization of HDP-IR in such data. Furthermore, since the SDR of the HDP-IR model retains complete information about the response, we are interested in assessing whether the SDR can be used as a measurement of the document to improve the explanatory models in future social science research.

Acknowledgments

This work was supported by the US National Science Foundation under Grant No. SES-1314631 and also under Grant No. DUE-1303362.

References

[1] D. Duan, Y. Li, R. Li, R. Zhang, X. Gu, and K. Wen, "LIMTopic: A framework of incorporating link based importance into topic modeling," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 10, pp. 2493–2506, Oct. 2014.

[2] A. Abbasi, H. Chen, S. Thoms, and T. Fu, "Affect analysis of web forums and blogs using correlation ensembles," *IEEE Trans. Knowl. Data Eng.*, vol. 20, no. 9, pp. 1168–1180, Sep. 2008.

[3] A. Abbasi, S. France, Z. Zhang, and H. Chen, "Selecting attributes for sentiment classification using feature relation networks," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 3, pp. 447–462, Mar. 2011.

[4] Y. Gao, Y. Xu, and Y. Li, "Pattern-based topics for document modelling in information filtering," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 6, pp. 1629–1642, Jun. 2015.

[5] H. Soleimani and D. J. Miller, "Parsimonious topic models with salient word discovery," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 3, pp. 824–837, Mar. 2015.

[6] Y. Zhuang, H. Gao, F. Wu, S. Tang, Y. Zhang, and Z. Zhang, "Probabilistic word selection via topic modeling," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 6, pp. 1643–1655, Jun. 2015.

[7] J. Zeng, Z.-Q. Liu, and X.-Q. Cao, "Fast online EM for big topic modeling," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 3, pp. 675–688, Mar. 2016.

[8] V. C. Cheng, C. H. C. Leung, J. Liu, and A. Milani, "Probabilistic aspect mining model for drug reviews," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 8, pp. 2002–2013, Aug. 2014.

[9] A. Bulut, "TopicMachine: Conversion prediction in search advertising using latent topic models," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 11, pp. 2846–2858, Nov. 2014.

[10] R. Arun, V. Suresh, C. V. Madhavan, and M. N. Murthy, "On finding the natural number of topics with latent dirichlet allocation: Some observations," in *Proc. Pacific-Asia Conf. Knowl. Discovery Data Mining*. Berlin, Germany: Springer, 2010, pp. 391–402.

[11] S. Lacoste-Julien, F. Sha, and M. I. Jordan, "Disclda: Discriminative learning for dimensionality reduction and classification," in *Proc. Advances Neural Inf. Process. Syst.*, 2009, pp. 897–904.

[12] M. D. Hoffman, D. M. Blei, C. Wang, and J. W. Paisley, "Stochastic variational inference," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 1303–1347, 2013.

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, no. Jan., pp. 993–1022, 2003.

[14] W. X. Zhao, J. Wang, Y. He, J.-Y. Nie, J.-R. Wen, and X. Li, "Incorporating social role theory into topic models for social media content analysis," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 4, pp. 1032–1044, Apr. 2015.

[15] J. D. Mcauliffe and D. M. Blei, "Supervised topic models," in *Proc. Advances Neural Inf. Process. Syst.*, 2008, pp. 121–128.

[16] D. Mimno and A. McCallum, "Topic models conditioned on arbitrary features with dirichlet-multinomial regression," in *Proc. 24th Annu. Conf. Uncertainty Artif. Intell.*, 2008, pp. 411–418.

[17] D. Ramage, D. Hall, R. Nallapati, and C. D. Manning, "Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora," in *Proc. Conf. Empirical Methods Natural Language Process.*, 2009, pp. 248–256.

[18] T. N. Rubin, A. Chambers, P. Smyth, and M. Steyvers, "Statistical topic models for multi-label document classification," *Mach. Learn.*, vol. 88, no. 1–2, pp. 157–208, 2012.

[19] J. Zhu, N. Chen, H. Perkins, and B. Zhang, "Gibbs max-margin topic models with data augmentation," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1073–1110, 2014.

[20] M. Rabinovich and D. M. Blei, "The inverse regression topic model," in *Proc. Int. Conf. Int. Conf. Mach. Learn.*, 2014, pp. 199–207.

[21] A. M. Dai and A. J. Storkey, "The supervised hierarchical dirichlet process," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 2, pp. 243–255, Feb. 2015.

[22] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical dirichlet processes," *J. Amer. Statistical Assoc.*, vol. 101, no. 476, pp. 1566–1581, 2006.

[23] R. D. Cook, "Fisher lecture: Dimension reduction in regression," *Statistical Sci.*, vol. 22, pp. 1–26, 2007.

[24] R. M. Neal, "Probabilistic inference using markov chain monte carlo methods," Dept. of Computer Science, Univ. Toronto, Tech. Rep. CRG-TR-93-1, Sept. 1993.

[25] A. Ghose and P. G. Ipeirotis, "Estimating the helpfulness and economic impact of product reviews: Mining text and reviewer characteristics," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 10, pp. 1498–1512, Oct. 2011.

[26] Y. W. Teh, "Dirichlet process," in *Encyclopedia of Machine Learning*. Berlin, Germany: Springer, 2011, pp. 280–287.

[27] R. Huang, G. Yu, Z. Wang, J. Zhang, and L. Shi, "Dirichlet process mixture model for document clustering with feature partition," *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 8, pp. 1748–1759, 2013.

[28] J. Sethuraman, "A constructive definition of dirichlet priors," *Statistica Sinica*, vol. 4, pp. 639–650, 1994.

[29] D. Blackwell and J. B. MacQueen, "Ferguson distributions via pólya urn schemes," *Ann. Statist.*, vol. 1, pp. 353–355, 1973.

[30] C. Wang, J. W. Paisley, and D. M. Blei, "Online variational inference for the hierarchical dirichlet process," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, 2011, pp. 752–760.

[31] K.-C. Li, "Sliced inverse regression for dimension reduction," *J. Amer. Statistical Assoc.*, vol. 86, no. 414, pp. 316–327, 1991.

[32] M. Taddy, "Multinomial inverse regression for text analysis," *J. Amer. Statistical Assoc.*, vol. 108, no. 503, pp. 755–770, 2013.

[33] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. Roy. Statistical Soc.*, pp. 267–288, 1996.

[34] L. Bottou, "Stochastic learning," in *Proc. Adv. Lectures Mach. Learn.*, 2004, pp. 146–168.

[35] M. Braun and J. McAuliffe, "Variational inference for large-scale models of discrete choice," *J. Amer. Statistical Assoc.*, vol. 105, no. 489, pp. 324–335, 2010.

[36] P. J. Bickel and K. A. Doksum, *Mathematical Statistics: Basic Ideas and Selected Topics*. Boca Raton, FL, USA: CRC Press, 2015, vol. 2.

[37] C. Wang and D. M. Blei, "Variational inference in nonconjugate models," *J. Mach. Learn. Res.*, vol. 14, no. 1, pp. 1005–1031, 2013.

[38] B. Pang and L. Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2005, pp. 115–124.

[39] "Yelp's academic dataset," Yelp, 2012. [Online]. Available: http://www.yelp.com/academic_dataset

[40] J. Blitzer, M. Dredze, and F. Pereira, "Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification," in *Proc. Annu. Meeting Assoc. Comput. Linguistics*, 2007, pp. 440–447.

[41] F. Wilcoxon, S. Katti, and R. A. Wilcox, "Critical values and probability levels for the wilcoxon rank sum test and the wilcoxon signed rank test," *Sel. Tables Math. Statist.*, vol. 1, pp. 171–259, 1970.

[42] K. S. Tai, R. Socher, and C. D. Manning, "Improved semantic representations from tree-structured long short-term memory networks," in *Proc. 53rd Annu. Meet. Assoc. Comput. Linguistics 7th Int. Joint Conf. Natural Language Process.*, 2015, pp. 1556–1566.

[43] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proc. Conf. Empirical Methods Natural Language Process.*, 2011, pp. 262–272.

Weifeng Li received the BS degree in management information systems from Shanghai Jiao Tong University, and the PhD degree in management information systems from the University of Arizona. He is an assistant professor in the Department of Management Information Systems, the University of Georgia. His research interests include machine learning, text mining, and social media analytics.

Junming Yin received the PhD degree in computer science from UC Berkeley, in 2011. He is an assistant professor in the Department of Management Information Systems, University of Arizona. Prior to that, he was a lane fellow in the Lane Center of Computational Biology, Carnegie Mellon University. His research focus is on statistical machine learning and its applications in business intelligence, digital marketing, healthcare systems, and computational biology.

Hsinchun Chen received the BS degree from National Chiao-Tong University (Taiwan), the MBA degree from the State University of New York System, Buffalo, and the MS and PhD degrees from New York University. He is the University of Arizona Regents' professor and Thomas R. Brown chair professor in management and technology. He recently served as the lead program director of the Smart and Connected (SCH) Program in the NSF (2014-2015), a multi-year multi-agency health IT research program of USA. He is author/editor of 20 books, 290 SCI journal articles, and 160 refereed conference articles covering digital library, data/text/web mining, business analytics, security informatics, and health informatics. His overall h-index is 91 (25,000 citations for 900 papers according to Google Scholar), among the highest in MIS and top 50 in computer science. He founded the Artificial Intelligence Lab with the University of Arizona in 1989, which has received more than $50M in research funding from NSF, NIH, NLM, DOD, DOJ, CIA, DHS, and other agencies (100 grants, 50 from NSF). He has served as editor-in-chief of major ACM/IEEE, and Springer journals and conference/program chair of major ACM/IEEE/MIS conferences in digital library, information systems, security informatics, and health informatics. He is internationally renowned for leading the research and development in the health analytics (data and text mining; health big data; DiabeticLink and SilverLink) and security informatics (counter terrorism and cyber security analytics; security big data; COPLINK, Dark Web, and Hacker Web) communities. See: http://ai.arizona.edu/hchen. He is a fellow of the ACM, IEEE, and AAAS.