Heterogeneous Supervised Topic Models

Dhanya Sridhar[◇] **and Hal Daumé III**[†] **and David Blei**[♠]

[◇]Université de Montréal and Mila-Quebec AI Institute, Canada
dhanya.sridhar@mila.quebec
[†]University of Maryland and Microsoft Research, USA
hal3@umd.edu
[♠]Columbia University, USA
david.blei@columbia.edu

Abstract

Researchers in the social sciences are often interested in the relationship between text and an outcome of interest, where the goal is to both uncover latent patterns in the text and predict outcomes for unseen texts. To this end, this paper develops the heterogeneous supervised topic model (HSTM), a probabilistic approach to text analysis and prediction. HSTMs posit a joint model of text and outcomes to find heterogeneous patterns that help with both text analysis and prediction. The main benefit of HSTMs is that they capture heterogeneity in the relationship between text and the outcome across latent topics. To fit HSTMs, we develop a variational inference algorithm based on the auto-encoding variational Bayes framework. We study the performance of HSTMs on eight datasets and find that they consistently outperform related methods, including fine-tuned black-box models. Finally, we apply HSTMs to analyze news articles labeled with pro- or anti-tone. We find evidence of differing language used to signal a pro- and anti-tone.

1 Introduction

Researchers in the social sciences are interested in modeling the relationship between text and an outcome of interest. In surveys about elections, how do respondents' open-ended responses relate to evaluations of their anger, fear, or concern (Roberts et al., 2014)? In news media, how do articles relate to the tone used by the writer towards the article's topic (Card et al., 2015)? On social media, how do tweets on mass shootings relate to the political ideology of the user (Demszky et al., 2019)? In these settings, the goal is to both uncover latent properties about the text and predict outcomes for unseen texts.

We develop the heterogeneous supervised topic model (HSTM), a probabilistic approach to modeling outcomes from text. The key innovation in HSTMs—the hetereogeneity—is the idea that individual words can be predictive of the outcome not just on their own, but also specifically in combination with latent topics in the text.

As a running example, consider news articles about US immigration labeled with the pro- or anti-tone of the writer towards US immigration (Card et al., 2015). Table 1 illustrates how an HSTM can be used to analyze patterns in text that relate to pro- and anti-immigration tones.

As is standard for topic models (Blei et al., 2003), HSTMs learn the hidden topics in the corpus, which are captured by the ''neutral'' words shown in each topic. For example, US immigration articles discuss legal aspects (first topic) and views of the administration (second topic).

However, HSTMs also analyze the per-topic, heterogeneous associations between the text and outcome. In Table 1, this is captured by the ''pro'' and ''anti'' words in each topic. For example, in articles about the former administration (second topic), phrases such as ''families'' and ''workers'' reflect a pro-immigration tone, and phrases such as ''border'' and ''control'' reflect an anti-immigration tone. In articles about legal aspects (first topic), the words ''children'' and ''citizen'' capture a pro-immigration tone while words such as ''terrorist'' appear in anti-immigration articles.

A fitted HSTM can also perform supervised prediction of the outcome from text by using its model of the outcome based on the interactions between the topics and words. For example, an HSTM fit to the immigration articles can predict the pro- or anti-immigration tone of unseen articles about US immigration.

Transactions of the Association for Computational Linguistics, vol. 10, pp. 732–745, 2022. https://doi.org/10.1162/tacl_a_00487
Action Editor: Mauro Cettolo. Submission batch: 10/2021; Revision batch: 2/2022; Published 6/2022.

Neutral:	deportation, order, court
Pro:	children, facing deportation, citizen
Anti:	immigration, terrorist, student
Neutral:	president, trump, donald trump
Pro:	families, voters, workers
Anti:	control, border, immigration

Table 1: An example of the analysis enabled by HSTM on US immigration articles labeled with the pro or anti tone towards immigration by the writer.

The HSTM contributes to the large body of work on modeling outcomes from text. One common approach is to fine-tune large language models such as BERT (Devlin et al., 2018). These black-box methods capture complex interactions between parts of text and the outcome, which often leads strong predictive performance. However, it is difficult to visualize the parts of text that drive predictions (Feng et al., 2018; Jacovi and Goldberg, 2020). Even explanation methods such as saliency maps, rationales, and analysis of attention mechanisms are not robust (Jain and Wallace, 2019; Kindermans et al., 2019; Serrano and Smith, 2019) and are often not faithful to the underlying prediction model (Jacovi and Goldberg, 2020).

Since it is difficult to explain the predictions of black-box methods, there has been a push towards more transparent models (Rudin, 2019). Probabilistic models of text such as topic models (Blei et al., 2003) tend to be interpretable to experts (Chang et al., 2009) and are a mainstay in social science research for understanding patterns in text data (Krippendorff, 2018; Grimmer and Stewart, 2013). Their supervised counterparts such as supervised topic models (McAuliffe and Blei, 2008), sparse additive models (Eisenstein et al., 2011), multinomial text regression (Taddy, 2013), and related variants (Card et al., 2017) retain this interpretability, enabling exploratory data analysis. However, these probabilistic methods do not capture heterogeneous relationships between text and outcomes, limiting their predictive performance.

The HSTM enjoys the interpretability of topic models while extending them to capture flexible relationships between text and outcomes. The idea behind HSTMs is that the observed correlation among words provides evidence for latent topics, and latent topics mediate the associations between words and the outcome of interest. To implement this idea, the HSTM posits a joint generative model that captures how latent topics drive both text documents and outcomes. We fit HSTMs with auto-encoding variational Bayes (Kingma and Welling, 2013).

We study the HSTM across eight datasets that range from product reviews to surveys about immigration.[1] First, we evaluate the predictive performance of several methods for supervised text prediction. We compare HSTMs to related topic-modeling approaches (Card et al., 2017; McAuliffe and Blei, 2008; Blei et al., 2003) and fine-tuned BERT models (Devlin et al., 2018). The topic modeling approaches are easy to visualize while the black-box BERT models are difficult to interpret. We find that the HSTM achieves significantly better predictive performance than its topic modeling counterparts across all settings, and performs competitively with BERT in five out of eight settings. Next, we perform an ablation study to understand the effect of different modeling choices made in designing the HSTM. We find that the heterogeneous model of outcomes improves text prediction. Finally, we use news articles to study the use of HSTMs for qualitative data exploration.

2 Background

Consider a corpus of n text documents and their associated outcomes $\mathcal{D} = \{(\mathbf{w}_1, y_1), \ldots, (\mathbf{w}_n, y_n)\}$. Each document $\mathbf{w}_i$ is a sequence of m word tokens $\mathbf{w}_i = \{w_{i1} \ldots w_{im}\}$ that come from a vocabulary of size V. The corpus can also be represented as a bag-of-words (BoW) with matrix x, where x_{iv} is the frequency (or occurrence count) of word v in document i. The outcome y_i can be real-valued, binary, or categorical.

This work builds on topic models, probabilistic models of latent variables and observed text. We extend a topic model based on products of experts (PoE) (Hinton, 2002). In a PoE topic model,

[1]The code and data to reproduce the experiments are available at: https://github.com/dsridhar91/hstm.

each word w_{ij} comes from a product of K (possibly unnormalized) distributions called experts,

$$\begin{aligned} p(w_{ij} \mid \theta, \beta) &= \text{Mult}(\sigma(\theta_i^\top \beta)) \\ &\propto \prod_v \exp\big(\sum\nolimits_k \theta_{ik}\beta_{kv}\big)^v \\ &\propto \prod_k \prod_v \exp\big(\theta_{ik}\beta_{kv}\big)^v, \end{aligned} \tag{1}$$

where $\sigma(\cdot)$ is the softmax function, and v is a word in the vocabulary. Srivastava and Sutton (2017) refer to this model as ProdLDA.

Each document is associated with a local latent variable θ_i on the K-dimensional simplex ($\triangle^K$).[2] Each entry θ_{ik} is the document's affinity for the k-th expert. The k-th expert is governed by the global variable $\beta_k \in \mathbb{R}^V$. Each entry β_{kv} is the k-th expert's affinity for word v.

The PoE topic model is different than latent Dirichl et allocation (LDA) (Blei et al., 2003), a mixed membership model. To draw a word from LDA, it must have high probability in at least one component. In contrast, in a PoE topic model, to draw a word, all K experts must have a high affinity for the word. Consequently, PoE topic models tend to result in more contrasting topics (Srivastava and Sutton, 2017).

3 The Heterogeneous Supervised Topic Model

We present the heterogeneous supervised topic model (HSTM). It combines a PoE topic model of words with a heterogeneous model of the outcomes. HSTMs capture how associations between words and the outcome vary across latent topics.

The HSTM is a generative model of a text and outcome pair $(\mathbf{w}_i, y_i)$. To generate the text $\mathbf{w}_i$, the HSTM uses the PoE topic model described in § 2, and involves the same global and local latent variables β and θ_i. Additionally, it includes a global intercept term $b \in \mathbb{R}^V$. The model of text is

$$w_{ij} \mid \theta, \beta, b \sim \text{Mult}(\sigma(b + \theta_i^\top \beta)). \tag{2}$$

Each entry b_v is an intercept that reflects the baseline popularity of the word v in the corpus. In including this intercept, we follow Eisenstein et al. (2011), Roberts et al. (2014), and Card et al. (2017).

The local latent variable θ_i is parameterized by a logistic-Normal prior,

$$\begin{aligned} r_i &\in \mathbb{R}^K \sim \mathcal{N}(0, \mathbb{I}) \\ \theta_i &\in \triangle^K = \sigma(r_i). \end{aligned} \tag{3}$$

The priors on the global latent variables are

$$\begin{aligned} \beta_k &\in \mathbb{R}^V \sim \mathcal{N}(0, \mathbb{I}) \\ b &\in \mathbb{R}^V \sim \text{Laplace}(0, \lambda\mathbb{I}). \end{aligned} \tag{4}$$

The variables β describe the latent structure of text. In Table 1, we visualize the neutral words by ranking the values of each vector β_k. In the topic about legal aspects of US immigration, the values of β_k were large for words such as ''deportation'' and ''court''.

After generating the document $\mathbf{w}_i$, the HSTM generates its outcome y_i using a generalized linear model (McCullough and Nelder, 1989). Each outcome y is from an exponential family $f(y\,;\,\eta(\cdot))$ (e.g., Bernoulli, Normal, Categorical) whose natural parameter $\eta(\cdot)$ is a linear function of latent variables and text x_i.

The outcome model involves the global variables β and local variables θ_i. Additionally, the outcome model includes global latent variables $a \in \mathbb{R}^K, \omega \in \mathbb{R}^V$, and $\gamma_k \in \mathbb{R}^V$, which are vectors of coefficients. Let $\Theta_i = \{x_i, \theta_i, \beta, \gamma, a, \omega\}$. The outcome model is

$$\begin{aligned} &y_i \mid \Theta_i \sim f(y\,;\,\eta(\Theta_i)), \\ &\text{where} \\ &\eta(\Theta_i) = \overbrace{a^\top \theta_i}^{\text{Linear in topics}} + \overbrace{\omega^\top x_i}^{\text{Linear in words}} \\ &\qquad + \underbrace{\sum\nolimits_v x_{iv}\big(\sum\nolimits_k \theta_{ik}\beta_{kv}\gamma_{kv}\big)}_{\text{Heterogeneity}}. \end{aligned} \tag{5}$$

The priors for the global latent variables in the outcome model are

$$\begin{aligned} a &\in \mathbb{R}^K \sim \mathcal{N}(0, \mathbb{I}) \\ \omega &\in \mathbb{R}^V \sim \text{Laplace}(0, \lambda\mathbb{I}) \\ \gamma_k &\in \mathbb{R}^V \sim \text{Laplace}(0, \tau\mathbb{I}). \end{aligned} \tag{6}$$

[2] This is not necessary to define a PoE but Srivastava and Sutton (2017) show that it results in higher quality topics.

The first term in Eq. 5 is $\theta_i^\top a$. It captures the relationship between the context of a document and the outcome, similar to supervised topic models (McAuliffe and Blei, 2008; Card et al., 2017; Roberts et al., 2014; Eisenstein et al., 2011).

The second term is $x_i^\top \omega$. The coefficients ω capture which words are predictive of the outcome, regardless of the context. When the value ω_v is negative, the vocabulary word v has a negative pull on the outcome. The interpretation is similar for positive values.

In the last term of Eq. 5, the word frequency x_{iv} is multiplied by a quantity that is factorized over the K experts. Each term $\theta_{ik}\beta_{kv}$ is the contribution of expert k to the unnormalized log probability of word v in document i. The variable γ modulates this term to predict the outcome.

In Table 1, the word ''deportation'' has a large value of $\theta_{ik}\beta_{kv}$ in articles about legal aspects of US immigration. However, it may not carry much signal about the pro- or anti-immigration tone. Thus, the corresponding value of γ_{kv} for ''deportation'' will be small. The word ''citizen'' has a large value of γ_{kv} since it has a positive pull on the tone taken towards immigration. The anti-immigration word ''terrorist'' has a large negative value of γ_{kv}.

Since the variables γ and β are both real-valued, we cannot immediately interpret positive (and negative) values of γ as pro and anti words. Table 1 visualizes the pro and anti words in each context by ranking the values of $\frac{\gamma_{kv}}{\beta_{kv}}$. When this rescaled variable is positive, the word v is associated with a pro-immigration tone.

4 Inference

Given a dataset $\mathcal{D}$ of documents and outcomes, we perform maximum a posteriori estimation of the global variables $\mu = \{\beta, \gamma, a, \omega\}$ and variational inference for the local latent variables r_i, which is θ_i on the logit scale (Eq. 3). We follow the work on auto-encoding variational Bayes (Kingma and Welling, 2013).

The local latent variable r_i comes from an amortized variational family $q_\phi(r_i \mid \mathbf{w}_i)$, a multivariate Normal distribution whose mean and diagonal covariance are parameterized by a neural network called the encoder. The encoder has weights ϕ and take the text $\mathbf{w}_i$ as input.

We maximize the evidence lower bound (ELBO),

$$\mathcal{L}(\phi, \mu) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{r_i}\Big[\log p(\mathbf{w}_i \mid r_i, \mu) + \log p(y_i \mid r_i, \mathbf{w}_i, \mu)\Big] - KL\big(q_\phi(r_i \mid \mathbf{w}_i)||p(r_i)\big) + \log p(\mu), \tag{7}$$

a bound on the log marginal likelihood of the documents, $\sum_{i=1}^{n} \log p(\mathbf{w}_i \mid \mu)$. The expectation is taken with respect to the variational distribution $q_\phi(r_i \mid \mathbf{w}_i)$. The term $\log p(\mu)$ corresponds to $L1$ and $L2$ penalty terms,

$$\log p(\mu) = \left(\lambda \sum_v |b_v| + |\omega_v|\right) + \left(\tau \sum_{k,v} |\gamma_{kv}|\right) + \left(\sum_{k,v} \beta_{kv}^2 + \sum_k a_k^2\right). \tag{8}$$

We maximize the ELBO with stochastic gradient descent. When taking gradients, $\nabla \mathbb{E}_{q_\phi(r_i \mid \mathbf{w}_i)}(\cdot)$ poses a challenge. We approximate this gradient with the reparameterization trick (Kingma and Welling, 2013). We approximate $\mathbb{E}_{q_\phi(r_i \mid \mathbf{w}_i)}(\cdot)$ with Monte Carlo and sample $r \sim q_\phi(r_i \mid \mathbf{w}_i)$ by drawing from $\mathcal{N}(0, 1)$ and applying the location-scale transformation with the mean and variance of $q_\phi(\cdot)$. Then $\nabla \mathcal{L}(\phi, \mu)$ can be computed with automatic differentiation. (We use `PyTorch`; implementation details are discussed below.)

After fitting HSTMs, we use the estimated values $\hat{\beta}, \hat{\omega}$, and $\hat{\gamma}$ to visualize the results in Table 1. To form predictions for a new document $\mathbf{w}_d$, we obtain its representation $\hat{\theta}_d$ by passing the words through the encoder. We make predictions using Eq. 5 with the fitted variables.

5 Related Work

HSTMs contribute to the work on extending topic models (Blei et al., 2003) for supervised prediction (McAuliffe and Blei, 2008; Lacoste-Julien et al., 2008; Eisenstein et al., 2011; Roberts et al., 2014; Card et al., 2017) and modeling covariates (Taddy, 2013) such as authorship (Rosen-Zvi et al., 2012), political ideology (Nguyen et al., 2013), voting behavior (Nguyen et al., 2015; Vafa

et al., 2020), or category tags (Ramage et al., 2009).

In this line of work, HSTMs are most closely related to the neural models for documents with metadata (i.e., outcomes) proposed by Card et al. (2017). Card et al. (2017) develop a general framework for topic models fit using auto-encoding variational Bayes. In their framework, either the outcomes are predicted from the inferred topics or interactions between the topics and outcomes are used to model text. In the latter task, the interaction effects learned by the neural models are similar to the heterogeneity that HSTMs capture. However, the neural model framework does not accommodate predicting outcomes and capturing heterogeneity with the same model. In contrast, HSTMs can simultaneously capture topic-driven heterogeneity and enable supervised prediction based on the heterogeneous patterns.

HSTMs also relate to the literature on neural topic models (Dieng et al., 2020; Miao et al., 2016; Benton and Dredze, 2018; Cao et al., 2015; Das et al., 2015; Nguyen et al., 2013; Srivastava and Sutton, 2017; Lau et al., 2017; He et al., 2017; Larochelle and Lauly, 2012). Within this line of work, the HSTM builds on Srivastava and Sutton, (2017) who proposed a PoE topic modzel and adapted auto-encoding variational Bayes (Kingma and Welling, 2013) for topic models. In contrast to their work, which proposes a generative model of text, HSTMs are joint models of text and outcomes, enabling supervised prediction and text analysis that reveals how text relates to the outcome of interest.

6 Empirical Studies

We empirically evaluate HSTMs with eight datasets. The code and data to reproduce the results are available at `https://github.com/dsridhar91/hstm`.

Our empirical analysis is driven by three key questions: 1) How does the predictive performance of HSTMs compare to that of related methods? 2) How do different modeling components of HSTMs contribute to their performance? 3) How do HSTMs help exploratory text analysis?

For predictive performance, we find that: 1) in all eight settings that we study, the predictive performance of HSTMs significantly improves upon the results of approaches that also use a bag-of-words approach to representing text, making them easy to visualize; 2) in five out of eight settings, HSTMs are competitive with fine-tuned BERT models, a state-of-the-art transformer based approach that is difficult to interpret (in contrast, HSTMs are easy to visualize); and 3) in four out of eight settings, the terms in Eq. 5 that capture heterogeneity in the outcome model significantly improve prediction.

For exploratory text analysis, we apply HSTMs to analyze news articles labeled with the writer's pro- or anti-tone. Using the media framing corpus (Card et al., 2015), we study articles on US immigration, same sex marriage, the death penalty and gun control. In all article subjects, we find evidence of differing pro-issue and anti-issue word choices across learned topics.

6.1 Datasets and Preprocessing

We studied eight text corpora with labeled outcomes.

Amazon Office Products. This dataset contains the text of Amazon reviews for office-related products and each review's corresponding rating (one to five stars).[3] We used a randomly selected subset of 20k reviews. We normalized the ratings to be $\in [0, 1]$, and studied the regression task of predicting the normalized rating from text.

Amazon Grocery Products. This dataset contains the text of Amazon reviews and their ratings for grocery-related products.[4] Here, we are interested in the binary classification task of predicting a positive or negative review from text. We again randomly selected 20k reviews. We discarded reviews that had a 3-star rating and binned the remaining reviews into two categories: positive reviews (4- and 5-star ratings) and negative reviews (1- and 2-star ratings).

Yelp Reviews. This dataset contains Yelp reviews.[5] We considered a subset of 20k reviews ranging across all businesses. Each review has an associated rating for the business, ranging from one to four stars. The reviews were again binned as positive and negative based on their ratings. We consider the binary classification of positive and negative reviews from text.

[3] `http://jmcauley.ucsd.edu/data/amazon/`.
[4] `http://jmcauley.ucsd.edu/data/amazon/`.
[5] `http://www.yelp.com/dataset_challenge`.

Method	Description
BoW	Regression with bag-of-words (BoW) frequencies as features
LDA	Regression with document representation from LDA (Blei et al., 2003)
Bow + LDA	Regression with BoW and document representation from LDA
PCA	Regression with PCA word embeddings, averaged over document
sLDA	Supervised LDA (McAuliffe and Blei, 2008) with auto-encoding VB inference
STM	Supervised PoE topic model (Card et al., 2017)
BERT	Pre-trained BERT embeddings fine-tuned for prediction (Devlin et al., 2018)
HSTM	Full model fit jointly (this paper)

Table 2: Description of the methods compared.

PeerRead. This dataset comes from PeerRead, (Kang et al., 2018), a corpus of computer-science papers. The dataset contains each paper's abstract and whether or not they were accepted to a conference. We consider a subset of papers posted to the arXiv under `cs.cl`, `cs.lg`, or `cs.ai` between 2007 and 2017 inclusive.[6] The dataset includes 11,778 papers, of which 2,891 are accepted. We study the binary classification of acceptance (or not) from abstracts.

Media Framing Corpus. The last four datasets come from the media framing corpus (Card et al., 2015), a corpus of news articles labeled with the pro or anti tone taken by the writer towards the article subject. We follow Card et al. (2017) in analyzing this dataset. We consider articles across four subjects: US immigration, same sex marriage, death penalty, and gun rights. Each subject consists of about 4k articles. We study the binary classification of pro/anti tone from articles.

Pre-processing. For each dataset, we constructed a vocabulary of unigrams that occurred in at least 0.07% and in no more than 90% of the documents, and bigrams that occurred in at least 0.7% of the documents as our vocabulary. We considered normalized frequencies as BoW features.

6.2 Experimental Setup

Table 2 describes all the methods that we compared in the empirical studies.

[6]The dataset only includes papers that are not cross listed with any non-`cs` categories and are within a month of the submission deadline for a target conference. The conferences are ACL, EMNLP, NAACL, EACL, TACL, NeurIPS, ICML, ICLR, and AAAI. A paper is marked as accepted if it appeared in one of the target venues. Otherwise, the paper is marked as rejected.

Implementation Details. For logistic regression, ridge regression, PCA, and LDA, we used Python's `sklearn` package with default settings.

We implemented the BERT method with the `transformers` Python library.[7] The BERT method uses the raw text as input, with a maximum token length of 128. To predict, we apply a linear map from the final hidden layer, averaged over all tokens, to the outcome. The BERT method was fine-tuned for 5 epochs. We use stochastic gradient-based optimization with Adam (Kingma and Ba, 2015), using a learning rate of 1×10^{-5}.

STMs, sLDA, and HSTM were all implemented with `PyTorch`. For sLDA, we fit the model proposed by McAuliffe and Blei (2008) but replace the Dirichlet priors with logistic Normal priors to enable stochastic optimization. For auto-encoding VB inference, we used an encoder with two hidden layers of size 300, ReLU activation, and batch-normalization after each layer. For stochastic optimization with Adam, we use automatic differentiation in `PyTorch`. We used a learning rate of 0.01 based on recommendations from Srivastava and Sutton (2017); Card et al. (2017). These methods and BERT were trained on Titan GPUs.

Hyperparameters. To fit HSTMs, STMs, sLDA, PCA, and LDA, we used 50 topics for PeerRead and Semantic Scholar, 30 for Amazon office products and Yelp, 20 for Amazon grocery products, and 10 for all media framing corpus subjects to fit LDA, PCA, and this paper's methods, chosen based on validation log likelihood.

For HSTMs, we selected values of the Laplace prior hyperparameters τ and λ based on cross-validated prediction error.

[7]`https://huggingface.co/transformers/`.

Amazon office		Yelp		PeerRead	
Method	MSE	**Method**	Accuracy	**Method**	Accuracy
BoW	0.80 (0.04)	BoW	0.83 (0.005)	BoW	0.76 (0.01)
BoW + LDA	0.78 (0.04)	BoW + LDA	0.84 (0.008)	BoW + LDA	0.79 (0.004)
LDA	0.90 (0.04)	LDA	0.82 (0.007)	LDA	0.79 (0.005)
PCA	0.87 (0.04)	PCA	0.78 (0.005)	PCA	0.78 (0.01)
sLDA	0.88 (0.04)	sLDA	0.78 (0.03)	sLDA	0.76 (0.01)
STM	0.93 (0.05)	STM	0.83 (0.005)	STM	0.78 (0.006)
HSTM	**0.70 (0.03)**	HSTM	**0.91 (0.005)**	HSTM	**0.81 (0.01)**
Amazon grocery		US immigration		Death penalty	
Method	Accuracy	**Method**	Accuracy	**Method**	Accuracy
BoW	0.89 (0.003)	BoW	0.62 (0.01)	BoW	0.66 (0.02)
BoW + LDA	0.90 (0.004)	BoW + LDA	0.71 (0.03)	BoW + LDA	0.70 (0.01)
LDA	0.89 (0.004)	LDA	0.70 (0.02)	LDA	0.69 (0.02)
PCA	0.88 (0.004)	PCA	0.71 (0.03)	PCA	0.70 (0.01)
sLDA	0.90 (0.003)	sLDA	0.67 (0.03)	sLDA	0.69 (0.01)
STM	0.90 (0.002)	STM	0.72 (0.02)	STM	0.69 (0.01)
HSTM	**0.93 (0.004)**	HSTM	**0.80 (0.01)**	HSTM	**0.78 (0.02)**
Same sex marriage		Gun rights			
Method	Accuracy	**Method**	Accuracy		
BoW	0.73 (0.01)	BoW	0.69 (0.01)		
BoW + LDA	0.75 (0.01)	BoW + LDA	0.69 (0.02)		
LDA	0.74 (0.01)	LDA	0.69 (0.02)		
PCA	0.76 (0.01)	PCA	0.70 (0.01)		
sLDA	0.75 (0.01)	sLDA	0.70 (0.01)		
STM	0.77 (0.01)	STM	0.69 (0.02)		
HSTM	**0.83 (0.01)**	HSTM	**0.78 (0.02)**		

Table 3: The HSTM achieves the best predictive performance across five datasets. Metrics are averages from cross validation across five folds (with standard errors in parenthesis). For regression, we report mean squared error (lower is better). For binary prediction, we report accuracy (higher is better). **Bolded** numbers indicate that the result shows a statistically significant improvement over the next best performing method (with a significance level of $\alpha = 0.01$).

Experimental Details. Following the work of Vafa et al. (2020), we initialize the variables β with a pre-trained model. Specifically, we fit LDA and use the log topics to initialize β. With this initialization, we reweight the KL term to be $2 \cdot D_{KL}(\cdot)$ following work on β-VAE (Burgess et al., 2018) to encourage the posterior to be closer to the prior. For a fair comparison with STMs (Card et al., 2017), we fit it using the same optimization strategies as we use for the HSTM. We also initialize the unnormalized topics using log topics from LDA as we do with the HSTM.

6.3 Predictive Performance

We investigate how HSTMs perform on prediction tasks relative to related methods. First, we compare HSTMs against the other methods that also use a bag-of-words approach to representing documents. This includes all the compared methods except for fine-tuned BERT models. The goal is to assess how HSTMs perform compared to other approaches that are also easy to visualize. Second, we compare HSTMs to fine-tuned BERT models, a state-of-the-art approach to supervised text prediction based on transformers. In contrast to HSTMs, BERT models are difficult to visualize.

Comparisons to BoW-based Approaches. Table 3 compares HSTMs to the related methods in Table 2 across the eight datasets. For binary prediction tasks, we report the average accuracy across five folds. For regression tasks, we report the average mean squared error (MSE).

Setting	HSTM	BERT
Amazon office	0.70 (0.03)	**0.53 (0.02)**
Yelp	0.91 (0.005)	**0.93 (0.002)**
PeerRead	0.81 (0.01)	0.78 (0.01)
Amazon grocery	0.93 (0.005)	**0.96 (0.006)**
US immigration	0.80 (0.01)	0.76 (0.03)
Death penalty	0.78 (0.02)	0.79 (0.01)
Same sex	0.83 (0.01)	0.83 (0.01)
Gun rights	0.78 (0.02)	0.77 (0.02)

Table 4: In five out of eight settings, the HSTM's predictive performance is not statistically significantly different (at a significance level of $\alpha = 0.01$) than the performance of fine-tuned BERT models, a state-of-the-art approach to supervised text prediction based on transformers. **Bolded** results indicate when BERT achieves significantly better performance. BERT is a black-box method, making it hard to interpret which parts of text drove prediction. In contrast, HSTMs are easy to visualize. We report MSE for the regression task in the Amazon office setting, and accuracy for the remaining binary classification tasks. The results are averaged across five folds.

Bolded results show a statistically significant improvement over the next best performing method (at significance level $\alpha = 0.01$).[8]

The results in Table 3 suggest that among approaches that are easy to visualize, HSTMs offer the best predictive performance. In all eight datasets, the HSTM achieves improvements in predictive performance that are statistically significant when compared to the next best performing method (and by extension, all other compared methods). There is variation in which method comes closest to the performance of HSTMs. In the Amazon office and PeerRead settings, the BoW + LDA method is the next best performing method after the HSTM. In the remaining settings, the STM approach, which relates to the framework from Card et al. (2017), is the second best.

Comparison to BERT-based Approach. Table 4 summarizes the comparison between HSTMs and fine-tuned BERT models. In five out of eight settings, the HSTM's predictive performance is not statistically significantly different (at significance level $\alpha = 0.01$) than the performance of BERT. For the regression task in the Amazon office setting, BERT offers drastic improvements in predictive performance. For the binary classification tasks in the Amazon grocery and Yelp settings, BERT again achieves significantly better performance than HSTMs.

We studied BERT and HSTMs further using the full Yelp reviews dataset, which contains 560,000 training examples and 38,000 test examples. This has become a well-studied benchmark dataset for binary classification from text.[9] The BERT model achieves a test-set accuracy of 0.96, which is consistent with state-of-the-art results.[10] Using the best performing hyper-parameters from the smaller Yelp setting, we trained the HSTM on this benchmark corpus. The HSTM achieves a test-set accuracy of 0.91. Taken together with the findings from Table 4, the results suggest that while BERT remains a competitive approach to supervised text prediction, when it is important to trade off performance against the ease of visualizing models, HSTMs might offer an attractive solution.

6.4 Ablation Study

Which components of the HSTM are important for its performance? We conduct an ablation study, removing the heterogeneity and linear-in-words terms from the outcome model to see the effect they have on prediction.

We compare the full HSTM to variants that omit terms from the outcome model in Eq. 5,

$$\begin{aligned} \text{HSTM-Het} \quad & \theta_i^\top \alpha + x_i^\top \omega \\ \text{STM} \quad & \theta_i^\top \alpha. \end{aligned} \tag{9}$$

For a fair comparison, we apply the initialization (using LDA) that the HSTM uses.

Table 5 summarizes the findings from this study. First, compared to STM, which only uses topics to model the outcome, HSTM-Het, which includes the linear-in-words term, substantially improves held out performance. Second, HSTM-Het performs better than its counterpart, BoW+

[8] We apply a Student's t-test to evaluate the null hypothesis that the methods' true mean results are the same. The test statistic follows a Student's t-distribution. For binary classification, accuracy reflects the probability of successfully predicting the true label in a Binomial distribution. However, since the Binomial distribution is well-approximated by a Normal distribution when the sample size is large, we use the t-test even for classification tasks.

[9] For example, see https://paperswithcode.com/sota/sentiment-analysis-on-yelp-binary.

[10] https://huggingface.co/textattack/bert-base-uncased-yelp-polarity.

Amazon office		PeerRead		Yelp	
Method	MSE	**Method**	Accuracy	**Method**	Accuracy
STM	0.93 (0.04)	STM	0.78 (0.006)	STM	0.83 (0.005)
HSTM-Het.	0.72 (0.04)	HSTM-Het.	0.80 (0.01)	HSTM-Het.	0.89 (0.008)
HSTM	0.70 (0.03)	HSTM	0.81 (0.01)	HSTM	**0.91 (0.005)**
Amazon grocery		US immigration		Death penalty	
Method	Accuracy	**Method**	Accuracy	**Method**	Accuracy
STM	0.90 (0.003)	STM	0.72 (0.02)	STM	0.69 (0.01)
HSTM-Het.	0.91 (0.005)	HSTM-Het.	0.77 (0.02)	HSTM-Het.	0.76 (0.01)
HSTM	**0.93 (0.005)**	HSTM	0.80 (0.01)	HSTM	0.78 (0.02)
Same sex marriage		Gun rights			
Method	Accuracy	**Method**	Accuracy		
STM	0.77 (0.01)	STM	0.69 (0.02)		
HSTM-Het.	0.81 (0.01)	HSTM-Het.	0.73 (0.01)		
HSTM	**0.83 (0.01)**	HSTM	**0.78 (0.02)**		

Table 5: An ablation study of the HSTM outcome model reveals that the heterogeneity term in Eq. 5 improves prediction. We report the average results across five folds. **Bolded** numbers indicate that the result shows a statistically significant improvement over the next best performing method (with a significance level of $\alpha = 0.01$).

zLDA (in Table 3). The difference between the methods is that HSTM-Het includes a sparsity-inducing L1 penalty for the linear-in-words term, suggesting that sparsity helps predictive performance in the settings we study.

In four out of the eight datasets, HSTMs see a statistically significant improvement (at significance level $\alpha = 0.01$) over the HSTM-Het variant, which only uses the linear-in-words and linear-in-topics terms. The finding suggests that the heterogeneity term is useful for text prediction. The gain over HSTM-Het varies, with the biggest gain in the media framing corpus setting with articles about gun control.

6.5 Exploratory Analysis

Following the example in Table 1, we use HSTMs to perform exploratory data analysis on the remaining article subjects in the media framing corpus (Card et al., 2015): death penalty, same sex marriage, and gun rights.

Table 6 visualizes the first four topics learned by HSTMs in each set of articles. As with the example in Table 1, the ‘‘neutral’’ words reflect the learned topics, the ‘‘pro’’ words capture language used to signal a pro tone (and similarly with ‘‘anti’’ words). As we describe in § 3, we identify the neutral, pro, and anti words using the fitted HSTM model parameters β and γ. For each subject in the media framing corpus, Table 6 suggests that HSTMs find relevant topics and heterogeneous associations between words and the outcome (i.e., tone) in the pro and anti words, as below.

In articles about the death penalty, different executions and court rulings appear as topics. In one ruling, pro death penalty language focuses on the perpetrators (‘‘Tamerlan’’) and the victims. In contrast, anti death penalty articles invoke phrases such as ‘‘evidence’’ and ‘‘pleas’’. Pro death penalty language invokes the crimes and their victims while anti death penalty language includes racially biased rulings, exoneration, and innocence.

In articles about same sex marriage, the HSTM topics include supreme court rulings, recently passed equality bills, and the views of the church. In articles about recently passed bills, pro same sex marriage is reflected by highlighting politicians and states that passed bills (‘‘Cuomo’’, ‘‘Maryland’’) while the anti tone is reflected by words such as ‘‘opposes’’ and ‘‘traditional.’’

In articles about gun rights, some discussed topics are about marches, background checks, school shootings, and gun laws. In articles about marches, being pro gun-rights is signaled with language about firearms while an anti gun-rights tone is signaled by invoking gun violence.

Death Penalty	
Neutral:	execution, court, state, supreme, electric, stay, convicted
Pro:	rejected, refused, executed, denied, killed, dead, wayne
Anti:	around, spared, flames, high, least, evidence, reprieve
Neutral:	penalty, jury, murder, guilty, said, trial, convicted
Pro:	victims, hasan, roof, head, victim, five, bolin
Anti:	parole, spared, life, deal, want, prison, said
Neutral:	penalty, cases, court, state, capital, justice, executions
Pro:	record, liberal, expected, executed, approved, defendant, california
Anti:	sentences, poor, racial, juries, bias, black, white
Neutral:	said, murder, penalty, charged, police, man, two
Pro:	said, executed, store, tamerlan, victims, could, charged
Anti:	plea, parole, without, evidence, fatal, possibility, williams
Same Sex Marriage	
Neutral:	new, york, cuomo, equality, gay, albany, bill
Pro:	equality, cuomo, legislation, maryland, push, first, making
Anti:	conservative, woman, christie, opposes, traditional, homosexual, race
Neutral:	society, editor, children, marriage, tax, moral, institution
Pro:	citizens, even, heterosexual, loving, threat, lesbians, parents
Anti:	society, moral, man, homosexual, homosexuality, definition, sexual
Neutral:	church, methodist, creech, united, methodists, bishops, bishop
Pro:	clergy, us, episcopalians, unitarian, rabbis, perform, allow
Anti:	stance, denomination, woman, catholic, bishops, homosexuality, texas
Neutral:	court, ban, ruling, supreme, federal, legal, state
Pro:	struck, supreme, equality, latest, five, last, indiana
Anti:	stay, whether, walker, thursday, courts, hold, dismissed
Gun Rights	
Neutral:	nra, march, said, children, washington, told, million
Pro:	nra, convention, rifle, target, executive, firearms, association
Anti:	violence, high, day, march, lost, demand, americans
Neutral:	background, system, check, purchases, shows, loophole, checks
Pro:	fbi, said, records, information, periods, national, using
Anti:	background, restraining, health, lobby, initiative, purchases, sensible
Neutral:	school, killed, high, said, ballot, columbine, colorado
Pro:	convention, rights, eight, outside, but, tech, lessons
Anti:	killed, children, denver, roberti, called, colorado, recall
Neutral:	illegal, charged, agents, crimes, said, police, federal
Pro:	firearm, period, find, banned, goods, bought, honest
Anti:	illegal, atf, found, firearms, charged, tobacco, head

Table 6: Topics learned by HSTMs from the media framing corpus. We visualize the top seven words across four topics. In each fitted topic, the ‘‘neutral’’ words capture the topic, ‘‘pro’’ words language is used to signal a pro tone towards the subject (and similarly for ‘‘anti’’ words).

Table 7 visualizes the overall pro and anti words in each article subject. As discussed in § 3, we use the inferred model parameters ω (i.e., linear-in-words term). For each subject, the overall pro-issue and anti-issue words are different than the words that have heterogenous associations with tone, depending on the topic (in Table 6).

Comparisons to Related Methods. To further investigate the benefits of HSTMs for exploratory text analysis, Table 8, visualizes the type of analysis enabled by LDA, STMs, and BoW regression on articles about same sex marriage. Table 8 reveals some qualitative contrasts between HSTM-based text analysis and the related analyses.

	Death penalty
Overall pro:	first, pronounced, died, victims, upholds, rejected, upheld
Overall anti:	spared, life, freed, exonerated, bias, indigent, cited
	Same sex marriage
Overall pro:	latest, toward, maryland, even, love, struck, equality
Overall anti:	woman, homosexual, define, homosexuality, definition, added, evangelical
	Gun rights
Overall pro:	means, owners, lawsuits, lawful, liberal, convention, defending
Overall anti:	sensible, lobby, found, lost, reasonable, handguns, dangerous

Table 7: Words with highest positive and negative ω weights learned by HSTMs for various subjects in the media framing corpus. These reflect words that reflect overall pro and anti tone irrespective of topic.

STM topics
new, york, marriage, civil, legalize
marriage, married, god, us, people
church, rev, methodist, unions, clergy
marriage, court, states, ruling, state
LDA topics
marriage, new, gay, state, york, civil, rights
marriage, gay, people, couples, but, one, right
church, said, unions, united, rev, methodist, bishop
marriage, court, state, gay, supreme, couples, states
BoW regression coefficients
Pro: benefits, equality, new, married, rights, gay, couples
Anti: amendment, said, woman, marriages, man, church, ban

Table 8: HSTMs reveal different insights compared to the analysis enabled by the STM, LDA, and BoW regression on articles about same sex marriage. First, we visualize topics 1 through 4 found by the STM and LDA. Second, we visualize the pro and anti words found by BoW regression.

After fitting LDA to the articles, most learned topics involve words such as ‘‘marriage’’ and ‘‘gay’’, making it harder to distinguish the themes that each topic capture. Further, from the LDA output alone, it can be difficult to distinguish between pro and anti language in each topic, especially with lack of knowledge of the domain.

Since we initialized both STMs and HSTMs with the log topics from LDA, it is not surprising that they find topic related to those from LDA. However, the topics found by the STM appear much more similar to LDA topics than the ones found by HSTM. As with LDA, ‘‘marriage’’ appears in three out of the four displayed STM topics.

As with LDA, the topics do not appear to reveal pro and anti language. For example, the third topic found by the STM includes the word ‘‘union’’, which indicates a pro tone towards same sex marriage, as well as the word ‘‘church,’’ which is associated more closely with the anti side.

The topics found by HSTM (shown in Table 6) from same sex marriage articles diverge from the topics found by both the STM and LDA. For example, the neutral words in the second topic include words such as ‘‘moral’’ and ‘‘institution’’, which were not ranked high in the corresponding LDA and STM topics.

Moreover, HSTM appears to be more effective at separating neutral words from pro- and anti-tone ones. For example, in the third topic about same sex marriage, the neutral words seem to capture religion generally. The pro words include ‘‘rabbis’’, ‘‘unitarian’’, and ‘‘episcopalian’’, denominations which are known to be more open to same sex marriage compared with other religious groups. The anti words include ‘‘catholic’’ and ‘‘bishop’’, ideologies that maintain an anti same sex marriage stance.

BoW regression infers pro and anti language that is similar to the overall pro-issue and anti-issue words found by HSTMs (Table 7), but (by design) does not cluster these into topics.

7 Discussion

We addressed the problem of modeling the relationship between text and an outcome of interest in service of text analysis and supervised prediction. We proposed the HSTM to jointly model the structure in text and capture heterogeneity in the relationship between text and outcomes based on the latent topics that drive the text.

We evaluated the HSTM on eight prediction settings and find that it outperforms related methods. We conducted an ablation study to examine how different components of the HSTM contribute to its performance. The study revealed that the HSTM's heterogeneous model of outcomes improves prediction. In addition to forming predictions, the HSTM lets us explore text datasets. We applied the HSTM to four corpora of news articles labeled with pro/anti tone, and discovered pro and anti wording choices.

The HSTM opens several avenues of future work. One area to explore is combining black-box language models such as BERT (Devlin et al., 2018) with the HSTM to leverage both word embeddings and topics. Another area is adapting the HSTM to produce dynamic topics that vary over time (Blei and Lafferty, 2006).

Acknowledgments

This work is supported by NSF grants IIS 2127869 and SaTC 2131508, ONR grants N00014-17-1-2131 and N00014-15-1-2209, the Simons Foundation, and the Sloan Foundation.

References

Adrian Benton and Mark Dredze. 2018. Deep Dirichlet multinomial regression. In *Proceedings of NAACL-HLT*. https://doi.org/10.18653/v1/N18-1034

David M. Blei and John D. Lafferty. 2006. Dynamic topic models. In *Proceedings of ICML*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022.

Christopher P. Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. 2018. Understanding disentangling in β-vae. *arXiv preprint arXiv:1804.03599*.

Ziqiang Cao, Sujian Li, Yang Liu, Wenjie Li, and Heng Ji. 2015. A novel neural topic model and its supervised extension. In *Proceedings of AAAI*.

Dallas Card, Amber Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of ACL*. https://doi.org/10.3115/v1/P15-2072

Dallas Card, Chenhao Tan, and Noah A. Smith. 2017. Neural models for documents with metadata. In *Proceedings of ACL*. https://doi.org/10.18653/v1/P18-1189

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L. Boyd-Graber, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Proceedings of NeurIPS*.

Rajarshi Das, Manzil Zaheer, and Chris Dyer. 2015. Gaussian LDA for topic models with word embeddings. In *Proceedings of ACL*.

Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Jesse Shapiro, Matthew Gentzkow, and Dan Jurafsky. 2019. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. In *Proceedings of NAACL-HLT*. https://doi.org/10.18653/v1/N19-1304

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.

Adji B. Dieng, Francisco J. R. Ruiz, and David Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Comxputational Linguistics*, 8:439–453. https://doi.org/10.1162/tacl_a_00325

Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. 2011. Sparse additive generative models of text. In *Proceedings of ICML*.

Shi Feng, Eric Wallace, Alvin Grissom II, Mohit Iyyer, Pedro Rodriguez, and Jordan Boyd-Graber. 2018. Pathologies of neural models make interpretations difficult. In *Proceedings of EMNLP*. https://doi.org/10.18653/v1/D18-1407

Justin Grimmer and Brandon M. Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3):267–297. https://doi.org/10.1093/pan/mps028

Junxian He, Zhiting Hu, Taylor Berg-Kirkpatrick, Ying Huang, and Eric P. Xing. 2017. Efficient correlated topic modeling with topic embedding. In *Proceedings of KDD*.

Geoffrey E. Hinton. 2002. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8):1771–1800. https://doi.org/10.1162/089976602760128018, PubMed: 12180402

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable nlp systems: How should we define and evaluate faithfulness? In *Proceedings of ACL*. https://doi.org/10.18653/v1/2020.acl-main.386

Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. In *Proceedings of NAACL-HLT*.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (peerread): Collection, insights and nlp applications. In *Proceedings of NAACL-HLT*. https://doi.org/10.18653/v1/N18-1149

Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T. Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. 2019. The (un) reliability of saliency methods, *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pages 267–280. Springer. https://doi.org/10.1007/978-3-030-28954-6_14

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR*.

Diederik P. Kingma and Max Welling. 2013. Auto-encoding variational Bayes. *arXiv preprint arXiv:1312.6114*.

Klaus Krippendorff. 2018. *Content Analysis: An Introduction to its Methodology*. Sage publications. https://doi.org/10.4135/9781071878781

Simon Lacoste-Julien, Fei Sha, and Michael I. Jordan. 2008. DiscLDA: Discriminative learning for dimensionality reduction and classification. In *Proceedings of NeurIPS*.

Hugo Larochelle and Stanislas Lauly. 2012. A neural autoregressive topic model. In *Proceedings of NeurIPS*.

Jey Han Lau, Timothy Baldwin, and Trevor Cohn. 2017. Topically driven neural language model. *Proceedings of ACL*.

Jon D. McAuliffe and David M. Blei. 2008. Supervised topic models. In *Proceedings of NeurIPS*.

P. McCullough and J. A. Nelder. 1989. *Generalized Linear Models*. Chapman and Hall.

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *Proceedings of ICML*.

Viet-An Nguyen, Jordan Boyd-Graber, Philip Resnik, and Kristina Miler. 2015. Tea party in the house: A hierarchical ideal point topic model and its application to republican legislators in the 112th Congress. In *Proceedings*

of ACL. https://doi.org/10.3115/v1/P15-1139

Viet-An Nguyen, Jordan L. Ying, and Philip Resnik. 2013. Lexical and hierarchical topic regression. In *Proceedings of NeurIPS*.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. 2009. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of EMNLP*. https://doi.org/10.3115/1699510.1699543

Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. 2014. Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4):1064–1082. https://doi.org/10.1111/ajps.12103

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2012. The author-topic model for authors and documents. *Proceedings of UAI*.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215. https://doi.org/10.1038/s42256-019-0048-x

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of ACL*.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. *Proceedings of ICLR*. https://doi.org/10.18653/v1/P19-1282

Matt Taddy. 2013. Multinomial inverse regression for text analysis. *Journal of the American Statistical Association*, 108(503):755–770. https://doi.org/10.1080/01621459.2012.734168

Keyon Vafa, Suresh Naidu, and David M. Blei. 2020. Text-based ideal points. In *Proceedings of ACL*.