Environment-adjusted Topic Models

Reviewed on OpenReview:

Editor:

Abstract

Probabilistic topic models are a powerful tool for extracting latent themes from large text datasets. However, when applied to data from diverse sources or environments, topic models can fail to capture consistent themes across different sources. Recognizing this limitation, we propose environment-adjusted topic models (EATMs) designed to uncover consistent topics across varying environments. EATMs are unsupervised probabilistic models that analyze text from multiple environments and can separate universal and environment-specific terms to learn consistent topics. Through extensive experimentation on a variety of political content, from ads to tweets and speeches, we show that EATMs produce interpretable global topics and separate environment-specific words. Importantly, EATMs retain higher performance on out-of-distribution data compared to strong baselines.

Keywords: Topic Models, ML for Social Science

1 Introduction

Topic models are a prominent tool for text analysis, offering a principled approach for extracting latent themes from vast amounts of text data. Their effectiveness and efficiency make them indispensable tools for text analysis, representation, and predictive tasks. (Blei et al., 2003; Blei and Lafferty, 2006; Mimno et al., 2009)

In the social sciences, these models have an additional application: estimating causal effects using interpretable latent variables uncovered from text (Roberts et al., 2014; Feder et al., 2022). For example, when studying voting behavior, researchers might want to understand the influence of political ads from various channels on election results (Ash et al., 2020). A common approach is to train a topic model on a sample of the ads and use the topic proportions in the model as a low dimensional interpretable representation of the text (Ash and Hansen, 2023).

However, a challenge arises when using topic proportions as treatment variables in such studies. Traditional topic models might inadvertently capture differences in language specific to some channels rather than the central themes of content. This can introduce bias when estimating causal effects.

To address this issue, this paper proposes environment-adjusted topic models (EATMs). EATMs are hierarchical probabilistic models designed to analyze text from different environments. To create environments, we leverage our knowledge about potential variables that lead to topic distributions learning irrelevant information within and across datasets, such as the source, style, and region where the text emerges. EATMs are constructed to effectively separate universal themes from environment-specific terms.

We test the performance of EATMs on a diverse set of political content, including ads, tweets, and speeches. Our results show that EATMs can consistently produce clear and interpretable topics while effectively filtering out environment-specific terms. Furthermore, when applied to out-of-distribution data, EATMs perform well compared to other models, showcasing their potential utility in a variety of text analysis scenarios.

Our contributions are (1) introducing two variants of environment-adjusted topic models (EATM), (2) demonstrating the benefit of using EATMs when learning from documents that include different ideologies and style and that originate from different sources, and (3) building three datasets that allow comparing topic models across multiple environments, including held-out, out-of-distribution environments.

2 Environment-adjusted Topic Models

The environment-adjusted topic model (EATM) is a probabilistic model that uncovers topics that occur across different text corpora, enabling us to learn a distribution of topics that is stable across different datasets or environments. It is tailored for capturing both global and environment-specific effects.

Consider a corpus of n text documents represented as $D = \{\mathbf{w}_1, \ldots, \mathbf{w}_n\}$. Each document $\mathbf{w}_i$ is a sequence of m word tokens, given by $\mathbf{w}_i = \{w_{i1}, \ldots, w_{im}\}$, that come from a vocabulary of size V. In topic modeling, each document is represented as a mixture of topics, with a local latent variable θ_i denoting the per-document topic intensities. Topics are denoted by β, and each β_k is a probability distribution over the vocabulary. To represent environments, we introduce an index $e \in \{1, 2, \ldots, E\}$. We introduce a new latent variable, $\gamma_{e,k,v}$, that is designed capture the effect that each environment has on the topic-word distribution, β. The graphical model for the environment-adjusted topic model is represented in Appendix A and the algorithm in Appendix B. In an environment-adjusted topic model, each document is assumed to have been generated through the following process:

1. For each document:
   (a) Retrieve corresponding environment index, $e \in \{1, 2, \ldots, E\}$
   (b) Draw topic proportions $\theta_i \sim \mathcal{N}(\cdot, \cdot)$
   (c) For each word m:
      i. Choose a topic assignment $Z \sim \mathrm{Mult}(\pi(\theta_i))$.
      ii. Choose a word $w_{ij} \sim \mathrm{Mult}(\pi(\beta_z + \gamma_{e,z}))$.

2.1 Sparse Priors

Our models build on the general topic model with the additional assumption that documents are generated from multiple environments. The goal is to separate global from environment-specific information. To do this, we introduced a new latent variable, γ, that is indexed by the environment $e \in E$ and topic $k \in K$. We further posit that environment effects on the global topic-word distribution β should be sparse. That is, if we have a dataset with tweets, speeches, and articles discussing political topics, many words will be shared across sources. We only want γ to place high density on words that are specific to a subset of environments. We experiment with two different approaches for enforcing sparsity on γ: the Horseshoe prior and the Automatic Relevance Determination (ARD) prior.

The Horseshoe prior. To enforce sparsity in EATM 1, we can employ a horseshoe prior for γ, which is defined as:

$$\gamma_{e,k,v} \mid \lambda_{ek}, \tau \sim \mathcal{N}(0, \lambda_{e,k}^2 \tau^2)$$

Here, $\lambda_{e,k}$ represents the local shrinkage parameter specific to each environment e and topic k, while τ is the global shrinkage parameter that applies to all γ variables. The horseshoe prior for $\lambda_{e,k}$ has the following characteristic form:

$$\begin{aligned} \lambda_{e,k} &\sim \mathcal{C}^+(0,1) \\ \tau &\sim \mathcal{C}^+(0,1) \end{aligned}$$

where $C^+(0,1)$ denotes the standard half-Cauchy distribution, which has a probability density function that is flat around zero and has heavy tails. As such, the prior is designed to retain strong signals in the data ($\gamma_{e,k}$ values far from zero) while pushing negligible effects towards zero. This structure encourages the majority of these environment-specific deviations to exhibit strong shrinkage, driving them towards zero, while allowing some to possess significant non-zero values, thereby highlighting truly influential environment-specific effects and allowing β to maintain its ability to capture topics across documents.

Another way to enforce sparsity is through the use of automatic relevant determination priors.

Automatic Relevance Determination (ARD) and Empirical Bayes. In many real world tasks, the input data contains a large number of irrelevant features Automatic Relevance Determination (ARD) is a tool used to determine the relevance of input features. (MacKay, 1992). Its basis is to assign independent Gaussian priors to the feature weights. Given the feature weights η, the ARD assigns priors as:

$$p(\eta|\alpha) = \prod_c \mathcal{N}(\eta_c|0, \alpha_c^{-1}) \tag{1}$$

Here, $\alpha = \{\alpha_c\}$ represents a vector of hyperparameters. Each hyperparameter α_i essentially controls how far its corresponding weight η_c is allowed to deviate from zero. A larger value of α_i will constrain its corresponding weight closer to zero, effectively indicating that the feature is less relevant.

Rather than fixing them a priori, ARD hyperparameters are learned from the data by maximizing the Bayesian evidence.

Empirical Bayes (EB) provides a systematic way to set these hyperparameters (Carlin and Louis, 2000). In the context of ARD, the EB method involves computing the marginal likelihood of the observed data, integrating out the main feature weights η given the current hyperparameter values:

$$p(\text{data}|\alpha) = \int p(\text{data}|\eta) \cdot p(\eta|\alpha)\, d\eta \tag{2}$$

Optimizing the marginal likelihood with respect to the hyperparameters α. This maximization procedure provides the "best" hyperparameter values in terms of fitting the observed data:

$$\hat{\alpha} = \arg\max_{\alpha} p(\text{data}|\alpha) \tag{3}$$

This process uses the observed data to find the hyperparameter values that maximize the likelihood of the data under the ARD model. The resulting values $\hat{\alpha}$ are then used in subsequent Bayesian analyses to compute the posterior distributions of the feature weights.

We now present a variant of our environment-adjusted topic model that draws on insights from work on automatic relevance determination (ARD) and empirical bayes. EATM 2 is identical to EATM 1 with the exception being that it deploys an ARD prior on γ, rather than a horseshoe prior.

$$\sigma_{e,k,v} \sim \text{Gamma}(a, b)$$
$$\gamma_{e,k,v} \sim \mathcal{N}(0, \sigma_{e,k,v}^{-1})$$

The ARD prior, formulated as $\gamma_{e,k,v} \sim \mathcal{N}(0, \sigma_{e,k,v}^{-1})$, enables a level of adaptiveness in determining the relevance of features, by associating a unique precision parameter to each feature. In the presented model, ARD is employed to the term $\gamma_{e,k,v}$. Each coefficient of this term is associated with a environment e, topic k, and vocabulary item v. These coefficients are drawn from a normal distribution with a variance controlled by their individual precision terms $\sigma_{e,k,v}$. These precision terms themselves are sampled from a Gamma distribution, allowing the model to adaptively learn and push irrelevant topic-environment-vocabulary relationships towards zero, achieving feature selection.

3 Inference

EATMs rely on multiple latent variables: topic-word distributions β, document-topic proportion θ, and environment-specific deviations on the topic-word distribution γ. Conditional on the text, we perform inference on these latents through the posterior distribution $p(\theta, \beta, \gamma | D)$. Calculating this posterior is intractable, so we rely on approximate inference.

We use mean-field variational inference to approximate the posterior distribution (Jordan et al., 1999; Blei et al., 2017). Set $\phi = (\theta, \beta, \gamma)$ as the variational parameters, and let $q_\phi(\theta, \beta, \gamma)$ be the family of approximate posterior distribution, indexed by the variational parameters. Variational inference aims to find the setting of ϕ that minimizes the KL divergence between q_ϕ and the posterior. Minimizing this KL divergence is equivalent to maximizing the evidence lower bound (ELBO):

$$\text{ELBO} = \mathbb{E}_{q_\phi}[\log p(\theta, \beta, \gamma) + \log p(x|\theta, \beta, \gamma) - \log q_\phi(\theta, \beta, \gamma)] \tag{4}$$

The ELBO sums the expectation of the log joint, which is broken up into the log prior and log likelihood, and the entropy of the variational distribution.

To approximate the posterior, we use the mean-field variational family, which results in our latent variables, θ, β, and γ being mutually independent and each governed by a distinct factor in the variational density. The mean-field family factorizes over the latent variables, where d is a document, k is a topic, and e is an environment:

$$q_\phi(\theta, \beta, \gamma) = \prod_{d,k,e} q(\theta_d) q(\beta_k) q(\gamma_e).$$

We employ Gaussian factors as our variational densities:

$$\theta_{d,k} \sim \mathcal{N}(\mu_\theta, \sigma_\theta^2) \tag{5}$$
$$\beta_{k,v} \sim \mathcal{N}(\mu_\beta, \sigma_\beta^2) \tag{6}$$
$$\gamma_{e,k,v} \sim \mathcal{N}(\mu_\gamma, \sigma_\gamma^2) \tag{7}$$

Our objective is to optimize the ELBO with respect to the variational parameters:

$$\phi = \{\mu_\theta, \sigma_\theta^2, \mu_\beta, \sigma_\beta^2, \mu_\gamma, \sigma_\gamma^2\}.$$

We maximize the ELBO using Black-Box Variational Inference (BBVI) with the reparameterization trick to handle the latent variables present in our model (Ranganath et al., 2014; Kingma and Welling, 2013). The model parameters are optimized using minibatch stochastic gradient descent in PyTorch by minimizing the negative ELBO. To achieve this optimization, we employ the Adam optimizer (Kingma and Ba, 2014).

Empirical Bayes Inference For the Empirical Bayes update, we focus on optimizing the hyperparameters of the Gamma distribution for σ. Given observed data $\mathcal{D}$ and hyperparameters a and b, the aim of Empirical Bayes is to estimate the hyperparameters by maximizing the marginal likelihood.

$$\hat{a}, \hat{b} = \arg\max_{a,b} p(\mathcal{D}|a, b) \tag{8}$$

We focus on γ, the latent variable with a Gaussian distribution, and its variance, which is determined by σ. σ is given by a Gamma distribution parameterized by hyperparameters a and b. Since we are learning with mean-field variational inference the empirical Bayes objective is to minimize the KL divergence between the approximate posterior of γ and its prior.

$$\mathrm{KL}(q(\gamma)||p(\gamma|a, b)) = \mathrm{KL}\left(\mathcal{N}(\gamma; \mu, \sigma)||\mathcal{N}(0, \sigma^{-1})\right) \tag{9}$$

The Empirical Bayes update adjusts the hyperparameters to minimize the KL divergence, effectively making the approximate posterior $q(\gamma)$ closer to the intended prior.

4 Related Work

Topic Models. Probabilistic topic models, especially Latent Dirichlet Allocation (LDA) (Blei et al., 2003), are crucial in text analysis for uncovering latent themes in large datasets. They are also the backbone for estimating causal effects with text data (Feder et al., 2022). Additionally, the Structural Topic Model (STM) by Roberts et al. (2014) integrated document metadata, boosting topic interpretability and relevance. Our environment-adjusted topic models (EATMs) advance these models by adapting to diverse environments, ensuring consistent topic discovery across various sources.

Sparse priors and empirical Bayes. Sparse priors, essential in Bayesian models for inducing sparsity and enhancing interpretability, are complemented by empirical Bayes methods, which focus on parameter estimation. The integration of sparse priors with empirical Bayes enhances performance in high-dimensional settings, as demonstrated by

Tipping (2001) in the development of the Relevance Vector Machine. The effectiveness of combining these approaches for robust estimation and feature selection is further evidenced in the work of Carvalho et al. (2010) on horseshoe estimators and Brown and Griffin (2010) with their exploration of normal-gamma priors. Efron (2012) provides a comprehensive overview of empirical Bayes methods provides valuable insights for large-scale inference, relevant to our EATMs.

Learning from multiple environments. There is a growing literature on invariant learning, which describes the problem of learning a representation that is generalizable across different data distributions (Peters et al., 2016; Arjovsky et al., 2019; Schölkopf et al., 2021). Our work considers a related problem of learning stable representations of text from multiple environments, focusing on a probabilistic approach.

5 Experiments and Results

We study the environment-adjusted topic model (EATM) on three datasets, each including documents from multiple environments. Table 3 in the Appendix presents the characteristics of each dataset. We compare both environment-adjusted topic models to the relevant baselines: Latent Dirichlet Allocation (LDA) (Blei et al., 2003, 2017), Vanilla topic model (VTM) represents the base version of our model without any environment-specific variations, the non-sparse environment-adjusted (nEATM) represents the EATM, but with a Normal distribution on the γ prior, and ProdLDA (Srivastava and Sutton, 2017; Hinton, 2002). Lastly, we evaluate nEATM, EATM 1, and 2 + γ. We want to evaluate how the performance shifts when including environment-specific information. To do this, we sum the environment-specific effects from a particular environment, $\gamma_{k,v}$, to $\beta_{k,v}$. For example, in Table 2 EATM 1 + γ represents the perplexity when using $\gamma_{k,v} + \beta_{k,v}$, rather than solely $\beta_{k,v}$, where the $\gamma_{k,v}$ is the learned article specific effects on the global topic-distribution, β. See Appendix D for more implementation details.

Our empirical analysis is driven by three key questions:

1. How stable is EATM's perplexity when tested on datasets from different environments?
2. How does stability change when we incorporate envirnoment specific information (γ) when calculating perplexity?
3. How does EATM's performance compare to other topic model variants?

For stability, we find that: (1) In all test settings, the predictive power of EATM is approximately stable across environments and significantly more stable than when incorporating environment-specific effects (γ) into the prediction. (2) When using environment-specific effects from an environment outside of the one we are currently testing (i.e., using article-specific effects to calculate perplexity for speeches), perplexity drops considerably (nearly 300 points and 400 points for EATM 2 and 1, respectively). (3) Compared to baselines, our EATM has better perplexity across all datasets.

5.1 Style Dataset

Our style dataset consists of news articles, senator tweets, and senate speeches related to U.S. immigration. The U.S. immigration articles emerge from the Media Framing Corpus (Card et al., 2015). We use all $4,052$ articles in the dataset. We augment the dataset used by Vafa et al. (2020). It is based on an open-source set of tweets of U.S. legislators from 2009–2017.

We create a list of keywords related to immigration and sample $4,052$ tweets that contain at least one of the keywords; we repeat the same process for Senate speeches from the 111-114th Congress (Gentzkow et al., 2018). The reason for using topics related to immigration is to ensure that the text style accounts for the majority of the variation between datasets and that the support of topics is the same. Using our style dataset, we perform two evaluations: (1) We train on all three environments and test on unseen samples from each environment in the training set. That is, we have three test sets, each one containing samples from only one environment. (2) we train on all speeches and articles and test on a held-out dataset of tweets.

Model Perplexity	Articles	Speeches	Tweets
VTM	1558	1590	1768
ProdLDA	3312	2057	11674
nEATM	1713	1785	1796
EATM 1	1575	1514	1593
EATM 2	1430	1417	1488

Table 1: The EATMs have lower perplexity than baseline models when trained on congressional senate speeches and news articles related to immigration and tested on tweets related to immigration from U.S. senators.

Table 1 represents the perplexity of our Vanilla topic model, ProdLDA, nEATM, EATM 1, and EATM 2 when trained on speeches and articles and tested on tweets. EATM 1 and 2 perform better on the in-distribution as well as out-of-distribution samples, highlighting our model's ability to capture relevant global information while simultaneously disregarding features that are predictive in one environment but not another. The EATMs are also more stable across different distributions than the VTM or ProdLDA baselines.

Table 2 represents the perplexity of gensim LDA, Vanilla topic model, ProdLDA, nEATM, EATM 1, and 2. It also includes the performance when using environment-specific information, γ, within our prediction. Here γ represents the article-specific effects on our topic-word distribution β. Notably, when using the article-specific effects for calculating perplexity on a test set consisting of only articles, the perplexity improves. Indicating that the article-specific effects captured in γ uncover information relevant to articles. However, when we use article-specific effects to calculate the perplexity on speeches, the perplexity declines considerably, whereas when we use only β, our perplexity remains stable across test sets, indicating that it captures a robust distribution of topics. See Appendix E for additional experiments and results.

5.2 EATM 1 vs EATM 2

Model criticism aims to identify the limitations of a model in a specific context and suggest areas for improvement (Blei, 2014; Gelman and Shalizi, 2012). Although EATM 1 and 2 exhibit commendable performance compared to other topic model variants, it is crucial to verify the expected behavior of the newly introduced γ parameter. Our evaluation of

Model Perplexity	Articles	Speeches	Tweets
Gensim	9344	3007	3.936×10^{12}
VTM	1345	1461	1584
ProdLDA	2757	2427	2000
nEATM	1762	1853	1868
nEATM + γ	1149	2256	1521
EATM 1	1388	1413	1393
EATM 1 + γ	1133	1526	1243
EATM 2	1293	1341	1351
EATM 2 + γ	1209	1450	1280

Table 2: Model perplexities training on all three sources and testing on an unseen test set from each environment. γ corresponds to article specific effects. ETM 1 and ETM 2 have stable performance across different styles. When we incorporate the γ specific effects the stability dissipates. VTM, ProdLDA, and LDA are less stable. In the case of LDA the drastic drop in perplexity can also be attributed to the fact that speeches are much longer than articles and tweets, indicating the model overfits to speeches.

γ prioritizes sparsity, implying that only a limited subset of features should be significant indicators of the environment. Additionally, we ensure that a given word w that is highly probable in a certain environment e_i and a specific topic k occurs more frequently in documents discussing topic k in an environment e_i than in documents discussing the same topic in a different environment e_j. We find that EATM 2 performs better on both of these criteria, motivating its use. See Appendix F for more experimental details.

6 Discussion

We addressed the problem of modeling data from multiple environments. We developed an environment-adjusted topic model that learns the environment-specific effects on the global topic distribution, resulting in a topic distribution that captures consistent topics across environments. The EATM has stable perplexity across different environments, captures meaningful information in the environment-specific latent variable, and performs better in and out of distribution compared to baseline models. The EATM opens several avenues for future work, such as integrating EATMs with word embeddings (Mikolov et al., 2013) and using the EATM in downstream tasks, such as measuring the causal effect of topics.

7 Reproducibility Statement

We provide all information about data preprocessing, setting hyperparameters, and our datasets in Appendix C. We provide code in an anonymous GitHub repository.

References

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Elliott Ash and Stephen Hansen. Text algorithms in economics. *Annual Review of Economics*, 15:659–688, 2023.

Elliott Ash, Sergio Galletta, Dominik Hangartner, Yotam Margalit, and Matteo Pinna. The effect of fox news on health behavior during covid-19. *Available at SSRN 3636762*, 2020.

Eli Bingham, Jonathan P. Chen, Martin Jankowiak, Fritz Obermeyer, Neeraj Pradhan, Theofanis Karaletsos, Rohit Singh, Paul Szerlip, Paul Horsfall, and Noah D. Goodman. Pyro: Deep Universal Probabilistic Programming. *Journal of Machine Learning Research*, 2018.

David M Blei. Build, compute, critique, repeat: Data analysis with latent variable models. *Annual Review of Statistics and Its Application*, 1:203–232, 2014.

David M Blei and John D Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120, 2006.

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

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Philip J Brown and Jim E Griffin. Inference with normal-gamma prior distributions in regression problems. 2010.

Dallas Card, Amber Boydstun, Justin H Gross, Philip Resnik, and Noah A Smith. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 438–444, 2015.

Bradley P Carlin and Thomas A Louis. Empirical bayes: Past, present and future. *Journal of the American Statistical Association*, 95(452):1286–1289, 2000.

Carlos M Carvalho, Nicholas G Polson, and James G Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.

Bradley Efron. *Large-scale inference: empirical Bayes methods for estimation, testing, and prediction*, volume 1. Cambridge University Press, 2012.

Amir Feder, Katherine A Keith, Emaad Manzoor, Reid Pryzant, Dhanya Sridhar, Zach Wood-Doughty, Jacob Eisenstein, Justin Grimmer, Roi Reichart, Margaret E Roberts, et al. Causal inference in natural language processing: Estimation, prediction, interpretation and beyond. *Transactions of the Association for Computational Linguistics*, 10:1138–1158, 2022.

Andrew Gelman and Cosma Shalizi. Philosophy and the practice of bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 2012.

Matthew Gentzkow, Jesse M Shapiro, and Matt Taddy. Congressional record for the 43rd-114th congresses: Parsed speeches and phrase counts. In *URL: https://data. stanford. edu/congress text*, 2018.

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

Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37:183–233, 1999.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

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

David JC MacKay. Bayesian interpolation. *Neural computation*, 4(3):415–447, 1992.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL `https://proceedings.neurips.cc/paper/2013/hash/9aa42b31882ec039965f3c4923ce901b-Abstract.html`.

David Mimno, Hanna Wallach, Jason Naradowsky, David A Smith, and Andrew McCallum. Polylingual topic models. In *Proceedings of the 2009 conference on empirical methods in natural language processing*, pages 880–889, 2009.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2016.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *Artificial intelligence and statistics*, pages 814–822. PMLR, 2014.

Margaret E Roberts, Brandon M Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G Rand. Structural topic models for open-ended survey responses. *American journal of political science*, 58(4): 1064–1082, 2014.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Towards causal representation learning. *CoRR*, abs/2102.11107, 2021. URL `https://arxiv.org/abs/2102.11107`.

Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models. *arXiv*, March 2017. URL `https://doi.org/10.48550/arXiv.1703.01488`.

Michael E Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of machine learning research*, 1(Jun):211–244, 2001.

Keyon Vafa, Suresh Naidu, and David M Blei. Text-based ideal points. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020.

Appendix A. Graphical Model

We present the graphical model an environment-adjusted topic model (EATM) in Algorithm 1.

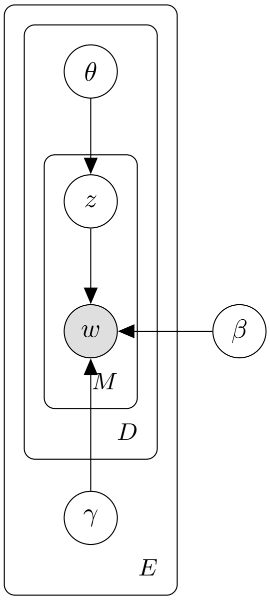

Figure 1: A graphical model for the Environment-adjusted topic model (EATM). In this model, there's an additional E plate denoting the different datasets or environments documents are drawn from. There's also an additional parameter γ which denotes the environment-specific weights for each word in the vocabulary.

Appendix B. Algorithm

We present the full algorithm for training an environment-adjusted topic model (EATM) in Algorithm 1.

Algorithm 1 Environment-adjusted topic model

```
Input: Number of topics $K$, number of words $V$, number of environments $E$
Output: Document intensities $\hat{\theta}$, global topics $\hat{\beta}$, environment-specific effects on global topics $\hat{\gamma}$
Initialize: Variational parameters $\mu_\theta$, $\sigma_\theta^2$, $\mu_\beta$, $\sigma_\beta^2$, $\mu_\gamma$, $\sigma_\gamma^2$ randomly
while the evidence lower bound (ELBO) has not converged do
    sample a document index $d \in \{1, 2, \ldots, D\}$
    For each document get the corresponding environment index, $e \in \{1, 2, \ldots, E\}$
    sample $z_\theta$, $z_\beta$, and $z_\gamma \sim \mathcal{N}(0, I)$          ▷ Sample noise distribution
    Set $\tilde{\theta} = \exp(z_\theta \odot \sigma_\theta + \mu_\theta)$          ▷ Reparameterize
    Set $\tilde{\beta} = \exp(z_\beta \odot \sigma_\beta + \mu_\beta)$          ▷ Reparameterize
    Set $\tilde{\gamma} = \exp(z_\gamma \odot \sigma_\gamma + \mu_\gamma)$          ▷ Reparameterize
    for $v \in \{1, \ldots, V\}$ do
        Set $w_{dv} = \sum_k \tilde{\theta}_{dk}(\tilde{\beta}_{kv} + \tilde{\gamma}_{ekv})$          ▷ Log-likelihood term
    end for
    Set $\log p(w_d|\tilde{\theta}, \tilde{\beta}, \tilde{\gamma}) = \sum_v \log p(w_{dv}|\tilde{\theta}, \tilde{\beta}, \tilde{\gamma})$          ▷ Sum over words
    Compute $\log p(\tilde{\theta}, \tilde{\beta}, \tilde{\gamma})$ and $\log q(\tilde{\theta}, \tilde{\beta}, \tilde{\gamma})$          ▷ Prior and entropy terms
    Set ELBO $= \log p(\tilde{\theta}, \tilde{\beta}, \tilde{\gamma}) + N \cdot \log p(w_d|\tilde{\theta}, \tilde{\beta}, \tilde{\gamma}) - \log q(\tilde{\theta}, \tilde{\beta}, \tilde{\gamma})$
    Compute gradients $\nabla_\phi$ELBO using automatic differentiation
    Update parameters $\phi$
end while
return approximate posterior means $\hat{\theta}$, $\hat{\beta}$, $\hat{\gamma}$ =0
```

Appendix C. Experimental Details

Dataset	**Style**	**Ideology**	**Political advertisements**
Focus of text	US Immigration	Politics	Politics
Environments	{Tweets from US Senators, US Senate speeches, news articles}	{Republican, Democrat} politicians	Channels from {Republican, Democrat} voting regions
Training set size	$4,052$ per environment	$12,941$ per environment	$12,446$ per environment

Table 3: A summary of the datasets we construct for testing topic models across multiple environments.

C.1 Style Dataset

The style dataset consists of 12156 samples, with an even amount of samples from each environment. We constructed a vocabulary of unigrams that occurred in at least 0.6% and in no more than 50% of the documents. We removed cities, states, and the names of politicians in addition to stopwords. For EATM 1, we set λ and τ, parameters used in the horseshoe prior, to be 0.4. For EATM 2, we set the hyperparameters of the gamma distribution, a and b, to be 5.4 and 0.04 respectively. These values were determined by training our model for 50 epochs, taking 2 gradient steps for updating a and b in the empirical bayes method for every 1 step for the rest of the model. This approach helps guarantee that hyperparameter updates are not overshadowed by the updates of the rest of the parameters in the model. We set the number of topics, k, to be 20 for all experiments in this paper.

C.2 Ideological Dataset

We constructed a vocabulary of unigrams that occurred in at least 0.6% and in no more than 40% of the documents. We removed cities, states, and the names of politicians in addition to stopwords. For EATM 1, we set λ and τ, parameters used in the horseshoe prior, to be 0.5. For EATM 2, we set the hyperparameters of the gamma distribution, a and b, to be 4.7 and 0.14 respectively. These values were determined by training our model for 15 epochs, taking 2 gradient steps for updating a and b in the empirical bayes method for every 1 step for the rest of the model.

C.3 Political Ads Dataset

The style dataset consists of 24892 samples, with an even amount of samples from each environment. We constructed a vocabulary of unigrams that occurred in at least 0.6% and in no more than 40% of the documents. We removed cities, states, and the names of politicians in addition to stopwords. For EATM 1, we set λ and τ, parameters used in the horseshoe prior, to be 0.4. For EATM 2, we set the hyperparameters of the gamma distribution, a and b, to be 3.1 and 0.1 respectively. These values were determined by training our model for 10 epochs, taking 2 gradient steps for updating a and b in the empirical bayes method for every 1 step for the rest of the model.

Appendix D. Baselines

We compare both environment-adjusted topic models to the relevant baselines:

- **LDA** - The Latent Dirichlet Allocation (LDA) model is a topic modeling tool that employs a variant of online Variational Bayes inference for learning (Blei et al., 2003, 2017).

- **Vanilla Topic Model** - Vanilla topic model (VTM) represents the base version of our model without any environment-specific variations.

$$\theta_d \sim \mathcal{N}(\cdot, \cdot) \tag{10}$$

$$\beta_k \sim \mathcal{N}(\cdot, \cdot) \tag{11}$$

$$w_{d,i} \sim Cat(\pi(\theta_d)\pi(\beta)) \tag{12}$$

- **Non-sparse EATM** - The non-sparse environment-adjusted (nEATM) represents the EATM, but with a Normal distribution on the γ prior.

- **ProdLDA** - In ProdLDA the distribution over individual words is a product of experts rather than the mixture model used in LDA (Srivastava and Sutton, 2017; Hinton, 2002). We use the standard implementation in Pyro (Bingham et al., 2018).

- **nEATM, EATM 1, and 2 + γ** - We want to evaluate how the performance shifts when including environment-specific information. To do this we sum the environment specific effects from a particular environment, $\gamma_{k,v}$, to $\beta_{k,v}$. For example, in Table 2 EATM 1 + γ represents the perplexity when using $\gamma_{k,v} + \beta_{k,v}$, rather than solely $\beta_{k,v}$, where the $\gamma_{k,v}$ is the learned article specific effects on the global topic-distribution, β.

Appendix E. Additional Experiments and Results

E.1 Style Dataset

Table 4 reflects how EATM 1 learns environment specific effects, and global topics when trained on the style dataset. In the topic related to finance, beta captures words that are appear across environments like 'banks' and 'credit' whereas words like 'regulatory' and 'rules' are predominant in senate speeches and 'fee' and 'fraud' are predominant in articles.

Source	**Top Words**
β	energy, oil, water, jobs, air card, credit, banks, financial, bank
γ: News Articles	canada, disaster, wind, property, construction card, cards, fee, fraud, investment
γ: Senate Speeches	national, infrastructure, country, projects, climate rules, consumers, industry, rates, regulatory
γ: Tweets	epa, climate, roll, environment, coal competition, settlement, consumers, exchange, regulate

Table 4: Top Words by Topics in the style dataset. The words in global topics appear in all environments when discussing a given topic, while the words that receive the top γ values predominately appear in one environment. We observe distinctive word choices in tweets, articles, and senate speeches, reflecting different communication styles.

E.2 Channels Dataset

The channels dataset consists of political advertisements run on TV channels across the United States. We create our two environments by splitting the original dataset and assigning channels from Republican voting regions to one environment, and channels from Democratic voting regions to the other. Table 5 presents a samples from an advertisement from a left and right leaning region respectively. The dataset has 24982 samples with an equal amount from each environment. Table 8 represents perplexity of LDA, VTM, ProdLDA, nEATM,

EATM 1, and 2. EATM 1, and 2 satisfy our desiderata: their predictive performance is consistent across environments, performance declines when using environment-specific effects γ_i from an environment i that differs from the environment of a test set j, and perplexity is better than alternative models.

Source	**Text**
Right (WKRG)	What does Governor Bob Riley call over 70,000 new jobs? A great start. His conservative leadership's given us the lowest unemployment in Alabama history, turning a record deficit into a record surplus. Now Governor Riley has delivered the most significant tax cuts in our history. The people get up every morning and work, they are the ones that allowed us to have the surplus. The only thing I'm saying, they should have some of it back. Governor Bob Riley, honest, conservative leadership.
Left (KSWB)	State budget cuts are crippling my classroom. So I can't believe the Sacramento politicians cut a backroom deal that will give our state's wealthiest corporations a new billion dollar tax giveaway. A new handout that can only mean larger class sizes and even more teacher layoffs. But passing Prop 24 can change all that. Prop 24 repeals the unfair corporate giveaway and puts our priorities first. Vote yes on Prop 24 because it's time to give our schools a break, not the big corporations. their corporate giveaway and puts their priorities first. Vote yes on Prop 24 because it's time to give our schools a break, not the big corporations.

Table 5: An example of advertisements from our dataset. KSWB is a San Diego based news channel, and WKRG is a station licensed to Mobile, Alabama.

Table 6 displays the terms that γ places high density on in each respective environment for EATM 1. We observe that in the γ distribution corresponding to left-leaning effects of the topic of taxation, there is a high density on words such as 'raise' and 'billion' while the γ distribution corresponding to right-leaning effects place high density on terms like 'wasteful'.

Source	**Top Words**
β	tax, taxes, spending, cut, budget
γ: Right	issue, cut, balanced, wasteful, tax
γ: Left	billion, taxes, raise, tax, raising

Table 6: Example of a learned topic using EATM 1 and its cross-channel variation. The γ effects are from right-leaning and left-leaning regions of the United States.

Source	**Top Words**
β	america, veterans, war, proud, iraq, military, troops
γ: Right	terror, liberties, isis, terrorism, freedom, terrorists, defeat
γ: Left	iraq, stay, guard, veterans, soldiers, port, home

Table 7: Example of a learned topic using EATM 2 and its cross-channel variation. The γ effects are from right-leaning and left-leaning regions of the United States.

Model Perplexity	Right	Left
VTM	1301	1380
ProdLDA	1582	1478
nEATM	911	917
nEATM + γ	576	812
EATM 1 + γ	578	668
EATM 1	660	669
EATM 2 + γ	575	698
EATM 2	677	691

Table 8: Perplexity performance across models trained on a dataset of political advertisements from channels in different regions of the America. γ represents right leaning effects.

E.3 Ideological Dataset

Our ideological dataset consists of US political advertisements from the last twenty years. We split the dataset by ideology and have an equal number of advertisements from right-leaning and left-leaning politicians. In the training set, there are 12941 samples from Republicans and 12941 from Democrats. We test on three held-out datasets: one from right-leaning politicians, one from left-leaning politicians, and a third that is an even mixture of advertisements from both left and right-leaning sources. Table 9 represents the perplexity of our baseline models and EATM variants. γ in Table 9 represents the right-leaning ideological effects. We see that EATM 1 and 2 perform significantly better on all test sets. Furthermore, we see that when using right-leaning ideological effects in the perplexity calculation for right-leaning text, we receive better perplexity than using only β; however, when we use right-leaning effects on the left-leaning test set, performance declines considerably. Indicating the information captured in γ is relevant to a specific environment, right-leaning ads, while irrelevant to left-leaning ads. Lastly, when using only the global topic distribution, the perplexity is nearly identical for each test set.

Model Perplexity	Right	Left	Balanced
Gensim	2766	3474	3149
VTM	1239	1231	1250
ProdLDA	1605	1541	1606
nEATM	899	886	902
nEATM + γ	540	704	627
EATM 1	687	686	695
EATM 1 + γ	578	731	661
EATM 2	602	587	600
EATM 2 + γ	569	633	607

Table 9: Perplexity performance across models trained on a dataset of political advertisements from right-leaning and left-leaning politicians. EATM 1 and 2 with γ represent a combination of the learned topic distribution β, where γ indicates the right-leaning deviations on each word distribution of β. Using EATM 1 and 2 with γ on the right-leaning test set improves perplexity. However, when deployed on the left-leaning test set, the perplexity worsens.

Additionally, we qualitatively analyze terms that γ places high density on. Table 10 displays the terms that γ places high density on in each respective environment for EATM 1. We observe that in the γ distribution corresponding to left-leaning effects of the topic of healthcare, there is a high density of words such as 'universal' and 'affordable' while the γ distribution corresponding to right-leaning effects place a high density on terms like 'debt.'

Source	**Top Words**
β	law, public, funding, helping, federal home, choice, war, iraq, military health, budget, debt, cost, costs
γ: Right	sanctuary, cities, control, federal, crimes terrorists, iran, terrorism, deal, isis takeover, debt, health, trillion, bureaucrats
γ: Left	public, helping, rape, safety, funding iraq, weapons, troops, assault, home health, affordable, healthcare, universal, medicaid

Table 10: When trained on the ideology dataset EATM 1 learns interpretable environment specific terms, while simultaneously uncovering meaningful global topics.

Appendix F. EATM 1 vs EATM 2

According to Occam's Razor principle, models with unnecessary complexity should not be preferred over simpler ones (MacKay, 1992). As indicated in Table 11 in the Appendix, EATM 1 is less sparse and exhibits greater uncertainty regarding its parameter values compared to EATM 2. Employing the ARD prior leads to a γ parameter that is not only more sparse but also more effective in capturing environment-specific terms. This is evident from EATM 2's superior performance on both in-distribution and out-of-distribution data. Besides having considerably lower perplexity, nEATM is also less sparse than both models.

Model	**Group**	**Perp.**	**Sparsity**	μ_γ	σ_γ
nEATM	Right	911	3.6%	7.1×10^{-3}	0.4
	Left	917	3.8%	6.7×10^{-3}	0.4
EATM 1	Right	665	41.64%	7.13×10^{-4}	0.2056
	Left	645	42.70%	-2.92×10^{-3}	0.2426
EATM 2	Right	578	79.95%	5.45×10^{-5}	0.02891
	Left	572	79.89%	1.37×10^{-4}	0.02853

Table 11: Model Comparison for Right and Left Leaning Groups. Sparsity is defined as any value less than 0.01.

To ensure that a given word w that is highly probable in a certain environment e_i and a specific topic k occurs more frequently in documents discussing topic k in environment e_i than in documents discussing the same topic in a different environment e_j, we introduce a metric: Count Opposite. It represents the number of words (from the top 10 γ words for each environment and each topic) that have a higher frequency in the test set environment opposite to the one they are associated with. For instance, if γ, in the context of a right-leaning environment, assigns a high probability to the word 'wasteful' occurring in discussions about taxation, this word should appear more frequently in a subset of right-leaning advertisements about taxation than in a subset of left-leaning advertisements on the same topic. Among the words receiving high γ values for a given environment and topic, these words are more likely to occur in the dataset corresponding to the environment represented by γ in EATM 2 than in EATM 1 for the same dataset. Motivating the use of the ARD prior. We find the median Count Opposite of the top 10 words for each topic and γ environment is 1.0 for EATM 2 and 2.0 for EATM 1.