Robust User Representations for Cross Domain Depression Classification using PLDA

Keith Harrigian
Johns Hopkins University
Baltimore, MD
kharrigian@jhu.edu

Abstract

Use of machine learning to infer mental health status based on social media has been a high value proposition explored by researchers in the last decade as a means to passively monitor population-level wellness and personalize individual psychiatric treatment. Unfortunately, recent studies have provided evidence that existing methods for detecting conditions such as depression fail to generalize to new data platforms and populations. In this study, we attempt to improve generalization by learning domain-invariant feature representations for depression classification using Partially-labeled Latent Dirichlet Allocation (PLDA). Experiments using synthetic datasets and two benchmarks from Twitter and Reddit provide an understanding of PLDA's effectiveness in comparison to Latent Dirichlet Allocation (LDA).

1 Introduction

An estimated 8.1% of adults in the United States are currently living with depression, while up to 16.2% of individuals will experience at least one major depressive episode during their lifetime [Kessler et al., 2003, Brody et al., 2018]. Unfortunately, social stigma towards discussing mental health and seemingly insurmountable socioeconomic barriers continue to drive a substantial portion of this population away from receiving adequate care [Gary, 2005]. Accordingly, there has been significant interest by clinicians and social workers alike to develop non-invasive tools that increase equity in mental health care, improve the efficiency of clinical treatment, and effectively monitor populational-level mental wellness at scale [Chen et al., 2019, Galea et al., 2020].

In response to this demand, computational researchers have devoted significant effort in the last decade towards developing systems that measure mental health using non-clinical data sources [De Choudhury et al., 2013, Jaques et al., 2015]. Proposed solutions have drawn insights from mobile activity sensors [Mohr et al., 2017], grade-school essays [Lynn et al., 2018], speech [Cummins et al., 2015], and most prominently, social media [Guntuku et al., 2017, Chancellor and De Choudhury, 2020]. Offering large amounts of user-generated language, social media platforms have provided an opportunity for researchers to translate preexisting psychological theory regarding the relationship between language usage and mental health into the digital world [Ramirez-Esparza et al., 2008].

As is common in computational domains, the primary focus of research has been maximizing performance in well-defined tasks, such as detecting depression in public posts [Lam et al., 2019, Sekulić and Strube, 2020] or estimating suicide risk in online conversations [Song et al., 2020]. To best harness the benefits of contemporary model architectures, researchers have curated large datasets using distantly-supervised data annotation mechanisms (e.g. activity patterns, self-disclosures) that map social media to an individual's mental health status [Coppersmith et al., 2014, Cohan et al., 2018]. Unfortunately, recent studies have provided evidence that computational methods based on these datasets fail to generalize to new data platforms and populations [Ernala et al., 2019, Harrigian et al., 2020].

In this study, we evaluate the use of Partially-labeled Latent Dirichlet Allocation (PLDA) [Ramage et al., 2011] for performing unsupervised domain adaptation in the context of a depression inference task. Specifically, we hypothesize PLDA's combined use of label-specific and label-invariant topics will encourage learning of user representations that are robust across domains. We explore this hypothesis first through the lens of synthetic data experiments, showing that PLDA outperforms vanilla Latent Dirichlet Allocation (LDA) [Blei et al., 2003] under certain data-generating regimes. Thereafter, we systematically compare domain-transfer abilities of LDA and PLDA using two real social media datasets from Twitter and Reddit.

Ethical Considerations. All data was collected and analyzed in accordance to protocols for mental health research enumerated in Benton et al. [2017]. Due to the sensitive nature of mental health data, we are unable to share any raw and/or processed forms of our datasets. However, we provide code for replicating our experiments and instructions for accessing datasets from their respective providers[1].

2 Related Work

The process of training statistical models that generalize from one distribution to another is often referred to as *domain adaptation* [Ben-David et al., 2006, Pan and Yang, 2009, Kouw and Loog, 2019]. Historically, research regarding generalization in models of mental health based on social media has been quite limited. De Choudhury and De [2014] and Ireland and Iserman [2018] explored language usage amongst individuals with depression as a function of the subreddits they posted comments within. Pirina and Çöltekin [2018] evaluated feasibility of detecting depression in individuals as a function of the manner in which control groups were defined, while Wolohan et al. [2018] compared depression classifier performance in the absence of data from explicitly-defined mental health forums.

Shen et al. [2018] was the first study to attempt transfer across social media platforms, ultimately suggesting that data from Twitter could be used to inform the prior for models deployed on Sina Weibo. However, later work from Ernala et al. [2019] and Harrigian et al. [2020] contradicted the optimism of these findings by providing evidence that models of schizophrenia and depression, respectively, transferred poorly between social media platforms as a result of underlying dataset biases. To date, no study has explicitly attempted to reduce error when transferring mental health classifiers from one platform or population to another.

Fortunately, the natural language processing (NLP) community has a rich foundation of domain adaptation research for non-mental-health applications [Blitzer et al., 2007, Daumé III, 2007, Dredze and Crammer, 2008, Daumé III, 2009]. Broadly, adaptation techniques can be categorized as being either feature-based — e.g. neural pretraining [Yang and Eisenstein, 2015, Dou et al., 2019], subspace mapping [Fernando et al., 2013, Sun and Saenko, 2015], invariant feature selection [Zhao et al., 2019, Subbaswamy et al., 2019] — or model-based — e.g. instance weighting [Jiang and Zhai, 2007, Wang et al., 2017], adversarial learning [Tzeng et al., 2017, Meng et al., 2018]. Unfortunately, many existing methods are poorly suited for application in the mental-health domain, either due to underlying hyperparameter sensitivity [Plank et al., 2014, Xia et al., 2018] or transformations that inhibit the interpretability required for health-related machine learning [Saria and Subbaswamy, 2019].

While they are less common within the NLP community, Bayesian methods have been used successfully in certain language-focused adaptation settings. For instance, Finkel and Manning [2009] presented a supervised adaptation method that makes use of a hierarchical Bayesian prior to perform named entity recognition and dependency parsing. Daumé III [2009] later proposed a Bayesian latent hierarchy model for inferring sentiment of reviews amongst different types of retail products. In both studies, the Bayesian perspective provided a natural fit for sharing covariance between domains.

Topic models have been the most widely used Bayesian domain adaptation approach used by the NLP community [Blei and Lafferty, 2009]. For example, Chen et al. [2012] used topics learned using Google search results to contextualize latent topics in microblog messages (i.e. Twitter, Sina Weibo). Later, Yang et al. [2019] presented a topic model that jointly aligns concepts across multiple language domains and thus improves downstream document classification. Serving as the foundation for our work, Bao et al. [2013] and Jing et al. [2018] constructed partially-supervised topic models that learn to distinguish between domain-specific and domain-invariant semantics. In doing so, they provide a

[1] https://github.com/kharrigian/topic-model-domain-adaptation

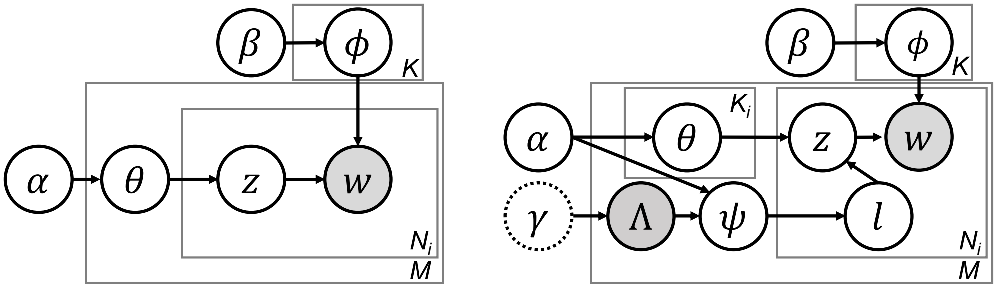

Figure 1: Comparison of LDA (left) and PLDA (right) sampling models. Shaded variables are observed. If we assume only latent topics (i.e. no label-specific topics) are present, the PLDA model reduces to LDA. Since labels Λ are observed, we can ignore the label prior γ.

mechanism for projecting user-generated text into a robust, low-dimensional feature space that can be used for domain adaptation [Baktashmotlagh et al., 2013]. We expand upon this idea further in §3.1.

Importantly, topic models are already familiar to those working on mental-health related applications of NLP and thus present an increased chance of adoption. For instance, Resnik et al. [2013, 2015a,b] demonstrated that LDA and its variants could be used to automatically decipher themes discussed by those with heightened levels of depression, neuroticism, and signs of post traumatic stress disorder. Researchers have used topic models for additional mental-health-related tasks such as identifying suicidal ideation [Huang et al., 2015, 2017] and monitoring wellness during the COVID-19 pandemic [Biester et al., 2020].

3 Methods

3.1 Topic Modeling

Topic models are a Bayesian generative extension to the matrix factorization procedure Latent Semantic Analysis, offering a mechanism for representing the semantic distribution of a document in a low-dimensional hyperspace [Blei et al., 2003, Blei and Lafferty, 2009]. In this study, we compare two formulations of the topic model — "vanilla" Latent Dirichlet Allocation (LDA) [Blei et al., 2003] and Partially-labeled Latent Dirichlet Allocation (PLDA) [Ramage et al., 2010, 2011].

Throughout this section and the remainder of the paper, we leverage the following notation. Let S and T represent the source and target data domains, respectively. We consider a dataset $\mathcal{D}$ of M documents (i.e. users), where each document $i = 1, ..., M$ has N_i words sampled from a vocabulary $\mathcal{V}$ of size V. We parameterize each topic model to have K topics, where each topic represents a probability distribution over the vocabulary. We assume each document i has a topic distribution θ_i of dimensionality K, where θ_{ik} denotes the proportion of words in the document sampled from latent topic z_k. We assume each topic distribution ϕ_k is a V-dimensional probability distribution over the vocabulary for $k = 1, ..., K$. We let w_{ij} represent the j-th word in document i.

LDA. "Vanilla" Latent Dirichlet Allocation was first presented by Blei et al. [2003] and has since been cited in over 30-thousand studies across a wide array of disciplines. We present a plate diagram for the sampling model in Figure 1 (left) and verbalize the sampling model below. We assume a Dirichlet prior for θ, parameterized by K-dimensional α, and a Dirichlet prior for ϕ, parameterized by V-dimensional β.

1. For $i = 1, ..., M$: $\theta_i \sim$ Dirichlet(α)
2. For $k = 1, ..., K$: $\phi_k \sim$ Dirichlet(β)
3. For $i = 1, ..., M$, For $j = 1, ..., N_i$:
    a) Sample Latent Topic $z_{ij} \sim$ Multinomial(θ_i)
    b) Sample Word $w_{ij} \sim$ Multinomial$(\phi_{z_{ij}})$

We are interested in learning the document-topic distributions θ, the topic-word distributions ϕ, and the word-topic assignments z. Though full conditional derivations can be derived for θ and ϕ, we note that by the structure of the generative model, both parameters can be computed using the topic indices z alone per Equation 1 [Darling, 2011]. Letting $n(i, z_k)$ represent the number of words in document i assigned to topic z_k and $n(z_k, v)$ represent the number of times word v was assigned to topic z_k across the entire corpus, we can derive θ and ϕ.

$$\theta_{i,k} = \frac{n(i, z_k) + \alpha}{\sum_{k=1}^{K} n(i, z_k) + \alpha}, \phi_{k,v} = \frac{n(z_k, v) + \beta}{\sum_{v=1}^{V} n(z_k, v) + \beta} \tag{1}$$

To implement a Gibbs sampler for z, we need the full conditional posterior for $z_{i,j}$ denoted in Equation 2.

$$p(z_{(i,j)} \mid z_{\neg(i,j)}, \mathcal{D}, \alpha, \beta) = \frac{p(z_{(i,j)}, z_{\neg(i,j)}, \mathcal{D} \mid \alpha, \beta)}{p(z_{\neg(i,j)}, \mathcal{D} \mid \alpha, \beta)} \propto p(z, \mathcal{D} \mid \alpha, \beta) \tag{2}$$

As shown in Carpenter [2010] and partially replicated below, we can derive this distribution by marginalizing over θ and ϕ.

$$\begin{aligned} p(z, \mathcal{D} \mid \alpha, \beta) &= \int \int p(z, \mathcal{D}, \theta, \phi \mid \alpha, \beta) d\theta d\phi \\ &= \int \int p(\phi \mid \beta) p(\theta \mid \alpha) p(z \mid \theta) p(\mathcal{D} \mid \phi_z) d\theta d\phi \\ &= \int \prod_{i=1}^{M} p(z_i \mid \theta_i) p(\theta_i \mid \alpha) d\theta \times \int \prod_{k=1}^{K} p(\phi_k \mid \beta) \prod_{i=1}^{M} \prod_{j=1}^{N_i} p(w_{i,j} \mid \phi_{z_{i,j}}) d\phi \\ &= \prod_{i=1}^{M} \frac{B(\sum_k n(i, z_k) + \alpha_k)}{B(\alpha)} \prod_{k=1}^{K} \frac{B(\sum_v n(z_k, v) + \beta_v)}{B(\beta)} \end{aligned} \tag{3}$$

where B is the multivariate Beta function. Note that the last step falls out naturally as a result of the Dirichlet prior being conjugate for the Multinomial distribution within each integral. For sampling $z_{(i,j)}$ (i.e. word j in document i with value $v = w_{i,j}$), we can use the above result to get a probability distribution over topics.

$$\begin{aligned} p(z_{(i,j)} \mid z_{\neg(i,j)}, \mathcal{D}, \alpha, \beta) &= \frac{p(\mathcal{D}, z)}{p(\mathcal{D}, z_{\neg(i,j)})} \\ \Rightarrow p(z_{(i,j)} = z_k \mid \cdot) &\propto \prod_{i=1}^{M} \frac{B(n(i, z_k) + \alpha_k)}{B(n_{\neg(i,j)}(i, z_k) + \alpha_k)} \prod_{k=1}^{K} \frac{B(n(z_k, v) + \beta_v)}{B(n_{\neg(i,j)}(z_k, v) + \beta_v)} \\ &\propto (n_{\neg(i,j)}(i, z_k) + \alpha_k) \frac{n_{\neg(i,j)}(z_k, v) + \beta_v}{\sum_{v\prime=1}^{V} n_{\neg(i,j)}(z_k, v\prime) + \beta_{v\prime}} \end{aligned} \tag{4}$$

We perform the computation above for each $k = 1, ..., K$ and then normalize to get a probability distribution over topics, from which we sample $z_{(i,j)}$ and update all existing counts. For inferring the topic distribution of a new document not seen during training, we can use the same Gibbs sampler; however, the topic assignments based on the training set remain fixed and updates are only made to the new document's topic-assignments.[2]

PLDA. While Latent Dirichlet Allocation has proven to be a powerful tool for summarizing large corpora of text [Park and Conway, 2017, Jelodar et al., 2019] and representing documents in a low-dimensional space [Henderson and Eliassi-Rad, 2009, Resnik et al., 2013], its unsupervised nature limits the degree to which researchers can take advantage of rich metadata that is often available with language data. Partially-labeled Latent Dirichlet Allocation (PLDA) was proposed by Ramage et al. [2010, 2011] as an extension of LDA that could use annotations to inform, but not fully restrict, the learning of topic distributions in a corpus. PLDA is structured very similarly to LDA, with the primary difference being that documents are generated from a mixture of latent topics general to the entire corpora and a subset of latent topics specific to the labels associated with each document.

We present the plate diagram for PLDA in Figure 1 (right) and describe the sampling model below. Note that we introduce the following additional notation: a set of labels in a corpora is denoted by

[2] To reduce computational expense, we use parallelized versions of LDA and PLDA implemented in the `tomotopy` Python package.

$\mathcal{L}$ with cardinality L, labels for document i are denoted by Λ_i, K_ℓ is the number of latent topics associated with label $l \in \mathcal{L}$, and ψ_i is the distribution of labels within document i. By convention, we have $\sum_{\ell \in \mathcal{L}} K_\ell = K$. In addition to the observed labels, we append a single "latent" label that is present for all documents.

1. For $i = 1, ..., M$, For $\ell \in \Lambda_i$: $\theta_{i,\ell} \sim \text{Dirichlet}(\alpha)$
2. For $\ell = 1, ..., L$, For $k = 1, ..., K_\ell$: $\phi_{\ell,k} \sim \text{Dirichlet}(\beta)$
3. For $i = 1, ..., M$: $\psi_i \sim \text{Dirichlet}(\vec{\alpha_L})$ [We define $\vec{\alpha_{L,\ell}} = \alpha K_\ell$ to simplify derivations]
4. For $i = 1, .., M$, For $j = 1, ..., N_i$:
   a) $\ell_{i,j} \sim \text{Multinomial}(\phi_i)$
   b) $z_{i,j} \sim \text{Multinomial}(\theta_{i,\ell_{i,j}})$
   c) $w_{i,j} \sim \text{Multinomial}(\phi_{z_{i,j}})$

We can re-index θ and ϕ to align closely with derivations for vanilla LDA's Gibbs sampler. That is, since each topic is only associated with one label by assumption, we can compute the posterior $p(z_{(i,j)} = k, l_{(i,j)} = q \mid \cdot)$ as a function of topic-assignments $z_{i,j}$ alone. The only difference in the MCMC procedure is that when we construct the probability distribution for sampling $z_{i,j}$, we are constrained to topic indices that are elements of Λ_i. Per Ramage et al. [2011], we have:

$$p(z_{(i,j)} = z_k, l_{(i,j)} = q \mid z_{\neg(i,j)}, l_{\neg(i,j)}, w_{(i,j)} = v, \alpha, \beta) \propto$$
$$I[q \in \Lambda_i; k \in 1, ..., K_q] \times (n_{\neg(i,j)}(i, z_k) + \alpha) \frac{\sum_{i=1}^{M} n_{\neg(i,j)}(i, z_k, v) + \beta}{\sum_{i=1}^{M} \sum_{v\prime=1}^{V} n_{\neg(i,j)}(i, z_k, v\prime) + \beta} \quad (5)$$

As with LDA, we compute this probability for each topic $k = 1, ..., K$ and then normalize to get a discrete sampling distribution. With this distribution formed, we sample and assign a new topic $z_{i,j}$ to word $w_{i,j}$ and update the global counts used for computing θ and ϕ.

3.2 Mental Health Inference

The primary aim of this pilot study is to evaluate user-representations derived using LDA in comparison to PLDA. Accordingly, we opt to use a non-Bayesian classifier for the downstream modeling task. We feed the user-representations inferred from each topic model into an l_2-regularized Logistic Regression classifier whose parameters are learned using the Scikit-learn implementation of Limited-memory BFGS [Liu and Nocedal, 1989, Pedregosa et al., 2011].

For PLDA, we define $\mathcal{L} = \{S, T, \text{Latent}\}$. Intuitively, this means that our model will learn topics that are specific to each domain, in addition to general latent topics that are present in each domain. Using this setup, we do not require any annotations in the target domain T. However, since PLDA assumes that each topic is only associated with a single label, we are restricted to representing the documents at inference time in the target domain using the general latent topics only. For consistency between LDA and PLDA, we re-normalize each document's distribution using an l_2 norm.

4 Data

4.1 Synthetic

Formulation of realistic synthetic text data is notoriously difficult given its high-dimensional, structured nature [Albuquerque et al., 2011, Shi et al., 2019]. For our purposes, we consider a vastly simplified data generating process that will still allow us to estimate LDA and PLDA's ability to represent documents from multiple domains. We describe the full sampling procedure with expected inputs and resulting outputs in Algorithm 1. We assume the presence of three latent topics; this allows us to consider the situation in which each domain has one unique topic *and* one general latent topic.

4.2 Observational

We consider two real-world social media datasets that map an individual's language usage to mental health status. The 2015 `CLPsych` Shared Task Twitter dataset was introduced by Coppersmith et al. [2015]. It contains up to 3000 tweets from 477 users who self-disclosed a depression diagnosis

Algorithm 1: Synthetic Text Generation

```
Input  : Sample Size $N$, Dirichlet Scaling $\sigma_0$, Domain Balance $p_D(T)$, Word Count Mean $\gamma$,
         Vocabulary Size $V$, Topic Concentrations $\theta[2 \times 3]$, Linear Coefficients $W[2 \times 3]$
Output : Document Term Matrix $X$, Depression labels $y$, Domain labels $D$
$\beta \leftarrow 1/V$; $\theta \leftarrow \sigma_0\theta$;
$\phi \sim$ Dirichlet($\beta$);
Initialize $X_{\text{latent}}[N \times 3], X[N \times V], D[N], p_y[N], y[N]$;
for $n = 1, ..., N$ do
    $D[n] \leftarrow T$ with probability $p_D(T)$, else $S$;
    $X_{\text{latent}}[n] \leftarrow$ Dirichlet($\theta_{D[n]}$);
    $n_d \leftarrow$ Poisson($\gamma$);
    for $i = 1, ..., n_d$ do
        $z_{n,i} \sim$ Multinomial($X_{\text{latent}}[n]$);
        $w_{n,i} \sim$ Multinomial($\phi_{z_{n,i}}$);
        $X[n, w_{n,i}]$ += 1
    end
end
for $d \in \{S, T\}$ do
    Standardize $X_{\text{latent}}[D == d]$;
    $p_y[D == d] \leftarrow$ Logit($W_d \cdot X_{\text{latent}}[D == d]$);
    $y[D == d] \leftarrow$ Depression with probability $p_y[D == d]$ else Control;
end
```

between Jan. 2011 and Dec. 2013, in addition to an equal number of age- and gender- matched controls with no known depression diagnosis. The Topic-Restricted Reddit dataset was first introduced by Wolohan et al. [2018] and then replicated by Harrigian et al. [2020] in a recent exploration of mental health model generalization abilities. It contains the entire comment history from 6,859 users who posted in the r/depression subreddit during October 2019, in addition to all historical comments made by an equal number of users who posted in the r/AskReddit subreddit (but not r/depression) during the same time period. The earliest comment is from 2010, but the majority of data is from after 2016. Accordingly, these datasets present variations in time, social media platform, and annotation mechanism.

To form "documents" for modeling, all posts made by an individual user are concatenated together and then tokenized using a modified version of Twokenizer [O'Connor et al., 2010]. Per the recommendations of Schofield et al. [2017], we remove the 250 most frequently occurring tokens from each corpus. We also ignore tokens used by less than 10 users throughout the training corpus.

5 Results

We begin with an analysis of our modeling procedure applied to synthetically-generated data, focusing primarily on understanding how the underlying data generation procedure (DGP) affects topic learning and downstream inference. Thereafter, we explore the effectiveness of LDA vs. PLDA at representing individuals in our real-world datasets. To accommodate space constraints, we maintain additional visualizations for each experiment discussed below in our digital supplement.[1].

5.1 Synthetic Data Experiments

For this pilot study, we focus primarily on the scenario in which there exists a covariate shift between domains (e.g. $p_S(y \mid x) = p_T(y \mid x)$, but $p_S(x) \neq p_T(x)$) [Sugiyama et al., 2007, Kouw and Loog, 2019]. In our case, this corresponds to each domain having different underlying topic-distributions, but fixed relationships between each topic and mental health status.

To simulate this scenario, we programmatically generate 10 3-dimensional binary classification datasets [Guyon, 2003, Pedregosa et al., 2011]. For each dataset, we fit a Logistic Regression classifier and cache the learned coefficient weights to use as an input W to our data generating process (i.e. Algorithm 1). We consider the following template for θ: $[[x_S, 1e-5, 1], [1e-5, x_T, 1]]$, where

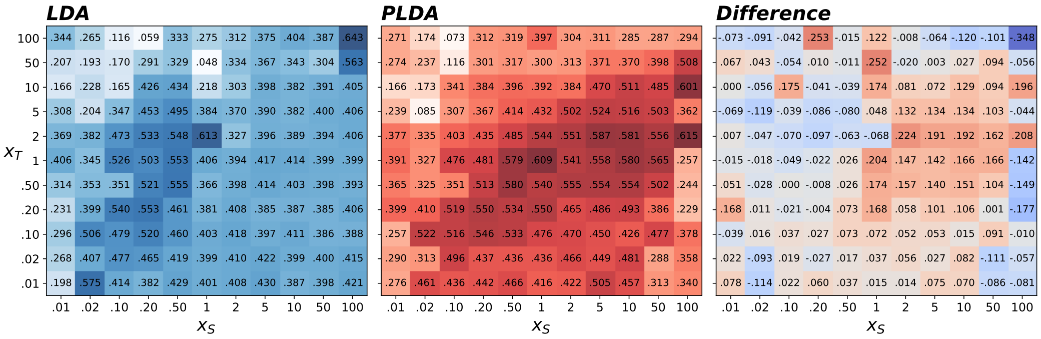

Figure 2: Synthetic performance (held-out F1) as a function of the underlying topic distribution. Right-most subplot shows the average within-trial F1 difference between topic modeling methods. PLDA is generally most effective (in comparison to LDA) when there exists a substantial proportion of source-specific language.

manipulating x_S and x_T changes the ratio of general to domain-specific topics for samples from the source and target domains, respectively.[3] For each coefficient weight W, we run simulations over the Cartesian product $x_S \times x_T$, where $x_d = \{0.01, 0.02, 0.1, 0.2, 0.5, 1, 2, 5, 10, 100\}$ for $d \in \{S, T\}$.

We present results under each combination of $\langle x_S, x_T \rangle$ in Figure 2, fixing $N = 1000$, $\sigma_0 = 1$, $p_D(T) = .5$, $\gamma = 20$, $V = 100$, and $\beta = 1/V$. On average PLDA outperforms LDA in the presence of underlying topic shifts. The most substantial improvements occur when x_S is large, suggesting that PLDA is most useful when selection bias results in the source domain having a high-proportion of domain-specific language. PLDA is less useful in the presence of source-specific noise when the target domain also has a significant presence of domain-specific language. Nonetheless, these trends are not fully consistent from one data generation parameterization to the next; results may become clearer by using a more finely-grained parameter search space with a higher number of simulations. Due to computational limitations, we leave exploration of this search space and interactions with V, β, and $p_D(T)$ for future work.

5.2 Observational Data Experiments

We begin our observational data experiments by downsampling the `Topic-Restricted` dataset to the size of the `CLPsych` dataset. We then establish train/development/test splits of size 518/124/300 users, respectively.[4] To most fairly compare LDA and PLDA, we first conduct a series of experiments that estimate the sensitivity of user-representation quality under perturbations to various topic model training parameters — α, β, and K. For each experiment, we run LDA and PLDA's Gibbs samplers for 1000 iterations. Per the recommendation of Nguyen et al. [2014], we average each document's inferred topic distributions over samples taken every 20th iteration after a burn-in period of 250 iterations to form a single representation for passing to the mental health classifier. For completeness, we perform each experiment under our two possible transfer scenarios: `CLPsych` $\rightarrow$ `Topic-Restricted` and `Topic-Restricted` $\rightarrow$ `CLPsych`. Ideally, we would select hyperparameters α, β, and K using a joint grid search and K-fold cross-validation; unfortunately, we remain constrained by computational expense and leave additional optimization for future work.

Prior Sensitivity. We vary α and β over all combinations from the following set: $\{0.001, 0.01, 0.1, 1, 5, 10, 100\}$ and examine F1 score achieved in the target domain's development set. For both datasets and models, we find downstream performance to be more sensitive to changes to β than changes to α. As with our synthetic data experiments, we do not observe a clear gradient in performance as we vary the prior hyperparameters. Ultimately, LDA achieves a maximum F1 of 0.587 in the `CLPsych` dataset and 0.737 in the `Topic-Restricted` dataset with (α, β) set to (100,10) and (100, 0.01), respectively. PLDA achieves a maximum F1 of 0.645 in the `CLPsych` dataset and 0.728 in the `Topic-Restricted` dataset with (α, β) set to (5,5) and (1, 0.01), respectively.

[3] For example $x_S = 0.1$ means there will be 10 times the prevalence of the general latent topic relative to the source-specific topic in samples from the source domain.

[4] We refrain from evaluating on the test set for this pilot analysis, so as not to bias results in ongoing research.

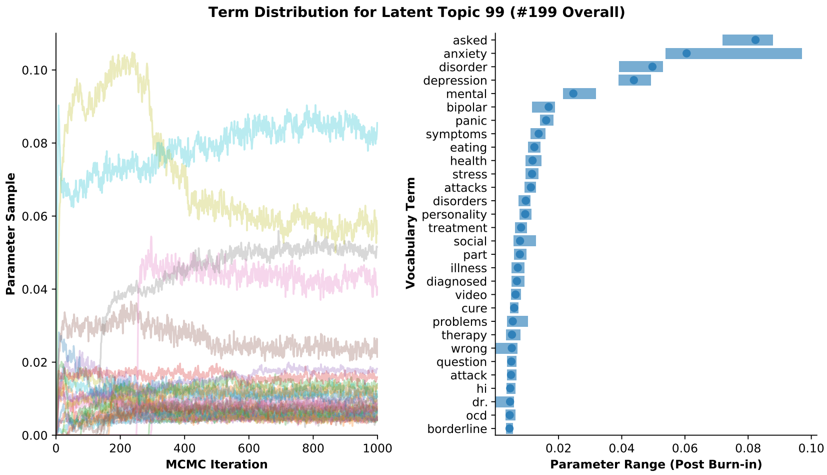

Figure 3: Trace plot and most associated terms for one latent topic from the "optimized" PLDA model. Chains achieve reasonable mixing after the burn-in period despite high dimensionality.

Topic Sensitivity. Next, we explore performance under different parameterizations of K. We consider $K \in \{10, 20, ..., 100, 150\}$ for both LDA and PLDA models, and $K_S = K_T \in \{10, 20, ..., 50\}$ for PLDA. We fix α and β each at 0.1 *a priori*. For both datasets and models, held-out performance tends to vary non-predictably as the number of topics is varied. For both datasets, we find downstream performance using representations from PLDA either outperforms or is equivalent to performance using LDA-based representations, regardless of K. For the `Topic-Restricted` dataset, LDA-based representations achieve a maximum F1 of 0.745 ($K = 80$), while PLDA-based representations achieve a maximum F1 of 0.764 ($K = 150$, $K_S = K_T = 50$). For the `CLPsych` dataset, LDA-based representations achieve a maximum F1 of 0.612 ($K = 60$), while PLDA-based representations also achieve a maximum F1 of 0.612 ($K = 30$, $K_S = K_T = 10$).

Qualitative Evaluation. As showcased in Figure 3, we note that LDA and PLDA parameters achieve reasonably good mixing over the MCMC sampling procedure. Moreover, the model reasonably identifies domain-specific, noisy topics that align with prior analysis from Harrigian et al. [2020]. The topics that have the highest influence in downstream mental health inference are those with a high proportion of health-related and interpersonal terms.

6 Discussion

In this paper, we evaluated the effectiveness of using Partially-labeled Latent Dirichlet Allocation to construct user-representations for cross-domain mental health status inference. Specifically, we showed that PLDA is able to separate domain-specific topics from domain-invariant topics and thus promote generalization under certain regimes. However, both LDA and PLDA's non-trivial sensitivity to hyperparameters makes drawing conclusions difficult.

We would be negligent not to recognize limitations of the current study. First, we have only considered two real-world datasets as part of our pilot analysis. Future work should explore the robustness of PLDA within several more datasets, ideally from a more diverse variety of social media platforms. Additionally, we have only explored transfer between datasets with a presence of depressed individuals that is non-representative of the general population (e.g. 50/50 group balance). Moving forward, researchers should consider using true random samples of data from the desired target domain to use for learning topic representations. Finally, as mentioned in §5, we were only able to explore a limited subset of hyperparameter choices and measure performance using a single data split, limiting statistical power of our analysis.

References

Ronald C Kessler, Patricia Berglund, Olga Demler, Robert Jin, Doreen Koretz, Kathleen R Merikangas, A John Rush, Ellen E Walters, and Philip S Wang. The epidemiology of major depressive disorder: results from the national comorbidity survey replication (ncs-r). *Jama*, 289(23):3095–3105, 2003.

Debra J Brody, Laura A Pratt, and Jeffery P Hughes. *Prevalence of depression among adults aged 20 and over: United States, 2013-2016*. US Department of Health and Human Services, Centers for Disease Control and Prevention, 2018.

Faye A Gary. Stigma: Barrier to mental health care among ethnic minorities. *Issues in mental health nursing*, 26(10):979–999, 2005.

Irene Y Chen, Peter Szolovits, and Marzyeh Ghassemi. Can ai help reduce disparities in general medical and mental health care? *AMA journal of ethics*, 21(2):167–179, 2019.

Sandro Galea, Raina M Merchant, and Nicole Lurie. The mental health consequences of covid-19 and physical distancing: The need for prevention and early intervention. *JAMA internal medicine*, 180(6):817–818, 2020.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. Predicting depression via social media. *Icwsm*, 13:1–10, 2013.

Natasha Jaques, Sara Taylor, Asaph Azaria, Asma Ghandeharioun, Akane Sano, and Rosalind Picard. Predicting students' happiness from physiology, phone, mobility, and behavioral data. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, pages 222–228. IEEE, 2015.

David C Mohr, Mi Zhang, and Stephen M Schueller. Personal sensing: understanding mental health using ubiquitous sensors and machine learning. *Annual review of clinical psychology*, 13:23–47, 2017.

Veronica Lynn, Alissa Goodman, Kate Niederhoffer, Kate Loveys, Philip Resnik, and H Andrew Schwartz. Clpsych 2018 shared task: Predicting current and future psychological health from childhood essays. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 37–46, 2018.

Nicholas Cummins, Stefan Scherer, Jarek Krajewski, Sebastian Schnieder, Julien Epps, and Thomas F Quatieri. A review of depression and suicide risk assessment using speech analysis. *Speech Communication*, 71:10–49, 2015.

Sharath Chandra Guntuku, David B Yaden, Margaret L Kern, Lyle H Ungar, and Johannes C Eichstaedt. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences*, 18:43–49, 2017.

Stevie Chancellor and Munmun De Choudhury. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ digital medicine*, 3(1):1–11, 2020.

Nairan Ramirez-Esparza, Cindy K Chung, Ewa Kacewicz, and James W Pennebaker. The psychology of word use in depression forums in english and in spanish: Texting two text analytic approaches. In *ICWSM*, 2008.

Genevieve Lam, Huang Dongyan, and Weisi Lin. Context-aware deep learning for multi-modal depression detection. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3946–3950. IEEE, 2019.

Ivan Sekulić and Michael Strube. Adapting deep learning methods for mental health prediction on social media. *arXiv preprint arXiv:2003.07634*, 2020.

Xingyi Song, Johnny Downs, Sumithra Velupillai, Rachel Holden, Maxim Kikoler, Kalina Bontcheva, Rina Dutta, and Angus Roberts. Using deep neural networks with intra-and inter-sentence context to classify suicidal behaviour. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1303–1310, 2020.

Glen Coppersmith, Mark Dredze, and Craig Harman. Quantifying mental health signals in twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality*, pages 51–60, 2014.

Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. Smhd: a large-scale resource for exploring online language usage for multiple mental health conditions. *arXiv preprint arXiv:1806.05258*, 2018.

Sindhu Kiranmai Ernala, Michael L Birnbaum, Kristin A Candan, Asra F Rizvi, William A Sterling, John M Kane, and Munmun De Choudhury. Methodological gaps in predicting mental health states from social media: Triangulating diagnostic signals. In *Proceedings of the 2019 chi conference on human factors in computing systems*, pages 1–16, 2019.

Keith Harrigian, Carlos Aguirre, and Mark Dredze. Do models of mental health based on social media data generalize? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 3774–3788, 2020.

Daniel Ramage, Christopher D Manning, and Susan Dumais. Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 457–465, 2011.

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

Adrian Benton, Glen Coppersmith, and Mark Dredze. Ethical research protocols for social media health research. In *Proceedings of the First ACL Workshop on Ethics in Natural Language Processing*, pages 94–102, 2017.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19:137–144, 2006.

Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

Wouter M Kouw and Marco Loog. A review of single-source unsupervised domain adaptation. *arXiv preprint arXiv:1901.05335*, 2019.

Munmun De Choudhury and Sushovan De. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Eighth international AAAI conference on weblogs and social media*, 2014.

Molly Ireland and Micah Iserman. Within and between-person differences in language used across anxiety support and neutral reddit communities. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, pages 182–193, 2018.

Inna Pirina and Çağrı Çöltekin. Identifying depression on reddit: The effect of training data. In *Proceedings of the 2018 EMNLP Workshop SMM4H: The 3rd Social Media Mining for Health Applications Workshop & Shared Task*, pages 9–12, 2018.

JT Wolohan, Misato Hiraga, Atreyee Mukherjee, Zeeshan Ali Sayyed, and Matthew Millard. Detecting linguistic traces of depression in topic-restricted text: Attending to self-stigmatized depression with nlp. In *Proceedings of the First International Workshop on Language Cognition and Computational Models*, pages 11–21, 2018.

Tiancheng Shen, Jia Jia, Guangyao Shen, Fuli Feng, Xiangnan He, Huanbo Luan, Jie Tang, Thanassis Tiropanis, Tat Seng Chua, and Wendy Hall. Cross-domain depression detection via harvesting social media. International Joint Conferences on Artificial Intelligence, 2018.

John Blitzer, Mark Dredze, and Fernando Pereira. Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 440–447, 2007.

Hal Daumé III. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic, June 2007. Association for Computational Linguistics. URL `https://www.aclweb.org/anthology/P07-1033`.

Mark Dredze and Koby Crammer. Online methods for multi-domain learning and adaptation. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 689–697, 2008.

Hal Daumé III. Bayesian multitask learning with latent hierarchies. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 135–142, 2009.

Yi Yang and Jacob Eisenstein. Unsupervised multi-domain adaptation with feature embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 672–682, 2015.

Zi-Yi Dou, Junjie Hu, Antonios Anastasopoulos, and Graham Neubig. Unsupervised domain adaptation for neural machine translation with domain-aware feature embeddings. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1417–1422, 2019.

Basura Fernando, Amaury Habrard, Marc Sebban, and Tinne Tuytelaars. Unsupervised visual domain adaptation using subspace alignment. In *Proceedings of the IEEE international conference on computer vision*, pages 2960–2967, 2013.

Baochen Sun and Kate Saenko. Subspace distribution alignment for unsupervised domain adaptation. In *BMVC*, volume 4, pages 24–1, 2015.

Han Zhao, Remi Tachet des Combes, Kun Zhang, and Geoffrey J Gordon. On learning invariant representation for domain adaptation. *arXiv preprint arXiv:1901.09453*, 2019.

Adarsh Subbaswamy, Peter Schulam, and Suchi Saria. Preventing failures due to dataset shift: Learning predictive models that transport. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3118–3127. PMLR, 2019.

Jing Jiang and ChengXiang Zhai. Instance weighting for domain adaptation in nlp. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 264–271, 2007.

Rui Wang, Masao Utiyama, Lemao Liu, Kehai Chen, and Eiichiro Sumita. Instance weighting for neural machine translation domain adaptation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1482–1488, 2017.

Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017.

Zhong Meng, Jinyu Li, Zhuo Chen, Yang Zhao, Vadim Mazalov, Yifan Gang, and Biing-Hwang Juang. Speaker-invariant training via adversarial learning. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5969–5973. IEEE, 2018.

Barbara Plank, Anders Johannsen, and Anders Søgaard. Importance weighting and unsupervised domain adaptation of pos taggers: a negative result. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 968–973, 2014.

Rui Xia, Zhenchun Pan, and Feng Xu. Instance weighting for domain adaptation via trading off sample selection bias and variance. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4489–4495. AAAI Press, 2018.

Suchi Saria and Adarsh Subbaswamy. Tutorial: safe and reliable machine learning. *arXiv preprint arXiv:1904.07204*, 2019.

Jenny Rose Finkel and Christopher D Manning. Hierarchical bayesian domain adaptation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 602–610, 2009.

David M Blei and John D Lafferty. Topic models. *Text mining: classification, clustering, and applications*, 10(71):34, 2009.

Yan Chen, Zhoujun Li, Liqiang Nie, Xia Hu, Xiangyu Wang, Tat-seng Chua, and Xiaoming Zhang. A semi-supervised bayesian network model for microblog topic classification. In *Proceedings of COLING 2012*, pages 561–576, 2012.

Weiwei Yang, Jordan Boyd-Graber, and Philip Resnik. A multilingual topic model for learning weighted topic links across corpora with low comparability. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1243–1248, 2019.

Yang Bao, Nigel Collier, and Anindya Datta. A partially supervised cross-collection topic model for cross-domain text classification. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 239–248, 2013.

Baoyu Jing, Chenwei Lu, Deqing Wang, Fuzhen Zhuang, and Cheng Niu. Cross-domain labeled lda for cross-domain text classification. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 187–196. IEEE, 2018.

Mahsa Baktashmotlagh, Mehrtash T Harandi, Brian C Lovell, and Mathieu Salzmann. Unsupervised domain adaptation by domain invariant projection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 769–776, 2013.

Philip Resnik, Anderson Garron, and Rebecca Resnik. Using topic modeling to improve prediction of neuroticism and depression in college students. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1348–1353, 2013.

Philip Resnik, William Armstrong, Leonardo Claudino, Thang Nguyen, Viet-An Nguyen, and Jordan Boyd-Graber. Beyond lda: exploring supervised topic modeling for depression-related language in twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 99–107, 2015a.

Philip Resnik, William Armstrong, Leonardo Claudino, and Thang Nguyen. The university of maryland clpsych 2015 shared task system. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 54–60, 2015b.

Xiaolei Huang, Xin Li, Tianli Liu, David Chiu, Tingshao Zhu, and Lei Zhang. Topic model for identifying suicidal ideation in chinese microblog. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 553–562, 2015.

Xiaolei Huang, Linzi Xing, Jed R Brubaker, and Michael J Paul. Exploring timelines of confirmed suicide incidents through social media. In *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 470–477. IEEE, 2017.

Laura Biester, Katie Matton, Janarthanan Rajendran, Emily Mower Provost, and Rada Mihalcea. Quantifying the effects of covid-19 on mental health support forums. *arXiv preprint arXiv:2009.04008*, 2020.

Daniel Ramage, Susan T Dumais, and Daniel J Liebling. Characterizing microblogs with topic models. *icwsm*, 10(1):16, 2010.

William M Darling. A theoretical and practical implementation tutorial on topic modeling and gibbs sampling. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pages 642–647, 2011.

Bob Carpenter. Integrating out multinomial parameters in latent dirichlet allocation and naive bayes for collapsed gibbs sampling. *Rapport Technique*, 4:464, 2010.

Albert Park and Mike Conway. Tracking health related discussions on reddit for public health applications. In *AMIA Annual Symposium Proceedings*, volume 2017, page 1362. American Medical Informatics Association, 2017.

Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li, and Liang Zhao. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11):15169–15211, 2019.

Keith Henderson and Tina Eliassi-Rad. Applying latent dirichlet allocation to group discovery in large graphs. In *Proceedings of the 2009 ACM symposium on Applied Computing*, pages 1456–1461, 2009.

Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

Georgia Albuquerque, Thomas Lowe, and Marcus Magnor. Synthetic generation of high-dimensional datasets. *IEEE transactions on visualization and computer graphics*, 17(12):2317–2324, 2011.

Hanyu Shi, Martin Gerlach, Isabel Diersen, Doug Downey, and Luis AN Amaral. A new evaluation framework for topic modeling algorithms based on synthetic corpora. *arXiv preprint arXiv:1901.09848*, 2019.

Glen Coppersmith, Mark Dredze, Craig Harman, Kristy Hollingshead, and Margaret Mitchell. Clpsych 2015 shared task: Depression and ptsd on twitter. In *Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 31–39, 2015.

Brendan O'Connor, Michel Krieger, and David Ahn. Tweetmotif: exploratory search and topic summarization for twitter. In *ICWSM*, pages 384–385, 2010.

Alexandra Schofield, Måns Magnusson, and David Mimno. Pulling out the stops: Rethinking stopword removal for topic models. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 432–436, 2017.

Masashi Sugiyama, Matthias Krauledat, and Klaus-Robert MÃžller. Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, 8(May):985–1005, 2007.

Isabelle Guyon. Design of experiments of the nips 2003 variable selection benchmark. In *NIPS 2003 workshop on feature extraction and feature selection*, volume 253, 2003.

Viet-An Nguyen, Jordan Boyd-Graber, and Philip Resnik. Sometimes average is best: The importance of averaging for prediction using mcmc inference in topic modeling. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1752–1757, 2014.