Mining Themes and Interests in the Asperger's and Autism Community

Yangfeng Ji, Hwajung Hong, Rosa Arriaga, Agata Rozga, Gregory Abowd, Jacob Eisenstein
School of Interactive Computing
Georgia Institute of Technology
{jiyfeng,hwajung,arriaga,agata,abowd,jacobe}@gatech.edu

Abstract

Discussion forums offer a new source of insight for the experiences and challenges faced by individuals affected by mental disorders. Language technology can help domain experts gather insight from these forums, by aggregating themes and user behaviors across thousands of conversations. We present a novel model for web forums, which captures both thematic content as well as user-specific interests. Applying this model to the Aspies Central forum (which covers issues related to Asperger's syndrome and autism spectrum disorder), we identify several topics of concern to individuals who report being on the autism spectrum. We perform the evaluation on the data collected from Aspies Central forum, including 1,939 threads, 29,947 posts and 972 users. Quantitative evaluations demonstrate that the topics extracted by this model are substantially more than those obtained by Latent Dirichlet Allocation and the Author-Topic Model. Qualitative analysis by subject-matter experts suggests intriguing directions for future investigation.

1 Introduction

Online forums can offer new insights on mental disorders, by leveraging the experiences of affected individuals — in their own words. Such insights can potentially help mental health professionals and caregivers. Below is an example dialogue from the Aspies Central forum,[1] where individuals who report being on the autism spectrum (and their families and friends) exchange advice and discuss their experiences:

- **User A**: *Do you feel paranoid at work? ... What are some situations in which you think you have been unfairly treated?*
- **User B**: *Actually I am going through something like that now, and it is very difficult to keep it under control...*
- **User A**: *Yes, yes that is it. Exactly ... I think it might be an Aspie trait to do that, I mean over think everything and take it too literally?*
- **User B**: *It probably is an Aspie trait. I've been told too that I am too hard on myself.*

Aspies Central, like other related forums, has thousands of such exchanges. However, aggregating insight from this wealth of information poses obvious challenges. Manual analysis is extremely time-consuming and labor-intensive, thus limiting the scope of data that can be considered. In addition, manual coding systems raise validity questions, because they can tacitly impose the pre-existing views of the experimenter on all subsequent analysis. There is therefore a need for computational tools that support large-scale *exploratory textual analysis* of such forums.

In this paper, we present a tool for automatically mining web forums to explore textual themes and user interests. Our system is based on Latent Dirichlet Allocation (LDA; Blei et al, 2003), but is customized for this setting in two key ways:

- By modeling sparsely-varying topics, we can easily recover key terms of interest, while retaining robustness to large vocabulary and small counts (Eisenstein et al., 2011).
- By modeling author preference by topic, we can quickly identify topics of interest for each user, and simultaneously recover topics that better distinguish the perspectives of each author.

The key technical challenge in this work lies in bringing together several disparate modalities into

[1] http://www.aspiescentral.com

a single modeling framework: text, authorship, and thread structure. We present a joint Bayesian graphical model that unifies these facets, discovering both an underlying set of topical themes, and the relationship of these themes to authors. We derive a variational inference algorithm for this model, and apply the resulting software on a dataset gathered from Aspies Central.

The topics and insights produced by our system are evaluated both quantitatively and qualitatively. In a blind comparison with LDA and the author-topic model (Steyvers et al., 2004), both subject-matter experts and lay users find the topics generated by our system to be substantially more coherent and relevant. A subsequent qualitative analysis aligns these topics with existing theory about the autism spectrum, and suggests new potential insights and avenues for future investigation.

2 Aspies Central Forum

Aspies Central (AC) is an online forum for individuals on the autism spectrum, and has publicly accessible discussion boards. Members of the site do not necessarily have to have an official diagnosis of autism or a related condition. Neurotypical individuals (people not on the autism spectrum) are also allowed to participate in the forum. The forum includes more than 19 discussion boards with subjects ranging from general discussions about the autism spectrum to private discussions about personal concerns. As of March 2014, AC hosts 5,393 threads, 89,211 individual posts, and 3,278 members.

AC consists of fifteen public discussion boards and four private discussion boards that require membership. We collected data only from publicly-accessible discussion boards. In addition, we excluded discussion boards that were website-specific (announcement-and-introduce-yourself), those mainly used by family and friends of individuals on the spectrum (friends-and-family) or researchers (autism-news-and-research), and one for amusement (forum-games). Thus, we focused on ten discussion boards (aspergers-syndrome-Autism-and-HFA, PDD-NOS-social-anxiety-and-others, obsessions-and-interests, friendships-and-social-skills, education-and-employment, love-relationships-and-dating, autism-spectrum-help-and-support, off-topic-discussion, entertainment-discussion, computers-technology-discussion), in which AC users discuss their everyday experiences, concerns, and challenges. Using the python library Beautiful Soup, we collected 1,939 threads (29,947 individual posts) from the discussion board archives over a time period from June 1, 2010 to July 27, 2013. For a given post, we extracted associated metadata such as the author identifier and posting timestamps.

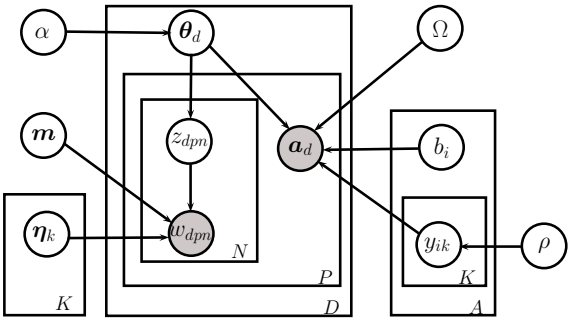

Figure 1: Plate diagram. Shaded notes represent observed variables, clear nodes represent latent variables, arrows indicate probabilistic dependencies, and plates indicate repetition.

3 Model Specification

Our goal is to develop a model that captures the preeminent themes and user behaviors from traces of user behaviors in online forums. The model should unite textual content with authorship and thread structure, by connecting these *observed variables* through a set of *latent variables* representing conceptual topics and user preferences. In this section, we present the statistical specification of just such a model, using the machinery of Bayesian graphical models. Specifically, the model descibes a stochastic process by which the observed variables are emitted from prior probability distributions shaped by the latent variables. By performing Bayesian statistical inference in this model, we can recover a probability distribution around the latent variables of interest.

We now describe the components of the model that generate each set of observed variables. The model is shown as a plate diagram in Figure 1, and the notation is summarized in Table 1.

3.1 Generating the text

The part of the model which produces the text itself is similar to standard latent Dirichlet allocation (LDA) (Blei et al., 2003). We assume a set of K latent topics, which are distributions over each word in a finite vocabulary. These topics are

Symbol	Description
D	number of threads
P_d	number of posts in thread d
N_p	number of word tokens in post p
α	parameter of topic distribution of threads
$\boldsymbol{\theta}_d$	the multinomial distribution of topics specific to the thread d
z_{dpn}	the topic associated with the nth token in post p of thread d
w_{dpn}	the nth token in post p of thread d
$\boldsymbol{a}_d$	authorship distribution for question post and answer posts in thread d respectively
y_{ik}	the topic-preference indicator of author i on topic k
b_i	the Gaussian distribution of author i's selection bias
$\boldsymbol{\eta}_k$	topic k in log linear space
$\boldsymbol{m}$	background topic
Ω	topic weights matrix
σ_λ^2	variance of feature weights
σ_b^2	variance of selection bias
ρ	prior probability of authors' preference on any topic

Table 1: Mathematical notations

shared among all D threads in the collection, but each thread has its own distribution over the topics.

We make use of the SAGE parametrization for generative models of text (Eisenstein et al., 2011). SAGE uses adaptive sparsity to induce topics that deviate from a background word distribution in only a few key words, without requiring a regularization parameter. The background distribution is written $\boldsymbol{m}$, and the deviation for topic k is written $\boldsymbol{\eta}_k$, so that $Pr(w = v|\boldsymbol{\eta}_k, \boldsymbol{m}) \propto \exp(m_v + \eta_{kv})$.

Each word token w_{dpn} (the n^{th} word in post p of thread d) is generated from the probability distribution associated with a single topic, indexed by the latent variable $z_{dpn} \in \{1 \dots K\}$. This latent variable is drawn from a prior $\boldsymbol{\theta}_d$, which is the probability distribution over topics associated with *all posts* in thread d.

3.2 Generating the author

We have metadata indicating the author of each post, and we assume that users are more likely to participate in threads that relate to their topic-specific preference. In addition, some people may be more or less likely to participate overall. We extend the LDA generative model to incorporate each of these intuitions.

For each author i, we define a latent preference vector $\boldsymbol{y}_i$, where $y_{ik} \in \{0, 1\}$ indicates whether the author i prefers to **answer** questions about topic k. We place a Bernoulli prior on each y_{ik}, so that $y_{ik} \sim \text{Bern}(\rho)$, where $\text{Bern}(y; \rho) = \rho^y(1-\rho)^{(1-y)}$. Induction of $\boldsymbol{y}$ is one of the key inference tasks for the model, since this captures topic-specific preference.

It is also a fact that some individuals will participate in a conversation regardless of whether they have anything useful to add. To model this general tendency, we add an "bias" variable $b_i \in \mathbb{R}$. When b_i is negative, this means that author i will be reluctant to participate even when she does have relevant interests.

Finally, various topics may require different levels of preference; some may capture only general knowledge that many individuals are able to provide, while others may be more obscure. We introduce a diagonal topic-weight matrix Ω, where $\Omega_{kk} = \omega_k \geq 0$ is the importance of preference for topic k. We can easily generalize the model by including non-zero off-diagonal elements, but leave this for future work.

The generative distribution for the observed author variable is a log-linear function of $\boldsymbol{y}$ and $\boldsymbol{b}$:

$$Pr(a_{di} = 1|\boldsymbol{\theta}_d, \boldsymbol{y}, \Omega, \boldsymbol{b}) = \frac{\exp(\boldsymbol{\theta}_d^\intercal \Omega \boldsymbol{y}_i + b_i)}{\sum_{j=1}^{A} \exp(\boldsymbol{\theta}_d^\intercal \Omega \boldsymbol{y}_j + b_j)} \tag{1}$$

This distribution is multinomial over authors; each author's probability of responding to a thread depends on the topics in the thread ($\boldsymbol{\theta}_d$), the author's preference on those topics ($\boldsymbol{y}_i$), the importance of preference for each topic (Ω), and the bias parameter b_i. We exponentiate and then normalize, yielding a multinomial distribution.

The authorship distribution in Equation (1) refers to a probability of user i authoring a single response post in thread d (we will handle question posts next). Let us construct a binary vector $\boldsymbol{a}_d^{(r)}$, where it is 1 if author i has authored any response posts in thread d, and zero otherwise. The probability distribution for this vector can be written

$$P(\boldsymbol{a}_d^{(r)}|\boldsymbol{\theta}_d, \boldsymbol{y}, \Omega, \boldsymbol{b}) \propto \prod_{i=1}^{A} \left(\frac{\exp(\boldsymbol{\theta}_d^\intercal \Omega \boldsymbol{y}_i + b_i)}{\sum_{j=1}^{A} \exp(\boldsymbol{\theta}_d^\intercal \Omega \boldsymbol{y}_j + b_j)} \right)^{a_{di}^{(r)}} \tag{2}$$

One of the goals of this model is to distinguish frequent responders (i.e., potential experts) from individuals who post questions in a given topic. Therefore, we make the probability of author i initiating thread d depend on the value $1 - y_{ki}$ for each topic k. We write the binary vector $\boldsymbol{a}_d^{(q)}$, where $a_{di}^{(q)} = 1$ if author i has written the question post, and zero otherwise. Note that there can only be one question post, so $\boldsymbol{a}_d^{(q)}$ is an indicator vector. Its probability is written as

$$p(\boldsymbol{a}_d^{(q)}|\boldsymbol{\theta}_d, \boldsymbol{y}, \Omega, \boldsymbol{b}) \propto \prod_{i=1}^{A} \left(\frac{\exp(\boldsymbol{\theta}_d^\intercal \Omega (\mathbf{1} - \boldsymbol{y}_i) + b_i)}{\sum_{j=1}^{A} \exp(\boldsymbol{\theta}_d^\intercal \Omega (\mathbf{1} - \boldsymbol{y}_j) + b_j)} \right)^{a_{di}^{(q)}} \tag{3}$$

We can put these pieces together for a complete distribution over authorship for thread d:

$$P(\boldsymbol{a}_d, |\boldsymbol{\theta}_d, \boldsymbol{y}, \Omega, \boldsymbol{b}) \propto \prod_{i=1}^{A} \left(\frac{\exp(\boldsymbol{\theta}_d^\intercal \Omega \boldsymbol{y}_i + b_i)}{\sum_{j=1}^{A} \exp(\boldsymbol{\theta}_d^\intercal \Omega \boldsymbol{y}_j + b_j)} \right)^{a_{di}^{(r)}} \cdot \prod_{i=1}^{A} \left(\frac{\exp(\boldsymbol{\theta}_d^\intercal \Omega (\mathbf{1} - \boldsymbol{y}_i) + b_i)}{\sum_{j=1}^{A} \exp(\boldsymbol{\theta}_d^\intercal \Omega (\mathbf{1} - \boldsymbol{y}_j) + b_j)} \right)^{a_{di}^{(q)}} \tag{4}$$

where $\boldsymbol{a}_d = \{\boldsymbol{a}_d^{(q)}, \boldsymbol{a}_d^{(r)}\}$. The probability $p(\boldsymbol{a}_d|\theta_d, \boldsymbol{y}, \Omega, \boldsymbol{b})$ combines the authorship distribution of authors from question post and answer posts in thread d. The identity of the original question poster does not appear in the answer vector, since further posts are taken to be refinements of the original question.

This model is similar in spirit to supervised latent Dirichlet allocation (sLDA) (Blei and McAuliffe, 2007). However, there are two key differences. First, sLDA uses point estimation to obtain a weight for each topic. In contrast, we perform Bayesian inference on the author-topic preference $\boldsymbol{y}$. Second, sLDA generates the metadata from the dot-product of the weights and $\bar{\boldsymbol{z}}$, while we use $\boldsymbol{\theta}$ directly. The sLDA paper argues that there is a risk of overfitting, where some of the topics serve only to explain the metadata and never generate any of the text. This problem does not arise in our experiments.

3.3 Formal generative story

We are now ready to formally define the generative process of our model:

1. For each topic k
   (a) Set the word probabilities $\boldsymbol{\beta}_k = \frac{\exp(\boldsymbol{m}+\boldsymbol{\eta}_k)}{\sum_i \exp(m_i+\eta_{ki})}$
2. For each author i
   (a) Draw the selection bias $b_i \sim \mathcal{N}(0, \sigma_b^2)$
   (b) For each topic k
      i. Draw the author-topic preference level $y_{ik} \sim \text{Bern}(\rho)$
3. For each thread d
   (a) Draw topic proportions $\boldsymbol{\theta}_d \sim \text{Dir}(\alpha)$
   (b) Draw the author vector $\boldsymbol{a}_d$ from Equation (4)
   (c) For each post p
      i. For each word in this post
         A. Draw topic assignment $z_{dpn} \sim \text{Mult}(\boldsymbol{\theta}_d)$
         B. Draw word $w_{dpn} \sim \text{Mult}(\boldsymbol{\beta}_{z_{dpn}})$

4 Inference and estimation

The purpose of inference and estimation is to recover probability distributions and point estimates for the quantities of interest: the content of the topics, the assignment of topics to threads, author preferences for each topic, etc. While recent progress in probabilistic programming has improved capabilities for automating inference and estimation directly from the model specification,[2] here we develop a custom algorithm, based on variational mean field (Wainwright and Jordan, 2008). Specifically, we approximate the distribution over topic proportions, topic indicators, and author-topic preference $P(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{y}|\boldsymbol{w}, \boldsymbol{a}, \boldsymbol{x})$ with a mean field approximation

$$q(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{y}|\gamma, \phi, \psi) = \prod_{i=1}^{A} \prod_{k=1}^{K} q(y_{ik}|\psi_{ik}) \prod_{d=1}^{D} \prod_{p=1}^{P_d} \prod_{n=1}^{N_{p,d}} q(z_{dpn}|\phi_{dpn}) \prod_{d=1}^{D} q(\theta_d|\gamma_d) \tag{5}$$

where P_d is the number of posts in thread d, K is the number of topics, and N_p is the number of word tokens in post P_d. The variational parameters of $q(\cdot)$ are γ, ϕ, ψ. We will write $\langle\cdot\rangle$ to indicate an expectation under the distribution $q(\boldsymbol{\theta}, \boldsymbol{z}, \boldsymbol{y})$.

We employ point estimates for the variables $\boldsymbol{b}$ (author selection bias), $\boldsymbol{\lambda}$ (topic-time feature weights), $\boldsymbol{\eta}$ (topic-word log-probability deviations), and diagonal elements of Ω (topic weights). The estimation of $\boldsymbol{\eta}$ follows the procedure defined in SAGE (Eisenstein et al., 2011); we explain the estimation of the remaining parameters below.

Given the variational distribution in Equation (5), the inference on our topic model can be formulated as constrained optimization of this bound.

$$\begin{aligned} \min\ & \mathcal{L}(\gamma, \phi, \psi; \boldsymbol{b}, \boldsymbol{\lambda}, \Omega) \\ \text{s.t.} & \gamma_{dk} \geq 0 \quad \forall d, k \\ & \boldsymbol{\phi}_{dpn} \geq 0, \sum_k \phi_{dpnk} = 1 \quad \forall d, p, n \\ & 0 \leq \psi_{ik} \leq 1 \quad \forall i, k \\ & \omega_k \geq 0 \quad \forall k \end{aligned} \tag{6}$$

The constraints are due to the parametric form of the variational approximation: $q(\boldsymbol{\theta}_d|\boldsymbol{\gamma}_d)$ is Dirichlet, and requires non-negative parameters;

[2] see `http://probabilistic-programming.org/`

$q(\boldsymbol{z}_{dpn}|\boldsymbol{\phi}_{dpn})$ is multinomial, and requires that $\boldsymbol{\phi}_{dpn}$ lie on the $K-1$ simplex; $q(y_{ik}|\psi_{ik})$ is Bernoulli and requires that ψ_{ik} be between 0 and 1. In addition, as a topic weight, ω_k should also be non-negative.

Algorithm 1 One pass of the variational inference algorithm for our model.

for $d = 1, \ldots, D$ **do**
  while not converged **do**
    for $p = 1, \ldots, P_d$ **do**
      for $n = 1, \ldots, N_{p,d}$ **do**
        Update ϕ_{dpnk} using Equation (7) for each $k = 1, \ldots, K$
      end for
    end for
    Update γ_{dk} by optimizing Equation (6) with Equation (10) for each $k = 1, \ldots, K$
  end while
end for
for $i = 1, \ldots, A$ **do**
  Update ψ_{ik} by optimizing Equation (6) with Equation (13) for each $k = 1, \ldots, K$
  Update $\hat{b}_i$ by optimizing Equation (6) with Equation (14)
end for
for $k = 1, \ldots, K$ **do**
  Update ω_k with Equation (15)
end for

4.1 Word-topic indicators

With the variational distribution in Equation (5), the inference on ϕ_{dpn} for a given token n in post p of thread d is same as in LDA. For the nth token in post p of thread d,

$$\phi_{dpnk} \propto \beta_{kw_{dpn}} \exp(\langle \log \theta_{dk} \rangle) \tag{7}$$

where $\boldsymbol{\beta}$ is defined in the generative story and $\langle \log \theta_{dk} \rangle$ is the expectation of $\log \theta_{dk}$ under the distribution $q(\theta_{dk}|\gamma_d)$,

$$\langle \log \theta_{dk} \rangle = \Psi(\gamma_{dk}) - \Psi(\sum_{k=1}^{K} \gamma_{dk}) \tag{8}$$

where $\Psi(\cdot)$ is the Digamma function, the first derivative of the log-gamma function.

For the other variational parameters γ and ψ, we can not obtain a closed form solution. As the constraints on these parameters are all convex with respect to each component, we employed a projected quasi-Newton algorithm proposed in (Schmidt et al., 2009) to optimize $\mathcal{L}$ in Equation (6). One pass of the variational inference procedure is summarized in Algorithm 1.Since every step in this algorithm will not decrease the variational bound, the overall algorithm is guaranteed to converge.

4.2 Document-topic distribution

The inference for document-topic proportions is different from LDA, due to the generation of the author vector $\boldsymbol{a}_d$, which depends on $\boldsymbol{\theta}_d$. For a given thread d, the part of the bound associated with the variational parameter $\boldsymbol{\gamma}_d$ is

$$\begin{aligned} \mathcal{L}_{\gamma_d} &= \langle \log p(\boldsymbol{\theta}_d|\alpha_d) \rangle + \langle \log p(\boldsymbol{a}_d|\boldsymbol{\theta}_d, \boldsymbol{y}, \Omega, \boldsymbol{b}) \rangle \\ &+ \sum_{p=1}^{P_d} \sum_{n=1}^{N_{p,d}} \langle \log p(z_{dpn}|\boldsymbol{\theta}_d) \rangle - \langle q(\boldsymbol{\theta}_d|\gamma_d) \rangle \end{aligned} \tag{9}$$

and the derivative of $\mathcal{L}_{\gamma_d}$ with respect to γ_{dk} is

$$\begin{aligned} \frac{d\mathcal{L}_{\gamma_d}}{d\gamma_{dk}} &= \Psi'(\gamma_{dk})(\alpha_{dk} + \sum_{p=1}^{P_d} \sum_{n=1}^{N_{p,d}} \phi_{dpnk} - \gamma_{dk}) \\ &- \Psi'(\sum_{k=1}^{K} \gamma_{dk}) \sum_{k=1}^{K} (\alpha_{dk} + \sum_{p=1}^{P_d} \sum_{n=1}^{N_{p,d}} \phi_{dpnk} - \gamma_{dk}) \\ &+ \frac{d}{d\gamma_{dk}} \langle \log p(\boldsymbol{a}_d|\boldsymbol{\theta}_d, \boldsymbol{y}, \Omega, \boldsymbol{b}) \rangle , \end{aligned} \tag{10}$$

where $\Psi'(\cdot)$ is the trigramma function. The first two lines of Equation (10) are identical to LDA's variational inference, which obtains a closed-form solution by setting $\gamma_{dk} = \alpha_{dk} + \sum_{p,n} \phi_{dpnk}$. The additional term for generating the authorship vector $\boldsymbol{a}_d$ eliminates this closed-form solution and forces us to turn to gradient-based optimization.

The expectation on the log probability of the authorship involves the expectation on the log partition function, which we approximate using Jensen's inequality. We then derive the gradient,

$$\begin{aligned} &\frac{\partial}{\partial \gamma_{dk}} \langle \log p(\boldsymbol{a}_d|\boldsymbol{\theta}_d, \boldsymbol{y}, \Omega, \boldsymbol{b}) \rangle \\ &\approx \omega_k \left(\sum_{i=1}^{A} a_{di}^{(r)} \psi_{ik} - A_d^{(r)} \sum_{i=1}^{A} \psi_{ik} \left\langle a_{di}^{(r)}|\boldsymbol{\theta}_d, \boldsymbol{y} \right\rangle \right) \\ &- \omega_k \left(\sum_{i=1}^{A} a_{di}^{(q)} \psi_{ik} - \sum_{i=1}^{A} \psi_{ik} \left\langle a_{di}^{(q)}|\boldsymbol{\theta}_d, \boldsymbol{y} \right\rangle \right) \end{aligned} \tag{11}$$

The convenience variable $A_d^{(r)}$ counts the number of distinct response authors in thread d; recall that there can be only one question author. The notation

$$\left\langle a_{di}^{(r)}|\boldsymbol{\theta}_d, \boldsymbol{y} \right\rangle = \frac{\exp(\left\langle \boldsymbol{\theta}^{\mathsf{T}} \right\rangle \Omega \langle \boldsymbol{y}_i \rangle + b_i)}{\sum_j \exp(\left\langle \boldsymbol{\theta}^{\mathsf{T}} \right\rangle \Omega \langle \boldsymbol{y}_j \rangle + b_j)},$$

represents the generative probability of $a_{di}^{(r)} = 1$ under the current variational distributions $q(\boldsymbol{\theta}_d)$ and $q(\boldsymbol{y}_i)$. The notation $\left\langle a_{di}^{(q)}|\boldsymbol{\theta}_d, \boldsymbol{y} \right\rangle$ is analogous, but represents the question post indicator $a_{di}^{(q)}$.

4.3 Author-topic preference

The variational distribution over author-topic preference is $q(y_{ik}|\psi_{ik})$; as this distribution is Bernoulli, $\langle y_{ik} \rangle = \psi_{ik}$, the parameter itself proxies for the topic-specific author preference — how much author i prefers to answer posts on topic k.

The part of the variational bound the relates to the author preferences is

$$\begin{aligned}\mathcal{L}_\psi &= \sum_{d=1}^{D} \langle \log p(\boldsymbol{a}_d|\boldsymbol{\theta}_d, \boldsymbol{y}, \Omega, \boldsymbol{b}) \rangle \\ &+ \sum_{i=1}^{A}\sum_{k=1}^{K} \langle p(y_{ik}|\rho) \rangle - \sum_{i=1}^{A}\sum_{k=1}^{K} \langle q(y_{ik}|\psi_{ik}) \rangle\end{aligned} \tag{12}$$

For author i on topic k, the derivative of $\langle \log p(\boldsymbol{a}_d|\boldsymbol{\theta}_d, \boldsymbol{y}, \Omega, \boldsymbol{b}) \rangle$ for document d with respect to ψ_{ik} is

$$\begin{aligned}&\frac{d}{d\psi_{ik}} \langle \log P(\boldsymbol{a}_d|\boldsymbol{\theta}_d, \boldsymbol{y}, \Omega, \boldsymbol{b}) \rangle \\ &\approx \langle \theta_{dk} \rangle \omega_k \left(a_{di}^{(r)} - \left\langle a_{di}^{(r)}|\boldsymbol{\theta}_d, \boldsymbol{y} \right\rangle - a_{di}^{(q)} + \left\langle a_{di}^{(q)}|\boldsymbol{\theta}_d, \boldsymbol{y} \right\rangle \right),\end{aligned} \tag{13}$$

where $\langle \theta_{dk} \rangle = \frac{\gamma_{dk}}{\sum_{k'} \gamma_{dk'}}$. Thus, participating as a respondent increases ψ_{ik} to the extent that topic k is involved in the thread; participating as the questioner decreases ψ_{ik} by a corresponding amount.

4.4 Point estimates

We make point estimates of the following parameters: author selection bias b_i and topic-specific preference weights ω_k. All updates are based on maximum a posteriori estimation or maximum likelihood estimation.

Selection bias For the selection bias b_i of author i given a thread d, the objective function in Equation (6) with the prior of $b_i \sim \mathcal{N}(0, \sigma_b^2)$ is minimized by a quasi-Newton algorithm with the following derivative

$$\begin{aligned}\frac{\partial}{\partial b_i} \langle \log P(\boldsymbol{a}_d|\boldsymbol{\theta}_d, \boldsymbol{y}, \Omega, \boldsymbol{b}) \rangle &\approx a_{d,i}^{(r)} - \\ \left\langle a_{di}^{(r)}|\boldsymbol{\theta}_d, \boldsymbol{y} \right\rangle &+ a_{d,i}^{(q)} - \left\langle a_{di}^{(q)}|\boldsymbol{\theta}_d, \boldsymbol{y} \right\rangle\end{aligned} \tag{14}$$

The zero-mean Gaussian prior shrinks b_i towards zero by subtracting b_i/σ_b^2 from this gradient. Note that the gradient in Equation (14) is non-negative whenever author i participates in thread d. This means any post from this author, whether question posts or answer posts, will have a positive contribution of the author's selection bias. This means that any activity in the forum will elevate the selection bias b_i, but will not necessarily increase the imputed preference level.

Topic weights The topic-specific preference weight ω_k is updated by considering the derivative of variational bound with respect to ω_k

$$\frac{\partial \mathcal{L}}{\partial \omega_k} = \sum_{d=1}^{D} \frac{\partial}{\partial \omega_k} \langle p(\boldsymbol{a}_d|\boldsymbol{\theta}_d, \boldsymbol{y}, \Omega, \boldsymbol{b}) \rangle \tag{15}$$

where for a given document d,

$$\begin{aligned}&\frac{\partial}{\partial \omega_k} \langle \log p(\boldsymbol{a}_d|\boldsymbol{\theta}_d, \boldsymbol{y}, \Omega, \boldsymbol{b}) \rangle \approx \langle \theta_{dk} \rangle \omega_k \cdot \\ &\sum_{i=1}^{A} \psi_{ik} \Big(a_i^{(r)} - a_i^{(q)} + \left\langle a_{di}^{(q)}|\boldsymbol{\theta}_d, \boldsymbol{y} \right\rangle \\ &- A_d^{(r)} \left\langle a_{di}^{(r)}|\boldsymbol{\theta}_d, \boldsymbol{y} \right\rangle \Big)\end{aligned}$$

Thus, ω_k will converge at a value where the observed posting counts matches the expectations under $\langle \log p(\boldsymbol{a}_d|\boldsymbol{\theta}_d, \boldsymbol{y}, \Omega, \boldsymbol{b}) \rangle$.

5 Quantitative Evaluation

To validate the topics identified by the model, we performed a manual evaluation, combining the opinions of both novices as well as subject matter experts in Autism and Asberger's Syndrome. The purpose of the evaluation is to determine whether the topics induced by the proposed model are more coherent than topics from generic alternatives such as LDA and the author-topic model, which are not specifically designed for forums.

5.1 Experiment Setup

Preprocessing Preprocessing was minimal. We tokenized texts using white space and removed punctuations at the beginning/end of each token. We removed words that appear less than five times, resulting in a vocabulary of the 4903 most frequently-used words.

Baseline Models We considered two baseline models in the evaulation. The first baseline model is latent Dirichlet allocation (LDA), which considers only the text and ignores the metadata (Blei et al., 2003). The second baseline is the Author-Topic (AT) model, which extends LDA by associating authors with topics (Rosen-Zvi et al., 2004; Steyvers et al., 2004). Both baselines are implemented in the Matlab Topic Modeling Toolbox (Steyvers and Griffiths, 2005).

Parameter Settings For all three models, we set $K = 50$. Our model includes the three tunable parameters ρ, the Bernoulli prior on topic-specific expertise; σ_b^2, the variance prior on use selection

bias; and α, the prior on document-topic distribution. In the following experiments, we chose $\rho = 0.2$, $\sigma_b^2 = 1.0$, $\alpha = 1.0$. LDA and AT share two parameters, α, the symmetric Dirichlet prior for document-topic distribution; β, the symmetric Dirichlet prior for the topic-word distribution. In both models, we set $\alpha = 3.0$ and $\beta = 0.01$. All parameters were selected in advance of the experiments; further tuning of these paramters is left for future work.

5.2 Topic Coherence Evaluation

To be useful, a topic model should produce topics that human readers judge to be *coherent*. While some automated metrics have been shown to cohere with human coherence judgments (Newman et al., 2010), it is possible that naive raters might have different judgments from subject matter experts. For this reason, we focused on human evaluation, including both expert and novice opinions. One rater, R1, is an author of the paper (HH) and a Ph.D. student focusing on designing technology to understand and support individuals with autism spectrum disorder. The remaining three raters are not authors of the paper and are not domain experts.

In the evaluation protocol, raters were presented with batteries of fifteen topics, from which they were asked to select the three most coherent. In each of the ten batteries, there were five topics from each model, permuted at random. Thus, after completing the task, all 150 topics — 50 topics from each model — were rated. The user interface of topic coherence evaluation is given in Figure 2, including the specific prompt.

We note that this evaluation differs from the "intrusion task" proposed by Chang et al. (2009), in which raters are asked to guess which word was randomly inserted into a topic. While the intrusion task protocol avoids relying on subjective judgments of the meaning of "coherence," it prevents expert raters from expressing a preference for topics that might be especially useful for analysis of autism spectrum disorder. Prior work has also shown that the variance of these tasks is high, making it difficult to distinguish between models.

Table 2 shows, for each rater, the percentage of topics were chosen from each model as the most coherent within each battery. On average, 80% of the topics were chosen from our proposed model. If all three models are equally good at discovering coherent topics, the average percentage across three models should be roughly equal. Note that the opinion of the expert rater R1 is generally similar to the other three raters.

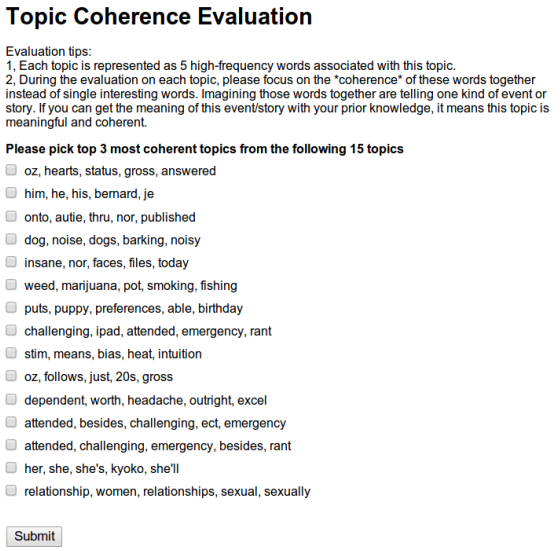

Figure 2: The user interface of topic coherence evaluation.

	Rater				
Model	R1	R2	R3	R4	Average
Our model	70%	93%	80%	77%	80%
AT	17%	7%	13%	10%	12%
LDA	13%	0%	7%	13%	8%

Table 2: Percentage of the most coherent topics that are selected from three different topic models: our model, the Author-Topic Model (AT), and latent Dirichlet allocation (LDA).

6 Analysis of Aspies Central Topics

In this section, we further use our model to explore more information about the Aspies Central forum. We want to examine whether the autism-related topics identified the model can support researchers to gain qualitative understanding of the needs and concerns of autism forum users. We are also interested in understanding the users' behavioral patterns on autism-related topics. The analysis task has three components: first we will describe the interesting topics from the autism domain perpective. Then we will find out the proportion of each topic, including autism related topics. Finally, in order to understand the user activity patterns on these autism related topics we will derive the topic-specific preference ranking of the users from our model.

Index	Proportion	Top keywords	Index	Proportion	Top keywords
1	1.7%	dont im organization couldnt construction	2	2.6%	**yah supervisor behavior taboo phone**
3	2.2%	game watched games fallout played	4	3.5%	**volunteering esteem community art self**
5	1.1%	**nobody smell boss fool smelling**	6	3.2%	firefox razor blades pc console
7	3.4%	doesn't it's mandarin i've that's	8	2.1%	**diagnosed facessenses visualize visual**
9	1.7%	obsessions bookscollecting library authors	10	2.6%	ptsd central cure neurotypical we
11	1.2%	stims mom nails lip shoes	12	1.8%	**classroom campus tag numbers exams**
13	1.6%	battery hawke charlie ive swing	14	1.9%	divorce william women marryrates
15	0.1%	chocolate pdd milk romance nose	16	5.8%	kinda holland neccesarily employment bucks
17	0.6%	eat burgers jokes memory foods	18	2.4%	dryer martial dream wake schedule
19	3.7%	**depression beleive christianity buddhism becouse**	20	1.4%	grudges pairs glasses museum frames
21	0.4%	alma star gods alien sun	22	2.6%	**facebook profiles befriend friendships friends**
23	0.4%	trilogy sci-fi cartoon iphone grandma	24	2.7%	flapping stuffed toes curse animal
25	1.5%	**empathy smells compassion emotions emotional**	26	1.7%	males evolution females originally constructive
27	0.5%	list dedicate lists humor song	28	4.6%	nts aspies autie qc intuitive
29	2.7%	captain i'm film anime that's	30	3.6%	homeless pic wild math laugh
31	3.3%	**shave exhausting during terrified products**	32	5.6%	you're you your yourself hiring
33	4.6%	dictionary asks there're offend fog	34	1.5%	grade ed school 7th diploma
35	1.0%	cave blonde hair bald disney	36	1.9%	**diagnosis autism syndrome symptoms aspergers**
37	1.3%	song joanna newsom rap favorites	38	1.8%	poetry asleep children ghosts lots
39	2.1%	heat iron adhd chaos pills	40	3.6%	bike zone rides zoning worrying
41	1.2%	uk maths team teams op	42	0.8%	book books read reading kindle
43	1.0%	husband narcissist husband's he hyper	44	1.1%	**songs guitar drums music synth**
45	1.3%	**autism disorder spectrum disorders pervasive**	46	0.7%	dog noise dogs barking noisy
47	0.6%	**relationship women relationships sexual sexually**	48	0.9%	**weed marijuana pot smoking fishing**
49	0.9%	him he his bernard je	50	2.0%	her she she's kyoko she'll

Table 3: 50 topics identified by our model. The "proportion" columns show the topic proportions in the dataset. Furthermore, 14 topics are highlighted as interesting topics for autism research.

Table 3 shows all 50 topics from our model. For each topic, we show the top five words related to this topic. We further identified fourteen topics (highlighted with **blue** color), which are particularly relevant to understand autism.

Among the identified topics, there are three popular topics discussed in the Aspies Central forum: topic 4, topic 19 and topic 31. From the top word list, we identified that topic 4 is composed of keywords related to psychological (e.g., self-esteem, art) and social (e.g., volunteering, community) well-being of the Aspies Central users. Topic 19 includes discussion on mental health issues (e.g., depression) and religious activities (e.g., believe, christianity, buddhism) as coping strategies. Topic 31 addresses a specific personal hygiene issue — helping people with autism learn to shave. This might be difficult for individuals with sensory issues: for example, they may be terrified by the sound and vibration generated by the shaver. For example, topic 22 is about making friends and maintaining friendship; topic 12 is about educational issues ranging from seeking educational resources to improving academic skills and adjusting to college life.

In addition to identifying meaningful topics, another capability of our model is to discover users' topic preferences and expertise. Recall that, for user i and topic k, our model estimates a author-topic preference variable ψ_{ik}. Each ψ_{ik} ranges from 0 to 1, indicating the probability of user i to answer a question on topic k. As we set the prior probability of author-topic preference to be 0.2, we show topic-author pairs for which $\psi_{ik} > 0.2$ in Table 4.

Topic	User index
5	USER_1, USER_2, USER_3, USER_4, USER_5
8	USER_1, USER_2, USER_6, USER_5, USER_7
12	USER_1, USER_2, USER_4, USER_8, USER_3
19	USER_1, USER_2, USER_3, USER_4, USER_7
22	USER_1, USER_2, USER_3, USER_9, USER_7
31	USER_1, USER_3, USER_2, USER_6, USER_10
36	USER_1, USER_2, USER_4, USER_3, USER_11
45	USER_1, USER_3, USER_4, USER_12, USER_13
47	USER_2, USER_14, USER_15, USER_16 , USER_6
48	USER_5, USER_4, USER_6, USER_9, USER_2

Table 4: The ranking of user preference on some interesting topics (we replace user IDs with user indices to avoid any privacy-related issue). USER_1 is the moderator of this forum. In total, our model identifies 16 user with high topic-specific preference from 10 interesting topics. For the other 4 interesting topics, there is no user with significantly high preference.

The dominance of USER_1 in these topics is explained by the fact that this user is the moderator of the forum. Besides, we also find some other users participating in most of the interesting topics, such as USER_2 and USER_3. On the other hand, users like USER_14 and USER_15 only show up in few topics. This observation is supported by their activities on discussion boards. Searching on the Aspies Central forum, we found most answer posts of user USER_15 are from the board "love-

relationships-and-dating".

7 Related Work

Social media has become an important source of health information (Choudhury et al., 2014). For example, Twitter has been used both for mining both public health information (Paul and Dredze, 2011) and for estimating individual health status (Sokolova et al., 2013; Teodoro and Naaman, 2013). Domain-specific online communities, such Aspies Central, have their own advantages, targeting specific issues and featuring more close-knit and long-term relationships among members (Newton et al., 2009).

Previous studies on mining health information show that technical models and tools from computational linguistics are helpful for both understanding contents and providing informative features. Sokolova and Bobicev (2011) use sentiment analysis to analyze opinions expressed in health-related Web messages; Hong et al. (2012) focus on lexical differences to automatically distinguish schizophrenic patients from healthy individuals.

Topic models have previously been used to mine health information: Resnik et al. (2013) use LDA to improve the prediction for neuroticism and depression on college students, while Paul and Dredze (2013) customize their *factorial* LDA to model the joint effect of drug, aspect, and route of administration. Most relevantly for the current paper, Nguyen et al. (2013) use LDA to discover autism-related topics, using a dataset of 10,000 posts from ten different autism commnities. However, their focus was on automated classification of communities as autism-related or not, rather than on analysis and on providing support for qualitative autism researchers. The applicability of the model developed in our paper towards classification tasks is a potential direction for future research.

In general, topic models capture latent themes in document collections, characterizing each document in the collection as a mixture of topics (Blei et al., 2003). A natural extension of topic models is to infer the relationships between topics and metadata such as authorship or time. A relatively simple approach is to represent authors as an aggregation of the topics in all documents they have written (Wagner et al., 2012). More sophisticated topic models, such as Author-Topic (AT) model (Rosen-Zvi et al., 2004; Steyvers et al., 2004) assume that each document is generated by a mixture of its authors' topic distributions. Our model can be viewed as one further extension of topic models by incorporating more metadata information (authorship, thread structure) in online forums.

8 Conclusion

This paper describes how topic models can offer insights on the issues and challenges faced by individuals on the autism spectrum. In particular, we demonstrate that by unifying textual content with authorship and thread structure metadata, we can obtain more coherent topics and better understand user activity patterns. This coherence is validated by manual annotations from both experts and non-experts. Thus, we believe that our model provides a promising mechanism to capture behavioral and psychological attributes relating to the special populations affected by their cognitive disabilities, some of which may signal needs and concerns about their mental health and social well-being.

We hope that this paper encourages future applications of topic modeling to help psychologists understand the autism spectrum and other psychological disorders — and we hope to obtain further validation of our model through its utility in such qualitative research. Other directions for future work include replication of our results across multiple forums, and applications to other conditions such as depression and attention deficit hyperactivity disorder (ADHD).

Acknowledgments

This research was supported by a Google Faculty Award to the last author. We thank the three reviewers for their detailed and helpful suggestions to improve the paper.

References

David M. Blei and Jon D. McAuliffe. 2007. Supervised Topic Models. In *NIPS*.

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

Jonathan Chang, Jordan L. Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading Tea Leaves: How Humans Interpret Topic Models. In Yoshua Bengio, Dale Schuurmans, John D. Lafferty, Christopher K. I. Williams, and Aron Culotta,

editors, *NIPS*, pages 288–296. Curran Associates, Inc.

Munmun De Choudhury, Meredith Ringel Morris, and Ryen W. White. 2014. Seeking and Sharing Health Information Online: Comparing Search Engines and Social Media. In *Procedings of CHI*.

Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. 2011. Sparse Additive Generative Models of Text. In *ICML*.

Kai Hong, Christian G. Kohler, Mary E. March, Amber A. Parker, and Ani Nenkova. 2012. Lexical Differences in Autobiographical Narratives from Schizophrenic Patients and Healthy Controls. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 37–47. Association for Computational Linguistics, July.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108. Association for Computational Linguistics.

A. Taylor Newton, Adam D.I. Kramer, and Daniel N. McIntosh. 2009. Autism online: a comparison of word usage in bloggers with and without autism spectrum disorders. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 463–466. ACM.

Thin Nguyen, Dinh Phung, and Svetha Venkatesh. 2013. Analysis of psycholinguistic processes and topics in online autism communities. In *Multimedia and Expo (ICME), 2013 IEEE International Conference on*, pages 1–6. IEEE.

Michael J. Paul and Mark Dredze. 2011. You Are What You Tweet: Analyzing Twitter for Public Health. In *ICWSM*.

Michael J. Paul and Mark Dredze. 2013. Drug Extraction from the Web: Summarizing Drug Experiences with Multi-Dimensional Topic Models. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 168–178, Atlanta, Georgia, June. Association for Computational Linguistics.

Philip Resnik, Anderson Garron, and Rebecca Resnik. 2013. 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*.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The Author-Topic Model for Authors and Documents. In *UAI*.

Mark Schmidt, Ewout van den Berg, Michael P. Friedlander, and Kevin Muphy. 2009. Optimizing Costly Functions with Simple Constraints: A Limited-Memory Projected Quasi-Netton Algorithm. In *AISTATS*.

Marina Sokolova and Victoria Bobicev. 2011. Sentiments and Opinions in Health-related Web messages. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 132–139, Hissar, Bulgaria, September. RANLP 2011 Organising Committee.

Marina Sokolova, Stan Matwin, Yasser Jafer, and David Schramm. 2013. How Joe and Jane Tweet about Their Health: Mining for Personal Health Information on Twitter. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 626–632, Hissar, Bulgaria, September. INCOMA Ltd. Shoumen, BULGARIA.

Mark Steyvers and Thomas Griffiths. 2005. Matlab Topic Modeling Toolbox 1.4.

Mark Steyvers, Padhraic Smyth, and Thomas Griffiths. 2004. Probabilistic Author-Topic Models for Information Discovery. In *KDD*.

Rannie Teodoro and Mor Naaman. 2013. Fitter with Twitter: Understanding Personal Health and Fitness Activity in Social Media. In *Proceedings of the 7th International Conference on Weblogs and Social Media*.

Claudia Wagner, Vera Liao, Peter Pirolli, Les Nelson, and Markus Strohmaier. 2012. It's not in their tweets: Modeling topical expertise of Twitter users. In *ASE/IEEE International Conference on Social Computing*.

Martin J. Wainwright and Michael I. Jordan. 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305.