Research Article

Applied Stochastic Models in Business and Industry

Received 6 August 2014, Revised 20 April 2015, Accepted 20 April 2015 Published online 4 June 2015 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/asmb.2123

A directed topic model applied to call center improvement

Theodore T. Allen[a*†], Hui Xiong[b] and Anthony Afful-Dadzie[c]

We propose subject matter expert refined topic (SMERT) allocation, a generative probabilistic model applicable to clustering freestyle text. SMERT models are three-level hierarchical Bayesian models in which each item is modeled as a finite mixture over a set of topics. In addition to discrete data inputs, we introduce binomial inputs. These 'high-level' data inputs permit the 'boosting' or affirming of terms in the topic definitions and the 'zapping' of other terms. We also present a collapsed Gibbs sampler for efficient estimation. The methods are illustrated using real world data from a call center. Also, we compare SMERT with three alternative approaches and two criteria. Copyright © 2015 John Wiley & Sons, Ltd.

Keywords: Bayesian modeling; Gibbs sampling; latent Dirichlet allocation

1. Introduction

In this paper, we consider the problem of modeling text corpora and other collections of discrete data. The goal is to find short descriptions of the members of the collection that enable classification, prioritization, novelty detection, and other basic tasks. Therefore, we consider the same problem as many researchers including Blei *et al.* [1], who proposed latent Dirichlet allocation (LDA).

Building on the success of LDA, researchers have created a class of related methods called 'topic models', with the goal of achieving greater interpretability and predictive power ([2–11]). The topic models generally identify definitions for the model clusters or 'topics' and also identify the specific data associated with these clusters. Many topic models are based on Bayesian estimation, and yet, the models involve priors intended to be general and not problem specific. The result has been that, while many of the cluster definitions offer appealing interpretability to users, almost inevitably, some of the cluster definitions are inaccurate in ways that users or subject matter experts (SMEs) can easily perceive.

In our own application of LDA ([12]), we generate topics that we desire to edit, but there is no way to make modifications. In that research, we study the Toyota unintended acceleration problem and desire to separate comments relating to praising the acceleration from negative comments. In this article, we propose a method that permits editing of the topic definitions, that is, a 'topic definition eraser' so that the edited topics can be interpreted and are useful to facilitate action.

Other research has investigated novel ways for SMEs to inject problem-specific information into model building. One way to do this, which is common in information retrieval, is through having the SME 'tag' or 'supervise' some of all of the data ([4]). This tagging is often time-consuming. It is also difficult in part, because each data might be on a different topic. Also, the topic or cluster definitions are difficult to define and could change after some or all of the documents have been tagged, necessitating rework.

Other research has solicited inputs from SMEs in various forms. These include eliciting 'cannot' or 'must' links between words, for example, the SME might declare that 'computer' and 'interface' must appear together in the context of a specific corpus. However, in the models presented so far, these constructions are to be applied without regard to specific clusters ([13]). Such general definitions are attractive, because the effort required from the SMEs is likely much less than that needed to tag even a fraction of the documents.

[a] *Integrated Systems Engineering, The Ohio State University, 1971 Neil Avenue - 210 Baker Systems, Columbus, Ohio 43221, U.S.A.*
[b] *Intel Corporation, 2501 NW 229th Ave, Hillsboro, Oregon 97124, U.S.A.*
[c] *University of Ghana Business School, P. O. Box LG 78, Legon, Accra, Ghana*
**Correspondence to: Theodore T. Allen, Integrated Systems Engineering, The Ohio State University, 1971 Neil Avenue - 210 Baker Systems, Columbus, OH 43221, U.S.A.*
[†] *E-mail: allen.515@osu.edu*

The 'cannot' and 'must' links have also been built into fully Bayesian formulations using 'forest priors' by [13]. These approaches preserve the ability to apply Bayesian estimation and Gibbs sampling for posterior estimation. Preserving the Bayesian status of model estimates is a subjectively pleasing property. Yet, because the 'cannot' and 'must' link information is derived without reference to specific clusters or topics, the burden on the SME could be too great. Specific rules might not apply to all topics and records in the database. Also, the SME inputs could be noisy. Because of the use of SME inputs as a prior and the specific formulation used, the effects of noisy inputs are difficult to comprehend.

Still, other researchers have solicited inputs from SMEs in the form of a background distribution over common words ([5]). Such distributions could, conceivably, be difficult to generate, because, as for forest priors, the background distributions do not relate to specific issues with derived topics. It might be more advantageous, perhaps, to ask the SME to critique the topics generated by an initial model, for example, LDA. Then, the SME might critique the derived topics affirming some words while recommending removal of others.

Further, we argue that additional attention to the process of generating inputs to any of the aforementioned models could be helpful. Among the questions to be addressed are whether users should be supposed to elicit forest priors or 'must' links before looking at initial analyses. Also, one analyst might apply input generated by SME A, while another uses tags from expert B. The choice of Bayesian formulations might imply that inputs are solicited before analysis, but that might not be desirable or needed. For every new analysis, all data inputs form part of the prior. In this article, we propose a new model formulation. In addition, we seek to further explore the process of modeling including elicitation of inputs from local subject SMEs and/or users.

The paper is organized as follows. In the next section, we introduce the notation and terminology building on work by [1]. Next, the SMERT model is presented and related to LDA. We also derive the collapsed Gibbs sampler for SMERT which permits posterior sampling. The case study application of SMERT to aid in call center issue prioritization is then presented along with a numerical experiment comparing SMERT with alternatives including LDA and fuzzy c clustering. Additional details are provided in an online supplement. Finally, conclusions and opportunities for future research are discussed.

2. Notation and terminology

This section describes the notation and terminology that focus on the application to text corpora. The intent is to establish an intuitive background for the presentation of subject matter expert refined topic (SMERT) models in the next section.

For the remainder of this paper, we will use the language of text collections following a combination of that used by [1] and [6]. It is important to note that neither LDA nor SMERT models are tied to text. They both have applications to other problems involving collections of data such as image analysis in manufacturing and information retrieval and bioinformatics.

2.1. General definitions

We use the following definitions:

- A *word* is a unit of discrete data which can take on any of WC whole number values in a dictionary. We represent words using the index in the dictionary, which we believe is simpler than the unit basis representation in [1].
- A *document* is a sequence of words indexed by d and denoted, $\{w_{d,1}, \ldots, w_{d,N_d}\}$, where N_d is the number of words in the dth document.
- A *corpus* is a collection of D documents. We use N to refer to a D-vector of N_d counts for all documents in a corpus.
- A *topic* is a cluster or grouping represented by the index $t = 1, \ldots, T$, where T is the number of topics. We use an index to represent topics, for example, $z_{d,n} = 4$ if the nth word in document d is the fourth topic.
- A *boost table* is a collection of $x_{t,c}$ of successes out of a possible $N_{t,c}$ Bernoulli trials for all topics $t = 1, , T$ and words $c = 1, \ldots, WC$. These data are called 'high-level' because they derive from local SMEs instead of the words in the 'low-level' documents. We use N_t to refer to a $T \times WC$ matrix of all Bernoulli trials for all topics and all words.

The goal of modeling the corpus is to accurately identify the topics and the fractions of each document related to each topic. Also, we seek to create and store valuable intellectual property of the individual or organization in the boost tables. In our examples, $N_{t,c} = 0$ for many combinations of words and topics so that the SMEs do not necessarily need to generate counts for all words in all topics.

2.2. Boosting and zapping

The generation of the boosting table could occur after an initial inspection of LDA results. Then, the SME enters boosting data to edit or critique the top words in each topic. We divide instances into two main types differentiated by a specific combination of a topic (t) and a word (c):

- *Boosting* occurs when $N_{t,c} \geqslant x_{t,c} > 0$. In this case, the SME is affirming that word c has a nonzero probability in topic t. For example, with a dictionary of 500 words, the condition $N_{t,c} = x_(t, c) = 3$ can be viewed as strong affirmation of the importance of the word because 3 out of 3 draws from the topic would be the word which would almost never happen in a usual text.
- *Zapping* is the case $N_{t,c} \geqslant x_{t,c} = 0$. Here, the SME is entering data indicating that word c likely does not belong in topic t, particularly if $N_{t,c} >> 0$. For example, with $N_{t,c} = 100,000$ and $x_{t,c} = 0$, the SME is entering data such that the word would likely never occur even in a long document with 100,000 words on the topic.

The data from boosting or zapping could, conceivably, come from an experiment in which the SME is asked to write $N_{t,c}$ words on a topic and count the number that are the word c. In our examples, we entered the $N_{t,c}$ and $x_{t,c}$ counts directly with reference to the top 10 words in each topic estimated using LDA. For the other words, we assume the null case, that is, $N_{t,c} = x_{t,c} = 0$.

As we show in Section 4, the collapsed Gibbs sampler for our model is analytically tractable for cases having zero, or a single instance of $N_{t,c} > x_{t,c}$ is greater than zero for each topic. Therefore, our posterior sampling is exact only for cases with a single zap per topic or none. We believe that these constraints still offer significant flexibility with which to refine topics. Also, for other cases, we describe how the estimation forms an approximation.

3. Subject matter expert refined topic models

In this section, we define SMERT models. We begin by describing the generative process for words in documents and also boost table counts. Next, we define the joint likelihood and the graphical model description.

3.1. Generative process for words and counts

A SMERT model is a generative probabilistic model of a corpus. Documents are represented as random mixtures over latent topics, and each topic is characterized by a distribution over the words. SMERT assumes the following generative process for document d in a corpus:

1. Choose $N_d \sim Poisson(\xi_1)$.
2. Choose $\theta_d \sim Dir(\alpha_d)$.
3. For each of the $j = 1, \ldots, N_d$ words, $w_{d,j}$:
   (a) Choose a topic for word $n, z_{d,n} \sim Categorical_T(\theta_d)$.
   (b) Choose a word $w_{d,j} \sim Categorical_{WC}(\phi_{z_{d,j}})$.

The vector, α , might be assumed to have all its elements equaling the same value, α_0. Then, prior to estimation, we would expect all the topics to have an equal probability. Simplifying assumptions are made in the basic model, including some that are removed in subsequent sections. First, the number of topics, T, is assumed to be known and fixed. Second, the word probabilities are parameterized by a WC vector ϕ_t for each topic, t, where $\phi_{t,c} = p(w = c|z = t)$, which we now treat as a fixed quantity that is to be estimated. Note that, as for LDA, the Poisson(ξ_1) assumption is not critical to anything that follows and more realistic document length distributions could be applied. It is assumed that N_d is independent of other data-generating variables (z and θ). We generally ignore the randomness of the N_d because they are ancillary variables. We also have a generative process for the boosting table:

1. For each topic t:
2. For each word c:
   (a) Choose a number of trials $N_{t,c} \sim Poisson(\xi_2)$.
   (b) Choose a success count $x_{t,c} \sim Binomial(N_{t,c}, \phi_{t,c})$, where $\phi_{t,c}$ is the same topic definition probability used in document generation.

One additional generation process relates to the word-topic probabilities, $\phi_{t,c}$. In many implementations of LDA, these are generated in the Bayesian prior using $\phi_{t,c} \sim Dir(\beta_t)$, where β_t is a vector like α which is both chosen to be non-informative and has all of its elements equal to the same value, β_0.

As for document generation, the Poisson assumption is ancillary to the boost-generation process so that Poisson (ξ_2) is playing no role in the modeling process. The two generative processes interact through the topic probability vectors, ϕ_t.

The SMERT model has been expressed with a combination of two specific types of inputs. The first are categorical random draws for the $z_{d,n}$ and $w_{d,n}$. The second are binomial draws for the $x_{t,c}$. Therefore, SMERT is a 'multi-response'

combination categorical and binomial model. Yet, the concept of SMERT is more general. The SMERT label can refer to any model in which some inputs derive from a usual generating process and others derive locally from SMEs.

3.2. Joint density and graphical model

The joint distribution or likelihood that defines the initial SMERT model is simply the product of the individual conditional densities:

$$P(z,w,x,\theta,\phi \mid N,N_t,\alpha,\beta) = \left[\prod_{t=1}^{T} P\left(\phi_t \mid \beta_t\right)\right]\left[\prod_{d=1}^{D} P\left(\theta_d \mid \alpha_d\right)\right] \times \left[\prod_{d=1}^{D}\prod_{j=1}^{N_d} P\left(z_{d,j} \mid \theta_d\right) P\left(w_{d,j} \mid \phi_{z_{d,j}}\right)\right] \times \left[\prod_{t=1}^{T}\prod_{c=1}^{WC} P\left(x_{t,c} \mid N_{t,c},\phi_{t,c}\right)\right] \tag{1}$$

where w and x are matrices of the data, and θ_d and ϕ_t are vector model parameters to be estimated. Specifically, θ_d has elements which are the topic proportions in document d, and ϕ_t has elements $\phi_{t,c}$ which represent the probability of word c in topic t. The vectors, α and β_t, contain prior parameters that might be assumed to have all its elements equaling the same values, α_0 and β_0. Also, effective constants include N the vector of document lengths with elements N_d and the $N_{t,c}$ are the SME counts for topic t and word c.

The constituent parts of Equation (1) are the Dirichlet, categorical, and binomial densities. Collecting all the parts, Equation (1) becomes:

$$P(z,w,x,\theta,\phi \mid N,N_t,\alpha,\beta) = \left[\prod_{t=1}^{T}\frac{\Gamma\left(\sum_{c=1}^{WC}\beta_{t,c}\right)}{\left(\prod_{c=1}^{WC}\Gamma(\beta_{t,c})\right)}\prod_{c=1}^{WC}\theta_{t,c}^{\beta_{t,c}-1}\right]\left[\prod_{d=1}^{D}\frac{1}{B(\alpha_d)}\prod_{t=1}^{T}\theta_{d,t}^{\alpha_{d,t}-1}\right] \times \left[\prod_{d=1}^{D}\prod_{t=1}^{T}\theta_{d,t}^{n^t_{d,(.)}}\right]\times\left[\prod_{t=1}^{T}\prod_{c=1}^{WC}\phi_{t,c}^{n^t_{(.),c}}\right] \times \left[\prod_{t=1}^{T}\prod_{c=1}^{WC}\binom{N_{t,c}}{x_{t,c}}\phi_{t,c}^{x_{t,c}}(1-\phi_{t,c})^{N_{t,c}-x_{t,c}}\right] \tag{2}$$

where

$$n^t_{d,(.)} = \sum_{j=1}^{N_d}\sum_{c=1}^{WC} I\left(z_{d,j}=t \;\; w_{d,j}=c\right), \tag{3}$$

$$n^t_{(.),c} = \sum_{d=1}^{D}\sum_{j=1}^{N_d} I\left(z_{d,j}=t \;\; w_{d,j}=c\right), \tag{4}$$

and $\Gamma(\ldots)$ is the gamma function. Notice that if $N_{t,c}=0$ for all $t=1,\ldots,T$ and $c=1,\ldots,WC$, then the last term becomes unity and the model reduces to the density for LDA.

The SMERT model is written graphically in Figure 1, which shows the LDA on the left-hand side. The right-hand side relates to the boost table. We call the right-hand side a 'handle' because it permits the SME to manipulate the model. The left-hand side is generally considerably larger than the right-hand side, because the number of documents could be large, for example, 10,000. Also, the number of words in each document could be large, for example, 1,000. Then, the left-hand side would include 10,000,000 words, w. At the same time, only a small number of words might be boosted or zapped for each topic, for example, 1. In this scenario and with $T=10$ topics, there would be only 100 nonzero $N_{t,c}$. Because of the effective access to the topic definitions, ϕ, the effect in shaping the topics could be substantial. Entering 100 pairs of numbers ($N_{t,c}$ and $x_{t,c}$) is a relatively small task compared with generating multiple documents and/or tagging a substantial fraction of 10,000 documents.

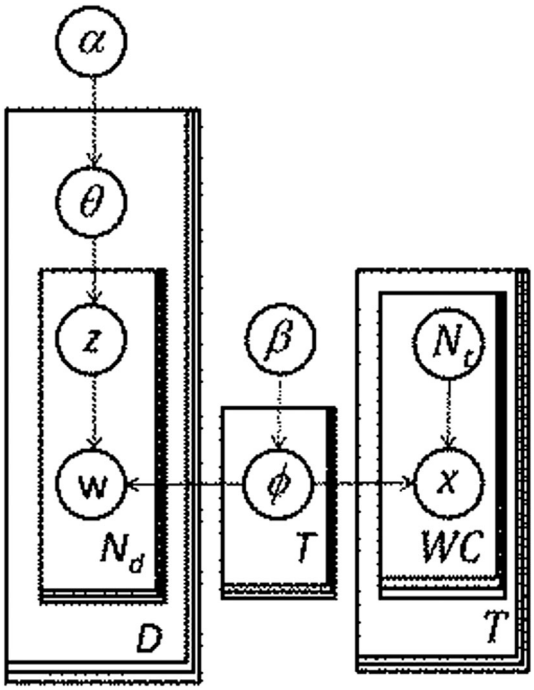

Figure 1. Graphical model representation of subject matter expert refined topic (SMERT). The boxes are 'plates' representing replicates. The left-hand side is latent Dirichlet allocation (LDA). The plates on the right-hand side are called a 'handle' because they permit the subject matter expert (SME) to control the model. The outer plates are repeated choices over documents and topics. The inner plates represent repeated choices over words in documents (left-hand side) and the dictionary (right-hand side).

3.3. Relationship of SMERT with latent Dirichlet allocation

As noted previously, the SMERT model is a generalization of LDA from [1]. LDA models and sampling formulas correspond to the special case with $N_{t,c} = 0$ for all topics t and words c. In the LDA special case, the handle on the right-hand side of Figure 1 is eliminated. Note that similar generalizations of other topic models are also possible and might also be called SMERT models.

Yet, the relationship between LDA and SMERT is more complicated than simple generalization. This follows because, in our experience, the SMEs cannot confidently provide the binomial data in the form of boosts and zaps until after inspecting the results of a preliminary analysis performed using LDA. Then, the 'high level' data provided becomes part of the prior for subsequent analyses.

Also, in certain situations the binomial data from the SMEs may be considered to be too subjective and not trustworthy. Then, the results from LDA might be regarded as preferable. Therefore, the SMERT and LDA are both alternative methods with advantages as well as methods that can work in combination. SMERT can be applied repeatedly after an initial modeling effort with LDA and rounds of experimentation involving SMEs. In the next section, we describe the real world application of LDA followed by a single round of data collection and SMERT modeling.

4. Inference using the collapsed Gibbs sampler

We have described the motivation behind SMERT and possible advantages in permitting the SME to influence the model with reasonable effort. In this section, we turn our attention to procedures for parameter estimation.

The collapsed Gibbs sampler offers a way to estimate the posterior mean of the model parameters using a sequential simulation approach involving a relatively small computational effort compared with ordinary Gibbs sampling. Griffiths and Steyvers [14] derived the collapsed Gibbs sampler for LDA. Our derivation follows [15].

In this section, we derive the collapsed Gibbs sampler for the SMERT model in Equations (1) and (2) for the case in which $N_{t,c} > x_{t,c}$ for only a single word (c) in each topic (t). We also describe the ability of the derived sampling equation for approximate posterior mean estimation in the more general case.

4.1. The exact sampler for the single zapping per topic case

Collapsed Gibbs sampling involves generating a single topic assignment, $z_{d,j}$, for the jth word in the dth document given all other variables in the model are assumed to be known. It is called 'collapsed', because the dependencies on the θ_d and ϕ_t variables are marginalized or integrated over. Given samples of the $z_{d,j}$, it is possible to estimate the posterior mean values for all the model parameters. The iterative sampling of the $z_{d,j}$ cycles from word to word and document to document until there is acceptable convergence for the posterior mean parameter estimates.

The collapsed Gibbs sampling process requires a formula for the probability of a topic assignment, $z_{d,j}$, for word j in document d, given all the other topic assignments and parameters are known. We use $z_{-(d,n)}$ and $w_{-(d,n)}$ to denote vectors including all topic assignments and words with the current topic assignment ($z_{d,n}$) and word ($w_{d,n}$) omitted respectively. We also use $w, \alpha, \beta, N_d, N_t$, and x to represent concisely all words in all documents model parameters.

Using the definition of conditional probability and assembling and rewriting:

$$\begin{aligned} P\left(z_{d,j} \mid z_{-(d,j)}, w, \alpha, \beta, N_d, N_t, x\right) &= \frac{P\left(z_{d,j}, z_{-(d,j)}, w \mid \alpha, \beta, N_d, N_t, x\right)}{P\left(z_{-(d,j)}, w \mid \alpha, \beta, N_d, N_t, x\right)} \\ &\propto P\left(z_{d,j}, z_{-(d,j)}, w \mid \alpha, \beta, N_d, N_t, x\right) \\ &= P\left(z, w \mid \alpha, \beta, N_d, N_t, x\right) \\ &= \int\int P\left(z, w, x, \theta, \phi \mid N, N_t, \alpha, \beta\right) d\theta d\phi \end{aligned} \tag{5}$$

The last line of Equation (5) derives from the rule of total probability. Therefore, the sampling probabilities are proportional to the marginal of the density in Equation (2). We only need to calculate $P(z_{d,j} \mid \ldots)$ up to a proportionality constant because we can derive the quantity for all topics and then normalize to calculate the probabilities. The integrals in Equation (5) are over unit simplexes for each of the relevant probabilities. Rearranging and moving the constant terms outside the integrals

$$\begin{aligned} P\left(z_{d,j} \mid z_{-(d,j)}, w, \alpha, \beta, N_d, N_t, x\right) = &\left[\prod_{d=1}^{D} \frac{1}{B(\alpha_d)}\right] \times \int\int \left[\prod_{d=1}^{D}\prod_{t=1}^{T} \theta_{d,t}^{n^t_{d,(.)}+\alpha_{d,t}-1}\right] d\theta \\ &\times \left[\prod_{t=1}^{T} \frac{\Gamma\left(\sum_{c=1}^{WC} \beta_{t,c}\right)}{\left(\prod_{c=1}^{WC} \Gamma(\beta_{t,c})\right.}\right] \\ &\times \int\int \prod_{t=1}^{T}\prod_{c=1}^{WC} \binom{N_{t,c}}{x_{t,c}} \phi_{t,c}^{\beta_{t,c}-1+n^t_{(.),c}+x_{t,c}} \\ &\times (1-\phi_{t,c})^{N_{t,c}-x_{t,c}} d\theta \end{aligned} \tag{6}$$

Here, we focus on the case in which there is only a single word, c_t, for each topic (t) which has $N_{t,c} > x_{t,c}$. For all other words for that topic, $N_{t,c} = x_{t,c}$. Therefore, in Equation (2), there is only a single term involving $(1-\phi_{t,c})$ for each t. With these assumptions the marginal becomes

$$\begin{aligned} &= \int\int P\left(z_{d,j} \mid z_{-(d,j)}, w, \alpha, \beta, N_d, N_t, x\right) d\theta d\phi \\ &\propto \int\int \left[\prod_{d=1}^{D}\prod_{t=1}^{T} \theta_{d,t}^{n^t_{d,(.)}+\alpha_{d,t}-1}\right] d\theta \\ &\quad \times \int\int \prod_{t=1}^{T} \binom{N_{t,c}}{x_{t,c}} (1-\phi_{t,c})^{N_{t,c}-x_{t,c}} \\ &\quad \times \prod_{c=1}^{WC} \phi_{t,c}^{\beta_{t,c}-1+n^t_{(.),c}+x_{t,c}} d\phi \end{aligned} \tag{7}$$

Note that the integral over $d\theta$ is unaffected by the boosting table. Therefore, the sampling probability is proportional to the same sum over topics as in LDA. Also, with a single zap per topic, the integral for each topic over the unit simplex is a classic integral with an anti-derivative. Additional details are available in [16].

The resulting sampling formula divides into two cases depending on the topic, t : (1)$w_{d,j} = c_t$ and (2)$w_{d,j} \neq c_t$. We use the delta function, δ, which equals 1 if the arguments are equal and zero otherwise. Substituting in the anti-derivatives and the sums ($n^{-(d,j)}_{z_{d,j},j,*}$ and $n^{-(t,c)}_{z_{d,j},*,w_{d,j}}$) over the topic assignments omitting the current one using the definitions in Equations (3) and (4), we derive for the topic assignment, $z_{d,j}$, for word j in document d:

$$\begin{aligned} &P\left(z_{d,j} \mid z_{-(d,j)}, w, \alpha, \beta, N_d, N_t, x\right) \\ &\propto \left(n^{-(d,j)}_{z_{d,j},j,*} + \alpha_{z_{d,j}}\right) \times \frac{\beta_{w_{d,j}} + n^{-(t,c)}_{z_{d,j},*,w_{d,j}} + x_{z_{d,j},c}}{N_{z_{a,b},c_{z_{a,b}}} - x_{z_{a,b},w_{z_{a,b}}} + \sum_{k=1}^{WC}\left(\beta_{z_{d,k}} + n^{-(t,c)}_{z_{d,j},*,w_{d,k}} + x_{z_{d,k},w_{d,k}}\right.} \\ &\times \left(1 + \frac{N_{z_{a,b},c_{z_{a,b}}} - x_{z_{a,b},c_{z_{a,b}}}}{-\beta_{w_{d,j}} - n^{-(t,c)}_{z_{d,j},*,w_{d,j}} - x_{z_{d,j},w_{d,j}} + \sum_{k=1}^{WC}\beta_k + n^{-t,c}_{z_{d,j},*,k} + x_{z_{d,j},k}}\right)^{1-\delta_{c_{z_{a,b}}=w_{a,b}}} \end{aligned} \tag{8}$$

We calculate the results in Equation (8) for all the topics and normalize and then sample $z_{d,j}$ from the categorical distribution. It can be checked that, for $N_{t,c} = x_{t,c} = 0$ case, the aforementioned reduces to the ordinary LDA collapsed Gibbs sampling formula ([14]).

4.2. Posterior mean values

After iterative sampling of the topic definition has achieved approximate convergence (discussed further subsequently) the posterior mean values that define the topic definitions, $E(\phi_{t,c} \mid x, w)$, and document assignments to topics, $E(\theta_d \mid x, w)$, can be estimated. As mentioned previously, the integral over $d\theta$ is unaffected by the boosting table. Therefore, the posterior mean topic probabilities or proportions are simply proportional to $\left(n^t_{(.),(.)} + \alpha\right)$, where $n^t_{(.),(.)}$ is the count of all words assigned to topic t and α is the prior parameter.

We next examine the posterior density function for the topic definitions, $\phi_{t,c}$, for $c = 1, \ldots, WC$ and $t = 1, \ldots, T$. The normalized joint posterior density function for a given topic, t, is:

$$\begin{aligned} f(\phi \mid t, N, x, w, \beta) = &\frac{\Gamma\left(\Delta_{t,c_t} + \sum_{c=1}^{WC} q_{t,c}\right)}{\prod_{c=1}^{WC}\Gamma(q_{t,c})} \times \frac{\Gamma\left(-q_{t,c_t} + \sum_{c=1}^{WC} q_{t,c}\right)}{\Gamma\left(\Delta_{t,c_t} - q_{t,c_t} + \sum_{c=1}^{WC} q_{t,c}\right)} \\ &\times \prod_{c=1}^{WC} \phi_{t,c}^{q_{t,c}-1}\left(1 - \phi_{t,c_t}\right)^{\Delta_{t,c_t}} \end{aligned} \tag{9}$$

The marginal for ϕ_{t,c_t} is the beta distribution derived by integrating over all $\phi_{t,c}$ other than ϕ_{t,c_t}:

$$\begin{aligned} f\left(\phi_{t,c_t} \mid t, N, x, w, \beta\right) = &\frac{\Gamma\left(\Delta_{t,c_t} + \sum_{c=1}^{WC} q_{t,c}\right)}{\Gamma(q_{t,c})\Gamma\left(\Delta_{t,c_t} - q_{t,c_t} + \sum_{c=1}^{WC} q_{t,c}\right)} \\ &\times \phi_{t,c}^{q_{t,c}-1}\left(1 - \phi_{t,c_t}\right)^{\Delta_{t,c_t} - q_{t,c_t} + \sum_{c=1}^{WC} q_{t,c}} \end{aligned} \tag{10}$$

Multiplying by ϕ_{t,c_t} and integrating over the support [0,1], the posterior mean for $\phi_{t,c}$ is, after rearranging and applying $x\Gamma(x) = \Gamma(x+1)$

$$E\left(\phi_{t,c_t} \mid t, N, x, w, \beta\right) = \frac{q_{t,c_t}}{\Delta_{t,c_t} - q_{t,c_t} + \sum_{c=1}^{WC} q_{t,c}} \tag{11}$$

which is similar to the LDA formula. This formula applies for all words in topics that are either the zapping word or any other word if there is no zapping word. Unfortunately, the marginal for $\phi_{t,c}$ with $c \neq c_t$ for topics having a zapping word is more complicated and is most easily expressed using so-called 'regularized hypergeometric'.

$$\begin{aligned} f\left(\phi_{t,c} \mid t, N, x, w, \beta\right) = &\frac{\Gamma\left(-q_{t,c_t} + \sum_{c=1}^{WC} q_{t,c}\right)}{\Gamma(q_{t,c})} \times \frac{\Gamma\left(\Delta_{t,c_t} + \sum_{c=1}^{WC} q_{t,c}\right)}{\Gamma\left(\Delta_{t,c_t} - q_{t,c_t} + \sum_{c=1}^{WC} q_{t,c}\right)} \\ &\times \phi_{t,c}^{q_{t,c}-1}\left(1 - \phi_{t,c_t}\right)^{Q_{t,c}-1} {}_2\tilde{F}_1\left(q_{t,c}, -\Delta_{t,c}; Q; 1 - \phi_{t,c}\right) \end{aligned}$$

Table I. First three records with the text included after the manual processing. Note that PH refers to policy holder or the end customer.

Time	Month	Type	Group	Category	Description	strResolution	Notes
Short	8	Home	Billing	Collection call	Collection issues	Resolved	PH called to wonder when his bill would be arriving to his house. PH was informed he was set up for auto draft. PH did not remember when he signed up. PH asked to change payment method to monthly. Policy was changed.
Short	8	Other	Policy management	Customer information	Update other	Resolved	Lienholder called to change the mortgage company on the PH's policy.
Short	8						Medical center calling to confirm insurance and check claim status.

where $Q_{t,c} = -q_{t,c} + \sum_{j=1}^{WC} q_{t,j}$ and where the ${}_?\tilde{F}_1$ is the regularized hypergeometric function. For simplicity, we quote the results assuming integral $q_{t,c}$ but the results generalize.

$$E\left(\phi_{t,c} \mid t, N, x, w, \beta\right) = \frac{q_{t,c}}{-q_{t,c_t} + \sum_{c=1}^{WC} q_{t,c}} \left(1 - \frac{q_{t,c}}{\Delta_{t,c_t} + \sum_{c=1}^{WC} q_{t,c}}\right) \tag{12}$$

It can be checked that the formula reduces to the $c \neq c_t$ case above if $N_{t,c} = x_{t,c}$.

4.3. Multiple zaps per topic cases

We have also studied cases for which more than a single word is zapped per topic. In these cases, the aforementioned formulas are not exact. Yet, in our computational experiments in the online supplement and [16], we have found that the aforementioned formulas can provide, at least in specific examples that we studied, reasonable estimates of the ground truth distribution.

5. Application of call center quality assurance data

In this section, we describe a real world case study involving 2,378 records of conversations between call center service representatives at an insurance company and callers. The purpose of applying SMERT methods is to prioritize caller issues and plan improvement projects. Therefore, the goals are both to identify accurate and interpretable topics and to estimate the proportions of these topics accurately. Additional details about this case study can be found in [17].

The data set includes the eight fields indicated by the first three records in Table I. To apply SMERT, we simply appended the field heading to each word. Therefore, `Time_Short` and `Notes_Short` are different words because the word 'short' is in different 'fields' or columns of the database. We applied natural language processing using methods from [18] to remove the stopping words, to create stemmed words, and to assign indices for the word stems. The results from natural language processing are indicated in Table II.

5.1. Latent Dirichlet allocations

To prepare our SME to make a boosting table, we applied LDA using the collapsed Gibbs sampling from [14] in Equation (8) with $N_{t,c} = 0$ for all topics t and words c. We left out 10% of the data and applied 5, 10, 15, and 20 topics. We found perplexity values of 205.1, 211.4, 222.9, and 198.7, respectively. This indicated noise from our Gibbs sampling and relative

Table II. The first three records from Table I after the application of natural language processing methods. The words are replaced with their associated number values in the dictionary. Only the first 10 word labels are shown.

Number of words	W1	W2	W3	W4	W5	W6	W7	W8	W9	W10
28	766	185	769	179	14	82	107	762	553	252
15	766	185	182	181	17	26	171	762	456	252
8	766	185	262	252	288	267	269	661		

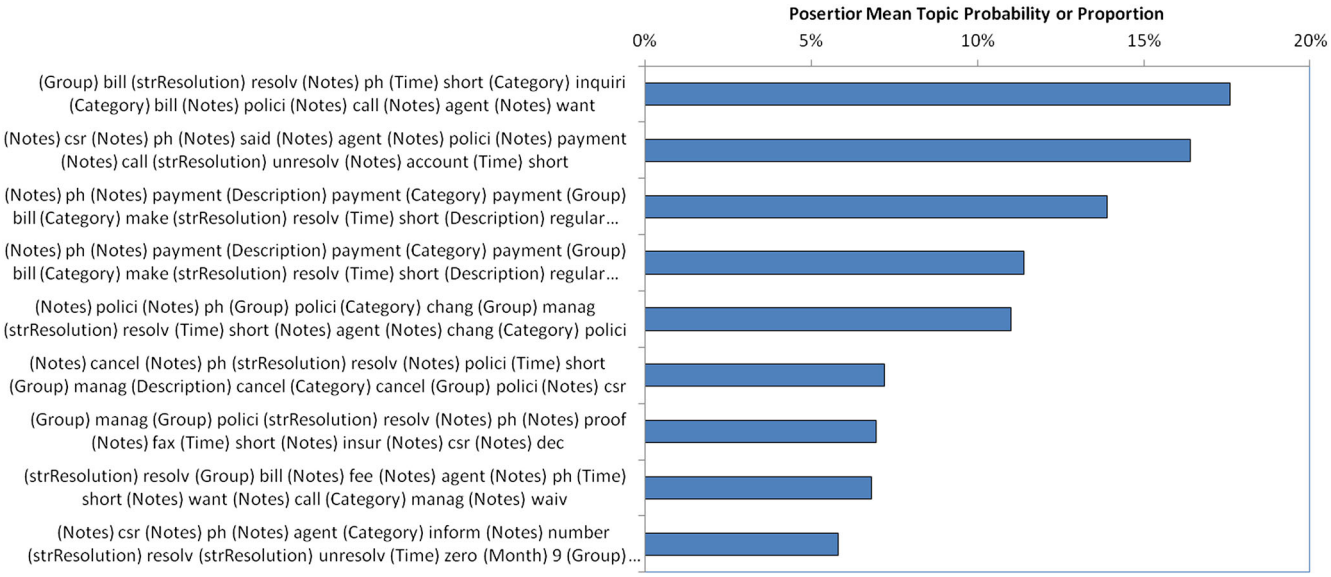

Figure 2. Pareto chart of latent Dirichlet allocation topic definitions sorted by estimated mean proportion.

stability. Therefore, with some arbitrariness, we used 10 topics. Also, we used the prior parameter settings from [12] or [16], that is, $\alpha = 0.15/T$ and $\beta = 2/WC = 0.0026$ because these were argued to foster relatively accurate topic proportions and definitions, which were our objectives.

After 1,000 iterations of topic assignments using Equation (18), approximate convergence was reached. Then, the posterior mean values for the topic definition probabilities, ϕ, and proportions, α, were estimated using Equation (11) and the $\left(n^{t}_{(.),(.)} + \alpha\right)$ proportionality, respectively. Figure 2 and Table III show the top 10 words for each topic corresponding to the highest estimated posterior mean probability. Figure 2 shows a 'Pareto chart' in which the topics are ordered in descending estimated probability. Allen and Xiong [12] argued that this information can be helpful in prioritizing improvement efforts because the topics can correspond to addressable system improvement issues.

We showed the results to our SME and asked for a possible interpretation of each topic. The descriptions are shown in Table III. We also asked the SME which words (if any) did not belong to specific topics.

5.2. *Subject matter expert refined topic models*

From inspection of the results in Table III, our SME developed the boosting results in Tables IV and V. The tables include 90 high-level data points. The SME explained the high-level data by concluding that there was an inaccurate confounding of the resolved and unresolved issues. This occurred because the topic definitions in Table III mixed the resolved and unresolved discussion words together such that interpreting specific topics as relating to either type is ambiguous. Specifically, topics 2, 9, and 10 involved high posterior probabilities for both resolved and unresolved fields. The SME believed that, based on knowledge of the domain, unresolved and resolved calls generally do not share characteristics. Moreover, separating the two types of calls is important from a system improvement point of view, because the unresolved calls were generally more relevant.

The data in Tables IV and V include boosts of eight words from each LDA topic definitions in Table III. The '`strResolution resolv`' words were zapped in topics 2 and 10 and the '`strResolution_unresolv`' words were zapped in all the other topics. This separated the topics, so that each relates either to a resolved or an unresolved issue.

Table III. The top 10 words (mean probability) for the topics and descriptions.

Topic Description (Top Words)	Probability	Possible interpretation
(Group) bill (strResolution) resolv (Notes) ph (Time) short (Category) inquiri (Category) bill (Notes) polici (Notes) call (Notes) agent (Notes) want	0.176	The policy holder or agent wanted to inquire about their bill which was shortly resolved.
(Notes) csr (Notes) ph (Notes) said (Notes) agent (Notes) polici (Notes) payment (Notes) call (strResolution) unresolv (Notes) account (Time) short	0.164	An agent said something about the policy holders' payment on their account leading often to an unresolved issue.
(Notes) ph (Notes) payment (Description) payment (Category) payment (Group) bill (Category) make (strResolution) resolv (Time) short (Description) regular (Notes) make	0.139	The policy holder had a regular payment issue on their bill which was shortly resolved.
(Notes) polici (Notes) ph (Group) polici (Category) chang (Group) manag (strResolution) resolv (Time) short (Notes) agent (Notes) chang (Category) polici	0.114	The policy holder changed the policy possibly involving an agent which was shortly resolved.
(strResolution) resolv (Group) manag (Group) polici (Category) coverag (Category) explain (Notes) verifi (Notes) polici (Notes) call (Time) short (Description) coverag	0.110	There was a policy explanation of coverage or verification which was shortly resolved.
(Notes) cancel (Notes) ph (strResolution) resolv (Notes) polici (Time) short (Group) manag (Description) cancel (Category) cancel (Group) polici (Notes) csr	0.072	The policy holder is canceling his or her policy which is resolved.
(Group) manag (Group) polici (strResolution) resolv (Notes) ph (Notes) proof (Notes) fax (Time) short (Notes) insur (Notes) csr (Notes) dec	0.070	There was a policy holder proof of insurance issue which was shortly resolved possibly with faxing.
(strResolution) resolv (Group) bill (Notes) fee (Notes) agent (Notes) ph (Time) short (Notes) want (Notes) call (Category) manag (Notes) waiv	0.068	The agent or possibly policy holder wanted a fee or other issue to be waived.
(Notes) csr (Notes) ph (Notes) agent (Category) inform (Notes) number (strResolution) resolv (strResolution) unresolv (Time) zero (Month) 9 (Group) administr	0.058	The agent and/or policy holder wanted customer information including numbers which was **often** unresolved largely in September (month 9).
(Notes) ph (Notes) claim (Notes) call (Time) short (Notes) transfer (Group) claim (strResolution) unresolv (Month) 9 (strResolution) resolv (Notes) need	0.031	There was a claim possibly involved in a transfer which was **often** unresolved.

Table IV. High-level data (HLD) for the subject matter expert refined topic (SMERT) application for topics 1–5.

Topic	Word index	x	N	Action	Word
1	179	2	2	Boost	(Group) bill
1	762	2	2	Boost	(strResolution) resolv
1	553	2	2	Boost	(Notes) ph
1	766	2	2	Boost	(Time) short
1	27	2	2	Boost	(Category) inquiri
1	7	2	2	Boost	(Category) bill
1	558	2	2	Boost	(Notes) polici
1	252	2	2	Boost	(Notes) call
1	763	0	1000000	Zap	(strResolution) unresolv
2	302	2	2	Boost	(Notes) csr
2	553	2	2	Boost	(Notes) ph
2	627	2	2	Boost	(Notes) said
2	214	2	2	Boost	(Notes) agent
2	558	2	2	Boost	(Notes) polici
2	545	2	2	Boost	(Notes) payment
2	252	2	2	Boost	(Notes) call
2	763	2	2	Boost	(strResolution) unresolv
2	762	0	1000000	Zap	(strResolution) resolv
3	553	2	2	Boost	(Notes) ph
3	545	2	2	Boost	(Notes) payment
3	130	2	2	Boost	(Description) payment
3	44	2	2	Boost	(Category) payment
3	179	2	2	Boost	(Group) bill
3	31	2	2	Boost	(Category) make
3	762	2	2	Boost	(strResolution) resolv
3	766	2	2	Boost	(Time) short
3	763	0	1000000	Zap	(strResolution) unresolv
4	558	2	2	Boost	(Notes) polici
4	553	2	2	Boost	(Notes) ph
4	182	2	2	Boost	(Group) polici
4	11	2	2	Boost	(Category) chang
4	181	2	2	Boost	(Group) manag
4	762	2	2	Boost	(strResolution) resolv
4	766	2	2	Boost	(Time) short
4	214	2	2	Boost	(Notes) agent
4	763	0	1000000	Zap	(strResolution) unresolv
5	762	2	2	Boost	(strResolution) resolv
5	181	2	2	Boost	(Group) manag
5	182	2	2	Boost	(Group) polici
5	16	2	2	Boost	(Category) coverag
5	21	2	2	Boost	(Category) explain
5	723	2	2	Boost	(Notes) verifi
5	558	2	2	Boost	(Notes) polici
5	252	2	2	Boost	(Notes) call
5	763	0	1000000	Zap	(strResolution) unresolv

We applied 1,000 iterations using Equation (8) sampling topic assignments for each word in the regular corpus. The high-level data entered into the model just as previous data would in Bayesian modeling. Therefore, the high-level data became part of the prior. The resulting posterior mean topic definition and proportion estimates are given in Table VI and Figure 4. The SME again analyzed the model output and generated the interpretations also shown in Table VI on the right-hand side. The SME then generated the following findings, which were communicated to representatives of the call center:

1. The most common topic, with 21.1% of posterior mean words, were related to agent or policy holder statements about their bill, payment, or account leading usually to an unresolved issue.
2. Bill, payment, and account issues accounted for virtually all unresolved issues.
3. Policy cancelation calls were generally short and resolved (topic 6 using the original, LDA, ordering).

Table V. High-level data (HLD) for the subject matter expert refined topic (SMERT) application for topics 6-10.

Topic	Word Index	x	N	Action	Word
6	255	2	2	Boost	(Notes) cancel
6	553	2	2	Boost	(Notes) ph
6	762	2	2	Boost	(strResolution) resolv
6	558	2	2	Boost	(Notes) polici
6	766	2	2	Boost	(Time) short
6	181	2	2	Boost	(Group) manag
6	76	2	2	Boost	(Description) cancel
6	9	2	2	Boost	(Category) cancel
6	763	0	1000000	Zap	(strResolution) unresolv
7	181	2	2	Boost	(Group) manag
7	182	2	2	Boost	(Group) polici
7	762	2	2	Boost	(strResolution) resolv
7	553	2	2	Boost	(Notes) ph
7	568	2	2	Boost	(Notes) proof
7	366	2	2	Boost	(Notes) fax
7	766	2	2	Boost	(Time) short
7	432	2	2	Boost	(Notes) insur
7	763	0	1000000	Zap	(strResolution) unresolv
8	762	2	2	Boost	(strResolution) resolv
8	179	2	2	Boost	(Group) bill
8	368	2	2	Boost	(Notes) fee
8	214	2	2	Boost	(Notes) agent
8	553	2	2	Boost	(Notes) ph
8	766	2	2	Boost	(Time) short
8	734	2	2	Boost	(Notes) want
8	252	2	2	Boost	(Notes) call
8	763	0	1000000	Zap	(strResolution) unresolv
9	302	2	2	Boost	(Notes) csr
9	553	2	2	Boost	(Notes) ph
9	214	2	2	Boost	(Notes) agent
9	26	2	2	Boost	(Category) inform
9	517	2	2	Boost	(Notes) number
9	762	2	2	Boost	(strResolution) resolv
9	763	0	1000000	Zap	(strResolution) unresolv
9	767	2	2	Boost	(Time) zero
9	186	2	2	Boost	(Month) 9
10	553	2	2	Boost	(Notes) ph
10	269	2	2	Boost	(Notes) claim
10	252	2	2	Boost	(Notes) call
10	766	2	2	Boost	(Time) short
10	696	2	2	Boost	(Notes) transfer
10	180	2	2	Boost	(Group) claim
10	763	2	2	Boost	(strResolution) unresolv
10	186	2	2	Boost	(Month) 9
10	762	0	1000000	Zap	(strResolution) resolv

4. Requests for waivers were generally short and resolved (topic 8 using the original, LDA ordering).
5. Approximately 12.4% of calls related to verifications and explanations.

Putting these findings together, we had a potential basis for recommending improvement activities. For example, a new program to automate verifications and explanations of basic information might be able to eliminate 12.4% of the call volume. Moreover, the relevance of percentage information and topic accuracy motivates our emphasis in the next sections on distributional accuracy measures. The Pareto chart in Figure 3 is a summary of possibility the most interesting information for improvement analysts from the database. The listing of words has been reinterpreted using the descriptions in Table VI.

Table VI. The subject matter expert refined topic (SMERT) model top 10 words for the topics and possible descriptions.

Topic description (top words)	Probability	Possible interpretation
(Group) polici (Group) manag (strResolution) resolv (Category) explain (Category) coverag (Notes) call (Notes) verifi (Time) short (Notes) polici (Description) coverag	0.124	An explanation or verification was provided about the policy and its coverage and shortly resolved.
(Notes) csr (Notes) ph (Notes) agent (Notes) said (Notes) polici (Group) bill (Notes) account (Notes) payment (Notes) call (Time) short (Month) 9 (Notes) refund (strResolution) unresolv	0.211	An agent or policy holder said an issue about their bill/payment/account which was usually unresolved.
(Notes) ph (Notes) payment (Group) bill (Category) payment (strResolution) resolv (Description) payment (Time) short (Notes) call (Notes) csr (Notes) make	0.07	The policy holder had a regular payment issue on their bill which was shortly resolved.
(Notes) ph (strResolution) resolv (Category) inquiri (Group) bill (Category) bill (Time) short (Notes) want (Notes) call (Month) 9 (Notes) know	0.103	The policy holder inquired about the bill wanting to know something particularly in September which was shortly resolved.
(Notes) ph (Notes) csr (Notes) polici (Time) short (strResolution) resolv (Notes) want (Notes) agent (Notes) call (Group) bill (Month) 9	0.08	The agent likely wanted a bill issue addressed which was shortly resolved.
(Notes) polici (Notes) cancel (strResolution) resolv (Notes) ph (Notes) agent (Time) short (Notes) call (Notes) csr (Month) 9 (Group) manag	0.113	The policy holder or agent is canceling his or her policy which is resolved.
(Notes) ph (strResolution) resolv (Notes) proof (Group) manag (Time) short (Group) polici (Notes) csr (Category) issu (Notes) insur (Group) administr	0.059	There was a policy holder proof of insurance issue which was shortly resolved possibly with faxing.
(strResolution) resolv (Notes) fee (Notes) ph (Notes) agent (Notes) call (Notes) waiv (Group) administr (Notes) csr (Time) short (Category) inform	0.051	The agent or possibly policy holder wanted a fee or other issue to be waived.
(Group) polici (Group) manag (Notes) polici (strResolution) resolv (Category) polici (Category) chang (Time) short (Notes) ph (Notes) agent (Notes) csr	0.076	The agent and/or policy holder wanted customer information including numbers which was resolved.
(Notes) ph (Notes) payment (Description) payment (Group) bill (Category) payment (Category) make (Description) regular (Time) short (Notes) make (Notes) call	0.113	There was a claim possibly involved in a transfer which was sometimes unresolved.

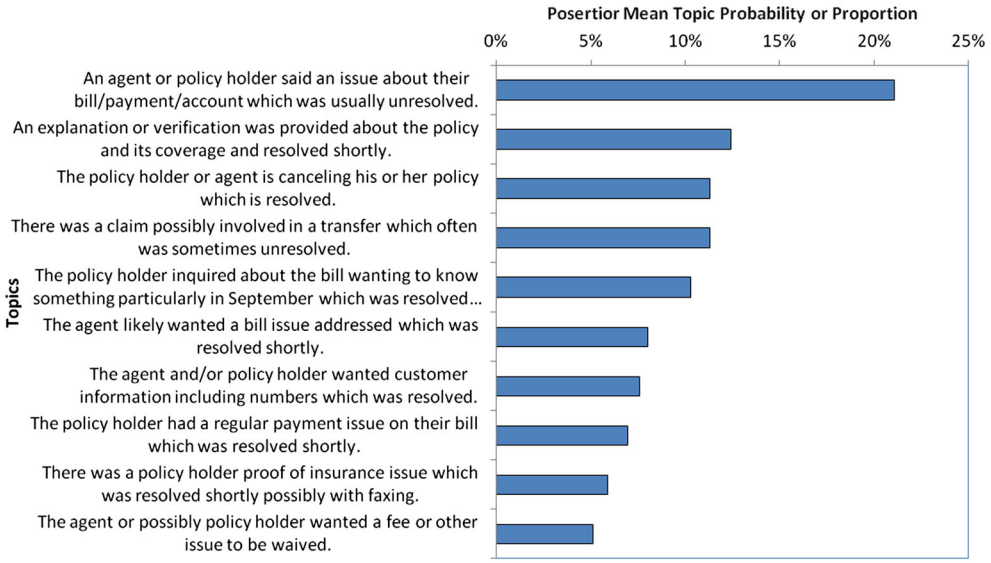

Figure 3. Pareto chart of subject matter expert refined topic (SMERT) interpreted topics.

5.3. Numerical comparison

In an online supplement, a computational experiment is described involving four cases or corpora, five alternative probabilistic clustering methods including two variations of SMERT models, and five performance measures or responses. The experiment focuses on two responses which relate to the accuracy of the fitted distribution both in relation to identifying topic definitions (ϕ) and topic proportions (α). Both measures relate to the minimum average Kullback Leibler divergence (AKLD), which is an average distributional distance between the best matched fitted distribution and the ground truth. The fitted model must be best matched because there are many possible orderings of the topics. Table VI in the supplement shows the experimental design and measured response values.

The results from analysis of variance models of the responses related to topic definitional accuracy, that is, $AKLD(\phi)$, are in Table VII. The results suggest that document diversity, topic overlap, and the method type main effects significantly influence topic definition accuracy. Also, the interaction between topic overlap and the method is significant. These results support the findings:

1. Document diversity significantly worsens the ability of all the methods to identify accurately cluster or topic definitions as illustrated in the interaction plots in Figure 5. Presumably, all the methods compared would perform better if documents were on only a single topic.
2. Topic overlap significantly improves the performance of the methods as shown in Figure 5. This occurs presumably because the methods are based on the assumption that clusters or topics will overlap. Specifically, the effect for LDA is actually greater than for the other methods. This likely relates to the influence of the prior on the modeling results. The non-informative prior in LDA implies an assumption of high topic overlap. We have not seen this effect presented in other articles.
3. The application of high-level data associated with SMERT methods results in an enhanced ability to define topics accurately, and more data yield further enhancement. Intuitively, methods based on additional accurate data should outperform other methods.
4. The fuzzy clustering methods and principal component analysis followed by fuzzy clustering methods both outperform LDA in average measures of topic accuracy. At least part of this is attributable to the high variation associated with applying LDA based on a single replicate noted by [19]. Apparently, fitting a fuzzy clustering method and then interpreting the results as a topic model results in improved accuracy compared with LDA.

The results from the analysis of variance on the measures of topic proportion accuracy, that is, $AKLD(\alpha)$, are shown in Table VIII. The only significant effects are document diversity and method. The results support the additional findings:

1. High levels of document diversity significantly harm the abilities of all the clustering methods to accurate estimate the proportions of words on specific topics. This effect is among the largest apparent in the interaction plots in the online supplement.

Table VII. ANOVA results relating to definition accuracy for AKLD(ϕ).

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Document Diversity	1	0.2607	0.2607	0.2607	25.7300	0.0000
Topic Overlap	1	0.0950	0.0950	0.0950	9.3700	0.0060
Method	4	1.2354	1.2354	0.3088	30.4900	0.0000
Document Diversity*Topic Overlap	1	0.0182	0.0182	0.0182	1.7900	0.1960
Document Diversity*Method	4	0.1095	0.1095	0.0274	2.7000	0.0600
Topic Overlap*Method	4	0.0674	0.0674	0.0169	1.6600	0.1980
Document Diversity*Topic Overlap	4	0.0185	0.0185	0.0046	0.4600	0.7670
Error	20	0.2026	0.2026	0.0101		
Total	39	2.0071				

Table VIII. ANOVA results relating to proportion accuracy, AKLD(α).

Source	DF	Seq SS	Adj SS	Adj MS	F	P
Document Diversity	1	0.0179	0.0179	0.0179	10.0900	0.0050
Topic Overlap	1	0.0003	0.0003	0.0003	0.1800	0.6790
Method	4	0.0182	0.0182	0.0046	2.5700	0.0690
Document Diversity*Topic Overlap	1	0.0004	0.0004	0.0004	0.2300	0.6340
Document Diversity*Method	4	0.0101	0.0101	0.0025	1.4200	0.2630
Topic Overlap*Method	4	0.0003	0.0003	0.0001	0.0500	0.9960
Document Diversity*Topic Overlap	4	0.0049	0.0049	0.0012	0.6900	0.6060
Error	20	0.0354	0.0354	0.0018		
Total	39	0.0876				

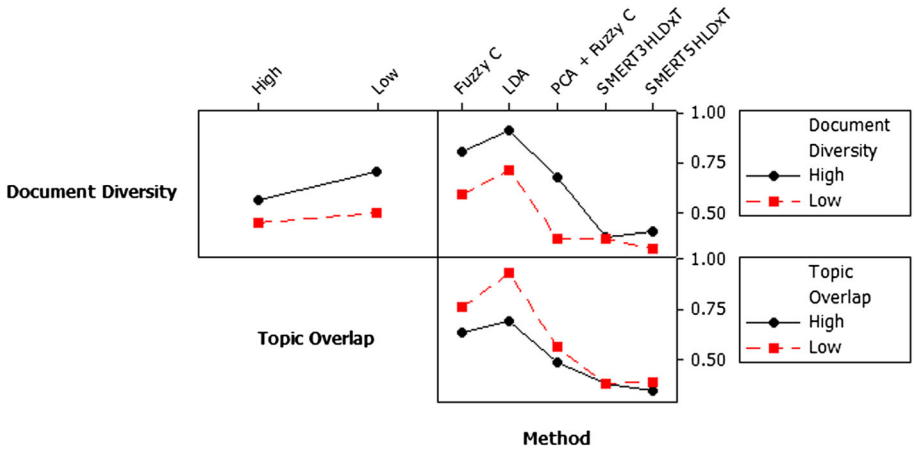

Figure 4. Interaction plot of the topic definition minimum average Kullback-Leibler divergence (AKLD).

2. The application of high-level data associated with SMERT methods results in an enhanced ability to estimate topic proportions accurately, and more data yields further enhancement. As for topic definitions, one would expect intuitively that methods based on additional accurate data should outperform other methods.
3. Fuzzy clustering methods and principal component analysis followed by fuzzy clustering methods both outperform LDA in average measures of topic proportion accuracy. Again, this likely relates to the noise from Gibbs sampling in LDA estimation.

As mentioned previously, the relative accuracy of SMERT models is expected. This follows because the models benefit from additional correct information about the ground truth distribution. Moreover, five high-level data per topic instead of three high-level data per topic results in a slight improvement. As noted previously, the LDA model is the SMERT model with zero high-level data per topic. The harmful and significant effect of document diversity on LDA performance in Figures 4 and 5 issomewhat surprising. Yet, the sampling formula in Equation (8) provides an indication of the importance of other words in the same document being on the same topic. If documents express a wide variety of topics, then LDA performance worsens. Adding correct high-level data, of course, mitigates this effect.

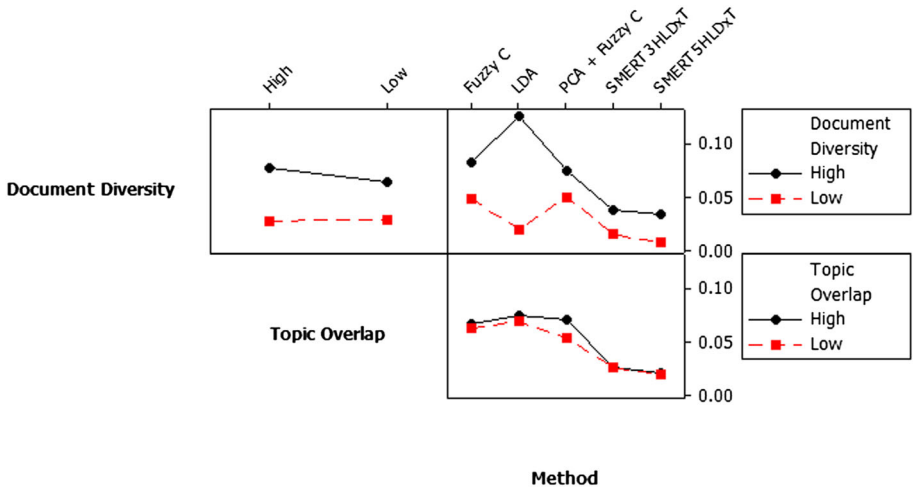

Figure 5. Interaction plot of the topic proportion minimum average Kullback-Leibler divergence (AKLD).

6. Discussion

We have proposed SMERT models, a generative probabilistic model for collections of discrete data which generalizes LDA. SMERT is based on supplementing the traditional corpus with results from binomial experiments involving a local SME. Therefore, SMERT models are topic models with both categorical and binomial inputs. We have also developed a collapsed Gibbs sampler for SMERT, which permits sampling from the Bayesian posterior distribution. Viewed as a Bayesian clustering technique, the high-level data is an informative prior.

Modeling involving informative priors may be regarded as overly subjective. Yet, we argue that 'boosting' or affirming words and 'zapping' or eliminating words from topic definitions are the type of data that SMEs are able to plausibly generate. Moreover, after inspecting the results from a preliminary LDA run, we have often found that, while many of the topics seemed appropriate and interpretable, others seemed to require editing. This motivated our development of SMERT methods. The inclusion of prior data makes SMERT similar in important ways to other informative clustering methods such as those in [13]. Yet, comparisons are complicated because of the varied forms that the additional information can take. We argue that having the SME provide information as a type of data offers some advantages in simplicity and transferability that other forms of information might not. We admit that, in our case study and applications, the high-level data were inputted with extreme values so that they functioned more like constraints than noisy data. Future implementations might, however, take different forms so that the high-level data are more like ordinary data, that is, involving counts much less than 1 million.

We have developed the collapsed Gibbs sampler for SMERT in an attempt to emulate the success of [14] methods for LDA. Yet, Gibbs sampling methods have computational limits and may be insufficient for corpora involving hundreds of thousands of records. Other approximate estimation methods including ones based on variational approaches such as those methods in [1] can also be developed for SMERT models.

We illustrated the application of the methods to the analysis of a call center in the insurance industry. We feel that this example provides evidence that the development of high-level data of relevance is possible for real world problems. Moreover, the goal of prioritizing actions based on topic proportion estimates motivated the evaluation of models based on accuracy measures.

We explored these accuracy measures using a computational experiment in which the ground truth for each of the four corpora was assumed to be known. We observed that document diversity, that is, having multiple topics in each document, worsened the performance of LDA. Also, predictably, the inclusion of correct high-level data into the SMERT model improved distributional accuracy and mitigated the effects of document diversity.

Compared with LDA or fuzzy 'c' clustering methods, SMERT models are relatively complicated. Yet, the idea of supplementing traditional corpora with data obtained from a local SME is simple and intuitive. Hypothetically, any type of distribution could be augmented using a 'handle' as we have done to LDA in creating the SMERT model. Therefore, extensions to LDA could become extensions to SMERT.

In addition, structures involving SME data such as the right-hand side in Figure 1 could be appended to document-topic proportion random variables (θ_d) instead or in addition to topic definition estimates (ϕ). If the 'handle' were added to the θ_d side of LDA, then the high-level data would relate to the number of words in a given document on a given topic. Adding handles to models results in losses in objectivity, of course. Yet, in our call center case study, the results seemed

relatively helpful, because we could divide out whether or not the customer issue was resolved. In general, permitting high-level data to edit estimated parameters promises to provide greater interpretability and usefulness for many types of probabilistic models.

Acknowledgements

We thank Soo Ho Lee for his insights and hard work. Also, inspiration and encouragement came from Tom Bishop, Emily Patterson, David Woods, and Ning Zheng.

References

1. Blei DM, Ng AY, Jordan MI. Latent Dirichlet allocation. *The Journal of Machine Learning Research* 2003; **3**:993–1022.
2. Rosen-Zvi M, Griffiths T, Steyvers M, Smyth P. The author-topic model for authors and documents. *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*. AUAI Press: Banff, Canada, 2004,487–494.
3. Blei DM, Lafferty JD. A correlated topic model of science. *The Annals of Applied Statistics* 2007:17–35.
4. Blei DM, McAuliffe JD. Supervised topic models. In *Advances in Neural Information Processing Systems*, J Platt, Koller Singer Y, R Roweis (eds). MIT Press: Cambridge, MA, 2008; 121–128.
5. Steyvers M, Smyth P, Chemuduganta C. Combining background knowledge and learned topics. *Topics in Cognitive Science* 2011; **3**(1):18–47.
6. Rosen-Zvi M, Chemudugunta C, Griffiths T, Smyth P, Steyvers M. Learning author-topic models from text corpora. *ACM Transactions on Information Systems (TOIS)* 2010; **28**(1):487–494.
7. Ramage D, Manning CD, Dumais S. Partially labeled topic models for interpretable text mining. *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM: San Diego, CA, 2011,457–465.
8. Rubin TN, Chambers A, Smyth P, Steyvers M. Statistical topic models for multi-label document classification. *Machine Learning* 2012; **88**(1-2):157–208.
9. Zhu J, Ahmed A, Xing EP. Medlda: maximum margin supervised topic models. *Journal of Machine Learning Research* 2012; **13**:2237–2278.
10. Blei DM. Probabilistic topic models. *Communications of the ACM* 2012; **55**(4):77–84.
11. Zheng N. Discovering interpretable topics in free-style text: diagnostics, rare topics, and topic. *Ph.D. Thesis*, The Ohio State University, 2008.
12. Allen TT, Xiong H. Pareto charting using multifield freestyle text data applied to toyota camry user reviews. *Applied Stochastic Models in Business and Industry* 2012; **28**(2):152–163.
13. Andrzejewski D, Zhu X, Craven M. Incorporating domain knowledge into topic modeling via dirichlet forest priors. *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM: Montreal, QC, 2009,25–32.
14. Griffiths TL, Steyvers M. Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America* 2004; **101**(Suppl 1):5228–5235.
15. Carpenter B. Integrating out multinomial parameters in latent Dirichlet allocation and naive Bayes for collapsed Gibbs sampling. *Rapport Technique 4*, 2010.
16. Xiong H. Combining subject expert experimental data with standard data in Bayesian mixture modeling. *Ph.D. Thesis*, The Ohio State University, 2011.
17. Lee SH. Comparison and application of probabilistic clustering methods for system improvement prioritization. *Ph.D. Thesis*, The Ohio State University, 2012.
18. Petterson J, Smola A, Caetano T, Buntine W, Narayanamurthy S. Word features for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems 23*, Lafferty J, Williams CKI, Shawe-Taylor J, Zemel RS, Culotta A (eds). NIPS, 2010; 1921–1929.
19. Marchette D. Inferential variability in topic models. *Technical Report*, Naval Surface Warfare Center: Washington Navy Yard, DC, 2012.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.