Infinite Dynamic Topic Models for Longitudinal Surveillance

Keith Harrigian
Department of Computer Science
Johns Hopkins University
kharrigian@jhu.edu

Abstract

The use of topic models to summarize large amounts of language data found on the web for population-level surveillance has become commonplace in computational and social sciences alike. However, many of these analyses obtain support from model architectures that are founded upon assumptions that are invalid for the data being analyzed. In this paper, we look to understand the effect that underlying assumptions regarding stationarity over time and constraints related to model specification have on a topic model's ability to faithfully summarize language data.

1 Introduction

Over the last two decades, the internet, and social media in particular, has emerged as a powerful medium for allowing individuals to voice personal concerns, describe life experiences, and discuss larger sociological challenges with individuals within their community and beyond [Shepherd et al., 2015, Mueller et al., 2021]. This democratization of speech has drawn significant interest from researchers in industry and academia alike, both of whom view online media as a viable alternative to traditional methods for gathering insight into human behavior (e.g. surveys, natural experiments). Towards this end, computational researchers have devoted significant effort to developing technical methods for efficiently summarizing the vast magnitude of thoughts, opinions, and reactions to current events presented online each second [Chua and Asur, 2013, Gao et al., 2018, Gorodnichenko et al., 2021]. Of these approaches, topic models remain one of the most widely utilized resource for extracting semantics in online text to provide a high-level snapshot of the world for researchers to use in downstream analyses [Resnik et al., 2015, Han and Lee, 2016, Lozano et al., 2017].

Unfortunately, topic models can be quite limited in their ability to capture linguistic nuance due to their underlying assumptions regarding the manner in which language is generated. For example, some of the most widely used topic models operate under assumptions of document exhangeability and topic distribution stationarity, which both critically ignore the fact that language usage evolves over time, often in a non-linear and unpredictable manner [Blei et al., 2003, Jelodar et al., 2019]. As a result, these topic models may struggle to adapt to changes in underlying semantics of words, shifts in the popularity of topics being discussed, or the introduction of new terms into the public's lexicon [Blei and Lafferty, 2006, Zhai and Boyd-Graber, 2013]. Such shortcomings can be particularly problematic for the analysis of web data, where linguistic shifts occur more rapidly than in other mediums of writing (e.g. newspaper articles, books) [Eisenstein et al., 2014].

The other major challenge that arises when using topic models to inform a qualitative analysis is their sensitivity to model specification and relatively weak evaluation metrics [Lau and Baldwin, 2016, Thompson and Mimno, 2018]. With respect to the former, we note that even Latent Dirichlet Allocation, one of the simplest topic model architectures, provides an infinite simplex of combinations over its document-topic prior, topic-word prior, and number of available components. Although several combinations can be ruled out from consideration using prior domain knowledge or experimentation,

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

there will remain a substantial number of model settings that can appear roughly equivalent in terms of topic quality. Analysts can use metrics such as topic coherence or perplexity to quantitatively score the model's ability to represent the data, but such metrics are not always in agreement with one another or with an analyst's subjective review of a model [Chang et al., 2009, Mimno et al., 2011].

In this study, we look to understand the degree to which the aforementioned assumptions and constraints affect our ability to accurately assess semantic changes in a population over time. Specifically, we assess the ability of four different topic model architectures to represent dynamically changing language data — Latent Dirichlet Allocation [Blei et al., 2003], Hierarchical Dirichlet Processes [Teh et al., 2006], Dynamic Topic Models [Blei and Lafferty, 2006], and Infinite Dynamic Topic Models [Ahmed and Xing, 2012]. Each of the latter three architectures extend the former to provide support for automatically learning an appropriate number of topics for the data at hand and/or learning non-stationary parameter distributions. In performing this evaluation, we construct recommendations for researchers to consider when using topic models to summarize and interpret large amounts of web-based language data over time.

2 Related Work

Topic models have an extensive history in both academic and industry research. Indeed, a search of "topic model" on Google Scholar returns over 5.5 million results, with the addition of "dynamic" to the query limiting the search space by a mere 50%. To review all of this prior work here would be frivolous; instead, we will focus our attention on the application of topic models for longitudinal surveillance using online media. For the curious reader interested in learning more about traditional topic models and their neural expansions, we recommend reviewing work from Alghamdi and Alfalqi [2015], Jelodar et al. [2019], and Zhao et al. [2021]. However, for the reader's convenience, we will briefly review the specific topic models considered in our work within §3.

With respect to surveillance applications, we note that topic models have been deployed for a wide range of purposes, such as monitoring political trends [Lucas et al., 2015, Greene and Cross, 2017, Stier et al., 2018], quantifying evolving public health concerns [Chen et al., 2014, Paul et al., 2016, Abdellaoui et al., 2018], and discovering current events in real time [Xia et al., 2015, Hong et al., 2016, Gao et al., 2018]. In some cases, researchers spend ample effort constructing models tailored to their specific use case [Dermouche et al., 2014, Jiang et al., 2015], while in others, researchers leverage off-the-shelf approaches with widely-available implementations to inform their downstream analysis [De Choudhury and De, 2014, Biester et al., 2020, Nobles et al., 2020]. From a scientific perspective, it is important to understand the trade-offs in efficiency and effectiveness that arise when opting to take one experimental approach over the other.

Yet, despite years of topic modeling research, no specific consensus has emerged regarding the importance of acknowledging a particular model's underlying assumptions prior to using it for downstream analysis. Generally, research proposing dynamic or metadata-informed topic models achieves increased data likelihoods, reduced perplexity, and more coherent learned topic distributions [Blei and Lafferty, 2006, Mimno et al., 2011, Han et al., 2018]. However, differences in these metrics are not always significant and, more importantly, these metrics themselves do not necessarily correlate well with human judgements of topic quality [Chang et al., 2009]. Opponents further argue that the extra model complexity, in terms of both increased runtimes and reduced posterior interpretability, are prohibitive for most use cases [Ramage et al., 2009, Zhou et al., 2017]. Our hope in this study is to provide further evidence that clarifies this existing debate.

3 Preliminaries

In this section, we will describe each of the four topic models evaluated within our study — Latent Dirichlet Allocation (LDA) [Blei et al., 2003], Hierarchical Dirichlet Processes (HDP) [Teh et al., 2006], Dynamic Topic Models (DTM) [Blei and Lafferty, 2006], and Infinite Dynamic Topic Models (iDTM) [Ahmed and Xing, 2012]. We will introduce sampling models (i.e. data generating procedure) and underlying assumptions for all models. For the latter model iDTM, we will also enumerate the Gibbs sampling steps associated with a single parameter update. Importantly, iDTM can be viewed as a generalization of the remaining three models given appropriately specified hyperparameters.

3.1 Latent Dirichlet Allocation

Introduced by Blei et al. [2003], Latent Dirichlet Allocation (LDA) has since been cited over 30 thousand times and remains the foundation to most topic modeling architectures. It can be viewed as a probabilistic extension to the matrix factorization approach latent semantic analysis [Dumais, 2004]. The data generating process a document $\vec{w_j}$ for LDA is defined as follows:

1. Draw $\theta_j \sim$ Dirichlet(α)
2. For each of the N_j words w_{jn}
   (a) Draw $Z_{jn} \sim$ Multinomial(θ_j)
   (b) Draw $w_{jn} \sim$ Multinomial($\phi_{Z_{jn}}$)

Generally, we also draw $\phi_k \sim$ Dirichlet(β) for each topic component k. This sampling model assumes that each word within a single document is exchangeable with other words in the document (i.e. bag of words assumption [Zhang et al., 2010]), and that each document is exchangeable with other documents. Also, the number of topics K is assumed to be known and static across all documents. The assumption that documents are exchangeable critically ignores the possibility that topics evolve over time, while the assumption that K is fixed and known ignores the practical challenges of knowing such a parameter *a priori* while failing to account for the possible introduction and removal of topics over time.

3.2 Hierarchical Dirichlet Process

The Hierarchical Dirichlet Process (HDP), introduced by Teh et al. [2006], generalizes LDA and removes the assumption that the number of components K is known *a priori*. The model accomplishes this feat by replacing the Dirichlet priors for ϕ and θ with Dirichlet Process priors. Specifically, assuming a global measure $G_0 \sim DP(\gamma_0, H)$, where $H \sim$ Dirichlet, each document $\vec{w_j}$ is generated as follows:

1. Draw $G_j \sim DP(\alpha_0, G_0)$
2. For each of the N_j words w_{jn}
   (a) Draw $\phi_{jn} \mid G_j \sim G_j$
   (b) Draw $w_{jn} \mid \phi_{jn} \sim F(\phi_{jn})$

We note that G_0 is a discrete distribution and accordingly introduces dependence between documents via atom-sharing. Like LDA, each word within a single document is assumed to be exchangeable, while documents are exchangeable with one another. The use of the Dirichlet Process prior allows the model to learn the number of components K based on data, though it still maintains the assumption that K is fixed across the entire corpus. In theory, HDP is expressive enough to learn topics that vary over time, either by learning unique topics for each time period or by nesting HDP models hierarchically based on time period. While the latter approach allows multiple time periods to share topic distributions, any notion of temporal ordering would be lost.

3.3 Dynamic Topic Model

To provide explicit support for modeling topic evolution over time, Blei and Lafferty [2006] introduced the Dynamic Topic Model (DTM). At a high level, the DTM can be thought of as a chained sequence of LDA models, with each time step t serving as the prior for the time step $t+1$. The Dirichlet priors used in the aforementioned architectures are replaced by Gaussian distributions, where the logistic transformation $\pi(\cdot)$ (e.g. softmax) is used to translate the prior parameters into the natural probability space for sampling individual words. Each time step t begins by sampling the topic proportions $\alpha_t \sim \mathcal{N}(\alpha_{t-1}, \delta^2 I)$ and topic-term distributions $\beta_t \sim \mathcal{N}(\beta_{t-1}, \sigma^2 I)$. To sample words in a single document $\vec{w_j}$ found in time step t, we assume the following generative procedure:

1. Draw $\eta_j \sim \mathcal{N}(\alpha_t, a^2 I)$
2. For each of the N_j words w_{jn}
   (a) Draw $Z_{jn} \sim$ Multinomial($\pi(\eta_j)$)
   (b) Draw $w_{jn} \sim$ Multinomial($\pi(\beta_{t,Z_{jn}})$)

The structure of generating a single word remains similar to the aforementioned models, but we note that documents in different time periods are no longer exchangeable with one another. DTM thus removes the assumption that topic distributions are fixed over time, but maintains the assumption that the number of components K is fixed and known *a priori* for each epoch. By sampling topic proportions α_t at each epoch, the model approximately supports the birth and death of topics over time; however, there remains a small probability that each topic is sampled throughout the entire time period. To explicitly model the birth and death of topics over time, we need to make a few more modifications.

3.4 Infinite Dynamic Topic Model

The Infinite Dynamic Topic Model (iDTM) introduced by Ahmed and Xing [2012] seeks to address the constraints imposed by assuming 1) a fixed and known number of components and 2) a static topic distribution over time. As was the case with aforementioned extensions, the generative process for iDTM follows a similar structure as LDA, but leverages different priors on the topic distributions, those that depend now on time. We assume a top-level measure $G_0 \sim DP(\gamma_0, H)$, where $H \sim \mathcal{N}(0, \sigma^2 I)$. Each time step has a specific and temporally-dependent measure G_0^t such that

$$G_0^t \mid \phi_{1:k}, G_0, \alpha \sim DP(\alpha + \sum_k m'_{kt}, \sum_k \frac{m'_{kt}}{\sum_l m'_{lt} + \alpha}\delta(\phi_k) + \frac{\alpha}{\sum_l m'_{lt} + \alpha} G_0)$$

where $m'_{kt} = \sum_{\delta=1}^{\Delta} \exp(\frac{-\delta}{\lambda}) m_{k,t-\delta}$ represents the prior weight of component k at epoch t. Here, m_{kt} can be thought of as the number of tables assigned to component k at time t (as in a Chinese Restaurant Franchise), while Δ and λ represent the width and decay factor of the time-decaying kernel. Integrating out the random measures $G_0^{1:T}$, the distributional parameters $\theta_{1:t}$ follow a polya-urn distribution with time-decay, known colloquially as the recurrent Chinese Restaurant Franchise [Ahmed and Xing, 2008]. The model collapses to a set of independent HDP models when $\Delta = 0$, and to a single HDP model when $\Delta = T$ and $\lambda = \infty$. Given this setup, we can sample words in a single document $\vec{w_j}$ found in time step t using what is essentially the same sampling model as the HDP:

1. Draw $G_j \sim DP(\alpha_0, G_0^t)$
2. For each of the N_j words w_{jn}
   (a) Draw $\phi_{jn} \mid G_j \sim G_j$
   (b) Draw $w_{jn} \mid \phi_{jn} \sim$ Multinomial$(\pi(\phi_{jn}))$

We note the necessity to perform the logistic transformation $\pi(\cdot)$ to the parameters ϕ_{jn} since the base measure H is assumed to be Gaussian. We also note that the parameters ϕ_k evolve in a Markovian fashion: $\phi_{t,k} \mid \phi_{t-1,k} \sim P(\cdot \mid \phi_{t-1,k})$. All together, this model successfully removes the assumption that components are fixed over time and have a cardinality known *a priori*. With this sampling model in mind, we can now define procedures for performing parameter inference.

3.4.1 Inference

We can think of iDTM using a similar mental model to that of the HDP, albeit with modifications to support temporal dependence. Under the recurrent Chinese Restaurant Franchise (RCRF), we consider each document $\vec{w_{td}}$ to be a restaurant d associated with some discrete time period t; each word w_{tdi} in a document can be viewed as a customer. Each customer enters the restaurant and can sit at an existing table b that has topic ψ_{tdb} with probability $\frac{n_{tdb}}{i-1+\alpha}$ or choose to start a new table b_{td}^{new} with probability $\frac{\alpha}{i-1+\alpha}$ and choose a new topic. If the customer sits at a new table, it can choose an already existing topic from the menu at epoch t with probability $\frac{m_{kt}+m'_{kt}}{\sum_{l=1}^{K_t} m_{lt}+m'_{lt}+\gamma}$ or choose a brand new topic sampled from the base distribution H with probability $\frac{\gamma}{\sum_{l=1}^{K_t} m_{lt}+m'_{lt}+\gamma}$. Note that if $m_{kt} = 0$, but $m'_{kt} > 0$, then w_{tdi} modifies the distribution of the topic: $\phi_{kt} \sim P(\cdot \mid \phi_{k,t-1})$. Putting this together, we have:

$$\theta_{tdi} \mid \theta_{td,1:i-1}, \alpha, \psi_{t-\Delta:t} \sim \sum_{b=1}^{b=B_{td}} \frac{n_{tdb}}{i-1+\alpha}\delta_{psi_{tdb}} + \frac{\alpha}{i-1+\alpha}\delta_{tdb^{\text{new}}}$$

$$\psi_{tdb^{\text{new}}} \mid \psi, \gamma \sim \sum_{k:m_{kt}>0} \frac{m_{kt}+m'_{kt}}{\sum_{l=1}^{K_t} m_{lt}+m'_{lt}+\gamma}\delta_{phi_{kt}} + \sum_{k:m_{kt}=0} \frac{m_{kt}+m'_{kt}}{\sum_{l=1}^{K_t} m_{lt}+m'_{lt}+\gamma}P(\cdot \mid \phi_{k,t-1}) + \frac{\gamma}{\sum_{l=1}^{K_t} m_{lt}+m'_{lt}+\gamma}H$$

We can use this definition to help derive Gibbs Sampling updates for the table-to-topic assignments k, word-to-table assignments b, and components ϕ. α and γ can be updated using the same methods as Escobar and West [1995] and are excluded here due to space constraints. Note that we adopt the standard notation of x^{-y} to denote that a particular item y has been excluded from calculation of x; additionally, we say $x^{-y\to x_0}$ is the same as x^{-y}, but also x is set equal to x_0. For sampling k_{tdb} and b_{tdi}, we iterate over time periods (i.e. epochs) sequentially to preserve the fact that documents in different time periods are no longer exchangeable.

Sampling Topic Assignments k_{tdb} for Table tdb: The conditional distribution for k_{tdb} is given as:

$$P(k_{tdb}=k \mid k^{-tdb}_{t-\Delta:t+\Delta}, b_{td}, \phi, w_t) \propto P(k_{tdb}=k \mid k^{-tdb}_{t-\Delta:t}, \phi, v_{tdb}) \prod_{\delta=1}^{\Delta} P(k_{t+\delta} \mid k^{-tdb\to k}_{t+\delta-\Delta:t+\delta-1})$$

where v_{tdb} is the frequency count of words sitting at table tdb. The first factor is the likelihood of each topic conditioned on the words assigned to the table thus far; the second factor represents the probability of future table assignments conditioned on re-assigning the table's dish k_{tdb} to each component k. In our experience, and as noted by Ahmed and Xing [2012], the second factor is expensive to compute and is ill-defined for future components that have not yet been created. For this reason, we approximate the conditional using the first term alone. Using the generative process of the RCRF described above, the proportionality thus becomes:

$$P(k_{tdb}=k \mid k^{-tdb}_{t-\Delta:t}, \phi, v_{tdb}) \propto \begin{cases} (m^{-tdb}_{kt}+m'_{kt})f(v_{tdb} \mid \phi_{kt}) & m^{-tdb}_{kt}>0 \\ m'_{kt}f(v_{tdb} \mid \phi_{kt}) & m'_{kt}>0, m^{-tdb}_{kt}=0, \\ & \phi_{kt} \sim P(\cdot \mid \phi_{k,t-1}) \\ \frac{\gamma}{Q}f(v_{tdb} \mid \phi^q_{kt}) & k \text{ is new}, \phi^q_{kt} \sim H(\cdot) \end{cases}$$

Note that here $f(\cdot \mid \phi_{kt})$ is a multinomial probability mass function and phi_{kt} is cast actually the parameters after being passed through the logistic transform π. The non-conjugacy of the base distribution H and the sampling probability f necessitates use of Algorithm 8 from Neal [2000], where we generate Q auxiliary samples from the base distribution.

Sampling Table Assignments b_{tdi} for Word w_{tdi}: The conditional distribution for b_{tdi} looks nearly identical to the conditional distribution of word to table assignments used in the HDP sampler:

$$P(b_{tdi}=b \mid b^{-tdi}_{td}, k_{t-\Delta:t+\Delta}, \phi, x_{tdi}) \propto \begin{cases} n^{-tdi}_{tdb}f(x_{tdi} \mid \phi_{k_{tjb},t}) & b \text{ exists} \\ \alpha P(k_{tjb^{\text{new}}}=k \mid k^{-tdi}_{t-\Delta:t+\Delta}, b^{-tdi}_{td}, \phi, x_{tdi}) & b^{\text{new}} \text{ is new} \end{cases}$$

As before, we consider the transformed parameters $\pi(\phi)$ above. We also note that the second case (for a new table) is a valid distribution—that is, it sums to 1 and thus the probability of generating a new table is always α.

Sampling Topic Components ϕ_k: The final critical sampling distribution is for our topic parameters ϕ. For a single component, we have the conditional $P(\phi_k \mid b, k, x) = P(\phi_k \mid v_k)$, where $v_k = \{v_{k,t}\}$ and $v_{k,t}$ is the frequency vector of words associated with component k at time t. As noted by Ahmed and Xing [2012], this is a state-space model with nonlinear emission and can be sampled as a block using a Metropolis-Hastings proposal. Specifically, if $q(\cdot)$ is the proposal distribution and ϕ^*_k is a sample from the proposal, we have the acceptance ratio $r = \min(1, u)$, where u equals the following:

$$\frac{H(\phi^*_{k,t_1}) \times \prod_t f(v_{kt} \mid \phi^*_{kt})P(\phi^*_{kt} \mid \phi^*_{k,t-1})}{H(\phi_{k,t_1}) \times \prod_t f(v_{kt} \mid \phi_{kt})P(\phi_{kt} \mid \phi_{k,t-1})} \times \frac{\prod_t q(\phi_{kt})}{\prod_t q(\phi^*_{kt})}$$

Ahmed and Xing [2012] construct a proposal based on a Kalman Smoother from Minka [1999]. In contrast, we leverage a Hidden Markov Model (HMM) with Gaussian emissions as our proposal distribution [Rabiner, 1989, Bilmes et al., 1998]. We find the HMM is relatively quick to learn and achieves a reasonably high acceptance ratio (~ 0.70). However, there are cases where the HMM fails to converge, even after several iterations of the expectation maximization algorithm. When this happens, the proposal samples tend to be rejected, or the components themselves only have minimal

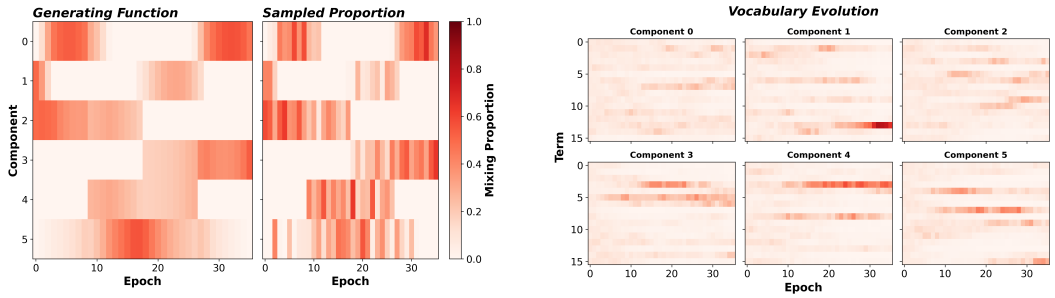

Figure 1: Synthetic dataset topic proportions and component evolution over time. We specify the existence of 6 components and a vocabulary size of 16.

data support and are dropped from the sampler in later MCMC iterations. We leave exploration of alternative proposal distributions that aren't as sensitive to parameter optimization as our HMM for future work.

It is important to note that, by nature of the iDTM architecture, topics can be alive for variable lengths of time. When a component is only used in one epoch, and has not been assigned to any tables in future epochs, we replace the HMM proposal with a multivariate Gaussian with $\phi^*_{kt} \sim \mathcal{N}(\phi_{kt}, \Sigma_q)$, where Σ_q is a hyperparameter for the proposal.

4 Data

4.1 Synthetic Data

To evaluate each model's ability to capture temporal dynamics, we construct an artificial dataset in which we know there exists longitudinal shifts in both topic proportion and composition over time (see Figure 1). Namely, we specify the existence of 6 unique topics for a vocabulary of size 16. Each topic proportion evolves over the course of 36 epochs (t) according to pre-specified functions, is normalized between 0 and 1, and then re-weighted so that the total component contribution at each epoch is 1. The equations and weights are as follows: $\sin(\frac{t}{4})$ $[w_0 = 0.15]$; $\cos(\frac{t+\pi}{4})$ $[w_1 = 0.15]$; 1 if $t \geq 18$, otherwise 0 $[w_2 = 0.1]$; 1 if $t < 18$, otherwise 0 $[w_3 = 0.1]$; 1 if $12 \leq t \leq 24$, otherwise 0 $[w_4 = 0.1]$; $\exp(\frac{t}{6})$ if $t < 18$, otherwise $\exp(-\frac{t-18}{6})$ $[w_5 = 0.3]$.

The proportions α generated using the above functions are multiplied by a concentration parameter γ and then used to parameterize independent Dirichlet distributions for sampling. Each topic component ϕ_k then evolves using the generating processes assumed by the DTM and iDTM models. That is:

1. For each component $k = 1, ..., 6$
   (a) $\phi_{k,1} \sim \mathcal{N}(0, \beta_0)$
   (b) For each time period $t = 2, ..., 36$
      i. $\phi_{k,t} \sim \mathcal{N}(\phi_{k,t-1}, \beta_1)$

A single document $\vec{w}_d$ occurring at time t is generated as follows:

1. Draw topic distribution $\theta_{td} \sim \text{Dirichlet}(\pi(\alpha_t))$
2. Draw word count $N_{td} \sim \text{Poisson}(\lambda_n)$
3. For each of the N_{td} words w_{tdi}
4. (a) Draw component $z_{tdi} \sim \text{Multinomial}(\theta_{td})$
   (b) Draw word $w_{tdi} \sim \text{Multinomial}(\pi(\phi_{z_{tdi},t}))$

For each epoch t, we generate $M_t \sim \text{Poisson}(\lambda_m)$ documents based on the logic above.

4.2 Real Data

While it allows us to evaluate our model implementation and quantitatively compare model architectures, model performance on synthetic data is not practically useful for other computational science

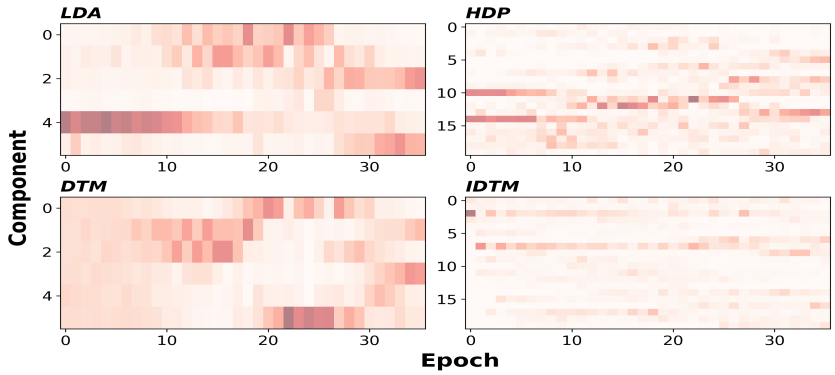

Figure 2: Estimated topic dynamics by each of the four topic modeling architectures. HDP and iDTM severely overestimate the latent dimensionality of the synthetic data.

and social science researchers interested in evaluating language dynamics. For this reason, we consider a simple case study in which temporally-aware topic models are often used. In particular, we decide to examine changes in language within the r/depression subreddit from January 1, 2020 through April 1, 2021 [Baumgartner et al., 2020]. Using the `retriever`[1] API wrapper, we acquire slightly over 2.9 million comments made in the subreddit during the aforementioned time period. An analysis of this type of language data has already been completed using LDA as the supporting topic model [Biester et al., 2020]. Our goal is to replicate prior findings using iDTM and identify differences that arise when using dynamic as opposed to static topic models.

5 Experiments

We hypothesize that iDTM will be at least as good at modeling dynamic language data as LDA, HDP, and DTM under an optimal parameterization.

5.1 Synthetic Data

To test our hypothesis, we first generate a synthetic dataset using the method described in §4 using the following parameters: $\beta_0 = 1e^{-1}, \beta_1 = 1e^{-1}, \lambda_n = 100, \lambda_m = 25, \gamma = 10$. We fit LDA, HDP, DTM, and iDTM models to the dataset and manually review the learned topic distrbutions and dynamics with reference to the ground truth. We leverage `tomotopy`'s implementations of collapsed Gibbs samplers for LDA, HDP, and DTM and implement iDTM ourselves, as no open-source implementation exists. We train LDA and HDP for 50k iterations, DTM for 100k iterations, and iDTM for 200 iterations with manual parameter tuning to optimize performance.[2]. To fairly compare LDA, HDP, and DTM with iDTM, we initialize the former models with 6 components.

We present a comparison of the learned topic dynamics as a function of topic model architecture in Figure 2. Immediately, we note LDA's ability to almost perfectly replicate the synthetic topic dynamics. HDP and iDTM generate too many topics, opting to split topics components that reoccur after several temporal epochs of dormancy. DTM struggles to converge with the given dataset, despite being run the longest. Indeed, we find DTM only converges when λ_m and λ_n are sufficiently high by construction. Given this fact and iDTM's Gaussian base measure, we suspect iDTM may also have not fully converged. Unfortunately, the computational expense associated with sampling components prohibits us from testing this hypothesis at this time.

5.2 Real Data

After preprocessing the Reddit data, we isolate the top 500 most frequent terms to use for our vocabulary. We feed a 10% sample of the dataset into LDA, HDP, DTM, and our iDTM model, using 3-month windows as the discrete temporal intervals. We train the LDA and HDP models for 10k

[1] `https://github.com/kharrigian/retriever`

[2] Parameters are included in our digital supplement at: `https://github.com/kharrigian/idtm`

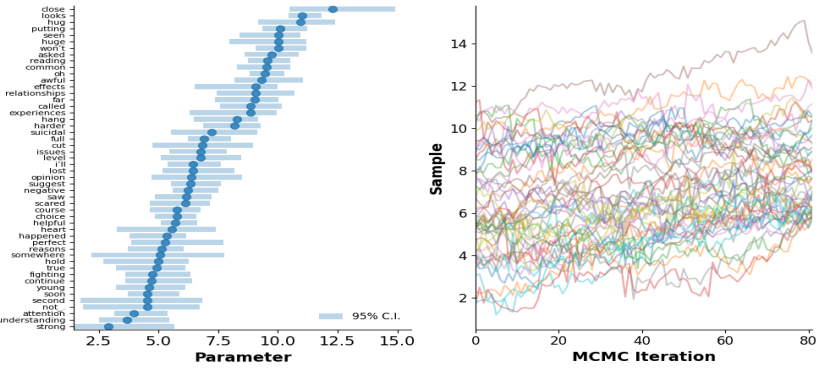

Figure 3: Trace plot for one of the iDTM real data topics. The model has yet to converge, despite several hours of sampling

iterations, the DTM model for 50k iterations, and the iDTM model for 5 iterations with a batch size of 250 samples (i.e. 405 updates). We provide a comparison of representative topics from each model below.

- LDA (Topic 6): *games play video game playing enjoy online watch*
- HDP (Topic 6): *meds medication side doctor effects psychiatrist brain*
- DTM (Topic 13, Epoch 3): *etc left hurt brain bed instead learn mom*
- iDTM (Topic 12, Epoch 1): *close, hug, looks, relationships, seen*

It becomes clear to us that both dynamic models have failed to converge. Given that DTM, even with access to a fast variational EM algorithm and several more sampling iterations, is unable to converge, we draw the conclusion that iDTM is likely ill-suited for real-world data of any reasonable size. Meanwhile, LDA and HDP both provide quality topics with intuitive dynamics (e.g. online hobbies increased at the start of the quarantine) while requiring significantly less computational overhead.

6 Conclusion

In this study, we evaluated the effectiveness of the Infinite Dynamic Topic Model (iDTM) in comparison to alternative topic modeling methods that require *a priori* knowledge of the number of topics in a dataset and/or do not allow topic components to evolve over time. Although iDTM offers more flexibility and requires fewer assumptions than LDA, HDP, or DTM topic models, it does not necessarily lead to different (or better) downstream results. Given the significant computational expense of running inference with iDTM, the model remains ill-suited for use by the general practitioner.

There are three major caveats in our analysis. First and foremost, we recall from §3.4.1 that we only approximated the conditional posterior for k_{tdb}, excluding the future transition probabilities due to computational expense and its convoluted definition in the original literature [Ahmed and Xing, 2012]. While the missing term could lead to different results in our analysis, we note that others have made similar approximations without issue [Ahmed and Xing, 2008, Beykikhoshk et al., 2016]. The second potential shortcoming is our limited hyperparameter tuning and analysis of proposal distributions for ϕ; it is possible that alternative settings may have allowed us to further optimize performance. Indeed, several other researchers have noted that iDTM is quite brittle when it comes to selecting parameters that will lead to appropriate convergence [Elshamy, 2013, Zhou et al., 2017]. iDTM's original authors even note that iDTM is extremely sensitive to initialization [Ahmed and Xing, 2012]. Finally, we note that we were unable to run an equal number of MCMC iterations for iDTM in comparison to the alternative methods studied in this paper. Given that DTM in particular required several more iterations of sampling than LDA and HDP, we hypothesize iDTM may not have been able to converge appropriately given our limited compute budget and difficulty identifying appropriate hyperparameters.

References

Andrew Shepherd, Caroline Sanders, Michael Doyle, and Jenny Shaw. Using social media for support and feedback by mental health service users: thematic analysis of a twitter conversation. *BMC psychiatry*, 15(1):1–9, 2015.

Aaron Mueller, Zach Wood-Doughty, Silvio Amir, Mark Dredze, and Alicia Lynn Nobles. Demographic representation and collective storytelling in the me too twitter hashtag activism movement. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1):1–28, 2021.

Freddy Chua and Sitaram Asur. Automatic summarization of events from social media. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 7, 2013.

Wei Gao, Peng Li, and Kareem Darwish. Joint topic modeling for event summarization across news and social media streams. In *Social Media Content Analysis: Natural Language Processing and Beyond*, pages 321–346. World Scientific, 2018.

Yuriy Gorodnichenko, Tho Pham, and Oleksandr Talavera. Social media, sentiment and public opinions: Evidence from# brexit and# uselection. *European Economic Review*, page 103772, 2021.

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

Jonghyun Han and Hyunju Lee. Characterizing the interests of social media users: Refinement of a topic model for incorporating heterogeneous media. *Information Sciences*, 358:112–128, 2016.

Marianela García Lozano, Jonah Schreiber, and Joel Brynielsson. Tracking geographical locations using a geo-aware topic model for analyzing social media data. *Decision Support Systems*, 99: 18–29, 2017.

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

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

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

Ke Zhai and Jordan Boyd-Graber. Online latent dirichlet allocation with infinite vocabulary. In *International conference on machine learning*, pages 561–569. PMLR, 2013.

Jacob Eisenstein, Brendan O'Connor, Noah A Smith, and Eric P Xing. Diffusion of lexical change in social media. *PloS one*, 9(11):e113114, 2014.

Jey Han Lau and Timothy Baldwin. The sensitivity of topic coherence evaluation to topic cardinality. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–487, 2016.

Laure Thompson and David Mimno. Authorless topic models: Biasing models away from known structure. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3903–3914, 2018.

Jonathan Chang, Jordan Boyd-Graber, Chong Wang, Sean Gerrish, and David M Blei. Reading tea leaves: How humans interpret topic models. In *Neural information processing systems*, volume 22, pages 288–296. Citeseer, 2009.

David Mimno, Hanna Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 262–272, 2011.

Yee Whye Teh, Michael I Jordan, Matthew J Beal, and David M Blei. Hierarchical dirichlet processes. *Journal of the american statistical association*, 101(476):1566–1581, 2006.

Amr Ahmed and Eric P Xing. Timeline: A dynamic hierarchical dirichlet process model for recovering birth/death and evolution of topics in text stream. *arXiv preprint arXiv:1203.3463*, 2012.

Rubayyi Alghamdi and Khalid Alfalqi. A survey of topic modeling in text mining. *Int. J. Adv. Comput. Sci. Appl.(IJACSA)*, 6(1), 2015.

He Zhao, Dinh Phung, Viet Huynh, Yuan Jin, Lan Du, and Wray Buntine. Topic modelling meets deep neural networks: A survey. *arXiv preprint arXiv:2103.00498*, 2021.

Christopher Lucas, Richard A Nielsen, Margaret E Roberts, Brandon M Stewart, Alex Storer, and Dustin Tingley. Computer-assisted text analysis for comparative politics. *Political Analysis*, 23(2): 254–277, 2015.

Derek Greene and James P Cross. Exploring the political agenda of the european parliament using a dynamic topic modeling approach. *Political Analysis*, 25(1):77–94, 2017.

Sebastian Stier, Arnim Bleier, Haiko Lietz, and Markus Strohmaier. Election campaigning on social media: Politicians, audiences, and the mediation of political communication on facebook and twitter. *Political communication*, 35(1):50–74, 2018.

Liangzhe Chen, KSM Tozammel Hossain, Patrick Butler, Naren Ramakrishnan, and B Aditya Prakash. Flu gone viral: Syndromic surveillance of flu on twitter using temporal topic models. In *2014 IEEE international conference on data mining*, pages 755–760. IEEE, 2014.

Michael J Paul, Abeed Sarker, John S Brownstein, Azadeh Nikfarjam, Matthew Scotch, Karen L Smith, and Graciela Gonzalez. Social media mining for public health monitoring and surveillance. In *Biocomputing 2016: Proceedings of the Pacific symposium*, pages 468–479. World Scientific, 2016.

Redhouane Abdellaoui, Pierre Foulquié, Nathalie Texier, Carole Faviez, Anita Burgun, and Stéphane Schück. Detection of cases of noncompliance to drug treatment in patient forum posts: topic model approach. *Journal of medical Internet research*, 20(3):e85, 2018.

Yunqing Xia, Nan Tang, Amir Hussain, and Erik Cambria. Discriminative bi-term topic model for headline-based social news clustering. In *The twenty-eighth international flairs conference*, 2015.

Lingzi Hong, Weiwei Yang, Philip Resnik, and Vanessa Frias-Martinez. Uncovering topic dynamics of social media and news: the case of ferguson. In *International Conference on Social Informatics*, pages 240–256. Springer, 2016.

Mohamed Dermouche, Julien Velcin, Leila Khouas, and Sabine Loudcher. A joint model for topic-sentiment evolution over time. In *2014 IEEE international conference on data mining*, pages 773–778. IEEE, 2014.

Shuhui Jiang, Xueming Qian, Jialie Shen, Yun Fu, and Tao Mei. Author topic model-based collaborative filtering for personalized poi recommendations. *IEEE transactions on multimedia*, 17(6): 907–918, 2015.

Munmun De Choudhury and Sushovan De. Mental health discourse on reddit: Self-disclosure, social support, and anonymity. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, 2014.

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

Alicia L Nobles, Eric C Leas, Mark Dredze, and John W Ayers. Examining peer-to-peer and patient-provider interactions on a social media community facilitating ask the doctor services. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 464–475, 2020.

Jun Han, Yu Huang, Kuldeep Kumar, and Sukanto Bhattacharya. Time-varying dynamic topic model: a better tool for mining microblogs at a global level. *Journal of Global Information Management (JGIM)*, 26(1):104–119, 2018.

Daniel Ramage, Evan Rosen, Jason Chuang, Christopher D Manning, and Daniel A McFarland. Topic modeling for the social sciences. In *NIPS 2009 workshop on applications for topic models: text and beyond*, volume 5, page 27, 2009.

Houkui Zhou, Huimin Yu, and Roland Hu. Topic evolution based on the probabilistic topic model: a review. *Frontiers of Computer Science*, 11(5):786–802, 2017.

Susan T Dumais. Latent semantic analysis. *Annual review of information science and technology*, 38 (1):188–230, 2004.

Yin Zhang, Rong Jin, and Zhi-Hua Zhou. Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1-4):43–52, 2010.

Amr Ahmed and Eric Xing. Dynamic non-parametric mixture models and the recurrent chinese restaurant process: with applications to evolutionary clustering. In *Proceedings of the 2008 SIAM International Conference on Data Mining*, pages 219–230. SIAM, 2008.

Michael D Escobar and Mike West. Bayesian density estimation and inference using mixtures. *Journal of the american statistical association*, 90(430):577–588, 1995.

Radford M Neal. Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265, 2000.

Tom Minka. From hidden markov models to linear dynamical systems. Technical report, Citeseer, 1999.

Lawrence R Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

Jeff A Bilmes et al. A gentle tutorial of the em algorithm and its application to parameter estimation for gaussian mixture and hidden markov models. *International Computer Science Institute*, 4(510): 126, 1998.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The pushshift reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 14, pages 830–839, 2020.

Adham Beykikhoshk, Dinh Phung, Ognjen Arandjelović, and Svetha Venkatesh. Analysing the history of autism spectrum disorder using topic models. In *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 762–771. IEEE, 2016.

Wesam Elshamy. Continuous-time infinite dynamic topic models. *arXiv preprint arXiv:1302.7088*, 2013.