Fully Bayesian Approaches to Topics over Time

Julián Cendrero [1,2,*], **Julio Gonzalo**[1], **and Ivar Zapata**[3]

[1] *Universidad Nacional de Educación a Distancia (UNED)*
[2] *Cohere*
[3] *mrHouston Tech Solutions*

Abstract

The Topics over Time (ToT) model captures thematic changes in timestamped datasets by explicitly modeling publication dates jointly with word co-occurrence patterns. However, ToT was not approached in a fully Bayesian fashion, a flaw that makes it susceptible to stability problems. To address this issue, we propose a fully Bayesian Topics over Time (BToT) model via the introduction of a conjugate prior to the Beta distribution. This prior acts as a regularization that prevents the online version of the algorithm from unstable updates when a topic is poorly represented in a mini-batch. The characteristics of this prior to the Beta distribution are studied here for the first time. Still, this model suffers from a difference in scale between the single-time observations and the multiplicity of words per document. A variation of BToT, Weighted Bayesian Topics over Time (WBToT), is proposed as a solution. In WBToT, publication dates are repeated a certain number of times per document, which balances the relative influence of words and timestamps along the inference process. We have tested our models on two datasets: a collection of over 200 years of US state-of-the-union (SOTU) addresses and a large-scale COVID-19 Twitter corpus of 10 million tweets. The results show that WBToT captures events better than Latent Dirichlet Allocation and other SOTA topic models like BERTopic: the median absolute deviation of the topic presence over time is reduced by 51% and 34%, respectively. Our experiments also demonstrate the superior coherence of WBToT over BToT, which highlights the importance of balancing the time and word modalities. Finally, we illustrate the stability of the online optimization algorithm in WBToT, which allows the application of WBToT to problems that are intractable for standard ToT.

Index Terms

topic modeling, latent dirichlet allocation, topics over time, bayesian probability, variational inference, beta distribution

1 Introduction

Topic modeling is widely acknowledged as a method for discovering the latent themes that characterize a collection of documents (for recent reviews with emphasis on applications, see Churchill and Singh 2021; Jelodar, Wang, Yuan, Feng, Jiang, Li and Zhao 2019). Mathematically, a topic is defined as a multinomial distribution over a collection of words that constitute a vocabulary. A document can be modeled as a multinomial distribution over topics, each of them representing the underlying semantic themes of the document. The seminal work on topic modeling, Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999), was later extended in Blei, Ng and Jordan (2003); Hoffman, Bach and Blei (2010) by making use of Bayesian methods to assemble the predominant Latent Dirichlet Allocation (LDA). The shortcomings of PLSA (namely, a large number of parameters and difficulties to handle new documents unseen during training) were solved in LDA thanks to the introduction of conjugate priors for each probability distribution.

LDA was the origin of an abundance of related models which extend its capabilities by including additional features like word embeddings (Bunk and Krestel, 2018; Dieng, Ruiz and Blei, 2020) and hierarchical structure (Viegas, Cunha, Gomes, Pereira, Rocha and Goncalves, 2020), or by adapting the generative process to fit specific types of documents (such as short texts, Li, Zhang and Ouyang 2019b).

Publication dates have been leveraged to investigate topic evolution in timestamped corpora. Modeling time-related characteristics of texts has been proven to be effective for many purposes, such as disaster detection, public opinion evaluation, and outbreak prediction (Asghari, Sierra-Sosa and Elmaghraby, 2020). There are two main strategies to take publication timestamps into account in topic modeling. First, one can see the topics as dynamic objects. This way, the documents in a given time slice are modeled by a set of topics that evolved from the documents of an adjacent time slice. This is the idea behind the Dynamic Topic Model (DTM) (Blei and Lafferty, 2006; Wang, Blei and Heckerman, 2008). On the other hand, one can assume that topics are static (like in LDA) and associate each topic with a continuous distribution over timestamps. This approach is the basis for the Topics over Time (ToT) model (Wang and McCallum, 2006), whose generative process models documents and timestamps jointly. The main advantage of such a continuous time model over DTM is the fact that there is no need to arbitrarily define the size of a discrete-time slice.

* *Corresponding author: Julián Cendrero (jcendrero3@alumno.uned.es).*
Work completed while the author was at mrHouston Tech Solutions and UNED.

The ToT model was designed for collections of timestamped documents that were published over a large period of time. In this type of corpus, topics that appear at the beginning of the period can be quite different from the ones that appear later: new trends emerge and others disappear through time. A time-unaware model such as LDA will extract topics that are present across the entire corpus. This is where the ToT model delivers an advantage: since this model associates each topic with a continuous distribution over timestamps, the projection of a document into the topic space will not only take into account word co-occurrences but also the document's timestamp. As a result, ToT will partition topics that are lexically close but show different time distributions.

Some recent applications of the ToT model include the detection of public concerns in mega infrastructure projects (Xue, Shen, Li, Wang and Zafar, 2020; Xue, Qiping Shen, Li, Han and Chu, 2021), time series analysis of supercomputer logs to predict node failures in high performance computing systems (Das, Mueller, Hargrove, Roman and Baden, 2018), and meeting summarization tools (Shi, Bryan, Bhamidipati, Zhao, Zhang and Ma, 2018). For us, such a variety of applications justify the interest in the ToT model and its potential improvements.

The main shortcoming of ToT is the fact that, unlike LDA, it is not a fully Bayesian model. Apart from the theoretical inconvenience of dealing with two different frameworks at the same time (a Bayesian approach for the text modality and a frequentist approach for the time modality), this results in stability problems. In particular, parameter estimation in an online fashion is susceptible to unstable updates when a topic is poorly represented in a mini-batch. The lack of a robust online optimization method makes this model unusable for many practical applications that require an online data stream mining (for example, content-based recommendation systems Al-Ghossein, Murena, Abdessalem, Barré and Cornuéjols 2018). This limitation supports the need for a fully Bayesian approach to Topics over Time. In this work, we solve this problem by introducing a conjugate prior to the timestamp probability distribution. We call this model Bayesian ToT (BToT).

However, as we will later show, we have found that this naive BToT suffers from a difference in scale between word and timestamp observations in a single document. In order to address this issue, we also introduce an extension of BToT, called Weighted Bayesian Topics over Time (WBToT), that dynamically lowers the number of time observations. This mechanism is similar to the introduction of a balancing hyperparameter in Wang and McCallum (2006); Fang, Macdonald, Ounis, Habel and Yang (2017). The WBToT model features a stable online optimization algorithm that makes it suitable for many applications where standard ToT is not applicable. The Bayesian formalism also simplifies many existing extensions of ToT (Poumay and Ittoo, 2021).

The structure of this paper is as follows. In Section 2, we will dive into the technical details of the ToT model and its extensions to motivate our Bayesian approach. In Section 3, we specify the contributions of our work. In Section 4, we introduce two original models, BToT and WBToT, and propose a variational approach to optimization (both in batch and online fashions). In Section 5 we describe the two datasets that will be employed to validate our models and the experiments that will be performed. In Section 6 we present the main results of our experiments and discuss their implications. We conclude with some final remarks in Section 7.

2 Literature Review

The use of Bayesian methods in topic modeling was introduced via the LDA model in Blei et al. (2003); Hoffman et al. (2010). For each of the multinomial distributions that constituted the original PLSA model, a conjugate prior (Dirichlet distribution) was added. Although the resulting posterior distribution was intractable for exact inference, an approximate inference was performed through variational inference. A family of lower bounds indexed by a set of free variational parameters is introduced, and its optimization could be easily achieved via an iterative fixed-point method. For reference, a detailed description of the LDA generative process and its optimization algorithms (batch and online) is presented in Appendix A. We will follow the same approach and will borrow most of its notation in order to turn ToT into a fully Bayesian model.

In the ToT model (Wang and McCallum, 2006), topics are created by looking not only at word co-occurrences but also at the documents' timestamps. In the generative process for ToT, a distribution of timestamps is postulated to come from a mixture of Beta distributions, one per topic, with the same mixture parameters per document as the one that generates the words. However, this model is not fully Bayesian because there is no prior for the parameters of the Beta distributions, so they have to be optimized by ordinary maximum likelihood. Gibbs sampling is used as the optimization algorithm for this model, and an approximation based on the method of moments simplifies the mathematics of the maximization.

Gibbs sampling is not the only possible way for ToT parameter inference. In Fang et al. (2017), optimization of the ToT log-likelihood was made via a variational ansatz which is identical to the one for LDA in Hoffman et al. (2010), together with an exact maximization of the evidence lower bound (ELBO) with respect to both variational and Beta distribution parameters. Note that, in this approach, the Beta distribution parameters are maximized directly, with no prior distribution involved. Therefore, although a variational inference approach is used instead of Gibbs sampling for the optimization, the ToT model is not actually modified. Notice that neither of the two works (Wang and McCallum, 2006; Fang et al., 2017) introduced an online optimization approach for the model. For reference, a detailed description of the ToT generative process and its optimization algorithm is presented in Appendix B.

Most extensions of ToT have focused on enhancing its capabilities by including additional contextual features, but none of them modified the basic generative process for the timestamps. For example, the author-topic over time model

(AToT) (Xu, Shi, Qiao, Zhu, Jung, Lee and Choi, 2014) includes the document's authorship into its generative model by the introduction of a uniform distribution over authors. This way, they are able to model how the authors' interests change over time, which has been successfully applied to real case scenarios like monitoring emerging technologies by using Twitter data mining (Li, Xie, Jiang, Zhou and Huang, 2019a). However, they still use an unregularized distribution for the timestamps, thus presenting the same problem as the original ToT in terms of stability.

The most recent extensions of ToT still suffer from this problem: the Hierarchical Topic Modeling Over Time (HTMOT) model (Poumay and Ittoo, 2021), which models topic temporality and hierarchy jointly, cannot be trained using variational inference due to the lack of a conjugate prior for the beta distribution. Due to the prohibitively long training time that standard Gibbs sampling takes for this model, this shortcoming forced the authors to develop a new implementation of Gibbs sampling. We believe that they could have employed a more straightforward optimization algorithm if the original ToT model were Bayesian to begin with.

Unlike these extensions, our work is an attempt to modify the core generative of ToT process to make it fully Bayesian. We will start from the basic generative process introduced in Wang and McCallum (2006) but, instead of using Gibbs sampling, we will consider the variational inference method described by Fang et al. (2017). We draw inspiration from Blei et al. (2003) and will introduce conjugate priors for each probability distribution. For the online version of the algorithm, we will closely resemble the approach introduced in Hoffman et al. (2010).

It must be noticed that our work is based on probabilistic topic models. Recently, alternative approaches that make use of large pre-trained language models (Vaswani, Shazeer, Parmar, Uszkoreit, Jones, Gomez, Kaiser and Polosukhin, 2017; Devlin, Chang, Lee and Toutanova, 2019) have also been proposed. The most popular framework, BERTopic (Grootendorst, 2022), projects documents into an embedding space by making use of Sentence-BERT (Reimers and Gurevych, 2019), then performs dimensionality reduction through UMAP (McInnes, Healy and Melville, 2018) and finally creates clusters with HDBSCAN (McInnes and Healy, 2017). Topic distributions are generated by using TF-IDF over such clusters. This approach reportedly yields more coherent topics than models based on the Bag of Words (BOW) assumption, but it does not leverage publication timestamps for event detection. Nevertheless, we will also include BERTopic in our experiments as an additional baseline.

3 Research Objectives and Contributions

The goal of the present paper is to extend the model introduced in Wang and McCallum (2006) and its subsequent variational approach in Fang et al. (2017) to make them fully Bayesian. In particular, our main contributions are:

1) Reformulation of ToT as a Bayesian model. This implies that a novel distribution (which we will call Beta-prior [1]) has to be used as a prior for the parameters of the timestamp Beta distributions. We call this model Bayesian Topics over Time (BToT).
2) Introduction of a modified version of BToT which dynamically lowers the number of time observations. This mimics the introduction of a balancing hyperparameter in Wang and McCallum (2006); Fang et al. (2017), which balances the relative influence of the two modalities (words and timestamps) in the log-likelihood. We call this model Weighted Bayesian Topics over Time (WBToT).
3) Calculation of a variational approach inference of parameters for both BToT and WBToT. In both cases, we introduce a batch and an online version.
4) Experimental validation on two datasets: a collection of over 200 years of US state-of-the-union addresses and a large-scale COVID-19 Twitter corpus.

4 Fully Bayesian Approaches to Topics over Time

In this section, we present two novel models and propose variational inference algorithms for parameter optimization: in Section 4.1 we introduce the naive Bayesian Topics over Time (BToT) model and in Section 4.2 we introduce the Weighted Bayesian Topics over Time (WBToT). For a summary of the notation employed in this section, see Table 1 - 2.

4.1 Bayesian Topics over Time (BToT)

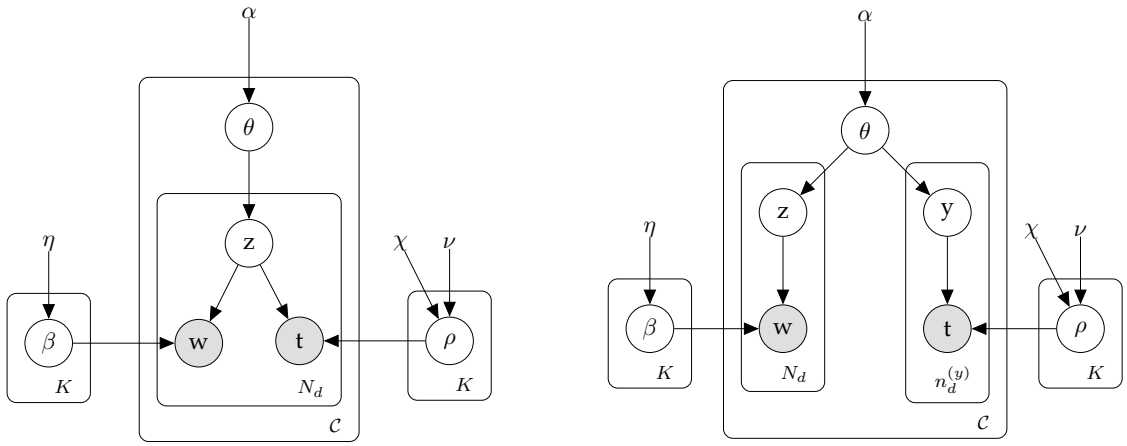
4.1.1 Description of the model

The ToT model introduced a Beta distribution of the form

$$p\left(t \,|\boldsymbol{\rho}_k\right) = B\left(\boldsymbol{\rho}_k\right)^{-1} t^{\rho_k^1 - 1} \left(1-t\right)^{\rho_k^2 - 1} \tag{1}$$

for the timestamp observation t in topic k, where B is the Beta function (employed here as a normalization constant). Notice that we have denoted $\boldsymbol{\rho}_k = \left(\rho_k^1, \rho_k^2\right)$ the 2D natural parameter of the Beta distribution, and hence $B\left(\boldsymbol{\rho}_k\right) = B\left(\rho_k^1, \rho_k^2\right)$. These two parameters must satisfy $\rho_k^1 > 0$ and $\rho_k^2 > 0$ to ensure that the distribution can be normalized. We will extend the ToT model by adding a prior for this distribution. To the best of our knowledge, no conjugate prior to the Beta distribution

1. Here, we call Beta-prior to the conjugate prior to the Beta distribution. Other works refer to Beta-prior as the Beta distribution seen as a prior to a multinomial distribution, to distinguish it from the resulting posterior, which is also a Beta distribution.



(a) Bayesian Topics over Time (BToT)

(b) Weighted Bayesian Topics over Time (WBToT)

Fig. 1: Directed acyclic graphs for Bayesian approaches to ToT

has been studied so far. However, since the Beta distribution is a member of the exponential family, one can always find a general conjugate prior (see (Bishop, 2006)) of the form

$$p\left(\boldsymbol{\rho}_k \left| \nu, \boldsymbol{\chi} \right.\right) = f\left(\nu, \boldsymbol{\chi}\right) \exp\left[\nu \left(\boldsymbol{\rho}_k \cdot \boldsymbol{\chi} - \log B\left(\boldsymbol{\rho}_k\right)\right)\right], \tag{2}$$

where $f\left(\nu, \boldsymbol{\chi}\right)$ is a normalization function. We will call the distribution in Eq. (2) Beta-prior. Note that, for simplicity of presentation, we use the same prior hyper-parameters for all topics, but the formulas could be generalized straightforwardly to include topic-dependent hyper-parameters. Since $\rho^i > 0$, integrability demands that $\nu > 0$. The $\boldsymbol{\chi}$ parameters are constrained by $\exp \chi^1 + \exp \chi^2 < 1$, which can be proved by transforming the normalization integral to polar coordinates, using the Stirling approximation in $\log B\left(\boldsymbol{\rho}\right)$ for $\rho \to \infty$ and then demanding that the exponential tends to zero at its largest asymptotic values (see Appendix C for a detailed explanation).

If one assumes that the Beta distribution parameters in the ToT model are now drawn from this Beta-prior, then the generative process can be described as:

1) Draw K multinomial parameters $\boldsymbol{\beta_k}$ from K Dirichlet distributions with parameters $\boldsymbol{\eta_k}$.
2) Draw K Beta distribution parameters $\boldsymbol{\rho}_k$ from K Beta-prior distributions with parameters $\nu, \boldsymbol{\chi}$.
3) For each document d, draw a set of multinomial parameters $\boldsymbol{\theta_d}$ from a symmetric Dirichlet distribution with parameters $\boldsymbol{\alpha}$. Then, for each word index $i \in \{1, \ldots, N_d\}$ in document d:
   a) Draw a topic z_{di} from a multinomial with parameters $\boldsymbol{\theta_d}$;
   b) Draw a word w_{di} from a multinomial with parameters $\boldsymbol{\beta_{z_{di}}}$;
   c) Draw a timestamp t_{di} from a Beta distribution with parameters $\boldsymbol{\rho}_{z_{di}}$.

See Fig. 1a for the directed acyclic graph representation of this generative process (it is useful to compare this figure with the corresponding graph representation of the classical ToT, which we have reproduced in Fig. 10a). Unlike ToT, this generative process is now fully Bayesian, hence the name Bayesian Topics over Time (BToT).

The log-likelihood for BToT can be written as

$$\mathcal{L}^{\text{BToT}} = \mathcal{L}^{\text{LDA}} + \sum_{d=1}^{D} \sum_{i=1}^{N_d} \log \underbrace{p\left(t_{di} | \boldsymbol{\rho}_{z_{di}}\right)}_{\text{Beta}} + \sum_{k=1}^{K} \log \underbrace{p\left(\boldsymbol{\rho}_k \left| \nu, \boldsymbol{\chi} \right.\right)}_{\text{Beta-prior}}, \tag{3}$$

where $\mathcal{L}^{\text{LDA}}$ is the log-likelihood of the standard LDA (see Eq. (39)). Notice that the only difference with respect to the classical ToT log-likelihood in Eq. (48) is the introduction of the Beta-prior. It is important to note that the two modalities (words and timestamps) of the BToT log-likelihood are decoupled, so they can be optimized separately. For reference, let us flesh out the terms in the log-likelihood that depend on timestamps and their priors (which we will denote $\mathcal{L}^{\text{BToT}}{}_{ts}$):

$$\begin{aligned} \mathcal{L}^{\text{BToT}}{}_{ts} = & \sum_{d=1}^{D} \sum_{i=1}^{N_d} \left\{ \left(\rho^1_{z_{di}} - 1\right) \log t_{di} + \left(\rho^2_{z_{di}} - 1\right) \log\left[1 - t_{di}\right] - \log B\left(\boldsymbol{\rho}_{z_{di}}\right) \right\} \\ & + \nu \sum_{k=1}^{K} \left\{ \boldsymbol{\rho}_k \cdot \boldsymbol{\chi} - \log B\left(\boldsymbol{\rho}_k\right) \right\} + K \log f\left(\nu, \boldsymbol{\chi}\right), \end{aligned} \tag{4}$$

where we have assigned the topic $z_{di} \in \{1, \ldots, K\}$ and timestamp $t_{di} \in (0, 1)$ to the word i on document d.

4.1.2 Variational Inference

It is well-known that the problem of computing the posterior distribution of the hidden variables given a document is intractable for LDA-based models. However, a variational approximation can be applied by considering a family of lower bounds. In the mean-field factorization approach, we consider a fully factorized distribution q^{BToT} which consists of the standard LDA ansatz q^{LDA} (see Eq. (40)) times K terms of the form $q_k(\boldsymbol{\rho}_k)$:

$$q^{\text{BToT}} = q^{\text{LDA}} \prod_k q_k(\boldsymbol{\rho}_k). \tag{5}$$

Notice that we are not imposing any explicit form for the $q_k(\boldsymbol{\rho}_k)$ distributions. The resulting evidence lower bound (ELBO) is

$$\mathcal{L}^{\text{BToT-ELBO}} = \left\langle \mathcal{L}^{\text{BToT}} - \log q^{\text{BToT}} \right\rangle_{q^{\text{BToT}}}. \tag{6}$$

The new terms that appear in the ELBO with respect to the standard LDA ELBO (see Eq. (41)), ${\mathcal{L}^{\text{BToT-ELBO}}}_{ts}$, read

$$\begin{aligned}
{\mathcal{L}^{\text{BToT-ELBO}}}_{ts} &= \left\langle {\mathcal{L}^{\text{BToT}}}_{ts} - \sum_{k=1}^{K} \log q_k(\boldsymbol{\rho}_k) \right\rangle_{q^{\text{BToT}}} \\
&= \sum_{dw} n_{dw} \sum_k \phi_{dwk} \left\{ \left(\left\langle \rho_k^1 \right\rangle_{q_k(\boldsymbol{\rho}_k)} - 1 \right) \log t_d + \left(\left\langle \rho_k^2 \right\rangle_{q_k(\boldsymbol{\rho}_k)} - 1 \right) \log\left[1 - t_d\right] \right. \\
&\quad \left. - \left\langle \log B(\boldsymbol{\rho}_k) \right\rangle_{q_k(\boldsymbol{\rho}_k)} \right\} \\
&\quad + \nu \sum_{k=1}^{K} \left\{ \left\langle \boldsymbol{\rho}_k \right\rangle_{q_k(\boldsymbol{\rho}_k)} \cdot \boldsymbol{\chi} - \left\langle \log B(\boldsymbol{\rho}_k) \right\rangle_{q_k(\boldsymbol{\rho}_k)} \right\} + K \log f(\nu, \boldsymbol{\chi}) \\
&\quad - \sum_{k=1}^{K} \left\langle \log q_k(\boldsymbol{\rho}_k) \right\rangle_{q_k(\boldsymbol{\rho}_k)},
\end{aligned} \tag{7}$$

where following the notation in Hoffman et al. (2010) we denote n_{dw} the number of occurrences of word w in document d and we write the mean field contribution to the topics attributions as $q\left(z_{dw}^k = 1 \,|\boldsymbol{\phi}_{dw}\right) = \phi_{dwk}$. We have assumed that there is only one timestamp t_d per document, so we have set $t_{di} = t_d$.

The extremization of Eq. (7) with respect to $q_k(\boldsymbol{\rho}_k)$, subject to integrability constraints $1 = \int\int d\boldsymbol{\rho}_k q_k(\boldsymbol{\rho}_k)$, is given by

$$\frac{\delta}{\delta q_k(\boldsymbol{\rho}_k)} \left\{ {\mathcal{L}^{\text{BToT-ELBO}}}_{ts} + \xi_k \sum_k \left(\int\int d\boldsymbol{\rho}_k q_k(\boldsymbol{\rho}_k)_k - 1 \right) \right\} = 0, \tag{8}$$

where we have introduced K Lagrange multipliers ξ_k. This leads (Bishop, 2006) to another Beta-prior distribution with parameters μ_k and $\boldsymbol{\psi}_k$ such that

$$\begin{aligned}
q_k(\boldsymbol{\rho}_k) &= f(\mu_k, \boldsymbol{\psi}_k) \exp\left[\mu_k \left(\boldsymbol{\psi}_k \cdot \boldsymbol{\rho}_k - \log B(\boldsymbol{\rho}_k)\right)\right] \\
\mu_k &= \nu + N_k \\
\boldsymbol{\psi}_k &= \mu_k^{-1} \left(N_k \boldsymbol{l}_k + \nu \boldsymbol{\chi}\right),
\end{aligned} \tag{9}$$

where

$$\begin{aligned}
N_k &= \sum_d N_{dk} \\
N_{dk} &= \sum_w n_{dw} \phi_{dwk},
\end{aligned} \tag{10}$$

are topic token average counts, and we have defined the vector of average log-timestamps as

$$\boldsymbol{l}_k = \left\langle \left(\log t_d, \log\left(1 - t_d\right)\right) \right\rangle_k := N_k^{-1} \sum_d N_{dk} \left(\log t_d, \log\left(1 - t_d\right)\right). \tag{11}$$

The update to the topic attributions multinomials is changed from Eq. (42) in standard LDA to

$$\begin{aligned}
\phi_{dwk} \propto \exp\Big[& \left\langle \log \theta_{dk} \right\rangle_{q^{\text{LDA}}} + \left\langle \log \beta_{kw} \right\rangle_{q^{\text{LDA}}} \\
& + \left(\left\langle \rho_k^1 \right\rangle_{q_k(\boldsymbol{\rho}_k)} - 1 \right) \log t_d + \left(\left\langle \rho_k^2 \right\rangle_{q_k(\boldsymbol{\rho}_k)} - 1 \right) \log\left(1 - t_d\right) - \left\langle \log B(\boldsymbol{\rho}_k) \right\rangle_{q_k(\boldsymbol{\rho}_k)} \Big].
\end{aligned} \tag{12}$$

Due to the fact that the normalization function $f(\mu_k, \boldsymbol{\psi}_k)$ in Eq. (9) cannot be calculated analytically, the averages $\left\langle \boldsymbol{\rho}_k \right\rangle_{q_k(\boldsymbol{\rho}_k)}$ and $\left\langle \log B(\boldsymbol{\rho}_k) \right\rangle_{q_k(\boldsymbol{\rho}_k)}$ that appear in this expression must be computed numerically. However, it can be useful to consider an asymptotic approximation of these integrals in order to build intuition and analyze the differences with respect to

TABLE 1: Observed data and other constants

Symbol	Description
$w_{di} \in \{1, \ldots, V\}$	Word corresponding to the ith token in document d
n_{dw}	Number of occurrences of word w in document d
N_d	Total number of tokens in document d
D	Total number of documents in the corpus
V	Total number of words in the vocabulary
$t_{di} = t_d \in (0, 1)$	Timestamp corresponding to the ith token in document d
$n_d^{(y)}$	Number of timestamp observations in document d (in WBToT)
K	Total number of topics

TABLE 2: Probability distribution parameters and latent variables

Symbol	Description
$\boldsymbol{\theta_d} = \{\theta_{dk}\}_{k=1,\ldots,K}$	Parameters of the document-topic multinomial distribution
$\boldsymbol{\alpha} = \{\alpha_k\}_{k=1,\ldots,K}$	Parameters of the Dirichlet prior for the topic-document multinomial distribution
$\boldsymbol{\beta_k} = \{\beta_{kw}\}_{w=1,\ldots,V}$	Parameters of the topic-word multinomial distribution
$\boldsymbol{\eta_k} = \{\eta_{kw}\}_{w=1,\ldots,V}$	Parameters of the Dirichlet prior for the topic-word multinomial distribution
$\boldsymbol{\rho}_k = \left(\rho_k^1, \rho_k^2\right)$	Parameters of the time-topic Beta distribution
$\nu, \boldsymbol{\chi} = \left(\chi^1, \chi^2\right)$	Parameters of the Beta-prior prior for the time-topic Beta distribution
$z_{di} \in \{1, \ldots, K\}$	Topic assignment to word token i in document d
$y_d \in \{1, \ldots, K\}$	Additional topic assignment to document d in WBToT
$\boldsymbol{\phi_{dw}} = \{\phi_{dwk}\}_{k=1,\ldots,K}$	Parameters of the variational multinomial distribution for topic assignment of word w in document d
$\boldsymbol{\gamma_d} = \{\gamma_{dk}\}_{k=1,\ldots,K}$	Parameters of the variational Dirichlet prior for the topic-document multinomial distribution
$\boldsymbol{\lambda_k} = \{\lambda_{kw}\}_{w=1,\ldots,V}$	Parameters of the variational Dirichlet prior for the topic-word multinomial distribution
$\mu_k, \boldsymbol{\psi}_k = \left(\psi_k^1, \psi_k^2\right)$	Parameters of the variational Beta-prior distribution
$\boldsymbol{\epsilon_d} = \{\epsilon_{dk}\}_{k=1,\ldots,K}$	Parameters of the variational multinomial distribution for additional topic assignment of document d in WBToT

standard ToT. To achieve this, notice that $N_k \gg 1$[2], so from Eqs. (9) - (11) it follows that $\mu_k \sim N_k$ and $\boldsymbol{\psi_k} \sim 1$ (we have assumed a moderate size for the model hyper-parameters, $\boldsymbol{\chi}, \nu \sim 1$, and that the order of magnitude notation "$\sim$" may include an arbitrary multiplicative constant). Therefore, one can use the leading Laplace approximation shown in Appendix D, Eqs. (65) - (66), which to order $O\left(N_k^{-1}\right)$ yields

$$\langle \boldsymbol{\rho}_k \rangle_{q_k} \simeq \boldsymbol{\rho}_{k0} \tag{13}$$

$$\langle \log B\left(\boldsymbol{\rho}_k\right) \rangle_{q_k} \simeq \log B\left(\boldsymbol{\rho}_{k0}\right), \tag{14}$$

where $\boldsymbol{\rho}_{k0}$ has been defined as

$$\boldsymbol{\rho}_{k0} = \mathrm{argmax}_{\boldsymbol{\rho_k}} \left\{ \boldsymbol{\rho_k} \cdot \boldsymbol{\psi_k} - \log \mathrm{B}\left(\boldsymbol{\rho_k}\right) \right\}, \tag{15}$$

and we have made use of standard big O notation. Note that Eq. (12) together with Eqs. (13) - (14) is identical to Eq. (50) in standard ToT. However, from its definition, it is immediate to see that $\boldsymbol{\rho}_{k0}$ now satisfies the equations

$$\frac{1}{1 + \nu/N_k} \left(\langle \log t_d \rangle_k + \frac{\nu}{N_k} \chi_1 \right) = \Psi\left(\rho_{k0}^1\right) - \Psi\left(\rho_{k0}^1 + \rho_{k0}^2\right) \tag{16}$$

$$\frac{1}{1 + \nu/N_k} \left(\langle \log\left(1 - t_d\right) \rangle_k + \frac{\nu}{N_k} \chi_2 \right) = \Psi\left(\rho_{k0}^2\right) - \Psi\left(\rho_{k0}^1 + \rho_{k0}^2\right), \tag{17}$$

where $\Psi(x) := d \log \Gamma(x) / dx$ is the digamma function. These equations constitute a regularized version of those in standard ToT (see Eqs. (51) - (52)), which can be recovered by setting $\nu = 0$ here. In the unregularized case, if $1 - \left(e^{\langle \log t_d \rangle_k} + e^{\langle \log(1 - t_d) \rangle_k} \right) \ll 1$ (which occurs for a topic k poorly represented in a mini-batch[3] or with a very peaked timestamp structure[4]), $\boldsymbol{\rho}_{k0}$ could grow by a large factor (Lau and Lau, 1991). However, within the present Bayesian approach, the hyper-parameters $\boldsymbol{\chi}$ and ν will automatically limit the growth of $\boldsymbol{\rho}_{k0}$, preventing instabilities. From a practical point of view, this is the main advantage of BToT with respect to ToT.

The other two variational parameters that appear in q^{LDA} (i.e. γ_{dk} and λ_{kw}) are the same as in standard LDA (see Eqs. (43) - (44) in Appendix A), since they are decoupled from the timestamp part of the log-likelihood. The complete algorithm for variational inference in BToT is shown in Algorithm 1.

4.1.3 Online optimization

Online optimization requires the computation of the natural gradient (Hoffman et al., 2010; Hoffman, Blei, Wang and Paisley, 2013) of the one-document contribution to the ELBO, l_d. From Eq. (7), we see that this contribution is given by

2. Unless one proposes a very large number of topics (or a small minibatch size in the case of online learning).
3. Very likely for a large number of topics and/or small mini-batch sizes.
4. Common sense indicates that this situation is probably marginal, unless the mini-batch come with very close timestamps.

Algorithm 1 Batch variational BToT

Initialize variational parameters $\gamma_{dk}, \lambda_{kw}, \mu_k, \boldsymbol{\psi}_k$.
repeat
  E-step:
  for $d = 1$ to D **do**
    repeat
      Set ϕ_{dwk} from Eq. (12).
      Set γ_{dk} from Eq. (43).
    until γ_{dk} converged.
  end for
  M-step:
  Set λ_{kw} from Eq. (44).
  Set $\mu_k, \boldsymbol{\psi}_k$ from Eq. (9).
until $\mathcal{L}$ converged.

$$l_d := \sum_{k=1}^{K} N_{dk} \left\{ \left(\langle \boldsymbol{\rho} \rangle_{q_k} - \mathbf{1} \right) \cdot \boldsymbol{lt}_d - \langle \log B(\boldsymbol{\rho}) \rangle_{q_k} \right\} \tag{18}$$
$$+ \frac{1}{D} \sum_{k=1}^{K} \left\{ \langle \boldsymbol{\rho} \rangle_{q_k} \cdot (\nu \boldsymbol{\chi} - \mu_k \boldsymbol{\psi}_k) - \langle \log B(\boldsymbol{\rho}) \rangle_{q_k} (\nu - \mu_k) + \log f(\nu, \boldsymbol{\chi}) - \log f(\mu_k, \boldsymbol{\psi}_k) \right\},$$

where we have defined

$$\boldsymbol{lt}_d := (\log t_d, \log [1 - t_d]), \tag{19}$$

and we have explicitly exhibited the dependence on μ_k and $\boldsymbol{\psi}_k$ by making use of Eq. (9). For convenience, we will make use of the "natural" variables $\mu'_k := \mu_k, \boldsymbol{\psi}'_k := \mu_k \boldsymbol{\psi}_k$. The natural gradients, $\hat{\nabla}$, in these coordinates are

$$\hat{\nabla}_{\mu'_k} l_d = N_{dk} + \frac{1}{D} (\nu - \mu'_k), \tag{20}$$

$$\hat{\nabla}_{\boldsymbol{\psi}'_k} l_d = N_{dk} \boldsymbol{lt}_d + \frac{1}{D} (\nu \boldsymbol{\chi} - \boldsymbol{\psi}'_k). \tag{21}$$

Given the values of the variables $\mu'_k(t)$ and $\boldsymbol{\psi}'_k(t)$ at iteration t, maximization leads to the following online updates after a mini-bath (MB) of S documents,

$$\mu'_k(t+1) = (1 - \rho_t) \mu'_k(t) + \rho_t \left(\nu + \frac{D}{S} \sum_{d \in \text{MB}} N_{dk} \right) \tag{22}$$

$$\boldsymbol{\psi}'_k(t+1) = (1 - \rho_t) \boldsymbol{\psi}'_k(t) + \rho_t \left(\nu \boldsymbol{\chi} + \frac{D}{S} \sum_{d \in \text{MB}} N_{dk} \boldsymbol{lt}_d \right), \tag{23}$$

where $\rho_t := (t + \tau)^{-\kappa}$ is the standard "mixing" parameter (Hoffman et al., 2010, 2013). Here, $\kappa \in (0.5, 1]$ regulates how previous values of the parameters are forgotten, and $\tau \geq 0$ slows down the first iterations.

The complete algorithm is shown in Algorithm 2.

4.2 Weighted Bayesian ToT (WBToT)

4.2.1 Description of the model

Notice that, although a single timestamp, t_d, is observed for each document d, the BToT model sees it N_d times during training, since we generate a timestamp for each word of the document. As we will empirically see in Section 6, this places excessive importance on timestamps while relegating semantic coherence of topics to a secondary place. To balance the relative weight between words and timestamps and rescale both modalities of the log-likelihood, a balancing hyperparameter is typically introduced (Wang and McCallum, 2006; Fang et al., 2017).

However, as exposed in Appendix E, it is difficult to follow this approach from a Bayesian perspective. To solve this problem, we slightly modify BToT to mimic the introduction of this balancing hyperparameter while keeping the Bayesian approach. We will call this model Weighted BToT (WBToT). In WBToT, an additional latent topic assignment variable, $y_d \in \{1, \ldots, K\}$, is introduced to account for the generation of timestamp observations from topics, instead of using the same z_{di} we employed for word generation. These new topic assignments are drawn from the same multinomial distributions as z, i.e. $p(y_d | \theta_d) = \theta_{dy_d}$. The observation of the document's timestamp through y is repeated $n_d^{(y)}$ times

Algorithm 2 Online variational BToT

Set $t = 0$.
Initialize variational parameters $\lambda_{kw}(t), \mu_k(t), \boldsymbol{\psi}_k(t)$.
repeat
  Set $t \leftarrow t + 1$.
  Set $\rho_t := (t + \tau)^{-\kappa}$.
  E-step:
  for $d = 1$ to S **do**
    Initialize variational parameters γ_{dk}.
    repeat
      Set ϕ_{dwk} from Eq. (12).
      Set γ_{dk} from Eq. (43).
    until γ_{dk} converged.
  end for
  M-step:
  Set $\lambda_{kw}(t + 1)$ from Eq. (47).
  Set $\mu_k(t + 1), \boldsymbol{\psi}_k(t + 1)$ from Eq. (22) and Eq. (23).
until $\mathcal{L}$ converged.

to mimic the effect of the balancing hyperparameter. This way, we can effectively control the relative weight of the two modalities of the log-likelihood.

The generative process for the WBToT model is as follows:

1) Draw K multinomial parameters $\boldsymbol{\beta_k}$ from K Dirichlet distributions with parameters $\boldsymbol{\eta_k}$.
2) Draw K Beta distribution parameters $\boldsymbol{\rho}_k$ from K Beta-prior distributions with parameters $\nu, \boldsymbol{\chi}$.
3) For each document d, draw a set of multinomial parameters $\boldsymbol{\theta_d}$ from a symmetric Dirichlet distribution with parameters $\boldsymbol{\alpha}$.
   a) Then, for each word index $i \in \{1, \ldots, N_d\}$ in document d:
      i) Draw a topic z_{di} from a multinomial with parameters $\boldsymbol{\theta_d}$;
      ii) Draw a word w_{di} from a multinomial with parameters $\boldsymbol{\beta_{z_{di}}}$;
   b) Draw $n_d^{(y)}$ topics y_d^i, with $i \in \{1, \ldots, n_d^{(y)}\}$, from a multinomial with parameters $\boldsymbol{\theta_d}$. For each of them:
      i) Draw a timestamp t_d^i from a Beta distribution with parameters $\boldsymbol{\rho}_{y_d^i}$.

See Fig. 1b for the directed acyclic graph representation of this generative process. The introduction of these repeated timestamp observations allows balancing the relative influence of words and timestamps along the inference process. For example, $n_d^{(y)} = 1$ corresponds strictly to the Bayesian version of the alternative ToT model depicted in Fig. 10b and $n_d^{(y)} = \delta N_d$ approximately corresponds to a Bayesian version of the ToT model with balance hyperparameter δ introduced in Fang et al. (2017).

The resulting log-likelihood for WBToT is

$$\mathcal{L}^{\text{WBToT}} = \mathcal{L}^{\text{LDA}} + \sum_{d=1}^{D} \sum_{i=1}^{n_d^{(y)}} \left\{ \log \underbrace{p\left(t_d^i | \boldsymbol{\rho}_{y_d^i}\right)}_{\text{Beta}} + \log \underbrace{p\left(y_d^i | \boldsymbol{\theta_d}\right)}_{\text{Multinomial}} \right\} + \sum_{k=1}^{K} \log \underbrace{p\left(\boldsymbol{\rho}_k | \nu, \boldsymbol{\chi}\right)}_{\text{Beta-prior}}. \tag{24}$$

For reference, let us flesh out the terms in the log-likelihood that depend on timestamps and their priors (which we will denote $\mathcal{L}^{\text{WBToT}}{}_{ts}$):

$$\begin{aligned} \mathcal{L}_{ts}^{\text{WBToT}} = & \sum_{d=1}^{D} \sum_{i=1}^{n_d^{(y)}} \left\{ \left(\rho_{y_d^i}^1 - 1\right) \log t_d^i + \left(\rho_{y_d^i}^2 - 1\right) \log\left[1 - t_d^i\right] - \log B\left(\boldsymbol{\rho}_{y_d^i}\right) + \log \theta_{d y_d^i} \right\} \\ & + \nu \sum_{k=1}^{K} \left\{ \boldsymbol{\rho}_k \cdot \boldsymbol{\chi} - \log B\left(\boldsymbol{\rho}_k\right) \right\} + K \log f\left(\nu, \boldsymbol{\chi}\right). \end{aligned} \tag{25}$$

4.2.2 Variational Inference

The variational ansatz is of the form

$$q^{\text{WBToT}} = q^{\text{LDA}} \prod_{d=1}^{D} \left(\prod_{i=1}^{n_d^{(y)}} \underbrace{q\left(y_d^i | \boldsymbol{\epsilon_d^i}\right)}_{\text{Multinomial}} \right) \prod_{k=1}^{K} \underbrace{q_k\left(\boldsymbol{\rho}_k | \mu_k, \boldsymbol{\psi}_k\right)}_{\text{Beta-prior}}, \tag{26}$$

where q^{LDA} is given by Eq. (40), and we have introduced a new set of multinomial variational parameters ϵ^i_{dk}. The resulting ELBO is

$$\mathcal{L}^{\text{WBToT-ELBO}} = \left\langle \mathcal{L}^{\text{WBToT}} - \log q^{\text{WBToT}} \right\rangle_{q^{\text{WBToT}}}. \tag{27}$$

The new terms that appear in the ELBO with respect to the standard LDA ELBO (see Eq. (41)), $\mathcal{L}^{\text{WBToT-ELBO}}{}_{ts}$, read

$$\begin{aligned}
\mathcal{L}^{\text{WBToT-ELBO}}{}_{ts} &= \left\langle \mathcal{L}^{\text{WBToT}}{}_{ts} - \sum_{d=1}^{D}\sum_{i=1}^{n_d^{(y)}} \log q\left(y_d^i\right) - \sum_{k=1}^{K} \log q_k\left(\boldsymbol{\rho}_k\right) \right\rangle_{q^{\text{WBToT}}} \\
&= \sum_{d=1}^{D}\sum_{i=1}^{n_d^{(y)}}\sum_{k}^{K} \epsilon^i_{dk} \Big\{ \Big(\langle \boldsymbol{\rho}_k \rangle_{q_k(\boldsymbol{\rho}_k)} - \mathbf{1} \Big) \cdot \boldsymbol{lt}_d - \langle \log B\left(\boldsymbol{\rho}_k\right) \rangle_{q_k(\boldsymbol{\rho}_k)} \\
&+ \langle \log \theta_{dk} \rangle_{q^{\text{LDA}}} - \log \epsilon^i_{dk} \Big\} \\
&+ \sum_{k=1}^{K} \Big\{ \langle \boldsymbol{\rho}_k \rangle_{q_k(\boldsymbol{\rho}_k)} \cdot \left(\nu \boldsymbol{\chi} - \mu_k \boldsymbol{\psi}_k\right) - \langle \log B\left(\boldsymbol{\rho}_k\right) \rangle_{q_k(\boldsymbol{\rho}_k)} \left(\nu - \mu_k\right) \\
&+ \log f\left(\nu, \boldsymbol{\chi}\right) - \log f\left(\mu_k, \boldsymbol{\psi}_k\right) \Big\},
\end{aligned} \tag{28}$$

where the average $\langle \log \theta_{dk} \rangle_{q^{\text{LDA}}}$ is given by Eq. (45), as in standard LDA. Notice that, due to the introduction of γ_{dk} through this term in the ELBO, the update equation for this variational parameter will be modified (this is a novelty with respect to BTOT, where γ_{dk} is the same as in LDA). On the other hand, ϕ_{dwk} has been replaced by ϵ_{dk} everywhere in this part of the ELBO.

Extremization of $\mathcal{L}^{\text{WBToT-ELBO}}$ (subject to the new constraint $\sum_k \epsilon^i_{dk} = 1$) leads to the following document specific variational parameters updates

$$\epsilon_{dk} := \epsilon^i_{dk} \propto \exp\left[\langle \log \theta_{dk} \rangle_{q^{\text{LDA}}} + \Big(\langle \boldsymbol{\rho}_k \rangle_{q_k(\boldsymbol{\rho}_k)} - \mathbf{1} \Big) \cdot \boldsymbol{lt}_d - \langle \log B\left(\boldsymbol{\rho}_k\right) \rangle_{q_k(\boldsymbol{\rho}_k)} \right] \tag{29}$$

$$\gamma_{dk} = \alpha_k + \underbrace{\sum_{w} n_{dw} \phi_{dwk}}_{N_{dk}} + n_d^{(y)} \epsilon_{dk}, \tag{30}$$

where we have explicitly noted that ϵ^i_{dk} does not depend on its index i because there is a single timestamp observation per document. From these formulae, it follows that the choice $n_d^{(y)} = 1$ behaves as a single word added to the document-topic pseudo-counts γ_{dk}, and therefore the timestamp influence will be small. Since ϕ_{dwk} is not present in $\mathcal{L}^{\text{WBToT-ELBO}}{}_{ts}$, its update equation will be the same as in LDA, Eq. (42).

Corpus-wide variational parameters obey the following self-consistency relations:

$$\mu_k = \nu + \sum_d n_d^{(y)} \epsilon_{dk} \tag{31}$$

$$\boldsymbol{\psi}_k = \mu_k^{-1} \left(\nu \boldsymbol{\chi} + \sum_d n_d^{(y)} \epsilon_{dk} \boldsymbol{lt}_d \right), \tag{32}$$

and the λ_{kw} variational parameter is exactly the same as in standard LDA (see Eq. (44) in Appendix A), since it is decoupled from the timestamp part of the log-likelihood.

It is interesting to note that the variational parameters satisfy $e^{\psi^1} + e^{\psi^2} < \left(e^{\chi^1} + e^{\chi^2} \right)^{\nu/\mu}$ (the k-dependence is implicit). This can be proved by using the generalized Hölder inequality. If we assume $n_d^{(y)} \gg 1$, Eqs. (31) - (32) imply that the inequality will be close to saturation. As a consequence, for non very strong Beta-prior parameters (namely $\nu \lesssim 1$), the Beta-prior variational parameters will be close to the boundary of their validity regions, $1 - \left(e^{\psi^1} + e^{\psi^2} \right) \ll 1$.

The updates to the Beta-prior hyper-parameters are the same as in BToT. The complete algorithm is shown in Algorithm 3.

4.2.3 Online optimization

In the case of online updates, it is immediate to see that one only has to substitute $N_{dk} \to n_d^{(y)} \epsilon_{dk}$ in Eqs. (22) - (23), that is:

$$\mu'_k\left(t+1\right) = \left(1 - \rho_t\right) \mu'_k\left(t\right) + \rho_t \left(\nu + \frac{D}{S} \sum_{d \in MB} n_d^{(y)} \epsilon_{dk} \right), \tag{33}$$

$$\boldsymbol{\psi}'_k\left(t+1\right) = \left(1 - \rho_t\right) \boldsymbol{\psi}'_k\left(t\right) + \rho_t \left(\nu \boldsymbol{\chi} + \frac{D}{S} \sum_{d \in MB} n_d^{(y)} \epsilon_{dk} \boldsymbol{lt}_d \right). \tag{34}$$

The complete algorithm for online optimization is shown in Algorithm 4.

Algorithm 3 Batch variational WBToT

Initialize variational parameters $\gamma_{dk}, \lambda_{kw}, \mu_k, \boldsymbol{\psi}_k, \epsilon_{dk}$.
repeat
  E-step:
  for $d = 1$ to D **do**
    repeat
      Set ϕ_{dwk} from Eq. (42).
      Set $\epsilon_{dk}, \gamma_{dk}$ from Eq. (29) and Eq. (30).
    until γ_{dk} converged.
  end for
  M-step:
  Set λ_{kw} from Eq. (44).
  Set $\mu_k, \boldsymbol{\psi}_k$ from Eq. (31) and Eq. (32).
until $\mathcal{L}$ converged.

Algorithm 4 Online variational WBToT

Set $t = 0$.
Initialize variational parameters $\lambda_{kw}(t), \mu_k(t), \boldsymbol{\psi}_k(t)$.
repeat
  Set $t \leftarrow t + 1$.
  Set $\rho_t := (t + \tau)^{-\kappa}$.
  E-step:
  for $d = 1$ to D **do**
    Initialize variational parameters $\gamma_{dk}, \epsilon_{dk}$.
    repeat
      Set ϕ_{dwk} from Eq. (42).
      Set $\epsilon_{dk}, \gamma_{dk}$ from Eq. (29) and Eq. (30).
    until γ_{dk} converged.
  end for
  M-step:
  Set $\lambda_{kw}(t + 1)$ from Eq. (47).
  Set $\mu_k(t + 1), \boldsymbol{\psi}_k(t + 1)$ from Eq. (33) and Eq. (34).
until $\mathcal{L}$ converged.

4.3 Time complexity analysis

In this section, we provide the time complexity analysis of the two models introduced in our paper. For reference, we also include LDA and standard ToT in our analysis.

Let N_{ud} be the number of unique tokens in document d, N_g the number of iterations in the E-step until γ_{dk} convergence, N_r the number of iterations to find roots in ToT E-step, and T the number of unique document timestamps.

It must be noted that N_g and N_r depend very weakly on the corpus and the number of topics. Strictly, they should not be included in the complexity computations. However, they usually amount to $10^1 - 10^3$, which can be quite significant, typically on the order or larger than the number of topics.

The complexities of variational inferences in the different models are:

1) LDA: each E-step requires setting ϕ_{dwk}, which needs $O(N_{ud} \times K)$ operations. This has to be iterated until convergence (of γ_{dk}), which requires N_g steps. The resulting complexity is $O\left(N_g \times D \times N_{ud} \times K\right)$. The M-step requires $O(D \times N_{ud} \times K + V \times K) = O(D \times N_{ud} \times K)$ operations, which is smaller than the E-step complexity.
2) ToT: on top of LDA complexity, there is another $O(N_r \times D \times N_{ud} \times K)$ contribution to complexity in the E- step. The contribution to the M-step can be neglected. The resulting complexity is $O((N_g + N_r) \times D \times N_{ud} \times K)$.
3) BToT and WBToT: there is a $O(T \times K)$ added complexity to LDA's complexity in E-Step. The resulting complexity is $O(N_g \times D \times N_{ud} \times K) + O(T \times K)$.

5 Datasets and experimental set-up

We will employ two datasets to validate our models. The datasets are very different in size and nature, and each of them will be used for different purposes. The first dataset consists of the annual State of the Union Addresses (SOTU) by US Presidents from 1790 to 2022, and will be used to compare different topic models between them. The second dataset is a massive corpus of tweets about the COVID-19 pandemic from March 2020 to June 2020, and will be employed to illustrate the use of the online algorithm in WBToT for a case that may otherwise be intractable in a batch setting due to memory constraints.

5.1 State of the Union Addresses

5.1.1 Dataset

As in the original work that introduced the ToT model (Wang and McCallum, 2006), we will use the State of the Union (SOTU) Addresses as our main dataset. The SOTU is an annual speech given by the president of the United States describing the economic, political, and social state of the country. We will use an updated dataset made of the 236 addresses given in the range 1790-2022 (from George Washington to Joe Biden). As in Wang and McCallum (2006), we split each address into 3-paragraph documents, which leaves us with a corpus of 7224 documents written in English. For each of these documents, the publication timestamp is the year in which the address was given.

We perform a standard text preprocessing on this corpus using the stanza package (Qi, Zhang, Zhang, Bolton and Manning, 2020): we tokenize the documents, remove stopwords and numbers, lemmatize the tokens, and filter out words that appear in less than 15 documents or in more than 50% of the documents. This leaves us with a vocabulary of 4086 words. Finally, we create a bag-of-words representation of the corpus. The dataset statistics are summarized in Table 3.

5.1.2 Experimental set-up

This is a relatively small dataset that we can employ to compare different topic models between them since training time is restrained. We will train LDA, BERTopic, BToT, and two WBToT models on this dataset (the two WBToT models will differ on the $n_d^{(y)}$ parameter, with one of them having $n_d^{(y)} = 0.1N_d$ and the other $n_d^{(y)} = \sqrt{N_d}$). In all cases, we will look for 50 topics (for BERTopic, this implies reducing the original number of topics by iterative merging).

In probabilistic topic models (i.e LDA, BToT, WBToT), we will optimize the standard hyper-parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\eta_k}$ (initialized to $1/K$), but not the novel ones here introduced, ν and $\boldsymbol{\chi}$ (which will be initialized to $1/K$ and $\log 0.45\,(1, 2)$ respectively). It is well known that topic models reach different optimal values depending on the parameter initialization. To alleviate this problem, for each model we will run 10 batch simulations with different parameter initializations, setting as our stopping criterion that the change in perplexity, $\mathcal{P}^t := e^{-\mathcal{L}^t/N_{tok}}$, is $\mathcal{P}^{t-1} - \mathcal{P}^t < 10^{-5}$ (where N_{tok} is the total number of tokens in the corpus and $\mathcal{L}^t$ is the ELBO after t iterations of the EM algorithm). We will also establish a maximum limit of 400 EM iterations, so the algorithm is stopped if the convergence criterion has not been satisfied by then. For each topic model, we will choose the simulation with the largest ELBO out of the 10 trials as our final result. In the WBToT models, the two options $n_d^{(y)} = 0.1N_d$ and $n_d^{(y)} = \sqrt{N_d}$ are chosen so that on average the timestamps contribute with equal weight (here $\langle N_d \rangle = 89.7$ and $\langle \sqrt{N_d} \rangle = 8.95$), but $n_d^{(y)} = \sqrt{N_d}$ weights timestamps more in documents with $N_d < 100$. For BERTopic, we will use the default parameters. The results for this experiment are presented in Section 6.1 - 6.2.

The absence of ν and $\boldsymbol{\chi}$ optimization requires some clarification. We first note that, as explained in Section 1, the main reasons to introduce the Beta-prior were to obtain a stable online variational inference algorithm and to provide a principled approach to regularizing large gradients. The other usual reason for introducing a prior, overfitting, is not relevant in our case because the Beta distribution won't overfit, as stated in Wang and McCallum (2006). When we tried to optimize ν and $\boldsymbol{\chi}$, we noted that both parameters migrated to the boundaries of their validity region. This behavior removes the regularizing effect that is required to handle large gradients. We guess that, because of the lack of overfitting, the system just tries to make the regularizing term as small as possible. A more detailed study of the optimization of ν and $\boldsymbol{\chi}$ is left for future research.

5.2 Large-Scale COVID-19 Twitter Chatter

5.2.1 Dataset

The Large-Scale COVID-19 Twitter Chatter dataset introduced by (Banda, Tekumalla, Wang, Yu, Liu, Ding, Artemova, Tutubalina and Chowell, 2021). This dataset consists of over 1.12 billion tweets related to COVID-19 chatter in different languages. Data was collected from the Twitter Stream API filtering by keywords such as "coronavirus", "2019nCoV", "COVD19", "CoronavirusPandemic", "COVID-19", "CoronaOutbreak", and other sensitive terms, plus additional collections of tweets provided by other researchers (we refer the reader to Banda et al. (2021) for a detailed description of the dataset). Misinformation in social media during the COVID-19 pandemic has attracted considerable attention from the ML community (Burel, Farrell and Alani, 2021), so this is a compelling dataset to test our model on. In particular, we have restricted ourselves to tweets written in English that are not retweets from 12 March 2020 to 12 June 2020, comprising a three months period. Furthermore, we have only sampled 10 million random tweets from this subset for efficiency purposes.

We have performed a standard text preprocessing on this corpus using the NLTK (Bird, Klein and Loper, 2009) TweetTokenizer class. As in the previous dataset, we have lemmatized the documents and removed stopwords and numbers. Additionally, we have removed Twitter usernames, hashtags, and URLs. Words that appear in less than 5000 tweets or more than 50% of the tweets are removed. This leaves us with a vocabulary of 2611 words. For each tweet, the publication timestamp is the day on which it was published. The dataset statistics are summarized in Table 3.

5.2.2 Experimental set-up

This is a large dataset that we will employ to illustrate the ability of WBToT to deal with massive corpora in an online fashion. Its large size makes it unsuitable for model comparisons, but it provides a good scenario to study the evolution of a model as successive batches of data arrive. The dataset will be shuffled and 10% of the corpus will be set apart

TABLE 3: Dataset Statistics

	SOTU addresses	Twitter
Dataset size	7224 documents	10^7 documents
Vocabulary size	4086 words	2611 words
Timespan	332 years	3 months
Granularity	1 year	1 day
$\langle N_d \rangle$	89.7 words	6.89 words

TABLE 4: Robust dispersion statistics for each model

Model	Median Absolute Deviation (MAD)	Interquartile Range (IQR)
WBToT($n_d^y = 0.1N_d$)	9.36 years	23.37 years
WBToT($n_d^y = \sqrt{N_d}$)	11.76 years	27.25 years
LDA	19.27 years	41.87 years
BERTopic	14.18 years	31.58 years

for testing purposes: the goal is to show that, for each optimization step taken after receiving a mini-batch of data, the held-out perplexity on this test set decreases. We will train a WBToT model with $n_d^{(y)} = \sqrt{N_d}$ and a batch size of 1 million documents (which implies that 9 optimizations will be performed for each iteration of the EM algorithm). As in the SOTU experiment, ν and $\boldsymbol{\chi}$ are not optimized. The number of topics will be 10 and only one EM iteration will be executed, since we are interested in studying the effect of successive batches of data, not the convergence of the EM algorithm itself.

Results for this experiment are presented in Section 6.3.

6 Results and Implications

In this section, we present the results of the experiments described above. Recall that the main advantage of introducing a time modality into the log-likelihood is the ability to capture events: we will study this property in Section 6.1. However, this time modality could degrade the model's word modeling performance: for this reason, we will study coherence in Section 6.2. Finally, we will illustrate the stability of WBToT in Section 6.3.

It must be noted BToT generates very low-quality topics, as already discussed in Section 4.2. We observed that, in many cases, the model generates completely uniform word distributions that are impossible to interpret. For this reason, we have not included BToT results in Section 6.1, and further motivation for this will be analyzed in the discussion about coherence in Section 6.2.

6.1 Event modeling

Due to the introduction of a time modality in the log-likelihood, events are better captured in WBToT than in time-unaware topic models (where no information about the timestamp is provided during training). This implies two things: first, that topic presence over time is more concentrated in WBToT, since topics describe particular events rather than occurred in a specific period of time; and second, that topics are semantically more distinct between them in WBToT, since a more singular vocabulary is necessary to describe specific events. Although this feature was already present in standard ToT (Wang and McCallum, 2006), it is important to note that we can reproduce these results using a fully Bayesian approach. To demonstrate this, we will use the topics extracted in the SOTU dataset (see Section 5.1) for LDA, BERTopic and the two versions of WBToT.

6.1.1 Topic concentration in time

Topic presence over time can be defined as the number of tokens assigned to each topic per unit of time (that is, $N_{kt} \coloneqq \sum_d^D \sum_w^V \delta_{tt_d} \theta_{dk} n_{dw}$, where t_d is the publication timestamp of document d and δ_{tt_d} is a Kronecker delta of t and t_d). In order to prove that topics are more concentrated in time in WBToT, we have to introduce a measure of the dispersion of N_{tk} for a fixed k. Due to the presence of outliers, which distorts the meaning of typical measures of variability such as the standard deviation, we will make use of robust metrics of statistical dispersion: the median absolute deviation (MAD) and the interquartile range (IQR). The MAD is defined as the median of the absolute deviations from the data's median $\tilde{N}_k = \text{median}\,(N_{tk})$,

$$\text{MAD} = \text{median}\left(\left|N_{tk} - \tilde{N}_k\right|\right). \tag{35}$$

The IQR is defined as the difference between the 75th percentile, Q_3, and the 25th percentile, Q_1,

$$\text{IQR} = Q_3 - Q_1. \tag{36}$$

For each model that we want to compare, we will compute these metrics for each of the K topics, and then we will take its mean value. Results are shown in Table 4.

It is clear that topic dispersion is greater in LDA than in both WBToT models. In particular, we observe that the MAD for the WBToT models is approximately a decade, while for LDA it goes up to almost two decades. The IQR comprises

TABLE 5: Mean semantic distance for each model

Model	Mean semantic distance
WBToT($n_d^y = 0.1N_d$)	13.31
WBToT($n_d^y = \sqrt{N_d}$)	13.23
LDA	12.74

between two and three decades for the WBToT models, while it goes beyond four decades for the LDA. These results imply that topics in LDA are almost half as concentrated in time as topics in WBToT models. BERTopic seems to be more concentrated in time than LDA, but it is still far from WBToT. Differences between the two WBToT models are not very significant, although it seems that the model with $n_d^y = 0.1N_d$ is slightly more concentrated in time than the model with $n_d^y = \sqrt{N_d}$.

6.1.2 Topic semantic distance

To prove this point, first, we have to introduce a notion of semantic distance between topics. We will employ the symmetric Kullback-Leibler divergence, defined as

$$d_{KL}^{\text{sym}}(P, Q) = D_{KL}(P||Q) + D_{KL}(Q||P), \tag{37}$$

where $D_{KL}(P||Q)$ is the Kullback-Leibler divergence of distribution P from distribution Q. We will compute the symmetric Kullback-Leibler divergence between the word distributions β_{kw} for all the possible pairs of topics for each of the three models (from which we obtain 1225 distances in each case) and then we will take its mean value.

Notice that this metric is not well-defined for BERTopic, since the Kullback-Leibler divergence requires that, for all x, if $Q(x) = 0$, then $P(x) = 0$. Probabilistic topic models exhibit a smooth topic distribution thanks to the introduction of priors, which prevents them from suffering from this problem. However, BERTopic makes use of TF-IDF to obtain word distributions, and this implies that the distributions may contain zeros. Therefore, we will exclude BERTopic from this analysis.

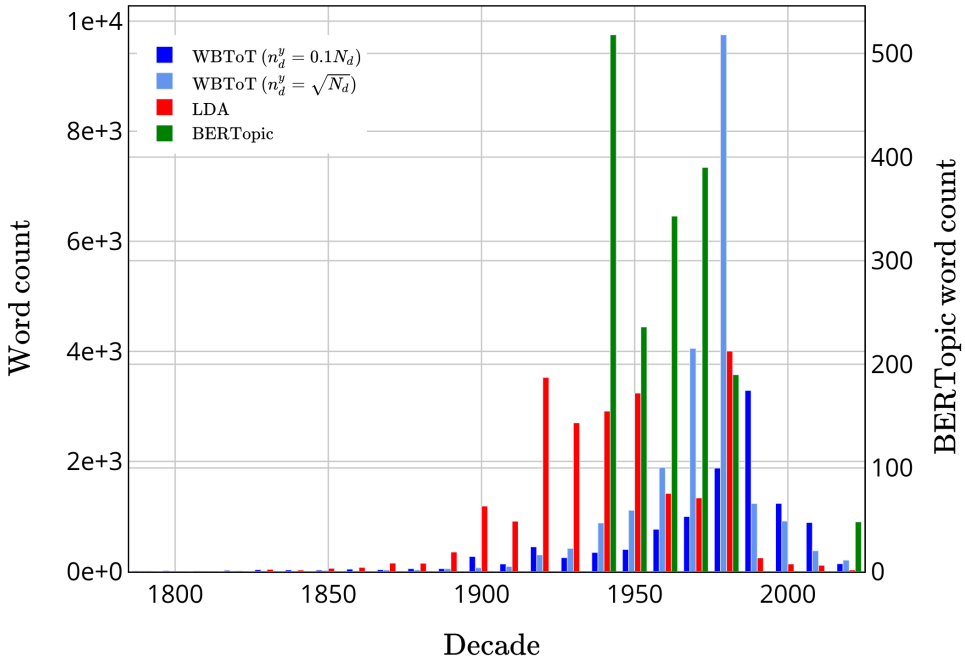
Results are presented in Table 5. Both WBToT models show a higher intra-model semantic distance between pairs of topics than LDA. This result is due to the fact that WBToT separates topics corresponding to different periods of time, and documents published during the same time span tend to share a similar vocabulary.

6.1.3 Some examples

In order to get an intuition of how well events are captured across the different models in this experiment (LDA, BERTopic, and the two WBToT models), let us sample some topics that show a clear link with historical events. For each topic in one model, we will identify the closest topics in the other models. The definition of "closest topics" is according to the symmetric KL divergence introduced in Eq. (37) for LDA and WBToT, and to word overlaps in the case of BERTopic (due to its problem with Kullback-Leibler divergence, as noted in the previous section). In each case, we will plot the presence of each topic over time (represented as a histogram over decades) and the top 20 words that characterize them. Results are shown in Fig. 2 - 5. We use two tones of blue to represent the two WBToT models, red for LDA and green for BERTopic.

In Fig. 2, we observe a set of topics associated with inflation. In particular, it seems that the WBToT models have captured the event known as “The Great Inflation”, a historical period of high inflation that peaked around the 1980s. The top words in these two models make clear reference to concepts such as “high”, “inflation”, “increase”, “percent”, and “cost”. None of these words appear in the LDA model, which seems to deal with more general economic issues. Notice that the topic presence over time is almost flat during most part of the 20th century for LDA, while histograms for both WBToTs models exhibit clear peaks around the 80s. To us, this is an indication that the WBToT models have captured a concrete period of American history characterized by high inflation, while the LDA is modeling economic issues in general. BERTopic correctly forms a topic about inflation, with a set of top words that resemble those of WBToT. However, this topic is not sufficiently populated in BERTopic: we have been forced to use two different y-axes in Fig. 2 (one specific for BERTopic) due to the small number of documents that have been assigned to this cluster. The peak is not clear either, since the topic distribution spans the entire second half of the 20th century.

In Fig. 3 a set of topics associated to the healthcare system reform has been discovered. In this case, LDA performs pretty well in comparison with both WBToT. However, the WBToT model with $n_d^{(y)} = 0.1N_d$ seems to capture better than any other model the debate around the Affordable Care Act (colloquially known as Obamacare), a polemic reform that was signed into law in 2010 and came into force during the rest of the decade. This WBToT model shows a clear peak in the 2010s with some important repercussions also in the 2020s and residual presence before that. The WBToT model with $n_d^{(y)} = \sqrt{N_d}$ actually shows a higher peak during the 2010s, although its presence is also noticeable in the preceding decades, so it has not been able to capture the event as precisely as the first one. The LDA model, on the other hand, shows a more dispersed histogram with two different peaks (one in the 2010s and the other in the 1990s) and tends to capture a broader vocabulary related to other types of reforms (notice that the keywords “health” and “care”, which are present in both WBToT models, do not appear in the LDA top words). BERTopic doesn’t exhibit the expected presence over time, with a peak around the 1990s. The top words are pretty cohesive and clearly describe a topic about healthcare, but its distribution over time seems generic.



The Great Inflation

WBToT($\mathbf{n_d^y = 0.1N_d}$) **top words**: high, inflation, growth, increase, plan, percent, government, people, cost, economy, american, reduce, spending, income, pay, rate, tax, cut, budget, deficit.

WBToT($\mathbf{n_d^y = \sqrt{N_d}}$) **top words**: deficit, inflation, job, economic, cut, spending, increase, federal, congress, economy, growth, reduce, budget, tax, program, percent, rate, administration, high, government.

LDA top words: private, provide, system, responsibility, administration, local, development, policy, program, national, economic, congress, problem, federal, continue, nation, effort, state, government, work.

BERTopic top words: inflation, price, percent, wage, increase, growth, rise, economy, control, cost, profit, hold, high, level, job, rate, economic, food, inflationary, goods.

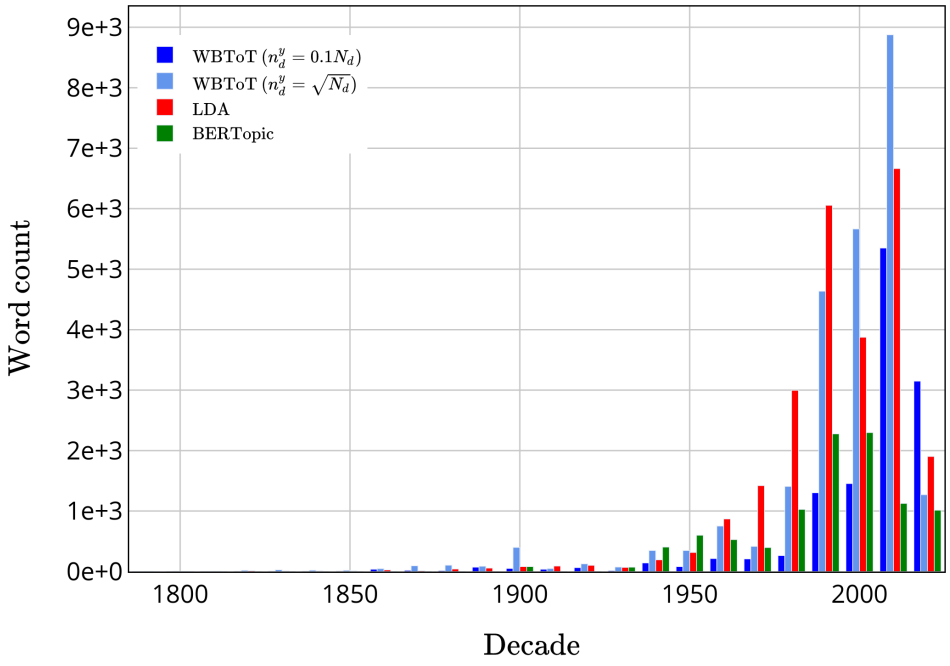
Fig. 2: Presence over time and 20 top words for the "The Great Inflation" topic.

In Fig. 4 we see a set of topics that captures some of the main conflicts that the incipient US faced on North American territory during the 19th century: in particular, conflicts with American Indians and with Mexico. It is interesting to note that the WBToT model with $n_d^{(y)} = 0.1N_d$ is more focused on the Mexican-American conflicts, with a clear peak in the 1840s (the Mexican-American war took place between April 1846 and February 1848, so it seems to capture this event in a very precise manner), while the WBToT model with $n_d^{(y)} = \sqrt{N_d}$ put more emphasis on American-Indian conflicts, with a peak in a much earlier period, around the 1810s. The LDA, however, exhibits a more disperse distribution that spans the entire 19th century (with three different peaks at different times), and includes other unrelated events such as the Spanish-American war in the final years of the century. In this case, BERTopic also exhibits a dispersed distribution that spans the entire 19th century, although its top words don't mix different topics.

In Fig. 5, a set of topics associated with the Cold War has been represented. Both WBToT models exhibit a clear peak in the second half of the 20th century (with the 1980s as the more populated decade) and vanishes quickly in the first half of the century. The LDA model also exhibits a strong presence during the second half of the century, but there is a noticeable presence during the first half of the century as well (probably due to the Second World War) and some residual presence even before the 20th century, which is clearly out of the period spanned by the Cold War. By looking at the top words of each model, we can see that LDA seems to exhibit a more general vocabulary that could be used to describe other types of wars and military tensions: in particular, it is interesting to note that both WBToT models feature "nuclear" as a top word (a concept that defined one of the key singularities of the Cold War with respect to other conflicts), while LDA lacks that word. BERTopic exhibits an excellent semantic representation of the topic, with very precise top words. However, its distribution over time exhibits a relative spike in the decades of the 2010s and the 2020s, long after the end of the conflict.

6.2 Coherence

The main motivation to introduce the WBToT model as an extension of the BToT model was to balance the excessive importance that BToT places on timestamps. As already stated in Section 4.2, in BToT the same timestamp, t_d, is generated for each word of document d. WBToT solves this problem by introducing an additional latent variable, y, that determines



Health care reform

WBToT($\mathbf{n_d^y = 0.1N_d}$) **top words**: health, create, pass, time, good, economy, reform, job, country, help, american, pay, america, plan, care, worker, business, work, people, family.

WBToT($\mathbf{n_d^y = \sqrt{N_d}}$) **top words**: reform, college, people, america, worker, good, health, tax, family, school, american, pay, job, child, education, business, help, work, high, care.

LDA top words: worker, plan, pass, deficit, congress, good, reform, business, family, america, budget, time, economy, cut, help, tax, people, american, job, work.

BERTopic top words: health, care, insurance, social, medical, coverage, security, medicare, family, cost, plan, drug, child, help, doctor, system, senior, reform, benefit, work.

Fig. 3: Presence over time and 20 top words for the "Health care reform" topic.

the number of times that the same timestamp is observed per document. This feature allows WBToT to generate topics that are more semantically coherent, instead of focusing almost exclusively on modeling publication timestamps as BToT does.

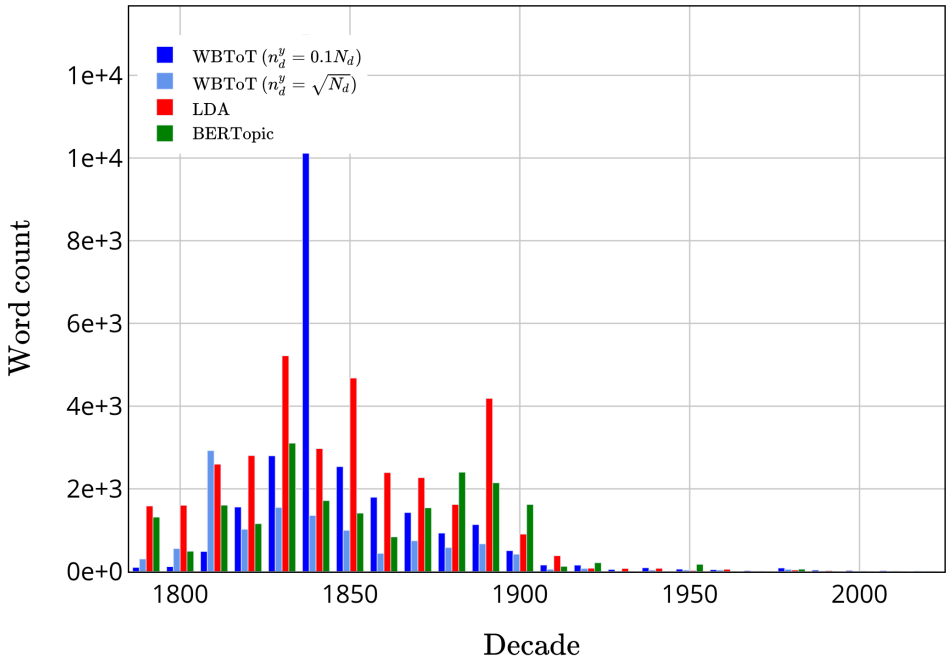
Topic coherence measures are typically used to assess topic's semantic quality and interpretability. We employ the framework described in Röder, Both and Hinneburg (2015) and adopt the coherence definition that exhibits the best overall performance in their experimental setting, C_V. This is a composition of a one-set segmentation of top words, a boolean sliding window, and an indirect confirmation measure that combines normalized pointwise mutual information and cosine similarity. In order to compute coherence, a list of words that identify each topic must be provided. The standard option is to rank words according to β_{kw} for each topic k and select a certain number of top words (we will select the 10 top words). Other scoring functions such as $r_{kw}^{\log}$,

$$r_{kw}^{\log} := \beta_{kw} \left(\log \beta_{kw} - \langle \log \beta_{kw} \rangle_k \right) \tag{38}$$

can be used as well. $r_{kw}^{\log}$ is a very common scoring function whose merit lies in demoting very frequent words, so the "essence" of the topic is better appreciated. In this section, we will compute coherence by scoring topic words both with $\beta_{kw}$and $r_{kw}^{\log}$.

For this comparison, we compute the coherence of each of the 50 topics found in the SOTU dataset (see Section 5.1) for LDA, BERTopic, BToT, and the two WBToT models. Since mean values might be affected by outliers, it is not a good idea to simply measure the average coherence for each collection of topics. Instead, it is more interesting to measure the number of tokens assigned to topics with less (or equal) coherence than a certain value. To achieve this, we will sort topics from less to more coherent (for each topic model). Then we compute the number of tokens assigned to each of these topics (that is, $\sum_d \theta_{dk} n_{dw}$) and we compute the cumulative sum of tokens for the coherence-sorted list of topics. Results are presented in Fig. 6, where coherence has been computed using β_{kw} for Fig. 6a and $r_{kw}^{\log}$ for Fig. 6b.

It is clear that the BToT model exhibits coherence values that consistently rank below those of the two WBToT models, especially in the midsection of the graphs. In fact, we discovered that 10 out of 50 topics in BToT were assigned a flat word probability (i.e. $\beta_{kw} \backsimeq \beta_{k0}$), which implies that they do not convey any semantic meaning at all. This result supports the introduction of an additional latent variable in WBToT, and therefore can be seen as an ablation study that proves the importance of balancing the different modalities in the log-likelihood.



American Indians and conflicts with Mexico

WBToT($\mathbf{n_d^y = 0.1N_d}$) **top words**: condition, time, tribe, interest, texas, foreign, policy, people, power, peace, indian, great, war, nation, territory, state, united, country, government, mexico.

WBToT($\mathbf{n_d^y = \sqrt{N_d}}$) **top words**: great, white, reservation, portion, treaty, peace, civilization, savage, settlement, policy, government, territory, country, united, limit, state, land, indian, frontier, tribe.

LDA top words: claim, nation, peace, spanish, hope, measure, treaty, country, power, united, spain, relation, time, receive, citizen, minister, government, war, state, subject.

BERTopic top words: indian, tribe, land, territory, reservation, savage, general, report, secretary, frontier, civilization, government, congress, settlement, alaska, mississippi, white, allotment, state, united.

Fig. 4: Presence over time and 20 top words for the "American Indians and conflicts with Mexico" topic.

Both versions of the WBToT model (depending on the $n_d^{(y)}$ parameter) perform similarly in terms of coherence, and it is not clear if one of them is better than the other.

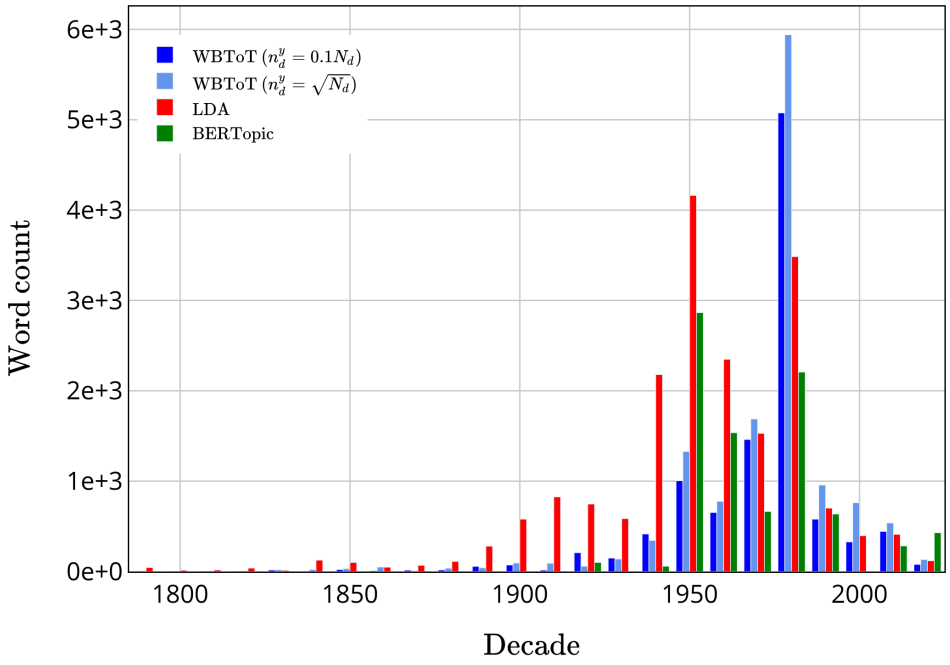
We also observe that the introduction of a time modality in WBToT does not degrade much the coherence with respect to LDA. In both figures, LDA coherence is similar to that of these two models, which implies that they preserve the quality of semantic meaning.

BERTopic coherence ranks above all the studied models when top words are ordered according to β_{kw}, but this is not so clear when we use $r_{kw}^{\log}$ instead. Furthermore, it must be noticed that BERTopic automatically removes documents that are classified as outliers, which naturally improves the coherence of topics. In WBToT, we could also remove ill-represented documents as a final step, but since coherence is not our priority, we leave this for future research.

6.3 Stability of WBToT

Standard ToT already captured events better than LDA (Wang and McCallum, 2006). Nevertheless, the main practical advantage of using a fully Bayesian model is the fact that instabilities are prevented when a topic is poorly represented in a mini-batch or exhibits a very peaked timestamp structure (recall Eqs. (16) - (17) and the subsequent discussion, which mathematically proves the instability of ToT online updates). Therefore, with WBToT one can make use of a stable online optimization approach which is not present in standard ToT.

To illustrate the stability of the online algorithm in WBToT, we will consider the experiment on the large-scale COVID-19 Twitter dataset described in Section 5.2. To begin with, we prove that perplexity on a held-out dataset decreases as successive mini-batches are analyzed. To do this, we evaluate perplexity on the test set (which is composed of 1 million random tweets) each time the model optimizes its parameters after receiving a new mini-batch. As we mentioned before, we have considered 9 mini-batches of 1 million tweets each. Results are presented in Fig. 7, where we have represented perplexity on the test set as a function of the number of mini-batches analyzed. Notice that only one iteration of the EM algorithm has been performed, but the perplexity seems to approach convergence quite fast.



Cold War

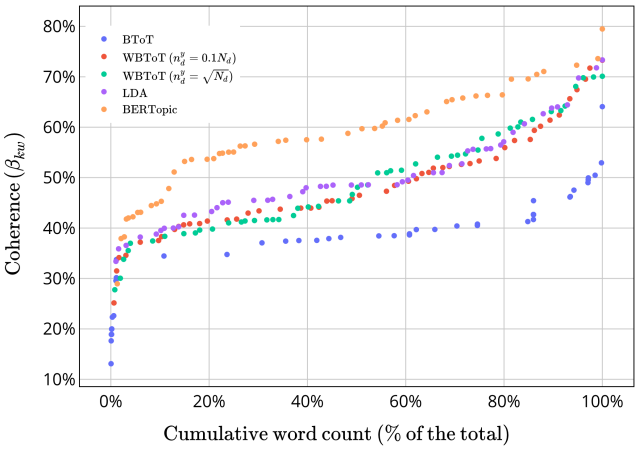
WBToT($\mathbf{n_d^y = 0.1N_d}$) top words: strong, arm, union, united, force, strategic, agreement, continue, nation, security, nuclear, policy, soviet, military, trade, america, economic, international, peace, defense.

WBToT($\mathbf{n_d^y = \sqrt{N_d}}$) top words: rights, east, war, union, nuclear, strong, soviet, america, force, future, continue, people, nation, peace, security, military, human, state, free, freedom.

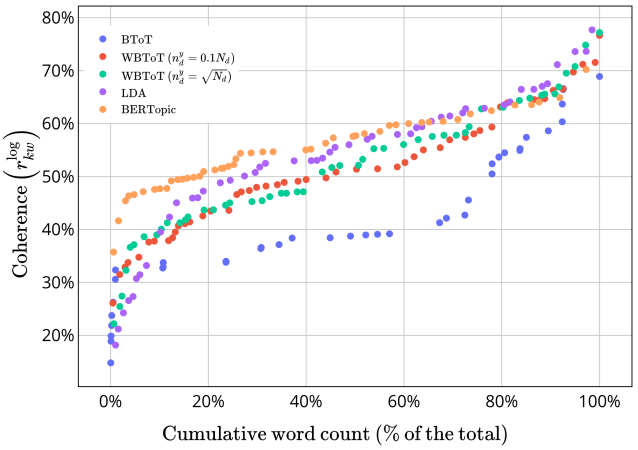
LDA top words: defense, effort, aggression, europe, continue, policy, help, strength, agreement, force, economic, international, united, free, soviet, peace, nation, military, war, security.

BERTopic top words: soviet, nuclear, communist, free, union, defense, military, force, threat, arm, ally, europe, korea, weapon, aggression, strategic, nation, agreement, security, peace.

Fig. 5: Presence over time and 20 top words for the "Cold War " topic.



(a) Using the β_{kw} rank



(b) Using the $r_{kw}^{\log}$ ranking function

Fig. 6: Model coherence as a function of cumulative word count

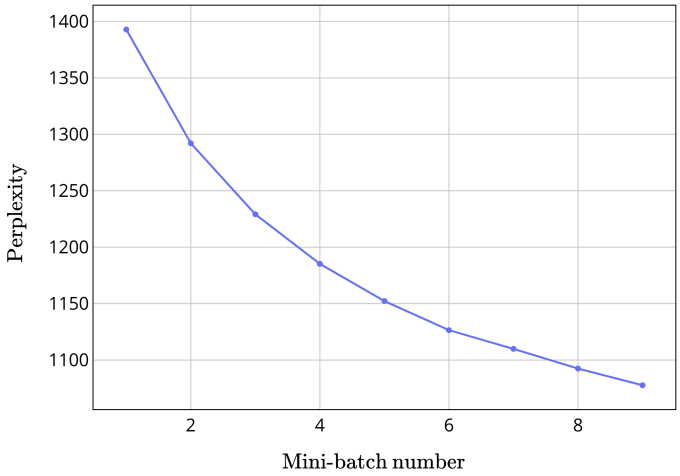
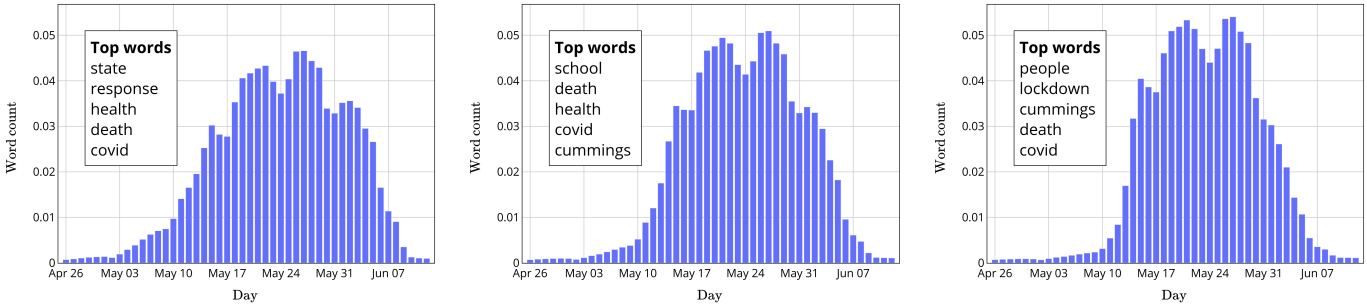


Fig. 7: Perplexity evolution with the number of analyzed mini-batches.



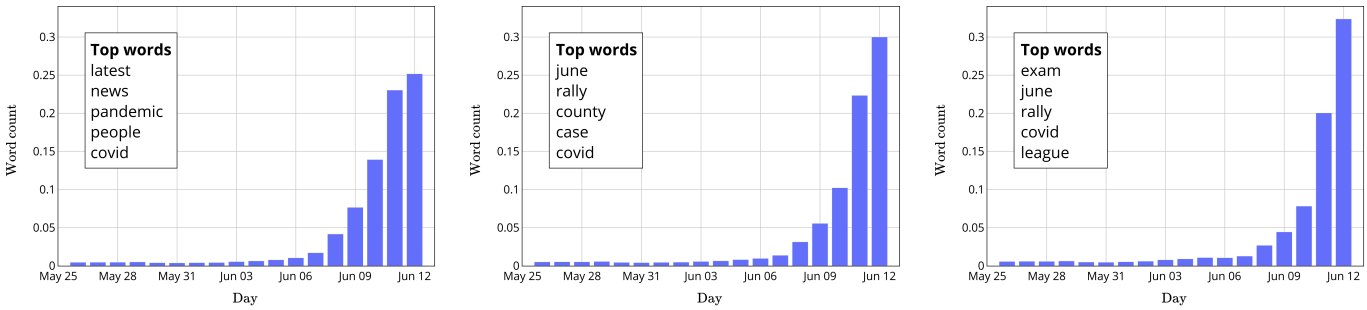
(a) After mini-batch 1 (b) After mini-batch 5 (c) After last mini-batch

Fig. 8: Evolution of topic regarding COVID-19 lockdowns.

Next, as an example, we present the evolution of a couple of topics discovered by the model. In Fig. 8 - 9 we plot the normalized presence of each topic over time and the top five words that describe them at three different stages of training: after analyzing the first mini-batch (left subfigures), after four mini-batches (central subfigures) and after all the mini-batches (right subfigures). Since the vocabulary is very similar for all the topics, we have ranked the top words according to $r_{kw}^{\log}$, as defined in Eq. (38), instead of β_{kw} to increase the readability of the topics.

In Fig. 8 we observe a topic that mainly deals with lockdowns and deaths associated to COVID-19. It starts as a relatively flat topic that comprises the entire month of May dealing with generic governments' responses to COVID-19. However, as training progresses, it becomes more and more peaked, and the vocabulary gravitates towards specific events that occurred at that time.

In Fig. 9 we observe a topic with a very sharp peak at the end of the period considered here (middle of June 2020). The top words after the first mini-batch are quite generic, but they become clearer as training progresses: the topic focuses on



(a) After mini-batch 1 (b) After mini-batch 5 (c) After last mini-batch

Fig. 9: Evolution of topic regarding June exams and Tulsa rally.

events that defined the month of June, such as the management of exams and political rallies for the first time since the start of the pandemic.

We tried to implement an online algorithm for standard ToT in an attempt to reproduce these results, but instabilities led to numerical errors. This result confirms the expected behavior that motivated our search for fully Bayesian models.

7 Conclusions

In this work, we reformulated the ToT model to make it fully Bayesian (BToT) by introducing a prior for the Beta distribution. In order to deal with the different scales of the two modes of the log-likelihood (words and timestamps), a modified version of this model, WBToT, was proposed. In WBToT, the number of timestamp observations is reduced by the introduction of an additional latent topic assignment variable.

Our experiments prove that WBToT captures events better than LDA while maintaining a similar level of coherence. Our experiments also illustrated that we can make use of an online optimization algorithm in WBToT, a feature that is not present in standard ToT due to stability issues (originating from the lack of a prior for the Beta distribution).

WBToT also exhibits better event modelization than embedding-based methods such as BERTopic. However, the superior coherence shown by BERTopic opens up an interesting line of research: can BERTopic be extended to take timestamps into account and better capture events? We will address this question in future works.

Notice that we have worked with news and tweets separately. In future works, we would like to investigate how our model performs on a collection of multiple and heterogeneous sources (news, tweets, RSS feeds...) to test if it's capable of recognizing events across different streams of data (Mele, Bahrainian and Crestani, 2019).

The WBToT model solves the main shortcoming of the popular ToT model. The introduction of a robust online optimization method allows using this model for practical applications that were problematic for standard ToT (see, for example, Al-Ghossein et al. 2018). Furthermore, we believe that our approach could also be integrated into any of the extensions of ToT described in Section 2, such as the AToT (Xu et al., 2014) and the HTMOT (Poumay and Ittoo, 2021) models.

Acknowledgment

This work was supported by the Spanish Centre for the Development of Industrial Technology (CDTI) under grant number IDI-20190525. The work of Julio Gonzalo was partially supported by the Spanish Ministry of Science and Innovation through Project FairTransNLP under Grant PID2021-124361OB-C32. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

Appendix A
Description of LDA

The main references for the results in this appendix are Blei et al. (2003); Hoffman et al. (2010). The generative process for LDA can be described as:

1) Draw K multinomial parameters $\boldsymbol{\beta_k}$ from K Dirichlet distributions with parameters $\boldsymbol{\eta_k}$.
2) For each document d, draw a set of multinomial parameters $\boldsymbol{\theta_d}$ from a symmetric Dirichlet distribution with parameters $\boldsymbol{\alpha}$. Then, for each word index $i \in \{1, \dots, N_d\}$ in document d:
   a) Draw a topic z_{di} from a multinomial with parameters $\boldsymbol{\theta_d}$;
   b) Draw a word w_{di} from a multinomial with parameters $\boldsymbol{\beta_{z_{di}}}$.

The corresponding log-likelihood has the form:

$$\mathcal{L}^{\text{LDA}} = \sum_{d=1}^{D} \left\{ \log \underbrace{p\left(\boldsymbol{\theta_d}|\boldsymbol{\alpha}\right)}_{\text{Dirichlet}} + \sum_{i=1}^{N_d} \left[\log \underbrace{p\left(z_{di}|\boldsymbol{\theta_d}\right)}_{\text{Multinomial}} + \log \underbrace{p\left(w_{di}|\boldsymbol{\beta_{z_{di}}}\right)}_{\text{Multinomial}} \right] \right\} + \sum_{k=1}^{K} \log \underbrace{p\left(\boldsymbol{\beta_k}|\boldsymbol{\eta_k}\right)}_{\text{Dirichlet}}. \tag{39}$$

In the variational inference approach, the true posterior distribution is approximated by a fully factorized distribution q^{LDA}, indexed by the variational parameters ϕ_{dwk}, γ_{dk} and λ_{kw}, of the form:

$$q^{\text{LDA}} = \prod_{k=1}^{K} \underbrace{q\left(\boldsymbol{\beta_k}|\boldsymbol{\lambda_k}\right)}_{\text{Dirichlet}} \prod_{d=1}^{D} \left[\underbrace{q\left(\boldsymbol{\theta_d}|\boldsymbol{\gamma_d}\right)}_{\text{Dirichlet}} \prod_{i=1}^{N_d} \underbrace{q\left(z_{di}|\boldsymbol{\phi_{dw_{di}}}\right)}_{\text{Multinomial}} \right]. \tag{40}$$

The evidence lower bound (ELBO) is

$$\mathcal{L}^{\text{LDA-ELBO}} = \left\langle \mathcal{L}^{\text{LDA}} - \log q^{\text{LDA}} \right\rangle_{q^{\text{LDA}}} \tag{41}$$

and its maximization leads to the standard VB LDA equations for the variational parameters

$$\phi_{dwk} \propto \exp\left[\left\langle \log \beta_{kw} \right\rangle_{q^{\text{LDA}}} + \left\langle \log \theta_{dk} \right\rangle_{q^{\text{LDA}}} \right], \tag{42}$$

Algorithm 5 Batch variational LDA

Initialize variational parameters $\gamma_{dk}, \lambda_{kw}$
repeat
  E-step:
  for $d = 1$ to D **do**
    repeat
      Set ϕ_{dwk}, γ_{dk} from Eq. (42) and Eq. (43).
    until γ_{dk} converged.
  end for
  M-step:
  Set λ_{kw} from Eq. (44).
until $\mathcal{L}$ converged.

Algorithm 6 Online variational LDA

Set $t = 0$.
Initialize variational parameters $\lambda_{kw}(t)$.
repeat
  Set $t \leftarrow t + 1$.
  Set $\rho_t := (t + \tau)^{-\kappa}$.
  E-step:
  for $d = 1$ to S **do**
    Initialize variational parameters γ_{dk}.
    repeat
      Set ϕ_{dwk}, γ_{dk} from Eq. (42) and Eq. (43).
    until γ_{dk} converged.
  end for
  M-step:
  Set $\lambda_{kw}(t + 1)$ from Eq. (47).
until $\mathcal{L}$ converged.

$$\gamma_{dk} = \alpha_k + \sum_w n_{dw}\phi_{dwk}, \tag{43}$$

$$\lambda_{kw} = \eta_{kw} + \sum_d n_{dw}\phi_{dwk}. \tag{44}$$

The Dirichlet averages appearing in Eq. (42) are:

$$\langle \log \theta_{dk} \rangle_{q^{\text{LDA}}} = \Psi(\gamma_{dk}) - \Psi\left(\sum_{k'} \gamma_{dk'}\right), \tag{45}$$

$$\langle \log \beta_{kw} \rangle_{q^{\text{LDA}}} = \Psi(\lambda_{kw}) - \Psi\left(\sum_{w'} \lambda_{kw'}\right), \tag{46}$$

where $\Psi(z) := d \log \Gamma(z)/dz$ is the digamma function.

In case of an online optimization, Eq. (44) is modified to

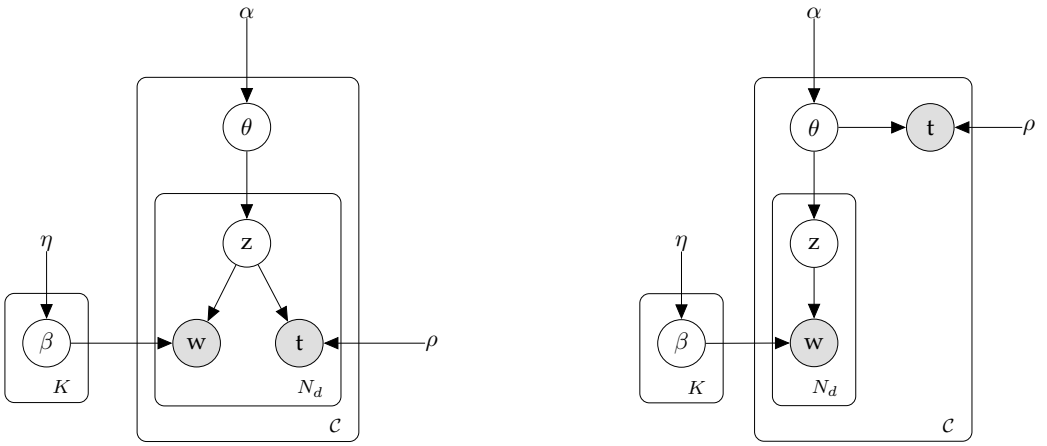
$$\lambda_{kw}(t + 1) = (1 - \rho_t)\lambda_{kw}(t) + \rho_t \left(\eta_{kw} + \frac{D}{S} \sum_{d \in \text{MB}} n_{dw}\phi_{dwk}\right). \tag{47}$$

where MB is a mini-batch of S documents. Both batch and online methods are summarized in Algorithm 5 - 6.

APPENDIX B
DESCRIPTION OF TOT

The main references for the results in this appendix are Wang and McCallum (2006); Fang et al. (2017). The generative process for ToT can be described as:

1) Draw K multinomial parameters $\boldsymbol{\beta_k}$ from K Dirichlet distributions with parameters $\boldsymbol{\eta_k}$.
2) For each document d, draw a set of multinomial parameters $\boldsymbol{\theta_d}$ from a symmetric Dirichlet distribution with parameters $\boldsymbol{\alpha}$. Then, for each word index $i \in \{1, \ldots, N_d\}$ in document d:



(a) Original view (b) Alternate view

Fig. 10: Directed acyclic graphs for Topics over Time

a) Draw a topic z_{di} from a multinomial with parameters $\boldsymbol{\theta_d}$;
b) Draw a word w_{di} from a multinomial with parameters $\boldsymbol{\beta_{z_{di}}}$;
c) Draw a timestamp t_{di} from a Beta distribution with parameters $\boldsymbol{\rho}_{z_{di}}$.

See Fig. 10a for the directed acyclic graph representation of this generative process. The corresponding log-likelihood has the form:

$$\mathcal{L}^{\text{ToT}} = \mathcal{L}^{\text{LDA}} + \sum_{d=1}^{D}\sum_{i=1}^{N_d} \log \underbrace{p\left(t_{di}|\boldsymbol{\rho}_{z_{di}}\right)}_{\text{Beta}}. \tag{48}$$

where $\mathcal{L}^{\text{LDA}}$ is given by Eq. (39).

Gibbs sampling was originally proposed as an approximate inference method in Wang and McCallum (2006). However, we will follow the variational inference method as described in Fang et al. (2017) since it more closely resembles our approach. The proposed evidence lower bound (ELBO) is

$$L^{\text{ToT-ELBO}} = \left\langle \mathcal{L}^{\text{ToT}} - \log q^{\text{LDA}} \right\rangle_{q^{\text{LDA}}}, \tag{49}$$

where q^{LDA} is given by Eq. (40). Notice that no variational distribution has been introduced to approximate the timestamp distribution, so the $\boldsymbol{\rho}_k$ parameters will be determined exactly (i.e. the variational ansatz q is the same as in LDA, q^{LDA}). The maximization of this ELBO leads to the following equations

$$\phi_{dwk} \propto \exp\left[\left\langle \log\theta_{dk}\right\rangle_{q^{\text{LDA}}} + \left\langle \log\beta_{kw}\right\rangle_{q^{\text{LDA}}} + \left(\rho_k^1 - 1\right)\log t_d + \left(\rho_k^2 - 1\right)\log\left[1 - t_d\right] - \log B\left(\boldsymbol{\rho}_k\right)\right], \tag{50}$$

$$\frac{\sum_{dw} n_{dw}\phi_{dwk}\log t_d}{\sum_{dw} n_{dw}\phi_{dwk}} = \Psi\left(\rho_k^1\right) - \Psi\left(\rho_k^1 + \rho_k^2\right), \tag{51}$$

$$\frac{\sum_{dw} n_{dw}\phi_{dwk}\log\left(1 - t_d\right)}{\sum_{dw} n_{dw}\phi_{dwk}} = \Psi\left(\rho_k^2\right) - \Psi\left(\rho_k^1 + \rho_k^2\right). \tag{52}$$

where $\Psi(x) := d\log\Gamma(x)/dx$ is the digamma function. The Dirichlet averages appearing in Eq. (50) are given by Eqs. (45) - (46), and the update equations for λ_{kw} and γ_{dk} are exactly the same as in standard LDA (see Eqs. (43) - (44)). The update equations for $\boldsymbol{\rho}_k$ cannot be solved directly due to the appearance of the digamma function in Eqs. (51) - (52), so numerical algorithms must be used.

The general optimization algorithm is summarized in Algorithm 7. Note that no online optimization version was proposed in Wang and McCallum (2006); Fang et al. (2017). This was expected, since Eqs. (51) - (52) would yield unstable mini-batch updates.

APPENDIX C
INTEGRABILITY ANALYSIS OF THE BETA-PRIOR DISTRIBUTION

Consider the normalization function $f(\nu, \boldsymbol{\chi})$ in the definition of the Beta-prior distribution,

$$f\left(\nu, \boldsymbol{\chi}\right)^{-1} := \int_0^\infty \int_0^\infty d^2\boldsymbol{\rho} \exp\left[\nu\left(\boldsymbol{\rho}\cdot\boldsymbol{\chi} - \log B\left(\boldsymbol{\rho}\right)\right)\right]. \tag{53}$$

Algorithm 7 Batch variational ToT

Initialize variational parameters γ_{dk}, λ_{kw} and $\boldsymbol{\rho}_k$.
repeat
  E-step:
  for $d = 1$ to D **do**
    repeat
      Set ϕ_{dwk}, γ_{dk} from Eq. (50) and Eq. (43).
    until γ_{dk} converged.
  end for
until $\mathcal{L}$ converged.
M-step:
Set λ_{kw}, $\boldsymbol{\rho}_k$ from Eqs. (44), (51) and (52).

Its expression in polar coordinates is

$$I_{g()}\left(\nu,\boldsymbol{\chi}\right)=\int_0^{\pi/2}d\theta\int_0^{\infty}g\left(\boldsymbol{\rho}\right)\exp\left[\nu\left(\rho\left(\chi^1\cos\theta+\chi^2\sin\theta\right)-\log B\left(\rho\cos\theta,\rho\sin\theta\right)\right)\right]\rho\,d\rho. \tag{54}$$

We will study the large ρ behavior of this integral. In order for Eq. (54) to be finite, the exponential term must satisfy:

$$\nu\rho\underbrace{\left[\cos\theta\left(\chi^1-\log\frac{\cos\theta}{\cos\theta+\sin\theta}\right)+\sin\theta\left(\chi^2-\log\frac{\sin\theta}{\cos\theta+\sin\theta}\right)\right]}_{F(\theta,\boldsymbol{\chi})}<0, \tag{55}$$

where we have made use of the Stirling approximation to write the leading part of the logarithm term in Eq. (54) as

$$\log B\left(\rho\cos\theta,\rho\sin\theta\right)\simeq\rho\left(\cos\theta\log\frac{\cos\theta}{\cos\theta+\sin\theta}+\sin\theta\log\frac{\sin\theta}{\cos\theta+\sin\theta}\right). \tag{56}$$

Let us denote θ' the angle for which $F\left(\theta,\boldsymbol{\chi}\right)$ takes its maximum value, i.e.

$$\frac{\partial F\left(\theta',\boldsymbol{\chi}\right)}{\partial\theta}=-\chi^1\sin\theta'+\chi^2\cos\theta'-\left(\cos\theta'\log\frac{\cos\theta'}{\cos\theta'+\sin\theta'}-\sin\theta'\log\frac{\sin\theta'}{\cos\theta'+\sin\theta'}\right)=0. \tag{57}$$

From Eq. (57) and the definition of $F\left(\theta,\boldsymbol{\chi}\right)$ in Eq. (55), it follows that

$$e^{\chi^1}+e^{\chi^2}=e^{\cos\theta' F(\theta')}\frac{\cos\theta'}{\cos\theta'+\sin\theta'}+\frac{\sin\theta'}{\cos\theta'+\sin\theta'}. \tag{58}$$

Therefore,

$$e^{\chi^1}+e^{\chi^2}<1\iff F\left(\theta',\boldsymbol{\chi}\right)<0. \tag{59}$$

APPENDIX D
ASYMPTOTIC EXPANSION OF BETA-PRIOR INTEGRALS

The normalization function $f(\nu,\boldsymbol{\chi})$ in the definition of the Beta-prior distribution, Eq. (53), cannot be calculated analytically. However, asymptotic methods can be applied to obtain an approximation. To do so, consider the large $\nu\gg1$ expansion of the various moments of the Beta-prior distribution,

$$I_{g()}\left(\nu,\boldsymbol{\chi}\right):=\int_0^{\infty}\int_0^{\infty}d^2\boldsymbol{\rho}\,g\left(\boldsymbol{\rho}\right)\exp\left[\nu\left(\boldsymbol{\rho}\cdot\boldsymbol{\chi}-\log B\left(\boldsymbol{\rho}\right)\right)\right], \tag{60}$$

where $g\left(\boldsymbol{\rho}\right)$ is an arbitrary function of $\boldsymbol{\rho}$. By the Laplace method (Gelman, Carlin, Stern, Dunson, Vehtari and Rubin, 2013), we obtain that

$$I_{g()}\left(\nu,\boldsymbol{\chi}\right)\simeq e^{\nu\omega(\boldsymbol{\rho}_0)}\frac{2\pi}{\nu\sqrt{\left|\det\left(\omega_{ij}\right)\right|}}\left(g\left(\boldsymbol{\rho}_0\right)+O\left(\nu^{-1}\right)\right), \tag{61}$$

where we have defined

$$\omega\left(\boldsymbol{\rho}\right):=\boldsymbol{\rho}\cdot\boldsymbol{\chi}-\log B\left(\boldsymbol{\rho}\right), \tag{62}$$

$$\boldsymbol{\rho}_0:=\mathrm{argmax}_{\boldsymbol{\rho}}\left\{\omega\left(\boldsymbol{\rho}\right)\right\}, \tag{63}$$

$$\omega_{ij}:=\frac{\partial^2\omega\left(\boldsymbol{\rho}_0\right)}{\partial\rho^i\partial\rho^j}. \tag{64}$$

Notice that the asymptotic expansion of the normalization integral $f(\nu,\boldsymbol{\chi})^{-1}$ can be obtained from Eq. (61) by setting $g(\boldsymbol{\rho})=1$. From this, the moments under the Beta-prior in Eq. (12) can now be estimated asymptotically as:

$$\begin{aligned}\langle\rho_i\rangle &= \frac{1}{\nu}\frac{\partial}{\partial\chi_i}\log f(\nu,\boldsymbol{\chi})^{-1}\\ &\simeq \rho_{0i}+O\left(\nu^{-1}\right)\end{aligned} \tag{65}$$

$$\begin{aligned}\langle\log B(\boldsymbol{\rho})\rangle &= -\frac{\partial}{\partial\nu}\log f(\nu,\boldsymbol{\chi})^{-1}+\boldsymbol{\chi}\cdot\langle\boldsymbol{\rho}\rangle\\ &\simeq -\omega(\boldsymbol{\rho}_0)+O\left(\nu^{-1}\right)+\boldsymbol{\chi}\cdot\langle\boldsymbol{\rho}\rangle\\ &= \log B(\boldsymbol{\rho}_0)+O\left(\nu^{-1}\right)\end{aligned} \tag{66}$$

APPENDIX E
BALANCING HYPER-PARAMETER IN TOT AND BTOT

Both the BToT model and the original ToT need some adjustments to balance the relative influence of words and timestamps in the log-likelihood, as described in Wang and McCallum (2006); Fang et al. (2017). Notice that there is a single timestamp per document, but in Fig. 1a and 10a its probability is repeated as many times as words are in the document. To deal with this issue, both references propose to change the log-likelihood in Eq. (4) by adding a "balancing hyperparameter" $0<\delta<1$. This new quasi-likelihood[5] now reads

$$\begin{aligned}\mathcal{L}^{\delta}{}_{ts} = \delta&\left[\sum_{d=1}^{D}\sum_{i=1}^{N_d}\left\{\left(\rho^1_{z_{di}}-1\right)\log t_{di}+\left(\rho^2_{z_{di}}-1\right)\log\left[1-t_{di}\right]-\log B\left(\boldsymbol{\rho}_{z_{di}}\right)\right\}\right]\\ &+\nu\sum_{k=1}^{K}\left\{\boldsymbol{\rho}_k\cdot\boldsymbol{\chi}-\log B(\boldsymbol{\rho}_k)\right\}+K\log f(\nu,\boldsymbol{\chi})\end{aligned} \tag{67}$$

Normalization is lost, but for the purpose of making inferences with the posterior, one can use this quasi-likelihood as an ordinary one because only quotients are involved (normalization can be recovered in a Gibbs sampling approach Wang and McCallum 2006). The same variational treatment described in Section 4.1.2 can be used, and results in the following modifications:

- The variational Beta-prior updates, Eq. (9), are changed to

$$\mu_k = \nu+\delta\,N_k \tag{68}$$

$$\boldsymbol{\psi}_k = \mu_k^{-1}\left(\nu\boldsymbol{\chi}+\delta\,N_k\boldsymbol{l}_k\right). \tag{69}$$

- The multinomial updates, Eq. (12), are changed to

$$\begin{aligned}\phi_{dwk}\propto\exp\Big[&\langle\log\beta_{kw}\rangle_{q^{\text{LDA}}}+\langle\log\theta_{dk}\rangle_{q^{\text{LDA}}}\\ &+\delta\left\{\left(\left\langle\rho^1_k\right\rangle_{q_k(\boldsymbol{\rho}_k)}-1\right)\log t_d+\left(\left\langle\rho^2_k\right\rangle_{q_k(\boldsymbol{\rho}_k)}-1\right)\log\left[1-t_d\right]-\left\langle\log B(\boldsymbol{\rho}_k)\right\rangle_{q_k(\boldsymbol{\rho}_k)}\right\}\Big]\end{aligned} \tag{70}$$

- The online updates, Eqs. (22) - (23), now read

$$\mu'_k(t+1) = (1-\rho_t)\,\mu'_k(t)+\rho_t\left(\nu+\frac{D}{S}\sum_{d\in MB}\delta\,N_{dk}\right) \tag{71}$$

$$\boldsymbol{\psi}'_k(t+1) = (1-\rho_t)\,\boldsymbol{\psi}'_k(t)+\rho_t\left(\nu\boldsymbol{\chi}+\frac{D}{S}\sum_{d\in MB}\delta\,N_{dk}\boldsymbol{l}\boldsymbol{t}_d\right). \tag{72}$$

Notice that, in most of these modifications, one simply changes $N_{dk}\to\delta\,N_{dk}$.

It is not difficult to show that one can absorb the hyperparameter δ into a re-scaling of the Beta-prior parameters $\boldsymbol{\chi}\to\boldsymbol{\chi}/\delta$ (if one is not too picky with respect to the domain of integration of $d^2\boldsymbol{\rho}$), so in order to impose effectively this hyperparameter it is necessary to strongly adjust the priors. Therefore, this model would not be appropriate if one wants to stick to a Bayesian approach.

In Wang and McCallum (2006) it is suggested that a natural value for δ would be the inverse of the average number of words per document in the collection, where it is also discussed a connection to a similar model in which timestamps are generated just once per document (see Fig. 10b). However, this model severely underestimates the influence of the time dimension.

5. The term quasi-likelihood is used with a different meaning in other contexts.

REFERENCES

Al-Ghossein, M., Murena, P.A., Abdessalem, T., Barré, A., Cornuéjols, A., 2018. Adaptive Collaborative Topic Modeling for Online Recommendation, in: Proceedings of the 12th ACM Conference on Recommender Systems, Association for Computing Machinery, Vancouver, British Columbia, Canada. pp. 338–346. doi:10.1145/3240323.3240363.

Asghari, M., Sierra-Sosa, D., Elmaghraby, A.S., 2020. A topic modeling framework for spatio-temporal information management. Information Processing & Management 57, 102340. URL: https://doi.org/10.1016/j.ipm.2020.102340, doi:10.1016/j.ipm.2020.102340.

Banda, J.M., Tekumalla, R., Wang, G., Yu, J., Liu, T., Ding, Y., Artemova, E., Tutubalina, E., Chowell, G., 2021. A Large-Scale COVID-19 Twitter Chatter Dataset for Open Scientific Research - An International Collaboration. Epidemiologia 2, 315–324. doi:10.3390/epidemiologia2030024.

Bird, S., Klein, E., Loper, E., 2009. Natural language processing with Python: analyzing text with the natural language toolkit. O'Reilly Media, Inc.

Bishop, C.M., 2006. Pattern Recognition and Machine Learning (Information Science and Statistics). 1 ed., Springer, New York, USA.

Blei, D.M., Lafferty, J.D., 2006. Dynamic Topic Models, in: Proceedings of the 23rd International Conference on Machine Learning, Association for Computing Machinery, Pittsburgh, Pennsylvania, USA. pp. 113–120. doi:10.1145/1143844.1143859.

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet Allocation. Journal of Machine Learning Research 3, 993–1022. doi:10.5555/944919.944937.

Bunk, S., Krestel, R., 2018. WELDA: Enhancing Topic Models by Incorporating Local Word Context, in: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries, pp. 293–302. doi:10.1145/3197026.3197043.

Burel, G., Farrell, T., Alani, H., 2021. Demographics and topics impact on the co-spread of COVID-19 misinformation and fact-checks on Twitter. Information Processing & Management 58, 102732. URL: https://doi.org/10.1016/j.ipm.2021.102732, doi:10.1016/j.ipm.2021.102732.

Churchill, R., Singh, L., 2021. The Evolution of Topic Modeling. ACM Computing Surveys doi:10.1145/3507900.

Das, A., Mueller, F., Hargrove, P., Roman, E., Baden, S., 2018. Doomsday: Predicting Which Node Will Fail When on Supercomputers, in: SC18: International Conference for High Performance Computing, Networking, Storage and Analysis, IEEE. pp. 108–121. doi:10.1109/SC.2018.00012.

Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference 1, 4171–4186. arXiv:1810.04805.

Dieng, A.B., Ruiz, F.J., Blei, D.M., 2020. Topic modeling in embedding spaces. Transactions of the Association for Computational Linguistics 8, 439–453. doi:10.1162/tacl_a_00325.

Fang, A., Macdonald, C., Ounis, I., Habel, P., Yang, X., 2017. Exploring Time-Sensitive Variational Bayesian Inference LDA for Social Media Data, in: Advances in Information Retrieval. ECIR 2017. Lecture Notes in Computer Science, Springer International Publishing, Cham. pp. 252–265. doi:10.1007/978-3-319-56608-5_20.

Gelman, A., Carlin, J.B., Stern, H.S., Dunson, D.B., Vehtari, A., Rubin, D.B., 2013. Bayesian Data Analysis. 3rd ed. ed., Chapman and Hall/CRC, New York. doi:10.1201/b16018.

Grootendorst, M., 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. URL: http://arxiv.org/abs/2203.05794, arXiv:2203.05794. arXiv preprint.

Hoffman, M.D., Bach, F., Blei, D.M., 2010. Online Learning for Latent Dirichlet Allocation, in: Advances in Neural Information Processing Systems, Curran Associates, Inc.. p. 856–864. doi:10.5555/2997189.2997285.

Hoffman, M.D., Blei, D.M., Wang, C., Paisley, J., 2013. Stochastic Variational Inference. Journal of Machine Learning Research 14, 1303–1347.

Hofmann, T., 1999. Probabilistic Latent Semantic Analysis, in: Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence, Morgan Kaufmann Publishers Inc., Stockholm, Sweden. pp. 289—-296. doi:10.5555/2073796.2073829.

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., Zhao, L., 2019. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimedia Tools and Applications 78, 15169–15211. doi:10.1007/s11042-018-6894-4.

Lau, H.S., Lau, A.H.L., 1991. Effective procedures for estimating beta distribution's parameters and their confidence intervals. Journal of Statistical Computation and Simulation 38, 139–150. doi:10.1080/00949659108811325.

Li, X., Xie, Q., Jiang, J., Zhou, Y., Huang, L., 2019a. Identifying and monitoring the development trends of emerging technologies using patent analysis and Twitter data mining: The case of perovskite solar cell technology. Technological Forecasting and Social Change 146, 687–705. doi:10.1016/j.techfore.2018.06.004.

Li, X., Zhang, J., Ouyang, J., 2019b. Dirichlet multinomial mixture with variational manifold regularization: Topic modeling over short texts. Proceedings of the AAAI Conference on Artificial Intelligence 33, 7884–7891. doi:10.1609/aaai.v33i01.33017884.

McInnes, L., Healy, J., 2017. Accelerated Hierarchical Density Based Clustering. IEEE International Conference on Data

Mining Workshops, ICDMW 2017-Novem, 33–42. doi:10.1109/ICDMW.2017.12, arXiv:arXiv:1705.07321v2.

McInnes, L., Healy, J., Melville, J., 2018. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. URL: http://arxiv.org/abs/1802.03426, arXiv:1802.03426. arXiv preprint.

Mele, I., Bahrainian, S.A., Crestani, F., 2019. Event mining and timeliness analysis from heterogeneous news streams. Information Processing & Management 56, 969–993. URL: https://doi.org/10.1016/j.ipm.2019.02.003, doi:10.1016/j.ipm.2019.02.003.

Poumay, J., Ittoo, A., 2021. HTMOT : Hierarchical Topic Modelling Over Time. arXiv preprint arXiv:2112.03104.

Qi, P., Zhang, Y., Zhang, Y., Bolton, J., Manning, C.D., 2020. Stanza: A Python Natural Language Processing Toolkit for Many Human Languages, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Association for Computational Linguistics, Online. pp. 101–108. doi:10.18653/v1/2020.acl-demos.14.

Reimers, N., Gurevych, I., 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference , 3982–3992doi:10.18653/v1/d19-1410, arXiv:1908.10084.

Röder, M., Both, A., Hinneburg, A., 2015. Exploring the Space of Topic Coherence Measures, in: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, Association for Computing Machinery, Shanghai, China. pp. 399–408. doi:10.1145/2684822.2685324.

Shi, Y., Bryan, C., Bhamidipati, S., Zhao, Y., Zhang, Y., Ma, K.L., 2018. MeetingVis: Visual Narratives to Assist in Recalling Meeting Context and Content. IEEE Transactions on Visualization and Computer Graphics 24, 1918–1929. doi:10.1109/TVCG.2018.2816203.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention Is All You Need, in: 31st Conference on Neural Information Processing Systems, p. 6000–6010. doi:10.1109/2943.974352.

Viegas, F., Cunha, W., Gomes, C., Pereira, A., Rocha, L., Goncalves, M., 2020. CluHTM - Semantic Hierarchical Topic Modeling based on CluWords, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online. pp. 8138–8150. doi:10.18653/v1/2020.acl-main.724.

Wang, C., Blei, D., Heckerman, D., 2008. Continuous Time Dynamic Topic Models, in: Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence, AUAI Press, Helsinki, Finland. pp. 579–586. doi:10.5555/3023476.3023545.

Wang, X., McCallum, A., 2006. Topics over Time: A Non-Markov Continuous-Time Model of Topical Trends, in: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, Philadelphia, PA, USA. pp. 424–433. doi:10.1145/1150402.1150450.

Xu, S., Shi, Q., Qiao, X., Zhu, L., Jung, H., Lee, S., Choi, S.P., 2014. Author-Topic over Time (AToT): A Dynamic Users' Interest Model, in: Mobile, ubiquitous, and intelligent computing, Springer Berlin Heidelberg, Berlin, Heidelberg. pp. 239–245. doi:10.1007/978-3-642-40675-1_37.

Xue, J., Qiping Shen, G., Li, Y., Han, S., Chu, X., 2021. Dynamic Analysis on Public Concerns in Hong Kong-Zhuhai-Macao Bridge: Integrated Topic and Sentiment Modeling Approach. Journal of Construction Engineering and Management 147, 04021049. doi:10.1061/(ASCE)CO.1943-7862.0002066.

Xue, J., Shen, G.Q., Li, Y., Wang, J., Zafar, I., 2020. Dynamic Stakeholder-Associated Topic Modeling on Public Concerns in Megainfrastructure Projects: Case of Hong Kong-Zhuhai-Macao Bridge. Journal of Management in Engineering 36, 04020078. doi:10.1061/(asce)me.1943-5479.0000845.