Neural Dynamic Focused Topic Model

Anonymous ACL submission

Abstract

Topic models and all their variants analyse text by learning meaningful representations through word co-occurrences. As pointed out by Williamson et al. (2010), such models implicitly assume that the probability of a topic to be active and its proportion within each document are positively correlated. This correlation can be strongly detrimental in the case of documents created over time, simply because recent documents are likely better described by new and hence rare topics. In this work we leverage recent advances in neural variational inference and present an alternative neural approach to the Focused Topic Model and its dynamic extensions. Indeed, we develop a neural model for topic evolution which exploits a compound Bernoulli structure in order to track the appearances of topics, thereby decoupling their activities from their proportions. On three different corpora namely, the `UN` general debates, the collection of `NeurIPS` papers, and the `ACL` Anthology dataset, our model outperforms competing neural variational topic models.

1 Introduction

Probabilistic topic models, the likes of Latent Dirichlet Allocation (LDA) (Blei et al., 2003), are generative models of word co-occurrence that analyse large document collections by learning latent representations (topics) encoding their themes. These models represent the documents of the collection as mixtures of latent topics, and group semantically related words into single topics by means of word-pair frequency information within the collection. Such a generic generative structure has been successfully applied to problems from information retrieval, visualization and multilingual modelling to linguistic understanding in fiction and non-fiction, scientific publications and political texts (see e.g. Boyd-Graber et al. (2017) for a review) and keeps being extended to new domains (Rezaee and Ferraro, 2020; Zhao et al., 2021).

Topic models implicitly assume that the documents within a given collection are exchangeable. Yet document collections such as magazines, academic journals, news articles and social media content not only feature trends and themes that change with time, but also employ their language differently as time evolves (Danescu-Niculescu-Mizil et al., 2013). The exchangeability assumption along the time component is hence inappropriate in these cases and topic models have been extended to account for changes in both topic (Blei and Lafferty, 2006; Wang et al., 2012; Jähnichen et al., 2018) and word (Bamler and Mandt, 2017; Rudolph and Blei, 2018; Dieng et al., 2019) distributions among documents collected over long periods of time.

It is easy to imagine, however, that if one analyses the collection's content as one moves forward in time, one would find that (some of) the topics describing those documents appear, disappear or reappear with time. This simple intuition entails that one should not only model the time- and document-dependent topic proportions, but also *the probabilities for the topics to be active*, and how such probabilities change with time. Previous work has already pointed out that existing topic models implicitly assume that the probability of a topic being active and its proportion within each document are positively correlated (Williamson et al., 2010; Perrone et al., 2016). This assumption is generally unwanted, simply because rare topics may account for a large part of the words in the few documents in which they are active. It is particularly detrimental (for both modelling and prediction) in a dynamic setting, because recent documents are likely better described by new and hence rare topics.

Indeed, whenever the topic distribution over documents is strongly skewed, topic models tend to learn the more general topics held by the big majority of documents in the collection, rather than the rare topics contained only by fewer documents

(Jagarlamudi et al., 2012; Tang et al., 2014; Zuo et al., 2014). Document collections that reflect evolving content typically feature skew topic distribution over its documents, with the newly added documents being well described by new, rare topics. Dynamic topic models that feature the topic proportion-activity coupling are then expected to perform badly, simply because these will not be able to infer the new topics characteristic of recent documents. To properly model such recent documents one should therefore allow rarely seen topics to be active with high proportion and frequently seem topics to be active with low proportion.

In this work we seek to decouple the probability for a topic to be active from its proportion with the introduction of a sparse Bernoulli variable, which selects the active topics for a given document at a particular instant of time. Earlier models attained such a decoupling via non-parametric priors, such as the Indian Buffet Process prior over infinite binary matrices, in both static (Williamson et al., 2010) and dynamic (Perrone et al., 2016) cases. Our construction follows the same logic and also deploys the Indian Buffet prior, but leverages the reparametrization trick to perform neural variational inference (Kingma and Welling, 2013). The result is a scalable model whose non-parametric nature allows the instantaneous number of active topics per document to fluctuate and infers the total number of topics in the collection directly from the data.

We introduce the Neural Dynamic Focused Topic Model (NDF−TM) which builds on top of Neural Variational Topic models (Miao et al., 2016), uses Deep Kalman Filters (Krishnan et al., 2015) to model the topic dynamics, and the stick-breaking Variation Autoencoder (Nalisnick and Smyth, 2016) to infer the Bernoulli variable selecting the active topics. We show below that NDF−TM explicitly decouples the topic proportion from its activity and outperforms competing neural models on different metrics.

2 Related Work

The NDF−TM model merges concepts from dynamic topic models, dynamic embeddings and neural topic models.

Dynamic topic models. The seminal work of Blei and Lafferty (2006) introduced the Dynamic Topic Model (DTM), which uses a state space model on the natural parameters of the distribution representing the topics, thus allowing the latter to change with time. The DTM methodology was first extended by Caron et al. (2007) to a nonparametric setting, via the correlation of Dirichlet process mixture models in time. Later Wang et al. (2012) replaced the discrete state space model of DTM with a Diffusion process, thereby extending the approach to a continuous time setting. Jähnichen et al. (2018) further extended DTM by introducing Gaussian process priors that allowed for a non-Markovian representation of the dynamics. Other recent work on dynamic topic models is that of Hida et al. (2018)

Dynamic embeddings. Rather than modelling the content evolution of document collections like DTM, other works focus on modelling how word semantics change with time (Bamler and Mandt, 2017; Rudolph and Blei, 2018). These works use continuous representation of words capturing their semantics (as e.g. those of Pennington et al. (2014)) and evolve such representation via diffusion processes. More recently, Dieng et al. (2019) represent topics as dynamic embeddings, and model words via categorical distributions whose parameters are given by the inner product between the static word embeddings and the dynamic topic embeddings. As such, this model corresponds to the dynamic extension of Dieng et al. (2020).

Neural topic models. Another line of research leverages neural networks to improve the performance of topic models, the so-called neural topic models (Miao et al., 2016; Srivastava and Sutton, 2017; Zhang et al., 2018; Dieng et al., 2020, 2019) which deploy neural variational inference (Kingma and Welling, 2013) for training.

Decoupling topic activity from its proportion. Williamson et al. (2010) noted the implicit and undesirable correlation between topic activity and proportion assumed by standard topic models and introduced the Focused Topic Model (FTM). FTM uses the Indian Buffet Process (IBP) to decouple across-data prevalence and within-data proportion in mixed membership models. Later Perrone et al. (2016) extended FTM to a dynamic setting by using the Poisson Random Fields model from population genetics to generate dependent IBPs, which allow them to model temporal correlations in data.

Both of these models are trained using complex sampling schemes, which can make the fast and accurate inference of their model parameters difficult (Miao et al., 2017). In what follows we propose

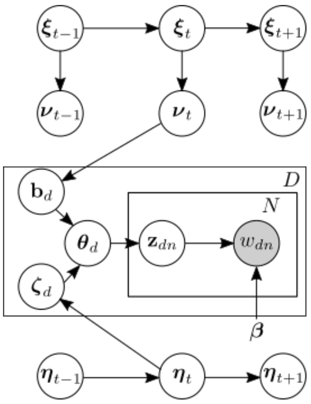

Figure 1: Graphical model representation of `NDF-TM`.

an alternative neural approach to the dynamic Focused Topic model of Perrone et al. (2016), trainable via backpropagation, which learns to decouple the dynamic topic activity from its dynamic topic proportion.

3 Neural Bernoulli-Beta Topic Model

Suppose we are given an ordered collection of corpora $\mathcal{D} = \{D_1, D_2, \dots, D_T\}$, so that the tth corpus D_t is composed of N_t documents, all received within the tth time window. Let $\mathbf{W}_t$ denote the Bag-of-word (BoW) representation for the whole document set within D_t and let $\mathbf{w}_{t,d}$ denote the BoW representation of the d-th document in D_t.

Let us now suppose that the corpora collection is described by a set of K unknown topics. We then assume there are two sequences of continuous hidden variables $\boldsymbol{\eta}_1, \dots, \boldsymbol{\eta}_T \in \mathbb{R}^{\dim(\eta)}$ and $\boldsymbol{\xi}_1, \dots, \boldsymbol{\xi}_T \in \mathbb{R}^{\dim(\xi)}$ which encode, respectively, how the topic proportions and the topic activities change among corpora as time evolves (i.e. as one moves from D_t to D_{t+1}). That is, $\boldsymbol{\eta}_t$ and $\boldsymbol{\xi}_t$ encode the *global dynamics* of semantic content. We also assume there are two *local* hidden variables, conditioned on the global ones, namely a continuous variable $\boldsymbol{\zeta}_{t,d} \in \mathbb{R}^K$ which encodes the content of the dth document in D_t, in terms of the available topics, and a binary variable $\mathbf{b}_{t,d} \in \{0,1\}^K$ which encodes which topics are active in the document in question. We combine these local variables to compute the topic proportions $\boldsymbol{\theta}_{t,d} \in [0,1]^K$ from which each document in D_t is generated.

3.1 Generation

Let us denote with ψ the set of parameters of our generative model. We are first of all interested in learning the number of active topics per document at each time step from the data directly. To do so, we model the time-dependent Bernoulli variable $\mathbf{b}_{t,d}$ using the stick-breaking prior of the Indian Buffet Process (IBP) (Teh et al., 2007). Explicitly, we generate $\mathbf{b}_{t,d}$ as follows

$$\begin{aligned}
\boldsymbol{\xi}_t &\sim \mathcal{N}\left(\boldsymbol{\mu}^{\xi}_{\psi}(\boldsymbol{\xi}_{t-1}), \boldsymbol{\delta}\,\mathbf{I}\right), && (1)\\
\mathbf{a}_t &= \alpha_0 \, \mathrm{Sigmoid}\left(\mathbf{W}_a \, \boldsymbol{\xi}_t + \mathbf{c}_a\right), && (2)\\
\boldsymbol{\nu}_t &\sim \mathrm{Beta}(\mathbf{a}_t, 1), && (3)\\
\pi_{t,k} &= \prod_{j=1}^{k} \nu_{t,k}, \quad k = \{1, \dots, K\}, && (4)\\
\mathbf{b}_{t,d} &\sim \mathrm{Bernoulli}(\boldsymbol{\pi}_t), && (5)
\end{aligned}$$

where equation 3–5 correspond to the stick-breaking construction of the IBT (Teh et al., 2007) and α_0 is a hyperparameter controlling the average number of active topics. Likewise $\mathbf{W}_a \in \mathbb{R}^{K \times \dim(\xi)}, \mathbf{c}_a \in \mathbb{R}^K \subset \psi$ are trainable parameters. Also note that, just as in Deep Kalman Filters (Krishnan et al., 2015), $\boldsymbol{\xi}_t$ is Markovian and evolves under a Gaussian noise with mean $\boldsymbol{\mu}^{\xi}_{\psi}$, defined via a neural network with parameters in ψ, and variance $\boldsymbol{\delta}$. We choose $\boldsymbol{\xi}_1 \sim \mathcal{N}(0,1)$.

We model $\mathbf{b}_{t,d}$ via the stick-breaking construction of the IBT to allow the number of active topics (of document d at time t) to be inferred directly from the data. As a consequence, the instantaneous number of topics per document is allowed to fluctuate and the total number of topics in the collection is allowed to grow with the collection's size (up to K).

Analogously, we generate the topic proportions $\boldsymbol{\theta}_{t,d}$ as

$$\begin{aligned}
\boldsymbol{\eta}_t &\sim \mathcal{N}\left(\boldsymbol{\mu}^{\eta}_{\psi}(\boldsymbol{\eta}_{t-1}), \boldsymbol{\delta}\,\mathbf{I}\right), && (6)\\
\boldsymbol{\zeta}_{t,d} &\sim \mathcal{N}\left(\mathbf{W}_\zeta \, \boldsymbol{\eta}_t + \mathbf{c}_\zeta, \, 1\right), && (7)\\
\boldsymbol{\theta}_{t,d} &= \frac{\mathbf{b}_{t,d} \odot \exp\left(\boldsymbol{\zeta}_{t,d}\right)}{\sum_k^K b^k_{t,d} \odot \exp\left(\zeta^k_{t,d}\right)}, && (8)
\end{aligned}$$

where $\mathbf{b}_{t,d}$ is defined in equation 5 and $\odot$ labels element-wise product, $\mathbf{W}_\zeta \in \mathbb{R}^{K \times \dim(\eta)}, \mathbf{c}_\zeta \in \mathbb{R}^K \subset \psi$ are trainable, and $\boldsymbol{\mu}^{\eta}_{\psi}$ is modelled via a neural network. Here $\boldsymbol{\eta}_t$ is also Markovian and we set $\boldsymbol{\eta}_1 \sim \mathcal{N}(0,1)$. Note that the topic proportion thus defined are *sparse vectors*.

Once we have $\boldsymbol{\theta}_{t,d}$ we generate the corpora sequence by sampling

$$\begin{aligned}
y_{t,d,n} &\sim \mathrm{Categorical}(\boldsymbol{\theta}_{t,d}), && (9)\\
w_{t,d,n} &\sim \mathrm{Categorical}(\boldsymbol{\beta}_{y_{t,d,n}}), && (10)
\end{aligned}$$

where $y_{t,d,n}$ is the time-dependent topic assignment for $w_{t,d,n}$, which labels the nth word in document $d \in D_t$, and $\boldsymbol{\beta} \in \mathbb{R}^{K \times V}$ is a learnable topic distribution over words. We define the latter as

$$\boldsymbol{\beta} = \mathrm{softmax}(\boldsymbol{\alpha} \otimes \boldsymbol{\rho}), \tag{11}$$

with $\boldsymbol{\alpha} \in \mathbb{R}^{K \times E}, \boldsymbol{\rho} \in \mathbb{R}^{V \times E}$ learnable topic and word embeddings, respectively, for some embedding dimension E, and $\otimes$ denoting tensor product.

The `NDF-TM` is summarized in Figure 1.

3.2 Inference

The generative model above involves two independent global hidden variables $\boldsymbol{\xi}_t, \boldsymbol{\eta}_t$, together with the intermediate (global) Beta variable $\boldsymbol{\nu}_t$ (see Eq. 3) and the local variables $\boldsymbol{\zeta}_{t,d}$ and $\mathbf{b}_{t,d}$. Our task is to infer the posterior distributions of all these variables. [1] Denoting with $\boldsymbol{\Gamma}_{t,d}$ the set $\{\boldsymbol{\xi}_t, \boldsymbol{\eta}_t, \boldsymbol{\nu}_t, \boldsymbol{\zeta}_{t,d}, \mathbf{b}_{t,d}\}$, we approximate the true posterior distribution of the model with a variational posterior of the form

$$\begin{aligned} q_\varphi(\boldsymbol{\Gamma}_{t,d}|\mathbf{w}_{t,d}, \mathbf{W}_{1:T}) &= \prod_t^T q_\varphi(\boldsymbol{\nu}_t|\mathbf{W}_t, \boldsymbol{\xi}_t) \\ &\times q_\varphi(\boldsymbol{\eta}_t|\boldsymbol{\eta}_{1:t-1}, \mathbf{W}_{1:T})\, q_\varphi(\boldsymbol{\xi}_t|\boldsymbol{\xi}_{1:t-1}, \mathbf{W}_{1:T}) \\ &\times \prod_d^{N_t} q_\varphi(\boldsymbol{\zeta}_{t,d}|\mathbf{w}_{t,d}, \boldsymbol{\eta}_t)\, q_\varphi(\mathbf{b}_{t,d}|\mathbf{w}_{t,d}, \boldsymbol{\nu}_t), \end{aligned} \tag{12}$$

where $\mathbf{W}_{1:T} = (\mathbf{W}_1, \ldots, \mathbf{W}_T)$ is the ordered sequence of BoW representations for the corpus collection and φ labels the variational parameters.

Local variables. The posterior distribution over the local variables $\boldsymbol{\zeta}_{t,d}, \mathbf{b}_{t,d}$ are chosen as Gaussian and Bernoulli, respectively, each parametrized by neural networks. Explicitly, we write

$$q_\varphi(\boldsymbol{\zeta}_{t,d}|\mathbf{w}_{t,d}, \boldsymbol{\eta}_t) = \mathcal{N}(\boldsymbol{\mu}^\zeta_\varphi, \boldsymbol{\sigma}^\zeta_\varphi), \tag{13}$$

where $\boldsymbol{\mu}^\zeta_\varphi$ and $\boldsymbol{\sigma}^\zeta_\varphi$ are both functions of $\mathbf{w}_{t,d}, \boldsymbol{\eta}_t$ and are modelled via neural networks. Likewise

$$q_\varphi(\mathbf{b}_{t,d}|\mathbf{w}_{t,d}, \boldsymbol{\xi}_t) = \mathrm{Bernoulli}(\pi_\varphi(\mathbf{w}_{t,d}, \boldsymbol{\xi}_t)), \tag{14}$$

where π_φ lives on the K-simplex and is modelled with a neural network with a `Softmax` function as output nonlinearity.

Global variables. The posterior distribution over the dynamic global variables $\boldsymbol{\xi}_t, \boldsymbol{\eta}_t$ are also Gaussian, but now depend not only on the previous latent variables at time $t-1$, but also on the entire sequence of BoW representations $\mathbf{W}_{1:T}$. This follows directly from the graphical model in Figure 1, as noted by Krishnan et al. (2015). We shall use LSTM networks (Hochreiter and Schmidhuber, 1997) to model these dependencies. Specifically let

$$q_\varphi(\boldsymbol{\xi}_t|\boldsymbol{\xi}_{t-1}, \mathbf{W}_{1:T}) = \mathcal{N}(\boldsymbol{\mu}^\xi_\varphi, \boldsymbol{\sigma}^\xi_\varphi), \tag{15}$$

where $\boldsymbol{\mu}^\xi_\varphi$, $\boldsymbol{\sigma}^\xi_\varphi$ are neural networks which take as input the pair $\boldsymbol{\xi}_{t-1}$, $\mathbf{h}^\xi_t$, with $\mathbf{h}^\xi_t$ a hidden representation encoding the sequence $\mathbf{W}_{1:T}$. Similarly

$$q_\varphi(\boldsymbol{\eta}_t|\boldsymbol{\eta}_{t-1}, \mathbf{W}_{1:T}) = \mathcal{N}(\boldsymbol{\mu}^\eta_\varphi, \boldsymbol{\sigma}^\eta_\varphi), \tag{16}$$

where $\boldsymbol{\mu}^\eta_\varphi, \boldsymbol{\sigma}^\eta_\varphi$, again neural networks, take as input the pair $\boldsymbol{\eta}_{t-1}, \mathbf{h}^\eta_t$, with $\mathbf{h}^\eta_t$ a second hidden representation also encoding $\mathbf{W}_{1:T}$.

These hidden representations $\mathbf{h}^i_t$, with $i = \{\xi, \eta\}$, correspond to the hidden states of LSTM networks whose update equation read

$$\mathbf{h}^i_t = f^i_\varphi(\mathbf{W}_t, \mathbf{h}^i_{t-1}). \tag{17}$$

Finally, since the Beta distribution does not have a non-centered parametrization, we follow Nalisnick and Smyth (2016) and choose the Kumaraswamy distribution (Kumaraswamy, 1980)

$$\mathrm{Kumar}(x; a, b) = abx^{a-1}(1-x^a)^{b-1}, \tag{18}$$

for $x \in (0,1)$ and $a, b > 0$, which has a closed-form CDF, as the posterior of $\boldsymbol{\nu}_t$. Explicitly we write

$$q_\varphi(\boldsymbol{\nu}_t|\mathbf{W}_t, \boldsymbol{\xi}_t) = \mathrm{Kumar}\left(\boldsymbol{\nu}_t; \mathbf{c}^{\boldsymbol{\nu}}_\varphi, \mathbf{d}^{\boldsymbol{\nu}}_\varphi\right), \tag{19}$$

where the functions $\mathbf{c}^{\boldsymbol{\nu}}_\varphi, \mathbf{d}^{\boldsymbol{\nu}}_\varphi$ take the pair $\mathbf{W}_t, \boldsymbol{\xi}_t$ as input and are each modelled with a neural network, with a `Softplus` function as output nonlinearity.

Note that we can sample Eq. 19 thus

$$\boldsymbol{\nu}_t = (1 - \mathbf{u}^{\frac{1}{\mathbf{d}^{\boldsymbol{\nu}}_\varphi}})^{\frac{1}{\mathbf{c}^{\boldsymbol{\nu}}_\varphi}}, \tag{20}$$

with $\mathbf{u} \sim \mathrm{Uniform}(0,1)$.

3.3 Training Objective

To optimize the model parameters $\{\psi, \varphi\}$ we minimize the variational lower bound on the logarithm of the marginal likelihood $p_\psi(w_{t,d,n}|\boldsymbol{\beta})$. Following standard methods (Bishop, 2006) the latter can readily be shown to be

[1] Note in passing that we do not need to perform inference of the latent topics $y_{t,d,n}$, simply because these can be integrated out.

	UN		NeurIPS		ACL	
Models	PPL-DC	P-NLL	PPL-DC	P-NLL	PPL-DC	P-NLL
`DTM*`	2393.5	-	-	-	4324	-
`DTM-REP`	3012 ± 14	8.334 ± 0.003	6107±907	8.5 ± 0.4	6503 ± 875	8.5 ± 0.5
`D-ETM`	**2275 ± 13**	**7.918 ± 0.002**	5404±418	9 ± 1	2733±109	7.99 ± 0.02
`NDF-TM`	2899 ± 24	8.192 ± 0.004	3768±223	**8.32 ± 0.02**	**2365±146**	**7.7 ± 0.6**
`NDF-TM-DE`	2644 ± 11	8.00 ± 0.03	**3665±312**	17 ± 5	2727±187	8.03 ± 0.08

Table 1: Perplexity on document completion (PPL-DC) and predictive negative log likelihood (P-NLL). **Lower is better**. PPL-DC is calculated by conditioning the model on the first half of the document and evaluate the perplexity on the second half of the document. P-NLL is estimated using equation 26). (*) Results are taken from (Dieng et al., 2019). All other results are obtained by training the models on 10 different random splits of the datasets.

$$
\begin{aligned}
\mathcal{L}[\boldsymbol{\beta}, \psi, \varphi] = &\sum_{t=1}^{T}\sum_{d=1}^{N_t}\sum_{n=1}^{N_d} \mathbb{E}_{\boldsymbol{\Gamma}}\Big\{ \log p_\psi(w_{t,d,n}|\boldsymbol{\beta}, \boldsymbol{\Gamma})\Big\} \\
&- \mathrm{KL}\left[q_\varphi(\boldsymbol{\eta}_1|\mathbf{W}_{1:T}); p(\boldsymbol{\eta}_1)\right] - \mathrm{KL}\left[q_\varphi(\boldsymbol{\xi}_1|\mathbf{W}_{1:T}); p(\boldsymbol{\xi}_1)\right] \\
&- \sum_{t=2}^{T} \mathrm{KL}\left[q_\varphi(\boldsymbol{\eta}_t|\boldsymbol{\eta}_{1:t-1}, \mathbf{W}_{1:T}); p_\psi(\boldsymbol{\eta}_t|\boldsymbol{\eta}_{t-1})\right] \\
&- \sum_{t=2}^{T} \mathrm{KL}\left[q_\varphi(\boldsymbol{\xi}_t|\boldsymbol{\xi}_{1:t-1}, \mathbf{W}_{1:T}); p_\psi(\boldsymbol{\xi}_t|\boldsymbol{\xi}_{t-1})\right] \\
&- \sum_{t=1}^{T} \mathbb{E}_{\boldsymbol{\xi}_t}\Big\{\mathrm{KL}\left[q_\varphi(\boldsymbol{\nu}_t|\mathbf{W}_t, \boldsymbol{\xi}_t); p_\psi(\boldsymbol{\nu}_t|\boldsymbol{\xi}_t)\right]\Big\} \\
&- \sum_{t=1}^{T}\sum_{d=1}^{N_t}\Bigg(\mathbb{E}_{\boldsymbol{\eta}_t}\Big\{\mathrm{KL}\left[q_\varphi(\boldsymbol{\zeta}_{t,d}|\mathbf{w}_{t,d}, \boldsymbol{\eta}_t); p_\psi(\boldsymbol{\zeta}_{t,d}|\boldsymbol{\eta}_t)\right]\Big\} \\
&+ \mathbb{E}_{\boldsymbol{\xi}_t}\Big\{\mathrm{KL}\left[q_\varphi(\mathbf{b}_{t,d}|\mathbf{w}_{t,d}, \boldsymbol{\xi}_t); p_\psi(\mathbf{b}_{t,d}|\boldsymbol{\xi}_t)\right]\Big\}\Bigg), \qquad (21)
\end{aligned}
$$

where KL labels the Kullback-Leibler divergence and $\boldsymbol{\beta}$ is given in equation 11. Note that to compute the KL between the Kumaraswamy posterior $q_\varphi(\boldsymbol{\nu}_t)$ and the Beta prior $p_\psi(\boldsymbol{\nu}_t)$ we approximate the infinite sum as done in Nalisnick and Smyth (2016).

4 Experiments

In this section we introduce our datasets and define our baselines. Details about preprocessing and experimental setup can be found in the Appendix. However two important parameters in our model are the maximum topic number K and the hyperparameter controlling the average number of active topics α_0. Both these hyperpameters are chosen via cross-validation, with $K = 50$ and $\alpha_0 = 10$ given the best results[2].

4.1 Datasets

We evaluate our model on three datasets, namely the collection of UN speeches, NeurIPS papers and the ACL Anthology. The UN[3] dataset (Baturo et al., 2017) contains the transcription of the speeches given at the UN General Assembly during the period between the years 1970 and 2016. It consists of about 230950 documents. The NeurIPS papers dataset contains the collection of papers published in NeurIPS[4] between the years 1987 and 2016. It consists of about of about 6562 documents. Finally, the ACL Anthology (Bird et al., 2008) contains a collection of computational linguistic and natural language processing papers published between 1973 and 2006. It consists of about 10514 documents.

4.2 Baselines

Our main aim is to study the effect of the topic proportion-activity decoupling in the performance of *dynamic topic models*[5] on data collections displaying evolving content. To do so we compare against three models:

(1) `DTM` — the Dynamic Topic Model (Blei and Lafferty, 2006), which uses Kalman Filters to model the topic dynamics.

(2) `DTM-REP` — the neural extension of DTM, fitted using neural variational inference (Dieng et al., 2019). This model uses a logistic-normal distribution, parametrized with feedforward neural networks, as posterior for the topic proportion distribution as in Miao et al. (2017). It also uses Kalman Filters to model the topic dynamics, but parametrizes the posterior distribution over the dynamic latent variables with LSTM networks, as in Deep Kalman Filters (Krishnan et al., 2015). As such, `DTM-REP` works as the dynamic extension of Miao et al. (2017). Comparing our model with `DTM-REP` should show the effect of adding the

[2] K was chosen from the set 30, 50 and 200. We found 50 to be the best value for all models, i.e. including the baselines

[3] https://www.kaggle.com/unitednations/un-general-debates

[4] https://www.kaggle.com/benhamner/nips-papers

[5] This means we do not consider static topic models

	UN		NeurIPS		ACL	
Models	TC	TD	TC	TD	TC	TD
DTM*	0.1317	0.0799	-	-	0.1429	0.5904
DTM-REP	0.108 ± 0.003	0.59 ± 0.001	-0.022±0.007	0.15±0.01	0.007 ± 0.008	0.55 ± 0.02
D-ETM	**0.201** ± **0.002**	**0.68** ±**0.006**	-0.019±0.008	0.28 ±0.05	0.137±0.004	0.61±0.02
NDF-TM	0.173 ± 0.002	0.62 ± 0.003	**0.01**±**0.02**	0.37±0.01	**0.20**±**0.02**	**0.82**±**0.01**
NDF-TM-DE	0.191 ± 0.007	0.51 ± 0.002	-0.071±0.034	**0.38**±**0.05**	0.135±0.009	0.64±0.03

Table 2: Topic coherence (TC) and Topic diversity (TD) for all models. **Higher is better**. TC is calculated by taking the average pointwise mutual information between two words drawn randomly from the same topic. TD is the percentage of unique words in the top 25 words of all topics. (*) Results taken from (Dieng et al., 2019). All other results are obtained by training the models on 10 different random splits of the datasets.

activity-coupling to neural topic models.

(3) D-ETM — the Dynamic Embedded Topic Model (Dieng et al., 2019), which captures the evolution of topics in such a way that both the content of topics and their proportions evolve over time. Thus, this model adds complexity to DTM-REP by modelling words via categorical distributions whose parameters are given by the inner product between the static word embeddings and the dynamic topic embeddings.

To better compare with D-ETM, we also allow NDF-TM to capture time-varying topic content by learning a posterior distribution over time-dependent embeddings. We label this model NDF-TM-DE. In practice this means our topic embeddings are now indexed by time $\boldsymbol{\alpha}_t \in \mathbb{R}^{K\times E}$ for $t = \{1, \ldots, T\}$.

Generation in NDF-TM-DE — To train the $\boldsymbol{\alpha}_t$ we augment the generative model, equations 1–8, with the following prior on the evolution of topic embeddings

$$\boldsymbol{\alpha}_{t,k} \sim \mathcal{N}(\boldsymbol{\alpha}_{t-1,k},\, \delta\mathbf{I}), \tag{22}$$

where $\boldsymbol{\alpha}_{1,k} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ for $k = \{1, \ldots, K\}$.

Inference in NDF-TM-DE — For *inference* we use a mean-field solution of the form

$$q_\varphi(\boldsymbol{\alpha}_{1:T,1:K}) = \prod_{t=1}^{T}\prod_{k=1}^{K}\mathcal{N}(\boldsymbol{\alpha}_{t,k},\, \delta\mathbf{I}), \tag{23}$$

where the $\boldsymbol{\alpha}_{t,k}$ are learnable. This last expression is to be multiplied to equation 12 above.

Training Objective in NDF-TM-DE — The extended model has an additional term added to its loss functions, namely the Kullback-Leibler divergence between the prior and posterior distribution of $\boldsymbol{\alpha}_t$. Explicitly we have

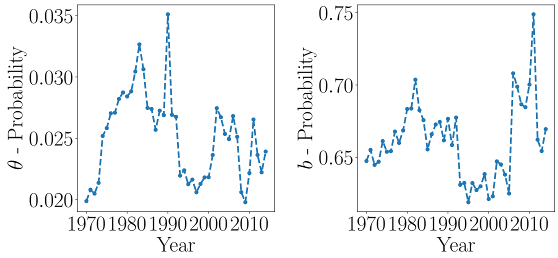

Figure 2: Evolution of topic proportion and activity probability for the topic *middle east* inferred from the UN dataset via NDF-TM-DE.

$$\mathcal{L}[\boldsymbol{\rho}, \psi, \varphi] = \mathbb{E}_{q_\varphi(\boldsymbol{\alpha}_{1:T})}\Big\{\mathcal{L}[\boldsymbol{\alpha}_{1:T}, \boldsymbol{\rho}, \psi, \varphi]\Big\} - \sum_{t=1}^{T}\sum_{k=1}^{K}\mathrm{KL}[q_\varphi(\boldsymbol{\alpha}_t); p(\boldsymbol{\alpha}_t)], \tag{24}$$

where $\boldsymbol{\rho}$ are the learnable word embeddings of above.

5 Results

In order to quantify the performance of our models, we first focus on two aspects, namely its prediction capabilities and its ability to generalize to unseen data. Later we also (qualitatively) discuss how the model actually performs the decoupling between topic activities and proportions.

(1) To test how well our models perform on a prediction task we compute the *predictive negative log likelihood* (P-NLL). Since to our knowledge the latter does not appear explicitly in the dynamic topic model literature, we briefly revisit how to estimate it in what follows.

In order to predict N steps into the future we rely on the generative process of our model, albeit conditioned on the past. Essentially, one must generate Monte Carlo samples from the posterior distribution and propagate the latent representations ($\boldsymbol{\xi}_t$

and $\boldsymbol{\eta}_t$ in our model) into the future with the help of the prior transition function (equations 1 and 6, respectively)[6]. This procedure is depicted on the conditional predictive distribution of our model

$$p(\mathbf{W}_{T+1}|\mathbf{W}_{1:T}) = \int p_\psi(\mathbf{W}_{T+1}|\boldsymbol{\Gamma}_{T+1}) \times p_\psi(\boldsymbol{\Gamma}_{T+1}|\boldsymbol{\Gamma}_T) q_\varphi(\boldsymbol{\Gamma}_{1:T}|\mathbf{W}_{1:T}) d\boldsymbol{\Gamma}_{1:T}, \quad (25)$$

where we replaced the true (intractable) posterior with the approximate posterior $q_\varphi(\boldsymbol{\Gamma}_{1:T}|\mathbf{W}_{1:T})$, and where $\boldsymbol{\Gamma}_{t,d}$ labels the set $\{\boldsymbol{\xi}_t, \boldsymbol{\eta}_t, \boldsymbol{\nu}_t, \boldsymbol{\zeta}_{t,d}, \mathbf{b}_{t,d}\}$ as before.

We can now define the predictive log likelihood as

$$\text{P-NLL} = \mathbb{E}_{p(\boldsymbol{\Gamma}_{T+1}|\boldsymbol{\Gamma}_T)}\mathbb{E}_{q(\boldsymbol{\Gamma}_{1:T}|\mathbf{W}_{1:T})}\Big\{ \log p_\psi(\mathbf{W}_{T+1}|\boldsymbol{\Gamma}_{T+1})\Big\}. \quad (26)$$

(2) To test generalization we use three metrics namely, *perplexity* (PPL) on document completion, *topic coherence* (TC) and *topic diversity* (TD).

The document completion PPL is calculated on the second half of the documents in the test set, conditioned on their first half (Rosen-Zvi et al., 2012).

The TC is calculated by taking the average pointwise mutual information between two words drawn randomly from the same topic (Lau et al., 2014) and measures the interpretability of the topic. In contrast, TD is the percentage of unique words in the top 25 words of all topics (Dieng et al., 2020). Note that one also often finds in the literature the *topic quality* metric (TQ), defined as the product of TC with TD.

5.1 Comparison with baselines

The results on both P-NLL and PPL tasks are shown in Table 1. Both our models (NDF-TM and NDF-TM-DE) outperformed all baselines on the NeurIPS and ACL datasets, but are only second and third to D-ETM in the UN dataset.

One could argue that NeurIPS and ACL feature more emergent and volatile topics (wrt. their activity), as compared to those characteristic of the UN dataset (see for example Table 5 in the Appendix, which shows six randomly sampled topics from each dataset as inferred by NDF-TM. Note how those inferred from the UN dataset seem to circle about war and peace).

It is easy to imagine that the more generic topics in the UN dataset (like war, climate, etc) have reached some type of equilibrium and thus display overall a less skewed distribution over the document collection. If this were the case, explicitly decoupling topic proportion from its activity would have little role on the effective modelling of the dataset. That is, rare topics would be less relevant in the UN dataset.

In sharp contrast, topic models trained on say NeurIPS typically infer topics about Neural Networks and their training, as well as about Reinforcement Learning (see e.g. Topic 1, 5 and 4 in Table 5 of the Appendix). Such topic easily display a strongly skewed distribution on the NeurIPS collection, which would explain the good predictive performance of our models.

Figure 3 shows the (Shannon) entropy of the topic distribution, averaged over documents as time evolves as inferred by all models [7]. Note how the entropy inferred by DTM-REP for UN is close to zero, meaning that DTM-REP usually describes the documents with few topics, whereas for NeurIPS the entropy of the average topic distribution is close to its maximum value ($\log(K=50) \approx 3.9$), meaning that it allocates almost equal probability for all K topics, as expected for a skew topic distributions. In contrast, NDF-TM uses the additional Bernoulli variable to redistribute the noise in the topic dynamics. See e.g. Figure 4 which shows the topic activity probability, average all documents as time evolves, as inferred from NDF-TM on the UN dataset. We refer the reader to the Appendix for further data visualization of the entropy dynamics for all our datasets and all models.

The results on both TC and TD shown in Table 2 reflect a similar story: NDF-TM and NDF-TM-DE perform well on NeurIPS and ACL, but are outperformed by D-ETM on the UN dataset.

Note that D-ETM learns different embeddings for each topic *at each time step* (i.e. $K * T$ embeddings in total). In comparison, NDF-TM learns only K topic embeddings, whereas NDF-TM-DE, although with $K * T$ available embeddings, has only about α_0 *active* embeddings (in average) at each time step. Interestingly enough NDF-TM, al-

[6] Note that one is effectively performing a sequential Monte Carlo sample (Speekenbrink, 2016), in which future steps are particles sampled from the posterior and propagated by the prior.

[7] The Shannon entropy of the topic distribution per document and time is defined here by $H_{t,d} = -\sum_i^K \theta_{t,d}^{(i)} \log \theta_{t,d}^{(i)}$, where $\theta_{t,d}^{(i)}$ is the ith component of $\boldsymbol{\theta}_{t,d}$.

beit with less capacity, is consistently better in both NeurIPS and ACL than NDF−TM−DE. In contrast, NDF−TM−DE is better than NDF−TM in the UN. Here again we can argue that topic embeddings are more useful for modelling the UN dataset than the topic activity-proportion decoupling.

5.2 Qualitative results

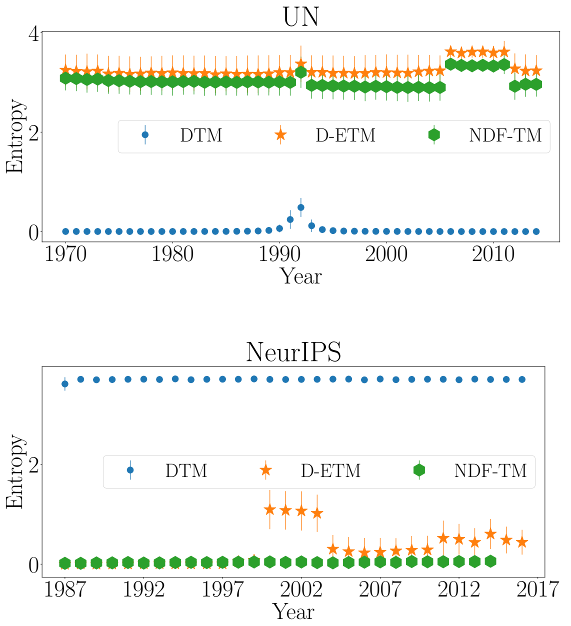

Figure 3: Entropy of topic distribution inferred by DTM−REP, D−ETM and NDF−TM, averaged over documents as time evolves. Values shown with one standard deviation for both UN (above) and NeurIPS (below) datasets. Note that the maximum entropy value is $\log(K = 50) \approx 3.9$.

One of our main claims is that decoupling topic activity from topic proportion helps the model better describe sequentially collected data. We have seen above this is indeed the case from a quantitative point of view. Nevertheless, one could ask whether (or how) this decoupling is effectively taking place as time evolves. To study how the model encodes the temporal aspects of the data, we track the time evolution of (i) the topic proportion, (ii) activity probabilities and (iii) most important words for some inferred topics. Figure 2 shows our results for the topic inferred from the UN dataset, namely *middle east*. Note, for example, that the topic proportion for *middle east* peaks in the year 1990, which coincides with the Gulf War (see also year 1990 in Table 4) to then drop right after. Such a drop is also reflected in the topic activity. Later, in 2011, the Syrian Civil War started. This event is captured by the topic activity which peaks at 2011, even though the topic proportion probability is *decreasing*. That is, even when the proportion of the *middle east* topic is low within the documents of that year, it must remain active to properly describe the data. Figure 5 in the Appendix shows similar behavior in the *climate* topic time series.

Also compare the dynamics of the topic *middle east* with that of all topic activities in Figure 4. Also note Table 4 in Appendix, which shows the evolution of the most important five words of the topic *middle east* over time. Clearly the topic remains coherent as time evolves and follows historical events (e.g. the 1974 coup, Turkish invasion and division of Cyprus).

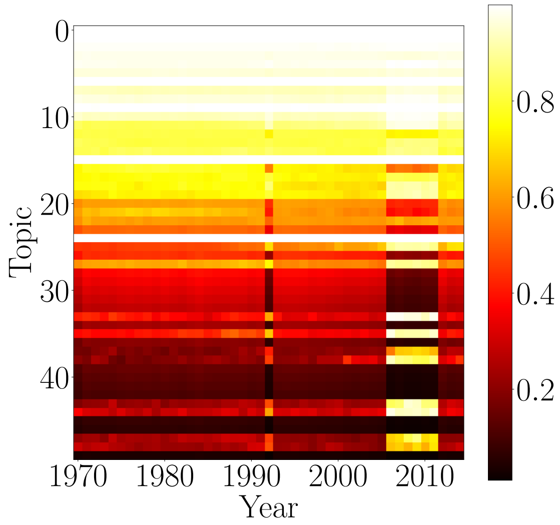

Figure 4: Average topic activity $\mathbf{b}_{t,d}$ as time evolves for all K topics in NDF−TM for UN dataset.

6 Conclusion

We have introduced the Neural Bernoulli-Beta topic model for sequentially collected data, which explicitly decouples the dynamic topic proportions from the topic activities through the addition of a Bernoulli variable modelled with a nonparametric prior. We have shown that our approach consistently yields coherent and diverse topics, which correctly capture historical events. Future work includes using NDF−TM together with Variational Autoencoders for topic-guided text generation or classification.

References

Robert Bamler and Stephan Mandt. 2017. Dynamic word embeddings. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 380–389. JMLR. org.

Alexander Baturo, Niheer Dasandi, and Slava J Mikhaylov. 2017. Understanding state preferences with text as data: Introducing the un general debate corpus. *Research & Politics*, 4(2):2053168017712821.

Steven Bird, Robert Dale, Bonnie J Dorr, Bryan Gibson, Mark Thomas Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir R Radev, and Yee Fan Tan. 2008. The acl anthology reference corpus: A reference dataset for bibliographic research in computational linguistics.

Christopher M Bishop. 2006. *Pattern recognition and machine learning*. springer.

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

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

Jordan L Boyd-Graber, Yuening Hu, David Mimno, et al. 2017. *Applications of topic models*, volume 11. Now Publishers Incorporated.

François Caron, Manuel Davy, and Arnaud Doucet. 2007. Generalized polya urn for time-varying dirichlet process mixtures. In *Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence*, UAI'07, page 33–40, Arlington, Virginia, USA. AUAI Press.

Cristian Danescu-Niculescu-Mizil, Robert West, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. No country for old members: User lifecycle and linguistic change in online communities. In *Proceedings of the 22nd International Conference on World Wide Web*, page 307–318, New York, NY, USA. Association for Computing Machinery.

Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2019. The dynamic embedded topic model. *arXiv preprint arXiv:1907.05545*.

Adji B Dieng, Francisco JR Ruiz, and David M Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Rem Hida, Naoya Takeishi, Takehisa Yairi, and Koichi Hori. 2018. Dynamic and static topic model for analyzing time-series document collections. *CoRR*, abs/1805.02203.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. 2012. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 204–213, Avignon, France. Association for Computational Linguistics.

Patrick Jähnichen, Florian Wenzel, Marius Kloft, and Stephan Mandt. 2018. Scalable generalized dynamic topic models. *arXiv preprint arXiv:1803.07868*.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Rahul G. Krishnan, Uri Shalit, and David Sontag. 2015. Deep kalman filters.

P. Kumaraswamy. 1980. A generalized probability density function for double-bounded random processes. *Journal of Hydrology*, 46(1):79–88.

Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539.

Luke Lefebure. 2018. Exploring the un general debates with dynamic topic models.

Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017. Discovering discrete latent topics with neural variational inference. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2410–2419. JMLR. org.

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *International conference on machine learning*, pages 1727–1736.

Eric Nalisnick and Padhraic Smyth. 2016. Stick-breaking variational autoencoders. *arXiv preprint arXiv:1605.06197*.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Valerio Perrone, Paul A Jenkins, Dario Spano, and Yee Whye Teh. 2016. Poisson random fields for dynamic feature models. *arXiv preprint arXiv:1611.07460*.

Mehdi Rezaee and Francis Ferraro. 2020. A discrete variational recurrent topic model without the reparametrization trick. In *Advances in Neural Information Processing Systems*, volume 33, pages 13831–13843. Curran Associates, Inc.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2012. The author-topic model for authors and documents. *arXiv preprint arXiv:1207.4169*.

Maja Rudolph and David Blei. 2018. Dynamic embeddings for language evolution. In *Proceedings of the 2018 World Wide Web Conference*, pages 1003–1011.

Maarten Speekenbrink. 2016. A tutorial on particle filters. *Journal of Mathematical Psychology*, 73:140–152.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*.

Jian Tang, Zhaoshi Meng, Xuanlong Nguyen, Qiaozhu Mei, and Ming Zhang. 2014. Understanding the limiting factors of topic modeling via posterior contraction analysis. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 190–198. PMLR.

Yee Whye Teh, Dilan Grür, and Zoubin Ghahramani. 2007. Stick-breaking construction for the indian buffet process. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, pages 556–563.

Chong Wang, David Blei, and David Heckerman. 2012. Continuous time dynamic topic models. *arXiv preprint arXiv:1206.3298*.

Sinead Williamson, Chong Wang, Katherine A Heller, and David M Blei. 2010. The ibp compound dirichlet process and its application to focused topic modeling.

Hao Zhang, Bo Chen, Dandan Guo, and Mingyuan Zhou. 2018. WHAI: Weibull hybrid autoencoding inference for deep topic modeling. In *International Conference on Learning Representations*.

He Zhao, Dinh Phung, Viet Huynh, Trung Le, and Wray Buntine. 2021. Neural topic model via optimal transport. In *International Conference on Learning Representations*.

Yuan Zuo, Jichang Zhao, and Ke Xu. 2014. Word network topic model: A simple but general solution for short and imbalanced texts.

A Dataset

A.1 Data preprocessing

We first tokenized and then removed all the numbers, punctuation and stop words in all datasets. Additionally, we removed all words with high document frequency (above 70%), as well as words with low frequency. The latter are defined as words that appear in less than 30 documents for UN and less than 10 documents for both NeurIPS and ACL. Furthermore, for UN we split the speeches into paragraph, and consider each paragraph as a document, as done in (Lefebure, 2018). For NeurIPS we took the first 120 sentence of each paper as a document. We then create ten (10) random splits of each dataset, where each split contains train (85%), validation (5%), and test (15%) set. Table 3 shows the statistics of each dataset. The preprocessing and the instruction how to run it is encoded in the source code provided as resources.

A.2 Experimental Setup

Choosing the hyperparameters was used grid search with manual tuning and used the perplexity as metric to choose the best ones. The search space for the different hyperparameters is the following: $K = [20, 50, 200]$ and $\alpha_0 = [10, 20, 50]$.

For all models we used $K = 50$ topics. We set $\alpha_0 = 10$ and $\boldsymbol{\delta} = 0.005$. We use 300 dimensional pretrained `GloVe` vectors (Pennington et al., 2014) as word embeddings.

To fairly compare with the models in Dieng et al. (2019) we set our parameter numbers similar to those of `D-EMB`. The inference models for $\boldsymbol{\theta}_{t,d}, \mathbf{b}_{t,d}$ and $\boldsymbol{\nu}_{t,d}$ consists of feed-forward neural networks with 2 hidden layers of size 800 and `ReLU` activation functions. The inference models for $\boldsymbol{\eta}_t$ and $\boldsymbol{\xi}_t$ consists of 4-layer LSTM networks with 400 hidden units per layer. Their bag-of-words $\mathbf{W}_t$ input is first map to a 400 dimensional space using a linear transformation. The output of the LSTMs is mapped to K-dimensional space to get the values of the means and log-variances for each $\boldsymbol{\eta}_t$ and $\boldsymbol{\xi}_t$.

We used SGD optimizer with learning rate of 0.001 and batch size of 200 for all datasets. We also applied gradient clipping with maximum norm of 2 and used early stopping during training.

The exact details can be also found in the code[8].

B Additional Results

In order to see the evolution of the important words per topic over time, we present in Table 4 the topic *middel east* with the top five most important words over time. Additionally, we present in Figure 5 the evolution of the θ topic proportion and the corresponding Bernoulli variable b for the *climate* topic over time. One can see that even though the topic proportion is decreasing (θ variable) in the period 1970 until 2000, the importance of the topic (b variable) is still high.

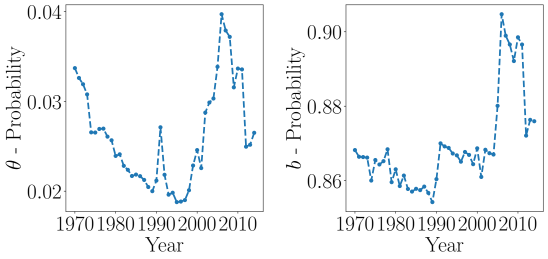

Figure 5: Evolution of topic proportion and activity probability for the topic *climate* inferred from the `UN` dataset via `NDF-TM-DE`.

In order to depict the change of the topic importance over time, we calculate the topics entropy per document per time-step. From there we calculate a histogram of the document entropy for each time-step (as well as aggregated entropy i.e. all the documents up until time step t) and we plot is as a time series of histogram (see Figure 6, 7 and 8).

[8] Uploaded as resources

Datasets	# Train Docs	# Val Docs	# Test Docs	# Time Steps	Vocabulary
UN	196,290	11,563	23,097	46	12,466
NeurIPS	5,249	329	984	30	6,958
ACL	8,936	527	1,051	31	35,108

Table 3: Dataset statistics.

1971	1975	1982	1985	1990	2014
solution	cyprus	east	east	kuwait	peace
problem	solution	solution	peace	arab	israel
settlement	problem	middle	israel	iraq	east
parties	settlement	peace	middle	security	palestinian
concerned	sovereignty	security	palestinian	east	state

Table 4: Time evolution of the top five most important words for the topic *middle east*, as inferred from the UN dataset via NDF-TM-DE. Note how the topic remains coherent as time evolves.

Datasets	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
UN	nuclear weapons disarmament treaty iraq	east middle israel arab palestinian	war conflicts violence tension power	terrorism acts fight drug crime	problems international crisis debt financial	peace region hope efforts stability
NeurIPS	layers gradient nets stochastic descent	graph nodes variables inference structure	game decision human player response	state action reward policy agent	gradient optimization stochastic convergence descent	ranking query pairwise search pages
ACL	speech recognition spoken speaker dialogue	annotation entity names annotated corpus	interpretation semantics representation sentence expressions	tag corpus pos tagging morphological	question answer correct knowledge candidate	speech prosodic phrase boundary pitch

Table 5: Top five words from six randomly sampled topics for each dataset. The topics are learned using NDF-TM.

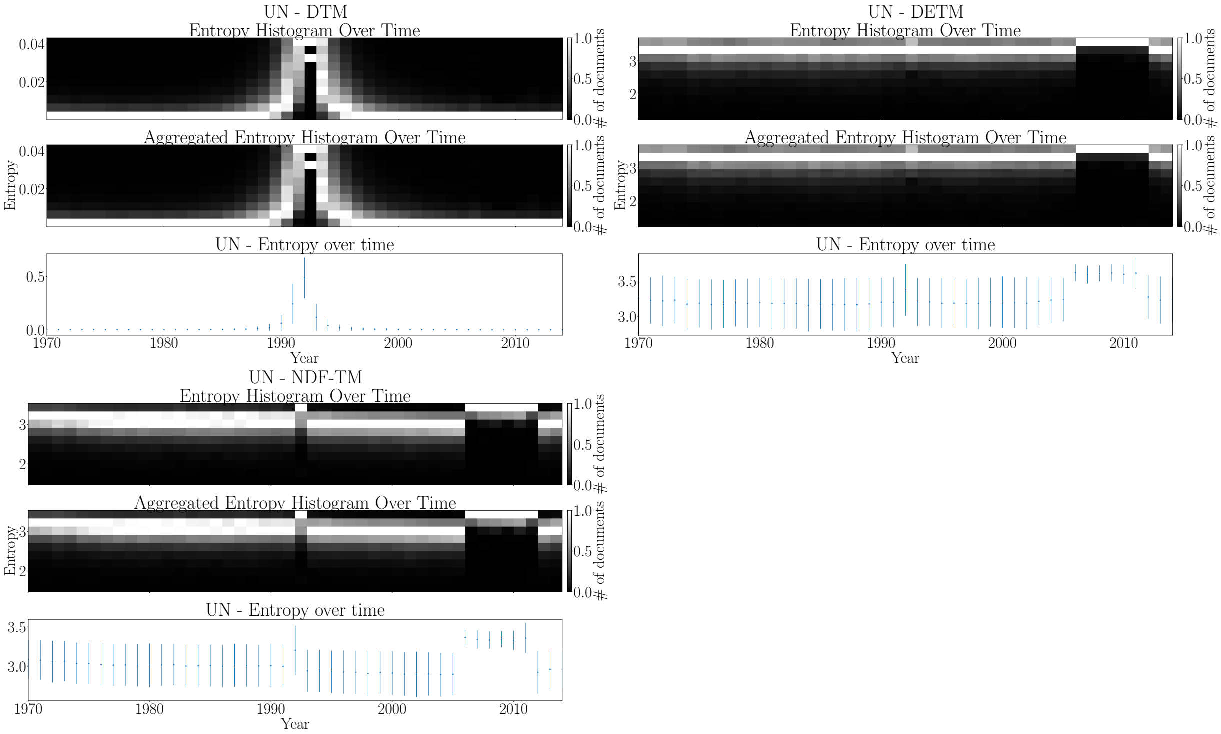

Figure 6: Histogram of document entropy per time step for the UN dataset.

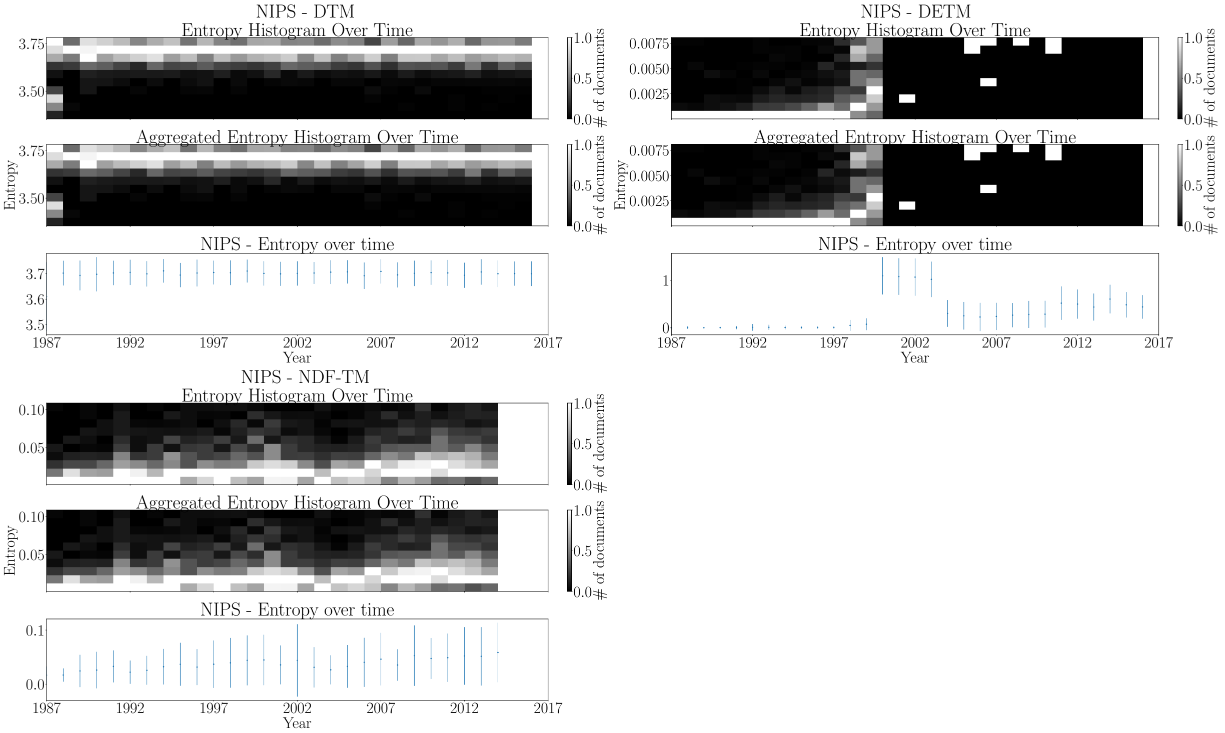

Figure 7: Histogram of document entropy per time step for the NIPS dataset.

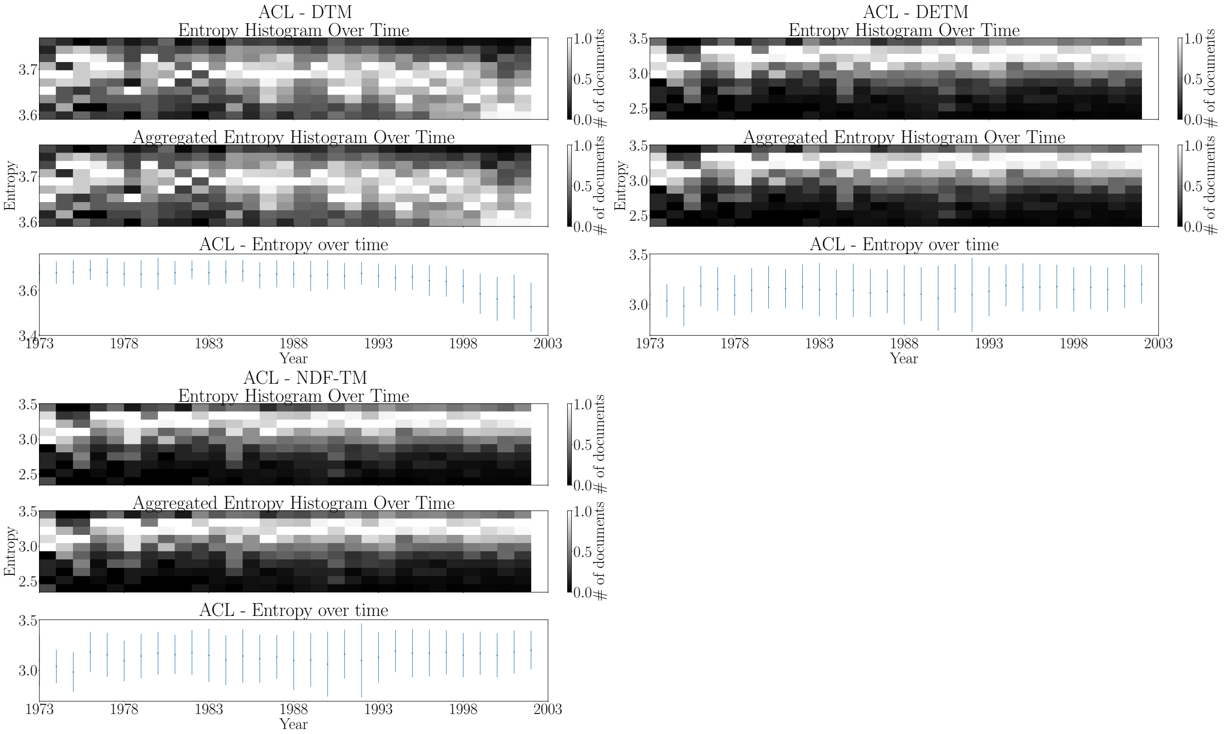

Figure 8: Histogram of document entropy per time step for the ACL dataset.