Stick-Breaking Embedded Topic Model with Continuous Optimal Transport for Online Analysis of Document Streams

Federica Granese
Université Côte d'Azur,
Inria (MARIANNE), Inria DEES,
France

Serena Villata
Université Côte d'Azur,
CNRS, I3S, Inria (MARIANNE),
France

Charles Bouveyron
Université Côte d'Azur,
CNRS, LJAD, Inria (MAASAI),
France

Abstract

Online topic models are unsupervised algorithms to identify latent topics in data streams that continuously evolve over time. Although these methods naturally align with real-world scenarios, they have received considerably less attention from the community compared to their offline counterparts, due to specific additional challenges. To tackle these issues, we present SB-**S**ETM, an innovative model extending the Embedded Topic Model (ETM) to process data streams by merging models formed on successive partial document batches. To this end, SB-**S**ETM *(i)* leverages a truncated stick-breaking construction for the topic–per-document distribution, enabling the model to automatically infer from the data the appropriate number of active topics at each timestep; and *(ii)* introduces a merging strategy for topic embeddings based on a continuous formulation of optimal transport adapted to the high dimensionality of the latent topic space. Numerical experiments show SB-**S**ETM outperforming baselines on simulated scenarios. We extensively test it on a real-world corpus of news articles covering the Russian–Ukrainian war throughout 2022–2023.

1 INTRODUCTION

Topic modeling aims to identify, in an unsupervised manner, latent topics in a collection of documents by analyzing word co-occurrence patterns. A topic is a probability distribution over the words of the vocabulary and can be interpreted through a representative set of words defining the topic. Each document can, in turn, be represented as a distribution over the topics (Blei et al., 2003). Beyond their ability to analyze various forms of textual data, ranging from scientific publications to political discourse (Boyd-Graber et al., 2017), topic models have proven effective in several real-world applications, including customer feedback analysis, content organization, and market research for sentiment analysis and trend detection (Sverdlik, 2025). Many of these applications are naturally aligned with streaming settings, where new data arrives continuously over time. This temporal dimension of topic modeling can be addressed in two main ways. The *dynamic* setting assumes access to the entire historical corpus, with topic evolution analyzed retrospectively (Blei and Lafferty, 2006; Dieng et al., 2019; Zhang and Lauw, 2022; Karakkaparambil James et al., 2024). In contrast, the *online* setting assumes that documents arrive sequentially, and the model must infer topics at time t using only the topics inferred at time $t-1$, without revisiting past data. Despite its closer alignment with real-world scenarios, the online setting has received less attention in the literature.

A recent contribution to this setting is StreamETM (Granese et al., 2025), which extends the Embedded Topic Model (ETM) (Dieng et al., 2020) by applying variational inference sequentially over consecutive batches of documents and leveraging unbalanced optimal transport to associate topics across time steps. While StreamETM performs well against strong baselines such as BERT-based models (Grootendorst, 2023, 2022a), it inherits two important limitations. First, like most topic models, it requires the number of topics K to be specified a priori. This assumption is particularly problematic in streaming contexts: too few topics may prevent the discovery of new ones, while too many can lead to topic explosion. Second, its merging strategy is based

Proceedings of the 29[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300.

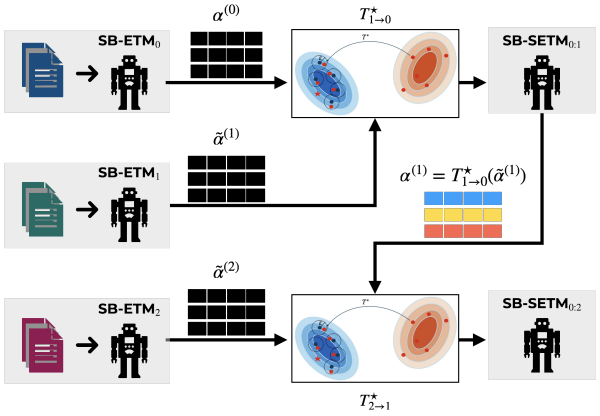


Figure 1: **The proposed SB-SETM**. At each time step, SB-ETM$_t$ receives a new corpus of documents and infers the topic embeddings. These embeddings are then projected into the latent space of the previous timestep via the transport map $T^\star_{t\to t-1}$.

on averaging topic embeddings, which may not lie in the same latent space. Although optimal transport identifies which topics should be merged, the subsequent averaging may overlook possible rotations or geometric inconsistencies in the embedding space. In this work, we address these limitations by introducing SB-**S**ETM, a novel online topic modeling framework with three main contributions:

From a theoretical perspective: **1)** we introduce a document-level truncated stick–breaking construction in place of the ETM's logistic–normal prior, enabling data-adaptive activation of topics without fixing K; we provide a reparameterizable variational family via a Gaussian latent $\mathbf{z}^{(d)}$ and Kumaraswamy sticks $\boldsymbol{\nu}^{(d)}$, yielding a tractable ELBO with a dedicated KL term that controls topic activation; and **2)** we recast the task of merging models estimated at consecutive timesteps as a continuous optimal transport problem; we rely for this on an efficient low-rank computation of the Monge map for high-dimensional spaces to transport the topic embeddings of time t into the latent space of time $t-1$, avoiding post-hoc averaging and preserving the geometry of the embeddings.
From an experimental perspective: **3)** we evaluate SB-**S**ETM on an expert-annotated news corpus covering the Russian–Ukrainian war (2022–2023), showing that topic-frequency peaks align with major events.

Related works. OnlineLDA (Hoffman et al., 2010) is one of the first algorithms for online topic modeling based on the Latent Dirichlet Allocation model (Blei et al., 2003). It considers a stochastic optimization algorithm based on a natural gradient step to optimize the variational Bayes lower bound as data arrive. As OnlineLDA is not suitable for document streams that cannot be stored, Amoualian et al. (2016) addresses this issue by extending LDA to document batches using copulas. Other existing solutions for the online setting rely on a variant of BERTopic (Grootendorst, 2022b). MergeBERT (Grootendorst, 2023) is a pseudo-online variant of BERTopic in which topic models are compared sequentially over time by computing the cosine similarity between topic embeddings at consecutive time steps. OnlineBERT (Grootendorst, 2022a) is the true online variant of the original BERTopic, preserving the embedding transformation and c-TF-IDF approach. However, it introduces online dimensionality reduction and a different clustering method, which may lead to the loss of non-linear relationships and less coherent topics in some cases. Recently, Granese et al. (2025) proposed StreamETM, an extension of ETM to the online setting. The method consists of aligning topic embeddings sequentially over time using Unbalanced Optimal Transport. The merge between the topic embeddings consists of the average of the matched embeddings. As our work is a direct extension of StreamETM, which clearly outperforms online BERT approaches, we benchmark against it.

Beyond their qualities, all of these online topic models share the same two critical issues.: *(i)* the number of topics must be fixed a priori, and *(ii)* the consistency between the latent spaces of topic embeddings is not guaranteed when they are compared. Regarding the first issue, we remind that the number of topics might be a critical parameter of a topic model and there is no agreement in the literature on the sequence of steps one must carry out to determine the best number of topics (Bulatov et al., 2023). Existing practice is to train multiple models with different values of K, evaluate a quality metric (often via cross-validation on held-out documents), and choose the best. This approach is not scalable in real-world online settings. Underestimating K can mask topics, while overestimating K tends to fragment them. Regarding the second issue, which is typical of MergeBERT and StreamETM, the topic embeddings at t and $t-1$ could not lie in the same latent space; therefore, a mapping based on cosine similarity (as in MergeBERT) could lead to unmeaningful results. Although StreamETM accounts for this mismatch via an OT mapping, it ultimately averages the matched embeddings, a step that can be geometrically inconsistent. Therefore, it only mitigates the issue rather than solves it.

2 THE PROPOSED APPROACH

Figure 1 provides a graphical overview of the proposed framework. At each time step t, a **S**tick-**B**reaking **E**mbedded **T**opic **M**odel (SB-ETM) takes as input a

corpus of documents and, for each discovered topic, produces a topic embedding $\tilde{\alpha}_i^{(t)}$. If $t > 0$, we compute a transport map $\underset{t \to t-1}{T^\star}$ to position $\tilde{\alpha}_i^{(t)}$ into the latent space of the topics at time $t-1$, yielding $\alpha_i^{(t)} = \underset{t \to t-1}{T^\star}\left(\tilde{\alpha}_i^{(t)}\right)$. To track topic evolution, the embeddings at $t-1$ serve as reference points: if $\alpha_i^{(t)}$ does not fall within the ε-neighborhood of any topic at $t-1$, we flag $\alpha_i^{(t)}$ as a new topic. The merged model obtained by combining time steps up to t, uses the transported embeddings as its topic embeddings.

Remark. We denote with SB-ETM the model at the single timestep before any merging, i.e., the model producing the raw topic embeddings $\tilde{\alpha}_i^{(t)}$. After merging, we refer to the resulting model as SB-**S**ETM (the additional S stands for **S**tream), which uses the transported embeddings $\boldsymbol{\alpha}^{(t)}$ in the reference space of $t-1$.

2.1 SB-ETM: The Model at Time t

At each time step t, the model M_t receives a corpus of documents $\mathcal{W} = \{\mathbb{W}^{(d)}\}_{d=1}^{D}$, where each document is a multiset of $N^{(d)}$ words. Each document is then mapped to a normalized bag-of-words (BoW) vector $\mathbf{W}^{(d)} \in \mathbb{R}^V$, representing the empirical word frequency over a fixed vocabulary $\mathcal{V} = \{v_1, \dots, v_V\}$ defined a priori based on the application domain.

Generative model. We define a probabilistic decoder that generates documents from latent topic proportions. Each topic $k \in \{1, \dots, K\}$ is represented by an embedding $\alpha_k \in \mathbb{R}^L$ and each word $v \in \mathcal{V}$ by an embedding $\rho_v \in \mathbb{R}^L$. The word distribution of topic k is defined as $\beta_k = \text{softmax}(\boldsymbol{\rho}^\top \alpha_k) \in \Delta^V$. A central limitation of classical topic models is the need to fix the number of topics K in advance. We address this issue by adopting a truncated stick-breaking construction of the topic per document distribution $\boldsymbol{\theta}^d \in \Delta^K$. Specifically, we first draw auxiliary variables $\nu_k^{(d)} \sim \text{Beta}(a, b), \quad k = 1, \dots, K-1, \quad a, b > 0$ and map them deterministically via the stick-breaking transform $f_{\text{SB}} : (0,1)^{K-1} \to \Delta^K$:

$$\begin{aligned}
\theta_1^{(d)} &= \nu_1^{(d)}, \\
\theta_k^{(d)} &= \nu_k^{(d)} \prod_{j=1}^{k-1} (1 - \nu_j^{(d)}) \quad k = 2, \dots, K-1, \\
\theta_K^{(d)} &= \prod_{j=1}^{K-1} (1 - \nu_j^{(d)}),
\end{aligned}$$

where $\theta_k^{(d)} \equiv f_{\text{SB}}(\nu_k^{(d)})$. Intuitively, the stick-breaking process consists of repeatedly breaking off and discarding a random fraction of a stick of unit length. $\nu_1^{(d)}$ determines the length of the first segment, which corresponds to the probability of the first topic; $\nu_2^{(2)}$ specifies what fraction of the remaining stick $1 - \nu_1^{(d)}$ is allocated to the second topic. Each $\nu_k^{(d)}$ therefore regulates how much probability mass is assigned to topic k, with later topics inheriting progressively smaller portions of the stick. In practice, we set K to a large value, allowing the model to automatically activate only as many topics as required by the data.

Given the topic proportions $\boldsymbol{\theta}^{(d)} = \left[\theta_1^{(d)}, \dots, \theta_K^{(d)}\right] \in \Delta^K$, for each token n in the document d: $\widehat{\tau}_n^{(d)} \sim \text{Multinomial}(\boldsymbol{\theta}^{(d)})$ and $\widehat{w}_n^{(d)} \sim \text{Multinomial}(\beta_{\widehat{\tau}_n^{(d)}})$. Let us denote the topic–word matrix $\boldsymbol{\beta} = (\beta_1, \dots, \beta_K) \in \mathbb{R}^{V \times K}$, marginalizing out $\widehat{\tau}_n^{(d)}$ we obtain:

$$p(\mathbb{W}^{(d)} \mid \boldsymbol{\theta}^{(d)}, \boldsymbol{\alpha}, \boldsymbol{\rho}) = \prod_{v \in \mathcal{V}} \sum_{k=1}^{K-1} \theta_k^{(d)} \, \text{softmax}(\rho_v^\top \alpha_k).$$

Inference model. The parameters of our model are the embeddings $\boldsymbol{\alpha}$, and $\boldsymbol{\rho}$. Our goal is therefore to maximize the marginal likelihood of the documents, which, however, is intractable due to the integral over the topic proportions

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\rho}) &= \sum_{d=1}^{D} \log p(\mathbb{W}^{(d)} \mid \boldsymbol{\alpha}, \boldsymbol{\rho}) \\
&\geq \sum_{d=1}^{D} \mathbb{E}_{q_\phi(\boldsymbol{\theta}^{(d)} \mid \mathbb{W}^{(d)})} \left[\log p(\mathbb{W}^{(d)} \mid \boldsymbol{\theta}^{(d)}, \boldsymbol{\alpha}, \boldsymbol{\rho})\right] \\
&\quad - \text{KL}\left(q_\phi(\boldsymbol{\theta}^{(d)} \mid \mathbb{W}^{(d)}) \,||\, p(\boldsymbol{\theta}^{(\boldsymbol{d})})\right). \qquad (1)
\end{aligned}$$

For completeness, we delegate in Appendix A.1 the derivation from the original loss to the ELBO.

Since the Beta distribution is not easily reparameterizable, we follow the strategy in Nalisnick and Smyth (2016) and approximate it with the Kumaraswamy distribution (Kumaraswamy, 1980). Crucially, this distribution has *(i)* the same support as the Beta, and for equivalent parameter values of (a, b), it resembles the Beta distribution, albeit with higher entropy; *(ii)* it admits a closed-form inverse CDF, which allows to generate differentiable samples by drawing $u \sim \mathcal{U}(0,1)$ and $\nu = (1 - (1-u)^{\frac{1}{b}})^{\frac{1}{a}})$; *(iii)* its KL-divergence from the Beta can be approximated in closed-form for ELBO computation (Nalisnick and Smyth, 2016).

We also introduce the auxiliary variable $\mathbf{z}^{(d)}$ and we define it $\mathbf{z}^{(d)} \sim \pi_\phi(\mathbf{z}^{(d)} \mid \mathbb{W}^{(d)}) \equiv \mathcal{N}\big(\mathbf{z}^{(d)}; \mu_\phi(\mathbb{W}^{(d)}), \sigma_\phi^2(\mathbb{W}^{(d)})\big)$, $\boldsymbol{\nu}^{(d)} \sim \pi_\psi(\boldsymbol{\nu}^{(d)} \mid \mathbf{z}^{(d)}) \equiv \prod_{k=1}^{K-1} \text{Kumaraswamy}\Big(\nu_k^{(d)}; a_\psi(\mathbf{z}^{(d)}), b_\psi(\mathbf{z}^{(d)})\Big)$.

Finally, our training objective, which is a lower bound of Eq. (1) (proof in Appendix A.2) is given by

$$\widetilde{\mathcal{L}}(\boldsymbol{\alpha}, \boldsymbol{\rho}) = \omega_R \cdot \underbrace{\sum_{d=1}^{D} \mathbb{E}_{q_\phi} \left[\log p(\mathbb{W}^{(d)} \mid \boldsymbol{\theta}^{(d)}, \boldsymbol{\alpha}, \boldsymbol{\rho}) \right]}_{\text{Reconstruction term}}$$
$$- \omega_G \cdot \underbrace{\mathrm{KL}\left(\pi_\phi(\mathbf{z}^{(d)} \mid \mathbb{W}^{(d)}) \,||\, \mathcal{N}(0, I) \right)}_{\text{Gaussian regulirizer term}}$$
$$- \omega_S \cdot \underbrace{\mathbb{E}_{\pi_\phi} \left[\mathrm{KL}\left(\pi_\psi(\boldsymbol{\nu}^{(d)} \mid \mathbf{z}^{(d)}) \,||\, \mathrm{Beta}(a, b) \right) \right]}_{\text{Stick-breaking term}}, \quad (2)$$

where $q_\phi \equiv q_\phi(\boldsymbol{\theta}^{(d)} \mid \mathbb{W}^{(d)})$. The reconstruction term in Eq. (2) plays the same role as in the classical ETM, ensuring that the model can accurately reconstruct the observed documents. The stick-breaking KL term regulates the effective number of active topics: for low values of ω_S, the model tends to maintain redundant topics, whereas for high values, the model may collapse using only a few topics. Finally, the Gaussian KL term acts as a regularizer, preventing the variational posterior from drifting too far from the prior. The relative weights ω_R, ω_G, ω_S depend also on the values of (a, b) used as prior for the Beta distribution. An analysis is provided in Appendix B. The learning algorithm is in Algorithm 1.

Links with related models. Let us quickly discuss the links and differences with related models. Regarding ETM Dieng et al. (2020), the main deviation lies in the prior over document-level topic proportions: instead of the logistic–normal prior, we adopt a truncated stick–breaking prior, which naturally promotes data-adaptive topic allocation. This modification (i) changes the generative process by replacing the logistic–normal mapping with the stick-breaking construction, and (ii) alters the training objective by introducing an additional KL term between the Kumaraswamy variational sticks and the Beta stick-breaking prior in the ELBO. Unlike Nalisnick and Smyth (2016), who introduce stick-breaking variational autoencoders in a general latent-variable modeling context, our approach specifically targets topic modeling and adapts the stick-breaking prior to document-level topic proportions. Moreover, we stress that our approach is conceptually distinct from hierarchical topic models, which aim to define a tree-structured organization of topics. In such models, a document is typically generated by sampling a path from the root to a leaf, selecting topics along this path, and generating words from the associated distributions (Griffiths et al., 2003). In contrast, our goal is not to uncover explicit parent-child relations among topics at time t, but rather to avoid fixing the number of topics a priori, a constraint that can significantly affect performance in the online settings. Finally, both our approach and Ouyang et al. (2024) use a Kumaraswamy stick-breaking construction. Unlike Ouyang et al., who map documents directly to (a, b) via an MLP with an LSTM-based recurrent stick-breaking in an offline setting, our model derives (a, b) from an intermediate Gaussian latent variable $z^{(d)}$, with an ELBO split into two KL terms.

2.1.1 Merging Consecutive Timesteps

We now focus on the way to align the models learned at two consecutive timesteps $t-1$ and t. Let us consider a stream of documents arriving as batches at discrete time steps, $\mathcal{W}_{[1:T]} = \{\mathcal{W}^{(1)}, \ldots, \mathcal{W}^{(t-1)}, \mathcal{W}^{(t)}, \mathcal{W}^{(t+1)}, \ldots, \mathcal{W}^{(T)}\}$, where each $\mathcal{W}^{(i)}$ in $\mathcal{W}_{[1:T]}$ represents a corpus of documents as defined in Section 2.1. Let us also assume that two SB-ETM models have been learned from the respective time batches, for which we aim at aligning the topic embeddings. Following the approach proposed in (Granese et al., 2025), we propose to rely on optimal transport to identify the correspondences between the identified topics of the two models and, when necessary, to attest to the discovery of new topics. In contrast to this seminal work, which used Discrete Optimal Transport (DOT), we propose here to rely on its Continuous version (COT) (Peyré et al., 2019). Crucially, instead of finding the point-to-point transport of the topic embeddings between the two models, the idea is to transport the distribution of the topic embeddings of time t over the distribution of their counterparts at time $t-1$. As the topic embeddings are living in a high-dimensional latent space and knowing that OT does not scale well when the dimensionality increases, we propose to use a recently proposed approach (Bouveyron and Corneli, 2025), allowing, in an efficient manner, the Continuous Optimal Transport between two high-dimensional Gaussian distributions.

Based on the modern formulation of Kantorovich (1942), standard OT is generally based on the Wasserstein distance. Given two random variables X_1 and X_2 supported on $\mathbb{R}^p$, with finite second moments and whose marginal cumulative distribution functions are denoted by μ_1 and μ_2, respectively, the squared 2-Wasserstein distance is defined as:

$$W_2^2(\mu_1, \mu_2) := \min_{\pi \in \Pi(\mu_1, \mu_2)} \mathbb{E}_{(X_1, X_2) \sim \pi} ||X_1 - X_2||_2^2, \quad (3)$$

where $\Pi(\mu_1, \mu_2)$ denotes the set of *joint* distributions with marginals μ_1 and μ_2, respectively. The joint distribution π^* minimizing Eq. (3) is known as an optimal coupling or optimal transport plan. Moreover, there exists a unique transport or *Monge* map $T^* : \mathbb{R}^p \to \mathbb{R}^p$ linked to the optimal transport plan π^* by the following relation:

$$\mathbb{E}_{(X_1, X_2) \sim \pi^*} \left[h(X_1, X_2) \right] = \mathbb{E}_{X_1 \sim \mu_1} \left[h(X_1, T^*(X_1)) \right],$$

holding for any continuous function $h : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. In the Gaussian case, both $W_2^2(\mu_1, \mu_2)$ and T^* have closed forms (Peyré et al., 2019). Yet, when the space dimensionality is large and, in particular, when the number of samples is small compared to the dimensionality, the computation errors of both quantities are extremely difficult. To tackle this issue, Bouveyron and Corneli (2025) proposed simplified closed-form expressions for the 2-Wasserstein distance and the Monge transport map, with efficient, robust calculation procedures based on a low-dimensional decomposition of empirical covariance matrices. In addition to analytical and computational advantages, this approach outperforms model-free methods in high dimensions, even for non-Gaussian distributions.

Let us assume that $\tilde{\alpha}^{(t)}$ and $\alpha^{(t-1)}$ are both distributed as high-dimensional Gaussian distributions, resp. $\tilde{\alpha}^{(t)} \sim \mathcal{N}_{HD}(m_t, U_t, \Lambda_t, \sigma_t^2, d)$ and $\alpha^{(t-1)} \sim \mathcal{N}_{HD}(m_{t-1}, U_{t-1}, \Lambda_{t-1}, \sigma_{t-1}^2, d)$, with respective intrinsic dimensionalities d_t and d_{t-1}. A L-dimensional random vector $X \in \mathbb{R}^L$ follows a HD-Gaussian distribution $\mathcal{N}_{HD}(m, U, \Lambda, \sigma^2, d)$ if it exists a low-dimensional latent random vector $Y \in \mathbb{R}^d$, of intrinsic dimensionality $d < L$ ($L = 300$ or $L = 800$ depending on the model in the experiments), and a L-dimensional noise random vector $\varepsilon \in \mathbb{R}^L$ such that:

$$\begin{aligned} \mathrm{X} &= UY + m + \varepsilon, \\ Y &\sim \mathcal{N}(0, \Lambda), \\ \varepsilon &\sim \mathcal{N}(0, \sigma^2 \mathrm{I}_L), \end{aligned}$$

where U is a $L \times d$ transformation matrix whose columns are orthonormal vectors, $m \in \mathbb{R}^L$ is the mean vector, $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$ and $\sigma^2 > 0$. In this case, Bouveyron and Corneli (2025) have shown that it is possible to transport the topic embeddings of t in the latent space of $t-1$ using the Monge map, $\forall \tilde{\alpha}^{(t)} \in \mathbb{R}^L$:

$$\underset{t \to t-1}{T^\star}\left(\tilde{\alpha}^{(t)}\right) = m_{t-1} + \Sigma_t^{-\frac{1}{2}} \mathbf{\Sigma}_t^{\frac{1}{2}} \Sigma_t^{-\frac{1}{2}} (\tilde{\alpha}^{(t)} - m_t),$$

where $\mathbf{\Sigma}_t \equiv \left[\Sigma_t^{\frac{1}{2}} \Sigma_{t-1} \Sigma_t^{\frac{1}{2}}\right]$ and both $\Sigma_t^{\frac{1}{2}}$ and $\Sigma_t^{-\frac{1}{2}}$ have the explicit closed-form formulations

$$\Sigma_t^{\frac{1}{2}} = \sigma_t I_L + U_t C_t U_t^t,$$

with $C_t = \mathrm{diag}(\sqrt{\delta_{t1}} - \sigma_1, ..., \sqrt{\delta_{td_t}} - \sigma_t) > 0$ and

$$\Sigma_t^{-\frac{1}{2}} = \frac{1}{\sigma_t}\left(I_L - U_t D_t U_t^t\right)$$

with $D_t = \mathrm{diag}\left(\frac{\sqrt{\delta_{t1}} - \sigma_1}{\sqrt{\delta_{t1}}}, \ldots, \frac{\sqrt{\delta_{td_t}} - \sigma_t}{\sqrt{\delta_{td_t}}}\right)$. See Theorem 2.10 of Bouveyron and Corneli (2025) and its proof for more details.

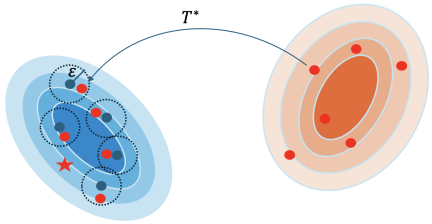


Figure 2: **Merging of topic embeddings between $t-1$ and t.** Latent space of $\alpha^{(t-1)}$ in blue and of $\tilde{\alpha}^{(t)}$ in orange. First, the topic embeddings at time t are transported using the Monge map $\underset{t \to t-1}{T^\star}$ between two high-dimensional Gaussian distributions; then topics that are not in the ε-neighborhood of an existing topic will be considered as new topics (denoted by a star).

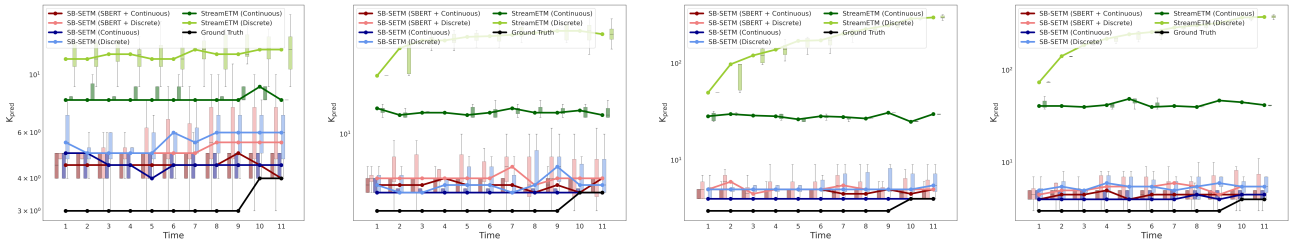
Tracing the topic evolution. The COT merging approach does not allow us to directly track the evolution of topics over time, as topics are projected into embedding spaces. To address this, we propose the following strategy. Let $\alpha^{(t)} = \underset{t \to t-1}{T^\star}\left(\tilde{\alpha}^{(t)}\right)$ denote the transported topic embeddings at time t in the latent space of the embeddings at time $t-1$. Once the transported embeddings $\alpha^{(t)}$ are computed, we match similar topics between consecutive time steps by finding the ε-neighbors of each $\alpha_j^{(t-1)}$, $j = 1, \ldots, K^{(t)}$: for $\varepsilon > 0$, $\alpha_j^{(t)}$ is an ε-neighbor of $\alpha_j^{(t-1)}$ if $||\alpha_j^{(t)} - \alpha_j^{(t-1)}||_2^2 \leq \varepsilon$. After this matching, the $\alpha_j^{(t)}$ that are not associated with any $\alpha_j^{(t-1)}$ are considered new topics. Figure 2 summarizes this merging process. Due to space limitations, additional details on the tracing strategy can be found in Appendix C.

3 EXPERIMENTAL SETTING

Datasets. We use the `20kNewsGroupOnline` a subset of `20kNewsGroup`[1] dataset designed by Granese et al. (2025) to mimic an online scenario. We focus on the CUSTOM setting therein, in which 5 (`autos`, `sport`, `medicine`, `space`, `religion`) of the 20 topics are chosen. At each time step, at most four out of five topics are *active*. Notably, at timestep 7, the topic on `space` disappears and the topic on `sport` appears. At timestep 10, the topic on `medicine` emerges. We also consider the `UkRuWarNews22-23` dataset[2]. The corpus comprises 4.211 news articles (in English and French) regarding the war Russian-Ukraine from 2022 to 2023. We focus on the 2.430 articles in English, including contributions from 113 different official news sources (e.g., president.gov.ua, Reuters, AP News, BBC, The

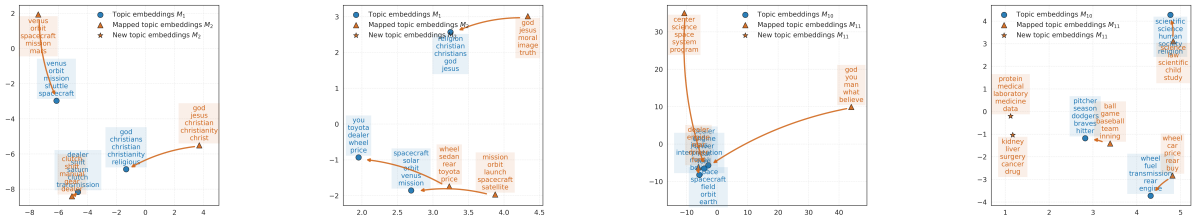
[1] `http://qwone.com/~jason/20Newsgroups/`

[2] The paper presenting the dataset is currently under submission, and cannot be referenced to ensure anonymity.



(a) $K_{\texttt{init}} = 15$ (b) $K_{\texttt{init}} = 25$ (c) $K_{\texttt{init}} = 50$ (d) $K_{\texttt{init}} = 75$

Figure 3: **Distribution of $K_{\texttt{pred}}$ (log-scale) at each timestep across 8 training runs at different $K_{\texttt{init}}$.** The solid lines connect the median values of the corresponding box plots.



(a) StreamETM (b) SB-**S**ETM (c) StreamETM (d) SB-**S**ETM

Figure 4: **PCA projections of topic embeddings** (at timesteps $t = 1$–2 for (a)-(b), and $t = 11$–12 for (c)-(d)). Arrows indicate associations between topics across consecutive timesteps. In (c), the topics in blue are: god/jesus/interpretation/moral/belief, space/spacecraft/field/orbit/earth, dealer/engine/driver/rear/ford.

Guardian, and CNN Edition). The articles have been categorised by three experts in political and communication sciences, over 39 distinct thematic categories on the conflict. Datasets' details are in Appendix B.

Evaluation metrics. We measure topic quality in terms of topic coherence (TC) (Mimno et al., 2011) and topic diversity (TD) (Dieng et al., 2020). We report the number of predicted (active) topics ($K_{\texttt{pred}}$), and compare it with the ground-truth ($K_{\texttt{real}}$) and the initialized value ($K_{\texttt{init}}$). Ideally, $K_{\texttt{pred}}$ remains approximately constant across different $K_{\texttt{init}}$ and satisfies $K_{\texttt{pred}} \approx K_{\texttt{real}}$. We consider $K_{\texttt{init}} \in \{15, 25, 50, 75\}$ and we let $e_i = |K^{(i)}_{\texttt{pred}} - K_{\texttt{real}}|$ be the topic-count error at initialization i. We define the error dispersion $\Delta = \max_i e_i - \min_i e_i$. We combine dispersion with topic quality via the metric $P = \Delta \cdot (1 - \mathrm{H}_{(\mathrm{TC},\,\mathrm{TD})})$ where $\mathrm{H}_{(\mathrm{TC},\,\mathrm{TD})}$ is the harmonic mean of TC and TD.

Sentence-BERT and Stick-Breaking modules. We modify StreamETM architecture by adding (i) document-level context via Sentence-BERT, SBERT for short, (`paraphrase-multilingual-mpnet-base-v2`) fused with BoW through self-attention, then projected to an 800-dimensional hidden space with residual blocks, batch norm, and dropout (0.1); (ii) a stick-breaking encoder with two linear layers (`fc_a`, `fc_b`) mapping $\mathbf{z}$ to Kumaraswamy parameters and with `softplus` to enforce $a_k, b_k > 0$. Additional details on the model, training procedure, and computational time are provided in Appendix B.

4 EXPERIMENTAL RESULTS

Table 1: Quantitative comparison of topic quality between the different compared approaches using the measure $P = \Delta \cdot (1 - \mathrm{H}_{(\mathrm{TC},\,\mathrm{TD})})$.

Model	SBERT	OT	$P \downarrow$
SB-SETM	✓	**Continuous**	**0.02**
SB-**S**ETM	✓	Discrete	0.05
SB-**S**ETM	×	Continuous	0.08
SB-**S**ETM	×	Discrete	0.08
StreamETM	×	Continuous	17.8
StreamETM	×	Discrete	156.88

Here, we distinguish between the models based on the Continuous Optimal Transport (COT-based) and Discrete Optimal Transport (DOT-based) merging strategies. Additionally, for SB-**S**ETM, we consider both the variant with the SBERT module and the one without

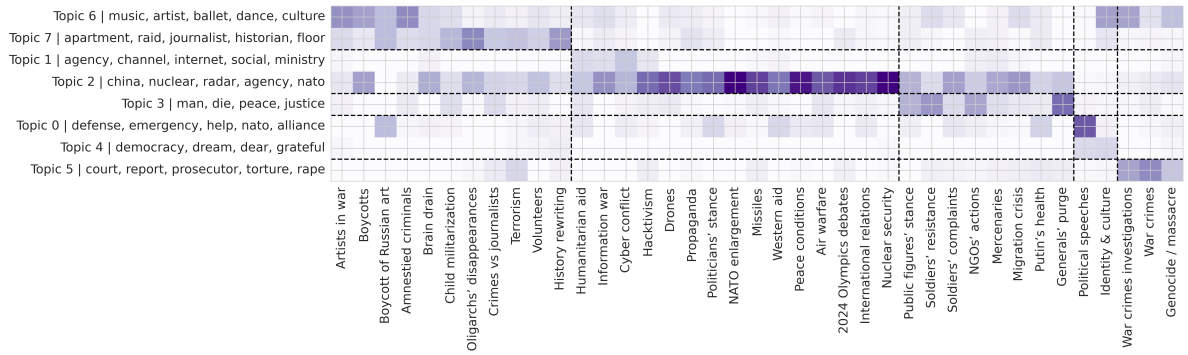


Figure 5: **Topic-term count matrix on `UkRuWarNews22-23`, SB-SETM ($K_{\texttt{init}} = 50$).** Each row is a predicted global topic, and each column is a dataset category. The entries indicate how many documents of a given class are assigned to each topic, rescaled with inverse column weights to account for class imbalance.

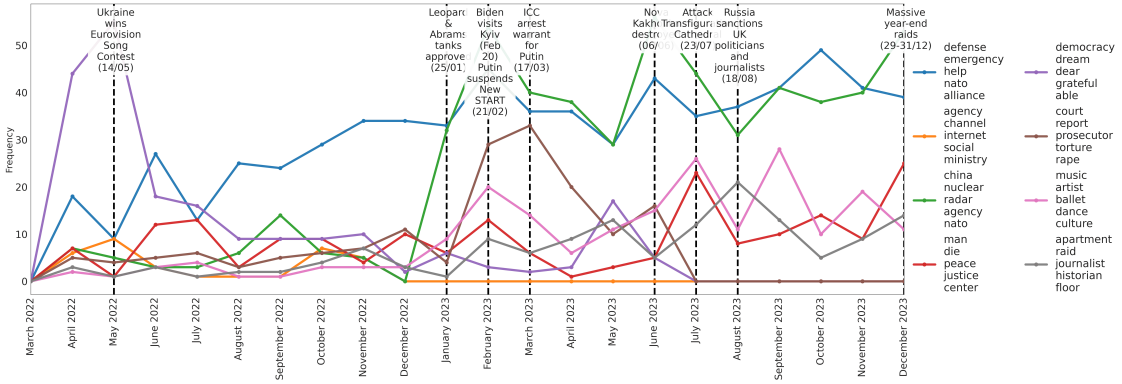


Figure 6: **Topic frequencies over time `UkRuWarNews22-23` dataset, SB-SETM model ($K_{\texttt{init}} = 50$).**

it. Full results for Table 3 and ablations (COT vs. DOT; SBERT contribution) are in Appendix C.

4.1 Results for `20kNewsGroupOnline`

Choice of the number of topics. In Figure 3 and Table 3, we present the analysis of the number of predicted topics at each timestep across the 8 training iterations. This section focuses purely on a quantitative perspective, without considering the semantic content of the topics. The latter is evaluated using $\mathrm{H}_{(\mathrm{TC},\,\mathrm{TD})}$ in Table 3 and further examined through a qualitative topic analysis. The black line indicates the true number of topics to be predicted. Ideally, a well-performing model should converge to approximately the same predicted number of topics regardless of the choice of $K_{\texttt{init}}$, indicating that it is correctly capturing the coherent semantic structure of the corpus. We observe that the SB-SETM family of models respects this property by consistently predicting ~ 5 topics, regardless of the initialization. In contrast, the baseline StreamETM is highly sensitive to this parameter, predicting up to 400 topics at the last timestep when initialized with $K_{\texttt{init}} = 75$. This behavior stems from the method's intrinsic nature, which, like most topic models in the literature, requires the number of topics to be fixed in advance. Consequently, at each timestep, the model starts with that predefined number of topics, which can only increase over time. This highlights the criticality of the choice of this parameter in an online setting. On the other hand, when the initialization matches the true number of topics (i.e., ~ 3 at each timestep), StreamETM behaves more reliably. However, we emphasize that having such information in advance is highly unrealistic in real-world scenarios, as it would be equivalent to having access to the true labels in a classification task. Finally, from Figure 3, we observe that the COT-based merging strategy consistently yields better performance across all models (indicated by the darker-colored lines). Indeed, while the DOT-based merging strategy acts primarily as a mapping, where the actual merging is performed by summing the embeddings of the associated topics, in the COT case, the merge is achieved directly by transporting the topics of the current timestep into the latent space of the previous one. This reduces the number of distinct topics, as redundant topics are absorbed into a single global representation.

On the quality of estimated topics. Table 1 reports the numerical results for the metric P, which combines the harmonic mean of TC and TD with the predicted number of topics (cf. Section 3). The lower, the better. From a numerical perspective, the SB-**S**ETM family of models behaves similarly to the variant enriched with SBERT, with the COT-based merging strategy yielding a slight improvement. The main contrast emerges with the baseline StreamETM, particularly under the DOT-based merging strategy, where performance is clearly affected by large values of $K_{\texttt{init}}$. Overall, from Table 3 it emerges that the COT-merging strategy yields fewer topics and a higher $\mathrm{H}_{(\texttt{TC},\ \texttt{TD})}$ no matter the underlying model. Clearly, there is a connection between TC, TD, and $K_{\texttt{pred}}$. Fewer topics may reduce TD but often increase TC, and the harmonic mean balances these two effects. It is worth noting, however, that within the SB-**S**ETM family, the difference in the number of predicted topics between COT and DOT is relatively small (less than 2 topics on average). By contrast, the difference in the harmonic mean is about 6 percentage points, indicating that the improvement is due to a genuine gain in topic quality.

Visualization of topic embeddings. Figure 4 shows PCA plots of topic embeddings, using a single projection computed from all embeddings for consistency. Here, we focus on the first and second timesteps and the last two. For the baseline, we use the best-performing configuration from Granese et al. (2025) ($K_{\texttt{init}} = 3$). For our proposed model, we select the variant yielding the best results in Table 1, i.e., the SBERT-enriched version with the COT merging strategy ($K_{\texttt{init}} = 50$). From the plots, we highlight three main observations: *(i)* topic embeddings that appear close in the Euclidean space can in fact be distant in the true latent space of the embeddings, since PCA inevitably distorts distances (e.g., in Figure 4b, wheel/sedan/rear is closer in Euclidean distance to spacecraft/solar/orbit, but COT correctly associates it with you/toyota/dealer); *(ii)* StreamETM often generates topics that reuse almost the same words as in the previous timestep, a direct consequence of the DOT-merging strategy (e.g., dealer/engine/driver appears in both $t = 10$ and $t = 11$, and venus/orbit/spacecraft in both $t = 1$ and $t = 2$); *(iii)* SB-**S**ETM discovers the novel topics related to medicine.

4.2 Results for `UkRuWarNews22-23`

On the topic-thematic association. In this section, we focus our analysis on the best-performing model identified in the previous section, namely SB-**S**ETM with the COT merging strategy ($K_{\texttt{init}} = 50$). The `UkRuWarNews22-23` datasets contain articles in English collected during the Russian-Ukrainian war between 2022 and 2023. In Figure 5, we visualize the reorganized topic-term count matrix, which indicates for each global topic obtained at the end, how many documents of a given category are included. The matrix has been obtained by applying the Spectral Coclustering algorithm (with 5 blocks). Because of dataset imbalance, these counts are rescaled by applying inverse column weights, obtained as the reciprocals of the total counts per category. From a semantic perspective, the associations within the main blocks are meaningful: terms related to culture consistently cluster with categories such as Artists in the war and Boycott of Russian art; legal and judicial vocabulary (e.g., torture, rape, prosecutor) aligns with War crimes and Genocide; and technical-military lexicon (e.g., nuclear, radar, NATO) maps onto Nuclear security and International relations. More transversal categories, such as Propaganda and Information warfare, display broader activation across multiple contexts, reflecting their pervasive role in the discourse rather than methodological noise. At the same time, some themes (e.g., Hacktivism) remain marginal, which likely corresponds to their limited representation in the corpus.

Distribution of the topics over time. Figure 6 shows the distributions of topic frequencies over time. Since the dataset also provides a chronological list of events, we align some of the most notable events of 2023 with the observed peaks in the distributions.

1. January: The United States announced the delivery of 31 Abrams tanks to Ukraine. This coincides with an increase in the blue topic, related to NATO/defense.
2. February: Russia suspended its participation in New START, the last major nuclear arms control treaty with the United States, which limits deployed strategic nuclear warheads and delivery systems. We observe a rise in the green topic (nuclear issues).
3. March: The International Criminal Court issued arrest warrants against Vladimir Putin and Maria Lvova-Belova for war crimes related to the deportation of Ukrainian children. This corresponds to a peak in the brown topic (reports and human rights violations).
4. July: Russia adopted a law severely restricting the right to gender transition for transgender people. Later that month, Odessa was struck by drone attacks that destroyed the Cathedral of the Transfiguration. The events coincide with a rise in the red/pink topic, related to human-centered issues.
5. August: Russia imposed sanctions on several British politicians and journalists in retaliation for UK support for Ukraine. This period aligns

with an increase in the gray topic, connected to journalism.

The plot also reveals that some topics, while highly salient in the early stages of the conflict, gradually lost relevance over time (e.g., the democracy/dream topic). Others emerged only later, such as the green topic related to nuclear issues and China, while some remained persistent throughout the entire period, such as the blue topic connected to defense and emergency.

5 CONCLUSION

This work considered the challenging problem of online topic modeling for document streams. To tackle issues of previously proposed approaches, we proposed an extension of ETM, named SB-SETM. SB-SETM first enables the automatic identification of the appropriate number of active topics in each timestep, by relying on a truncated stick-breaking construction for the topic–per-document distribution. Second, SB-SETM leverages an efficient low-rank formulation of continuous optimal transport in the high-dimensional latent space of the topic embeddings to merge similar topics between two consecutive timesteps or identify new unobserved topics. SB-SETM has been shown to be more accurate both for estimating the number of topics and for providing meaningful topics on simulated and real-world scenarios.

Acknowledgements

This work has been supported by the French government through the 3IA Côte d'Azur Investments, which are managed by the National Research Agency (ANR) with the reference number ANR-23-IACL-0001. This work was performed using HPC resources from GENCI-IDRIS.

References

Amoualian, H., Clausel, M., Gaussier, E., and Amini, M.-R. (2016). Streaming-lda: A copula-based approach to modeling topic dependencies in document streams. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 695–704.

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

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

Bouveyron, C. and Corneli, M. (2025). Scaling optimal transport to high-dimensional gaussian distributions. *HAL preprint n. 04930868*.

Boyd-Graber, J., Hu, Y., Mimno, D., et al. (2017). Applications of topic models. *Foundations and Trends® in Information Retrieval*, 11(2-3):143–296.

Bulatov, V., Alekseev, V., and Vorontsov, K. (2023). Determination of the number of topics intrinsically: Is it possible? In *International Conference on Analysis of Images, Social Networks and Texts*, pages 3–17. Springer.

Cover, T. M. (1999). *Elements of information theory*. John Wiley & Sons.

Dieng, A. B., Ruiz, F. J., and Blei, D. M. (2019). The dynamic embedded topic model. *arXiv preprint arXiv:1907.05545*.

Dieng, A. B., Ruiz, F. J., and Blei, D. M. (2020). Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453.

Granese, F., Navet, B., Villata, S., and Bouveyron, C. (2025). Merging embedded topics with optimal transport for online topic modeling on data streams. In *European Conference on Machine Learning (ECML), Porto, Portugal, Sept. 2025*.

Griffiths, T., Jordan, M., Tenenbaum, J., and Blei, D. (2003). Hierarchical topic models and the nested chinese restaurant process. *Advances in neural information processing systems*, 16.

Grootendorst, M. (2022a). BERTopic Documentation: Online Topic Modeling. `https://maartengr.github.io/BERTopic/getting_started/online/online.html`. "Year" corresponds to the date of the earliest commit in the repository's GitHub history.

Grootendorst, M. (2022b). Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Grootendorst, M. (2023). BERTopic Documentation: Merge Multiple Fitted Models. `https://maartengr.github.io/BERTopic/getting_started/merge/merge.html`. "Year" corresponds to the date of the earliest commit in the repository's GitHub history.

Hoffman, M., Bach, F., and Blei, D. (2010). Online learning for latent dirichlet allocation. *advances in neural information processing systems*, 23.

Kantorovich, L. V. (1942). On the translocation of masses. In *Dokl. Akad. Nauk. USSR (NS)*, volume 37, pages 199–201.

Karakkaparambil James, C., Nagda, M., Haji Ghassemi, N., Kloft, M., and Fellenz, S. (2024). Evaluating dynamic topic models. In Ku, L.-W., Martins,

A., and Srikumar, V., editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 160–176, Bangkok, Thailand. Association for Computational Linguistics.

Kumaraswamy, P. (1980). A generalized probability density function for double-bounded random processes. *Journal of hydrology*, 46(1-2):79–88.

Mimno, D., Wallach, H., Talley, E., Leenders, M., and McCallum, A. (2011). Optimizing semantic coherence in topic models. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 262–272.

Nalisnick, E. and Smyth, P. (2016). Stick-breaking variational autoencoders. *arXiv preprint arXiv:1605.06197*.

Ouyang, J., Wang, T., Cao, J., and Wang, Y. (2024). Applying kumaraswamy distribution on stick-breaking process: a dirichlet neural topic model approach. *Neural Computing and Applications*, 36(22):13731–13744.

Peyré, G., Cuturi, M., et al. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Sverdlik, D. (2025). Topic modeling for business (part 3): Business applications and case studies. `https://xenoss.io/blog/topic-modeling-business-applications-and-case-studies`.

Zhang, D. C. and Lauw, H. (2022). Dynamic topic models for temporal document networks. In *International Conference on Machine Learning*, pages 26281–26292. PMLR.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes, cf. Section 2]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes Appendix B]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [No]
2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes, cf. Section 2]
   (b) Complete proofs of all theoretical results. [Yes, cf. Appendices A.1 and A.2]
   (c) Clear explanations of any assumptions. [Yes, cf. Section 2]
3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes cf. Section 3 and Appendix B]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes, cf. Appendix B]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes cf. Section 3 and Appendix B]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes cf. Appendix B]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator, if your work uses existing assets. [Yes cf. Section 3 and Appendix B]
   (b) The license information of the assets, if applicable. [The model StreamETM and the dataset `20kNewsGroup` are all publicly available. The dataset `UkRuWarNews22-23` is currently under submission, and cannot be referenced to ensure anonymity]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Not Applicable]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Materials

A ADDITIONAL MATERIAL TO SECTION 2

A.1 Proof of Equation 1

We recall from Section 2 the probability associated to a document $\mathbb{W}^{(d)}$ given the topic embeddings $\boldsymbol{\alpha}$ and the word embeddings $\boldsymbol{\rho}$ is

$$p(\mathbb{W}^{(d)} \mid \boldsymbol{\theta}^{(d)}, \boldsymbol{\alpha}, \boldsymbol{\rho}) = \prod_{v \in \mathcal{V}} \sum_{k=1}^{K-1} \theta_k^{(d)} \operatorname{softmax}(\rho_v^\top \alpha_k) = \prod_{v \in \mathcal{V}} \sum_{k=1}^{K-1} \theta_k^{(d)} \beta_{v,k}.$$

Our goal is therefore to maximize the marginal likelihood of the documents:

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\alpha},\boldsymbol{\rho}) &= \sum_{d=1}^{D} \log \int_{\Theta} p(\mathbb{W}^{(d)} \mid \boldsymbol{\theta}^{(d)}, \boldsymbol{\alpha}, \boldsymbol{\rho}) p(\boldsymbol{\theta}^{(d)}) d\boldsymbol{\theta}^{(d)} \\
&= \sum_{d=1}^{D} \log \int_{\Theta} p(\mathbb{W}^{(d)}, \boldsymbol{\theta}^{(d)} \mid \boldsymbol{\alpha}, \boldsymbol{\rho}) d\boldsymbol{\theta}^{(d)} \\
&= \sum_{d=1}^{D} \log \int_{\Theta} \frac{p(\mathbb{W}^{(d)}, \boldsymbol{\theta}^{(d)} \mid \boldsymbol{\alpha}, \boldsymbol{\rho}) q_\phi(\boldsymbol{\theta}^{(d)} \mid \mathbb{W}^{(d)})}{q_\phi(\boldsymbol{\theta}^{(d)} \mid \mathbb{W}^{(d)})} d\boldsymbol{\theta}^{(d)} \\
&= \sum_{d=1}^{D} \log \mathbb{E}_{q_\phi(\boldsymbol{\theta}^{(d)} \mid \mathbb{W}^{(d)})} \left[\frac{p(\mathbb{W}^{(d)}, \boldsymbol{\theta}^{(d)} \mid \boldsymbol{\alpha}, \boldsymbol{\rho})}{q_\phi(\boldsymbol{\theta}^{(d)} \mid \mathbb{W}^{(d)})} \right] \\
&\geq \sum_{d=1}^{D} \mathbb{E}_{q_\phi(\boldsymbol{\theta}^{(d)} \mid \mathbb{W}^{(d)})} \left[\log \frac{p(\mathbb{W}^{(d)}, \boldsymbol{\theta}^{(d)} \mid \boldsymbol{\alpha}, \boldsymbol{\rho})}{q_\phi(\boldsymbol{\theta}^{(d)} \mid \mathbb{W}^{(d)})} \right] \qquad \text{(Jensen inequality)} \\
&= \sum_{d=1}^{D} \mathbb{E}_{q_\phi(\boldsymbol{\theta}^{(d)} \mid \mathbb{W}^{(d)})} \left[\log p(\mathbb{W}^{(d)}, \boldsymbol{\theta}^{(d)} \mid \boldsymbol{\alpha}, \boldsymbol{\rho}) \right] - \mathbb{E}_{q_\phi(\boldsymbol{\theta}^{(d)} \mid \mathbb{W}^{(d)})} \left[\log q_\phi(\boldsymbol{\theta}^{(d)} \mid \mathbb{W}^{(d)}) \right] \\
&= \sum_{d=1}^{D} \mathbb{E}_{q_\phi(\boldsymbol{\theta}^{(d)} \mid \mathbb{W}^{(d)})} \left[\log p(\mathbb{W}^{(d)}, \boldsymbol{\theta}^{(d)} \mid \boldsymbol{\alpha}, \boldsymbol{\rho}) \right] + \mathcal{H}\left(q_\phi(\boldsymbol{\theta}^{(d)} \mid \mathbb{W}^{(d)}) \right) \\
&= \sum_{d=1}^{D} \mathbb{E}_{q_\phi(\boldsymbol{\theta}^{(d)} \mid \mathbb{W}^{(d)})} \left[\log p(\mathbb{W}^{(d)} \mid \boldsymbol{\theta}^{(d)}, \boldsymbol{\alpha}, \boldsymbol{\rho}) \right] + \mathbb{E}_{q_\phi(\boldsymbol{\theta}^{(d)} \mid \mathbb{W}^{(d)})} \left[\log p(\boldsymbol{\theta}^{(d)}) \right] + \mathcal{H}\left(q_\phi(\boldsymbol{\theta}^{(d)} \mid \mathbb{W}^{(d)}) \right) \\
&= \sum_{d=1}^{D} \mathbb{E}_{q_\phi(\boldsymbol{\theta}^{(d)} \mid \mathbb{W}^{(d)})} \left[\log p(\mathbb{W}^{(d)} \mid \boldsymbol{\theta}^{(d)}, \boldsymbol{\alpha}, \boldsymbol{\rho}) \right] - \mathrm{KL}\left(q_\phi(\boldsymbol{\theta}^{(d)} \mid \mathbb{W}^{(d)}) \,||\, p(\boldsymbol{\theta}^{(d)}) \right),
\end{aligned}$$

where $\mathcal{H}(\cdot)$ represents the entropy and $\mathrm{KL}(\cdot)$ is the Kullback-Leibler divergence.

A.2 Proof of Equation 2

We first recall the stick-breaking transformation $f_{\mathrm{SB}} : (0,1)^{K-1} \to \Delta^{K-1}$, which deterministically maps a sequence of auxiliary variables $\boldsymbol{\nu}^{(d)} = (\nu_1^{(d)}, \ldots, \nu_{K-1}^{(d)}) \in (0,1)^{K-1}$ into a valid topic proportion vector $\boldsymbol{\theta}^{(d)} \in \Delta^{K-1}$. Intuitively, each $\nu_k^{(d)}$ specifies the fraction of the remaining stick that is assigned to the k-th component, while the last component $\theta_K^{(d)}$ collects the residual mass. Hence,

$$\boldsymbol{\theta}^{(d)} = f_{\mathrm{SB}}(\boldsymbol{\nu}^{(d)}).$$

The latent variable $\mathbf{z}^{(d)}$ does not directly enter the transformation, but it controls the distribution of $\boldsymbol{\nu}^{(d)}$ through a conditional variational distribution. Concretely, we define

$$\mathbf{z}^{(d)} \sim \pi_\phi(\mathbf{z}^{(d)} \mid \mathbb{W}^{(d)}) \equiv \mathcal{N}\big(\mathbf{z}^{(d)}; \mu_\phi(\mathbb{W}^{(d)}), \sigma^2_\phi(\mathbb{W}^{(d)})\big),$$
$$\boldsymbol{\nu}^{(d)} \sim \pi_\psi(\boldsymbol{\nu}^{(d)} \mid \mathbf{z}^{(d)}) \equiv \prod_{k=1}^{K-1} \text{Kumaraswamy}\Big(\nu_k^{(d)}; a_\psi(\mathbf{z}^{(d)}), b_\psi(\mathbf{z}^{(d)})\Big).$$

Since $\boldsymbol{\theta}^{(d)}$ is a deterministic function of $\boldsymbol{\nu}^{(d)}$, by the data-processing inequality (Cover, 1999), it follows that

$$\text{KL}\Big(q_\phi(\boldsymbol{\theta}^{(d)} \mid \mathbb{W}^{(d)}) \,\|\, p(\boldsymbol{\theta}^{(d)})\Big) \;\leq\; \text{KL}\Big(\pi_{(\phi,\psi)}(\mathbf{z}^{(d)}, \boldsymbol{\nu}^{(d)} \mid \mathbb{W}^{(d)}) \,\|\, p(\mathbf{z}^{(d)}, \boldsymbol{\nu}^{(d)})\Big),$$

where the joint variational distribution factorizes as $\pi_{(\phi,\psi)}(\mathbf{z}^{(d)}, \boldsymbol{\nu}^{(d)} \mid \mathbb{W}^{(d)}) = \pi_\phi(\mathbf{z}^{(d)} \mid \mathbb{W}^{(d)})\, \pi_\psi(\boldsymbol{\nu}^{(d)} \mid \mathbf{z}^{(d)})$, and the prior as $p(\mathbf{z}^{(d)}, \boldsymbol{\nu}^{(d)}) = p(\mathbf{z}^{(d)})\, p(\boldsymbol{\nu}^{(d)})$.

Applying the chain rule of the KL divergence (Cover, 1999), we obtain

$$\begin{aligned}
&\text{KL}\left(q_\phi(\boldsymbol{\theta}^{(d)} \mid \mathbb{W}^{(d)}) \,||\, p(\boldsymbol{\theta}^{(d)})\right) \\
&\leq \text{KL}\left(\pi_{(\phi,\psi)}(\mathbf{z}^{(d)}, \boldsymbol{\nu}^{(d)} \mid \mathbb{W}^{(d)}) \,||\, p(\mathbf{z}^{(d)}, \boldsymbol{\nu}^{(d)})\right) \\
&= \mathbb{E}_{\pi_{(\phi,\psi)}}\left[\log \frac{\pi_{(\phi,\psi)}(\mathbf{z}^{(d)}, \boldsymbol{\nu}^{(d)} \mid \mathbb{W}^{(d)})}{p(\mathbf{z}^{(d)}) p(\boldsymbol{\nu}^{(d)})}\right] \\
&= \mathbb{E}_{\pi_{(\phi,\psi)}}\left[\log \frac{\pi_\phi(\mathbf{z}^{(d)} \mid \mathbb{W}^{(d)}) \pi_\psi(\boldsymbol{\nu}^{(d)} \mid \mathbf{z}^{(d)})}{p(\mathbf{z}^{(d)}) p(\boldsymbol{\nu}^{(d)})}\right] \\
&= \mathbb{E}_{\pi_{(\phi,\psi)}}\left[\log \frac{\pi_\phi(\mathbf{z}^{(d)} \mid \mathbb{W}^{(d)})}{p(\mathbf{z}^{(d)})} + \log \frac{\pi_\psi(\boldsymbol{\nu}^{(d)} \mid \mathbf{z}^{(d)})}{p(\boldsymbol{\nu}^{(d)})}\right] \\
&= \mathbb{E}_{\pi_\phi}\left[\log \frac{\pi_\phi(\mathbf{z}^{(d)} \mid \mathbb{W}^{(d)})}{p(\mathbf{z}^{(d)})}\right] + \mathbb{E}_{\pi_\phi}\left[\mathbb{E}_{\pi_\psi}\left[\log \frac{\pi_\psi(\boldsymbol{\nu}^{(d)} \mid \mathbf{z}^{(d)})}{p(\boldsymbol{\nu}^{(d)})}\right]\right] \\
&= \text{KL}\left(\pi_\phi(\mathbf{z}^{(d)} \mid \mathbb{W}^{(d)}) \,||\, p(\mathbf{z}^{(d)})\right) + \mathbb{E}_{\pi_\phi}\left[\text{KL}\left(\pi_\psi(\boldsymbol{\nu}^{(d)} \mid \mathbf{z}^{(d)}) \,||\, p(\boldsymbol{\nu}^{(d)})\right)\right] \\
&= \underbrace{\text{KL}\left(\pi_\phi(\mathbf{z}^{(d)} \mid \mathbb{W}^{(d)}) \,||\, \mathcal{N}(0, I)\right)}_{\text{KL(Gaussian||Normal)}} + \mathbb{E}_{\pi_\phi}\left[\underbrace{\text{KL}\left(\pi_\psi(\boldsymbol{\nu}^{(d)} \mid \mathbf{z}^{(d)}) \,||\, \text{Beta}(a, b)\right)}_{\text{KL(Kumaraswamy||Beta)}}\right].
\end{aligned}$$

The KL between a Kumaraswamy and a Beta distribution has a closed form (Nalisnick and Smyth, 2016). The outer expectation over π_ϕ is estimated by Monte Carlo sampling. Finally, plugging these components into the evidence lower bound, the used loss takes the form

$$\begin{aligned}
\mathcal{L}(\boldsymbol{\alpha}, \boldsymbol{\rho}) &\geq \widehat{\mathcal{L}}(\boldsymbol{\alpha}, \boldsymbol{\rho}) \\
&\equiv \sum_{d=1}^{D} \mathbb{E}_{q_\phi(\boldsymbol{\theta}^{(d)} \mid \mathbb{W}^{(d)})}\left[\log p(\mathbb{W}^{(d)} \mid \boldsymbol{\theta}^{(d)}, \boldsymbol{\alpha}, \boldsymbol{\rho})\right] - \text{KL}\left(q_\phi(\boldsymbol{\theta}^{(d)} \mid \mathbb{W}^{(d)}) \,||\, p(\boldsymbol{\theta}^{(d)})\right) \\
&\geq \sum_{d=1}^{D} \mathbb{E}_{q_\phi(\boldsymbol{\theta}^{(d)} \mid \mathbb{W}^{(d)})}\left[\log p(\mathbb{W}^{(d)} \mid \boldsymbol{\theta}^{(d)}, \boldsymbol{\alpha}, \boldsymbol{\rho})\right] - \text{KL}\left(\pi_\phi(\mathbf{z}^{(d)} \mid \mathbb{W}^{(d)}) \,||\, \mathcal{N}(0, I)\right) \\
&\quad - \mathbb{E}_{\pi_\phi}\left[\text{KL}\left(\pi_\psi(\boldsymbol{\nu}^{(d)} \mid \mathbf{z}^{(d)}) \,||\, \text{Beta}(a, b)\right)\right].
\end{aligned}$$

A.3 Learning Algorithm

We present in Algorithm 1 the pseudocode of the learning algorithm at time step t for the SB-ETM model. In our experiments, we set $s = 1$.

Algorithm 1 Topic modeling at time t with stick-breaking variational inference

Initialize model and variational parameters
for iteration $i = 1, 2, \ldots$ **do**
  Compute $\beta_k = \text{softmax}(\boldsymbol{\rho}^\top \alpha_k)$ for each topic k
  Choose a minibatch $\mathcal{B}$ of documents
  for all each document $\mathbb{W}^{(d)} \in \mathcal{B}$ **do**
    Get normalized bag-of-word $\mathbf{W}^{(d)}$
    Compute $\boldsymbol{\mu}^{(d)} = NN(\mathbf{W}^{(d)}; \phi_\mu)$
    Compute $\boldsymbol{\Sigma}^{(d)} = NN(\mathbf{W}^{(d)}; \phi_\Sigma)$
    for iteration $s = 1, \ldots, S$ **do**
      $\epsilon \sim \mathcal{N}(0, I)$
      $\mathbf{z}^{(d)} \leftarrow \boldsymbol{\mu}^{(d)} + \boldsymbol{\sigma}^{(d)} \odot \epsilon$
      Compute $a^{(d)} = NN(\mathbf{z}^{(d)}; \psi_a)$
      Compute $b^{(d)} = NN(\mathbf{z}^{(d)}; \psi_b)$
      for each topic $k \in \{1, \ldots, K-1\}$ **do**
        $u_k^{(d)} \sim \mathcal{U}(0, 1)$
        $\nu_k^{(d)} \leftarrow \left(1 - (1 - u_k^{(d)})^{1/b_k^{(d)}}\right)^{1/a_k^{(d)}}$
      end for
      Compute $\theta_1^{(d)} \leftarrow \nu_1^{(d)}$
      for iteration $k = 2, \ldots, K-1$ **do**
        Compute $\theta_k^{(d)} \leftarrow \nu_k^{(d)} \cdot \prod_{j=1}^{k-1}(1 - \nu_j^{(d)})$
      end for
      Compute $\theta_K^{(d)} \leftarrow \prod_{j=1}^{K-1}(1 - \nu_j^{(d)})$
      Compute $\widehat{\mathbf{W}}_s^{(d)} \leftarrow \text{softmax}(\boldsymbol{\theta}^{(d)} \boldsymbol{\beta}^\top)$
    end for
    Compute $\widehat{\mathbf{W}}^{(d)} \leftarrow \frac{1}{S} \sum_s \widehat{\mathbf{W}}_s^{(d)}$
    Estimate the ELBO in eq. (2) and its gradient (backpropagation)
    Update model parameters $\boldsymbol{\alpha}$
    Update variational parameters $(\phi_\mu, \phi_\Sigma, \psi_a, \psi_b)$
  end for
end for

B ADDITIONAL MATERIAL TO SECTION 3

B.1 Datasets

`20kNewsGroupOnline`. Figure 10a provides the overall picture of the topic frequencies over time as constructed in (Granese et al., 2025). Following their strategy, we randomly draw 8 times approximately 5k samples from the total datasets. We partition the 5k samples into 500 sample batches to simulate an 11-step scenario.

`UkRuWarNews22-23`. The corpus comprises 4.211 news articles on the Russia–Ukraine war from 2022 to 2023. We focus on the 2.430 articles written in English, which span 39 thematic categories The dataset comprises 840 articles in 2022 and 1590 in 2023. For each dataset entry, the article URL, publication (or last-modified) date, and cleaned article text are included. All articles were manually reviewed to correct errors and fill missing fields, and the text was thoroughly cleaned to remove non-content elements (e.g., advertisements, newsletter prompts, image captions, alternative text, and cookie banners). The average number of tokens per article is ∼426 (max 4038, min 7). The average number of characters per article is ∼3.125 (max 31.657, min 46). The English corpus covers 113 news outlets, with president.gov.ua, Reuters, AP News, BBC, The Guardian, and CNN Edition among the most represented sources. The articles have been categorised by three experts in political and communication sciences. The reference to the dataset will be provided upon acceptance to ensure anonymity.

Text preprocessing. We process the articles with a uniform pipeline.

- Lemmatization: documents are lemmatized with `spaCy` large models (`en_core_web_lg`), disabling

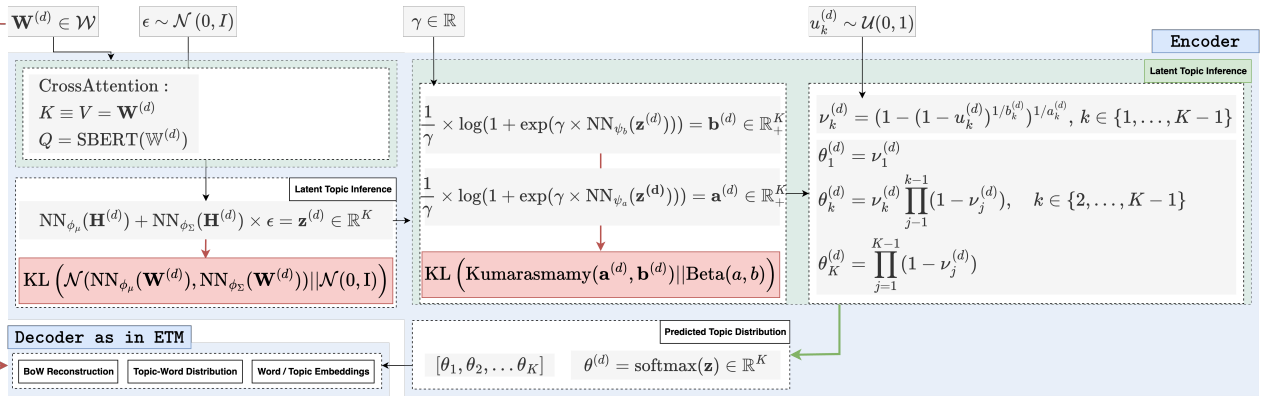


Figure 7: **Overall structure of SB-ETM at timestep t.** Green components denote the additional modules introduced beyond the original StreamETM/ETM design.

parser/NER to retain only lemmatization.

- Normalization and tokenization: we lowercase, replace apostrophes with spaces, strip punctuation, and tokenize with NLTK's `word_tokenize`.
- Token filtering: we keep alphabetic tokens of length > 2 and remove language-specific stopwords (NLTK lists), augmenting the English list with a small domain set {`notoc`, `coxnet`}.
- Vocabulary pruning: we remove tokens with frequency ≤ 1 and drop extremely frequent tokens using a document-frequency threshold (appear in $> 70\%$ of documents).

The preprocessing pipeline is applied *independently at each timestep* t: pruning of low- and high-frequency tokens is computed within the current batch only, not over the full corpus. The same procedure is used for both `20kNewsGroupOnline` and `UkRuWarNews22-23`. For `20kNewsGroupOnline`, we additionally strip e-mail headers (contact fields) and subject lines before processing.

B.2 Model Architecture and Training Procedure

We extend the StreamETM architecture, which builds upon the Embedded Topic Model, by integrating two additional modules. The base architecture consists of two embedding matrices: a `word embedding matrix`, initialized from pre-trained `GloVe` vectors of dimension 300 and kept fixed during training, and a `topic embedding matrix` (`Xavier` initialization), of the same dimensionality, which remains fully trainable. These matrices form a shared latent space connecting topics and words.

The encoder block of the main body architecture consists of a feedforward bag-of-words network that maps each normalized document vector into a dense hidden representation, combined with a Cross-Attention Encoder (cf. Sentence-BERT module) that integrates sentence-level representations extracted from a multilingual Sentence-BERT model (`paraphrase-multilingual-mpnet-base-v2`). The fused embedding is normalized through LayerNorm and combined with the BoW representation via a residual connection. Two fully connected latent heads ($\text{NN}_{\phi_\mu}(\cdot)$, $\text{NN}_{\phi_\Sigma}(\cdot)$) produce the mean and log-variance of the Gaussian posterior, each followed by Batch Normalization for stability. The reparameterized latent sample is then processed by a Stick-Breaking Encoder (cf. Stick-Breaking module), which maps it into document–topic proportions. The decoder reconstructs the document's word distribution by projecting topic proportions onto the shared embedding space, followed by Batch Normalization and a softmax layer.

Optimization uses Adam with an initial learning rate of 0.01 (One-Cycle schedule, linearly increasing during the first 10% of the total steps and then decaying following a cosine annealing strategy), weight decay of 0.006, batch size of 1024, and a scheduler with a step size of 5000, over 2600 epochs. All experiments were run on NVIDIA H100 GPUs.

Sentence-BERT (SBERT) module. The Cross-Attention Encoder fuses lexical and contextual information by aligning bag-of-words and sentence-level embeddings in a shared latent space. It consists of three components:

Table 2: **Training time for each method considered in the paper**. The reported values correspond to the average training time per timestep; therefore, the total training time for a complete execution should be multiplied by 11 timesteps for `20kNewsGroupOnline`. The training time of SB-**S**ETM with the COT-based merging strategy and $K_{\texttt{init}} = 50$ on `UkRuWarNews22-23` was 9.39 ± 1.65 minutes per timestep (to be multiplied by 22, the number of timesteps). We mark with a $*$ the cases where additional computational time was incurred due to server-related overhead.

Model	SBERT	OT	$K_{\texttt{init}}$	Time (minutes)
SB-**S**ETM	✓	Continuous	15	15.11 ± 3.47
SB-**S**ETM	✓	Continuous	25	15.04 ± 3.54
SB-**S**ETM	✓	Continuous	50	18.42* ± 4.31
SB-**S**ETM	✓	Continuous	75	15.47 ± 3.69
SB-**S**ETM	✓	Discrete	15	15.27 ± 3.48
SB-**S**ETM	✓	Discrete	25	16.44 ± 3.57
SB-**S**ETM	✓	Discrete	50	15.33 ± 3.65
SB-**S**ETM	✓	Discrete	75	16.58 ± 3.65
SB-**S**ETM	×	Continuous	15	6.31 ± 0.79
SB-**S**ETM	×	Continuous	25	7.38 ± 1.11
SB-**S**ETM	×	Continuous	50	7.19 ± 1.1
SB-**S**ETM	×	Continuous	75	6.75 ± 1.0
SB-**S**ETM	×	Discrete	15	6.52 ± 0.99
SB-**S**ETM	×	Discrete	25	6.25 ± 0.93
SB-**S**ETM	×	Discrete	50	7.06 ± 1.03
SB-**S**ETM	×	Discrete	75	6.58 ± 0.98
StreamETM	×	Continuous	3	0.85 ± 0.1
StreamETM	×	Continuous	15	0.85 ± 0.1
StreamETM	×	Continuous	25	0.86 ± 0.1
StreamETM	×	Continuous	50	0.94 ± 0.11
StreamETM	×	Continuous	75	0.91 ± 0.11
StreamETM	×	Discrete	3	**0.88** ± 0.1
StreamETM	×	Discrete	15	5.21* ± 20.22
StreamETM	×	Discrete	25	0.91 ± 0.1
StreamETM	×	Discrete	50	0.9 ± 0.12
StreamETM	×	Discrete	75	0.93 ± 0.12

- Projection layers: two linear mappings project the BoW vector and the SBERT embeddings into the same latent dimensionality.
- Multi-head attention: a single-head MultiheadAttention module (batch-first) takes the SBERT embeddings as queries and the projected BoW representation as both keys and values.
- Aggregation: the attention matrix is passed through a LayerNorm operation, producing the final document embedding. This embedding is then added to the original BoW embedding computed by the feedforward bag-of-words network.

Stick-Breaking module. The Stick-Breaking Encoder converts the latent Gaussian variables into normalized document–topic proportions through a differentiable stick-breaking process parameterized by the Kumaraswamy distribution. The module comprises two parallel linear layers, each mapping the latent vector to per-topic shape parameters ($\text{NN}_{\psi_a}(\cdot)$, $\text{NN}_{\psi_b}(\cdot)$). The outputs are activated through softplus and stabilized by a small additive constant. Stick variables are sampled via a deterministic Kumaraswamy sampler using clamped uniform noise to ensure numerical robustness. Topic weights are then obtained through cumulative products over the remaining stick lengths, resulting in a simplex-valued vector. We apply a warm-starting from the model at the previous timestep. Overlapping weights are copied, while new parameters are initialized through Xavier initialization; bias terms for newly added topics are adjusted to emphasize their activation.

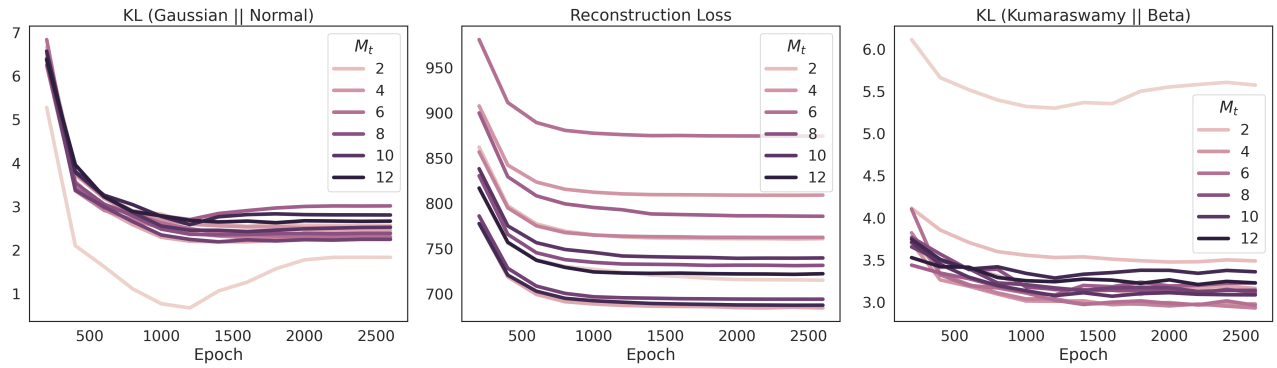
Table 3: **Ablation study.**

Model	SBERT	OT	$\mathrm{H}_{(\mathrm{TC,TD})}$	$K_{\texttt{pred}}$	$K_{\texttt{init}}$	$K_{\texttt{real}}$
SB-**S**ETM	✓	Continuous	0.88 ± 0.04	4.54 ± 0.68	15	3.18 ± 0.39
SB-**S**ETM	✓	Continuous	0.87 ± 0.04	4.53 ± 0.58	25	3.18 ± 0.39
SB-**S**ETM	✓	Continuous	0.88 ± 0.03	4.64 ± 0.56	50	3.18 ± 0.39
SB-**S**ETM	✓	Continuous	0.88 ± 0.04	4.51 ± 0.56	75	3.18 ± 0.39
SB-**S**ETM	✓	Discrete	0.86 ± 0.06	5.57 ± 1.97	15	3.18 ± 0.39
SB-**S**ETM	✓	Discrete	0.87 ± 0.06	5.71 ± 2.29	25	3.18 ± 0.39
SB-**S**ETM	✓	Discrete	0.84 ± 0.07	5.42 ± 1.62	50	3.18 ± 0.39
SB-**S**ETM	✓	Discrete	0.84 ± 0.08	5.53 ± 1.97	75	3.18 ± 0.39
SB-**S**ETM	×	Continuous	0.88 ± 0.03	4.55 ± 0.6	15	3.18 ± 0.39
SB-**S**ETM	×	Continuous	0.87 ± 0.03	4.24 ± 0.43	25	3.18 ± 0.39
SB-**S**ETM	×	Continuous	0.89 ± 0.03	3.89 ± 0.35	50	3.18 ± 0.39
SB-**S**ETM	×	Continuous	0.88 ± 0.03	4.29 ± 0.69	75	3.18 ± 0.39
SB-**S**ETM	×	Discrete	0.86 ± 0.04	5.6 ± 1.31	15	3.18 ± 0.39
SB-**S**ETM	×	Discrete	0.88 ± 0.04	5.2 ± 1.79	25	3.18 ± 0.39
SB-**S**ETM	×	Discrete	0.84 ± 0.05	5.67 ± 1.71	50	3.18 ± 0.39
SB-**S**ETM	×	Discrete	0.83 ± 0.07	5.66 ± 1.67	75	3.18 ± 0.39
StreamETM	×	Continuous	0.9 ± 0.04	2.02 ± 0.14	3	3.18 ± 0.39
StreamETM	×	Continuous	0.7 ± 0.13	8.39 ± 0.7	15	3.18 ± 0.39
StreamETM	×	Continuous	0.42 ± 0.14	14.22 ± 1.14	25	3.18 ± 0.39
StreamETM	×	Continuous	0.35 ± 0.28	28.0 ± 2.29	50	3.18 ± 0.39
StreamETM	×	Continuous	0.25 ± 0.26	42.8 ± 3.84	75	3.18 ± 0.39
StreamETM	×	Discrete	0.93 ± 0.04	2.01 ± 0.1	3	3.18 ± 0.39
StreamETM	×	Discrete	0.65 ± 0.1	12.2 ± 2.22	15	3.18 ± 0.39
StreamETM	×	Discrete	0.3 ± 0.08	45.59 ± 10.3	25	3.18 ± 0.39
StreamETM	×	Discrete	0.05 ± 0.02	196.61 ± 83.19	50	3.18 ± 0.39
StreamETM	×	Discrete	0.03 ± 0.01	281.21 ± 85.52	75	3.18 ± 0.39

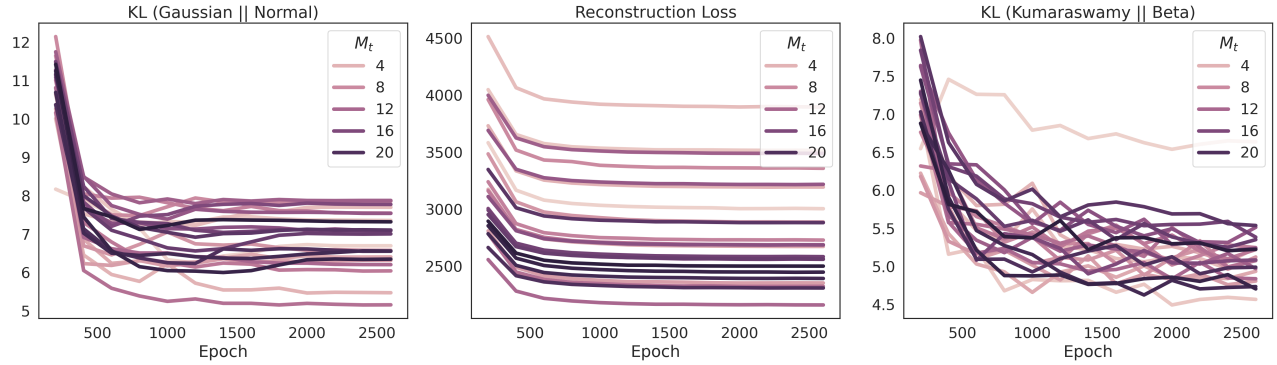
We provide in Figure 7 an overview of the SB-ETM model at time t.

$\omega_R, \omega_G, \omega_S$ and the Beta distribution parameter during the training. We observe from Table 5 that the best-performing configurations correspond to cases where the value of ω_S is small compared to ω_G, or more generally, when $\omega_S < 0.1$. Regarding the Beta parameterization, the most coherent results are obtained when $a = b = 0.5$, i.e., when the Beta distribution takes a U-shaped form. This behaviour aligns well with our stick-breaking formulation: at the beginning of the process, a higher probability mass is assigned to the first topics, while the model is progressively encouraged to allocate more weight to the newly introduced ones. Based on these findings, for all the experiments presented in this paper, we adopt the configuration highlighted in bold in the table. Moreover, we provide in Figure 8 the evolution of the 3 loss components during the training for each timestep on `20kNewsGroupOnline`. Reconstruction loss decreases and stabilizes near 700–750, while both KL terms converge to moderate values (≈2–3). This range reflects balanced regularization: the variational posteriors remain close to their priors, preventing collapse and preserving informative latent topics. Note that the reconstruction loss values are relatively high because they represent the sum of token-level log-likelihoods per document; they therefore scale with document length and corpus size, not with model inefficiency. This difference in scale is intrinsic to the ELBO formulation. A similar behaviour is also observed in `UkRuWarNews22-23`.

Training time analysis. Table 2 reports the training time for each method considered in the paper. We observe that the SBERT module is the most computationally expensive component, as its inclusion approximately doubles the training time per timestep. Notably, neither $K_{\texttt{init}}$ nor the COT-based merging strategy introduces a significant computational overhead. In contrast, the Stick-Breaking module increases the training time by approximately six minutes compared to the original model (StreamETM - Discrete - $K_{\texttt{init}} = 3$).



(a) `20kNewsGroupOnline`.



(b) `UkRuWarNews22-23`.

Figure 8: **Losses evolution during training**. SB-**S**ETM (COT-based merging strategy and $K_{\texttt{init}} = 50$).

Table 5: Quantitative results on `20kNewsGroupOnline` when SB-**S**ETM (COT-based merging strategy and $K_{\texttt{init}} = 50$) has been trained using different parameter values.

ω_R	ω_G	ω_S	a	b	$\mathrm{H}_{(\mathrm{TC},\ \mathrm{TD})}$	$K_{\texttt{pred}}$	$K_{\texttt{init}}$	$K_{\texttt{real}}$
1	**1**	**0.05**	**0.5**	**0.5**	0.97 ± 0.04	4.64 ± 0.57	50	3.18 ± 0.39
1	1	0.05	1	2	0.98 ± 0.05	4.82 ± 0.75	50	3.18 ± 0.39
1	1	0.05	1	1	0.94 ± 0.08	5.0 ± 0.77	50	3.18 ± 0.39
1	1	0.05	1	0.5	0.94 ± 0.08	5.0 ± 0.77	50	3.18 ± 0.39
1	1	0.05	2	2	0.95 ± 0.06	4.0 ± 0.0	50	3.18 ± 0.39
1	0.5	0.1	0.5	0.5	0.92 ± 0.05	5.36 ± 1.21	50	3.18 ± 0.39
1	1.5	0.01	0.5	0.5	0.97 ± 0.03	5.0 ± 0.0	50	3.18 ± 0.39
1	0.1	0.2	0.5	0.5	0.95 ± 0.05	4.91 ± 0.3	50	3.18 ± 0.39
1	0.5	0.005	0.5	0.5	0.98 ± 0.03	5.91 ± 0.3	50	3.18 ± 0.39

C ADDITIONAL MATERIAL TO SECTION 4

C.1 Tracing the Topic Evolution

The COT merging approach does not allow us to directly track the evolution of topics over time, as topics are directly projected into embedding spaces. We provide in Algorithm 2 the pseudo code of the algorithm we used to trace the topic evolution over time. To compute the transport map in the algorithm, we use the Cosine Distance for the cost map. The transport map is computed using the Python function `ot.unbalanced.mm_unbalanced`, with KL divergence and marginal relaxation at 0.09. In the experiments, we set $\varepsilon = 0.01$.

Algorithm 2 Association between of topic embeddings $\boldsymbol{\alpha}^{(t-1)}$ and $\boldsymbol{\alpha}^{(t)}$ at consecutive timesteps via transport-based matching at threshold ε

Compute $X \leftarrow \begin{bmatrix} \boldsymbol{\alpha}^{(t-1)} \\ \boldsymbol{\alpha}^{(t)} \end{bmatrix}$
Compute $\Sigma \leftarrow \mathrm{Cov}(X) + \epsilon_{\text{ridge}} \cdot \mathbf{I}_d$ ▷ where $d = K_{\texttt{pred}}^{(t-1)} + K_{\texttt{pred}}^{(t)}$ and $\epsilon_{\text{ridge}} = 1e-6$
Let $\lambda_1 \leq \cdots \leq \lambda_d$ be the eigenvalues of Σ; set $\lambda \leftarrow \lambda_d$
Compute threshold $T \leftarrow \sqrt{\lambda} \cdot \varepsilon$
Compute transport matrix $\mathrm{OT} \leftarrow \texttt{UnbalancedOT}(\boldsymbol{\alpha}^{(t)}, \boldsymbol{\alpha}^{(t-1)})$ ▷ $\mathrm{OT} \in \mathbb{R}^{K^{(t)} \times K^{(t-1)}}$
for each topic $i \in \{1, \ldots, K_{\texttt{pred}}^{(t)}\}$ **do**
    $w_i \leftarrow \max_j \mathrm{OT}[i, j]$
    $\mathrm{index}_i \leftarrow \arg\max_j \mathrm{OT}[i, j]$
    if $w_i \geq T$ **then**
        Assign topic i to class index_i
    else
        Flag topic i as a new topic
    end if
end for

C.2 Ablation Study

In Table 3, we report the complete set of quantitative results, evaluating the performance of the proposed model across all possible configurations—namely, with and without the SBERT module, and using both COT- and DOT-based merging strategies. The main trend that emerges concerns the DOT-based merging strategy, which generally leads to lower $\mathrm{H}_{(\mathrm{TC},\ \mathrm{TD})}$ values compared to the COT-based strategy. As discussed earlier, this result is not surprising, since the COT strategy explicitly projects the topic embeddings at time t into the latent space at time $t-1$. When comparing versions with and without the SBERT module, we do not observe a clear quantitative advantage from its inclusion. This outcome is expected, as TC and TD are not directly sensitive to topic semantics: even if topics become semantically more coherent, these metrics only measure whether words co-occur within a given context window in the text. For this reason, we complement this analysis with a qualitative evaluation in the next section.

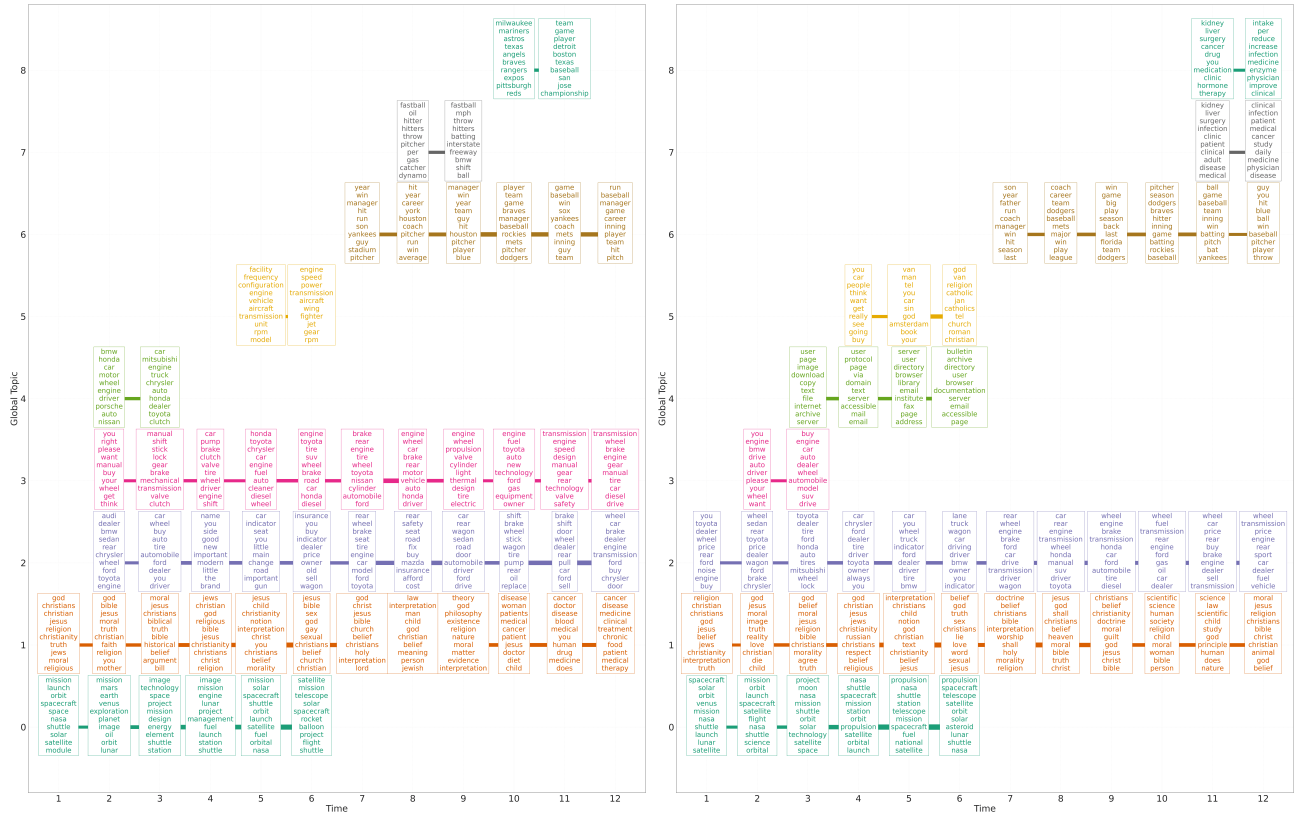
C.3 Qualitative Results

Topic frequency distributions on `20kNewsGroupOnline`. Figure 10 shows the topic frequency distributions for SB-**S**ETM (SBERT-enabled, COT-based merging, and $K_{\texttt{init}} = 50$) and for StreamETM in its best-performing configuration (i.e., as in the seminal paper by Granese et al. (2025), with DOT-based merging and $K_{\texttt{init}} = 3$). Ideally, a well-functioning model should be able to reproduce the distribution in Figure 10a. We observe that while both methods correctly capture the topics related to `science, space`, `autos`, and `talk, religion` up to timestep 7, StreamETM begins to confuse topics afterward, as the `science, space` category, in the ground truth reference, disappears, leaving room for the `sport, baseball` topic. This behavior, however, is correctly handled by our model, which also succeeds at timestep 10 in identifying the topic related to `science, medicine`, a topic that is instead absorbed into others by StreamETM.

SBERT module as a semantic regularizer. Figure 9 compares the models trained with and without the SBERT module, both using the COT-based merging strategy. For each timestep, we plot the top words describing each topic and apply Algorithm 2 to link the most similar topics over time.

From the plots, we observe that *(i)* at timestep 10, in the model without SBERT, the topic related to religion (the orange one, index 1) gradually drifts in meaning and eventually becomes associated with medicine-related terms. In contrast, in the SBERT-enabled model, the religious topic remains stable across time, while a distinct topic emerges for medicine (indexes 7 and 8). *(ii)* The model without SBERT also produces two parallel topics about automobiles—described by different sets of words but semantically overlapping—whereas this redundancy does not occur in the SBERT-based model.

Overall, these differences highlight the role of the SBERT module in enhancing the semantic consistency and



(a) SB-**S**ETM without SBERT module (COT-merging)

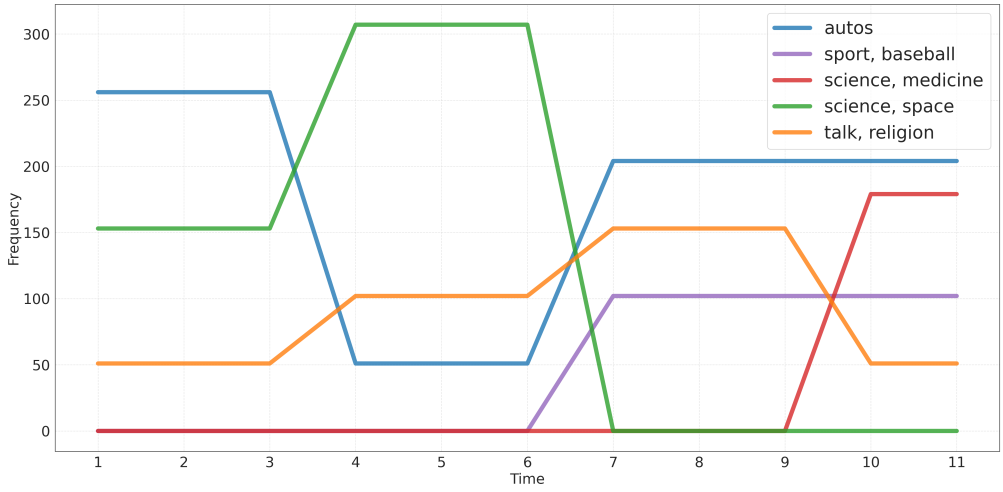
(b) SB-**S**ETM with SBERT module (COT-merging)

Figure 9: **Qualitative results on `20kNewsGroupOnline`**. Both SB-**S**ETM models were trained under identical conditions with $K_{\texttt{init}} = 50$, sharing the same seed and document sets. The line thickness between top-word boxes indicates the similarity between topics according to the tracing strategy.

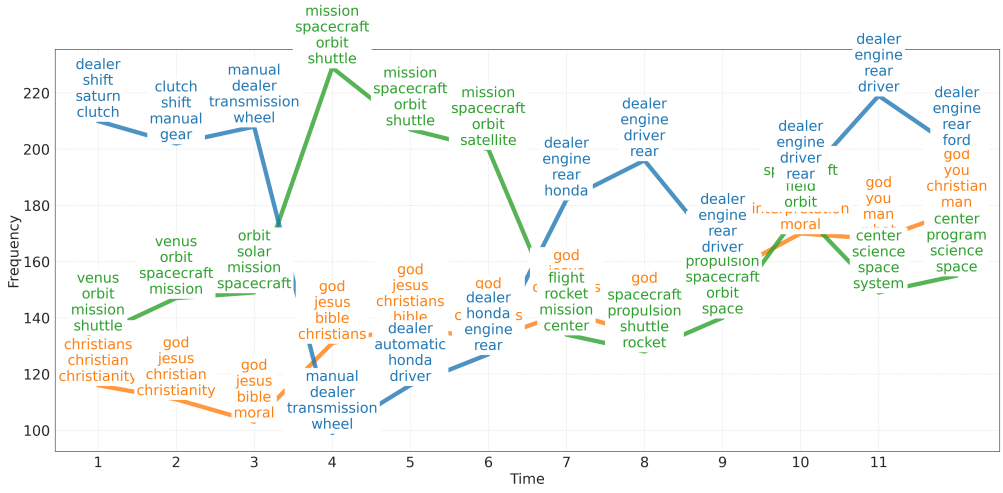
temporal coherence of topics. By grounding the topic representations in a semantic embedding space, SBERT reduces noisy drift and prevents concept splitting or merging that occurs solely due to lexical variation. The resulting trajectories are smoother and more interpretable, with topics preserving their meaning over time while allowing the emergence of genuinely new semantic areas (e.g., the separation of religion and medicine).

Table 4: **Thematic categories in** `UkRuWarNews22-23`.

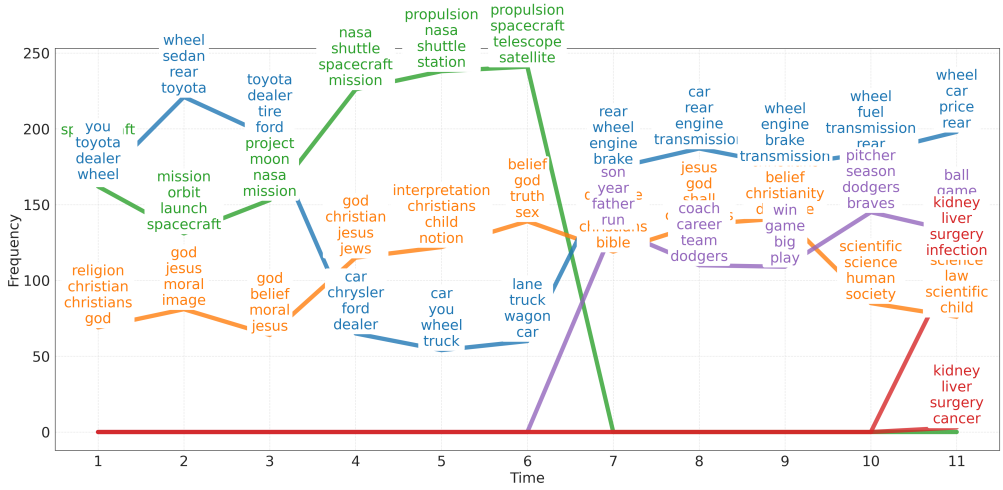
Code	Description	N. Samples
IDT	Identity, nationality and culture	33
CDG	War crimes (including population displacements, indiscriminate bombings, treatment of bodies, civilian hostages, etc.)	149
ENQU	Investigations into war crimes	55
GEN	Massacre / genocide	60
REEC	Rewriting of history (including the issue of Russian territory and the question of Nazism)	52
TERR	Terrorism	28
PRO	Propaganda	49
NUC	Nuclear security / management of energy resources	60
OCC	Western aid	50
MIGR	Migration crisis and "quality" of migrants	60
OTAN	NATO enlargement	57
POL	Positioning of politicians	39
PPU	Positioning of public figures	18
ONG	Positioning and actions of NGOs	27
PAIX	Question of the conditions of peace	65
INFO	Information warfare	41
REL	International relations	26
HUMA	Humanitarian aid	34
FUIT	Brain drain and artists' flight	20
ART	Artists in the war	35
BOYC	Question of the boycott of Russian art	8
BOY	Boycotting	13
CIO	Debates on the 2024 Olympics	34
CYBR	Cyberconflict	29
HACK	Hacktivism	35
MERC	Mercenaries	30
VOL	Volunteers	37
AIR	Aerial warfare	36
MISS	Missiles	38
DRO	Drones	35
DISC	Political speeches	993
PURG	Purge of generals	21
CRIM	Return of criminals amnestied by their engagement in Ukraine	24
OLIG	Disappearances of Russian oligarchs	32
JOUR	Atrocities against journalists on the front	25
EDUC	Militarization of children	22
SANT	Health of Vladimir Putin	10
RESI	Individual acts of resistance by Russian soldiers (refusal to fight and desertion)	27
PLAI	Complaints by Russian soldiers and their families	23



(a) Ground Truth



(b) StreamETM



(c) SB-SETM

Figure 10: **Topic frequencies over time,** `20kNewsGroupOnline`. In (b) StreamETM as in Granese et al. (2025): DOT-based merging, $K_{\texttt{init}}$=3. In (c) SB-SETM: SBERT-enabled, COT-based merging, $K_{\texttt{init}}$=50.