DRIFT: Constrained Dynamic Topic Models for Weak-Signal Trend Discovery

Anonymous authors
Paper under double-blind review

Abstract

Standard topic models recover a corpus's dominant topics, and existing extensions add either temporal dynamics or domain-guided discovery, but rarely both. This gap matters in settings such as mental-health discourse in social media, where the trends of real interest are rare, evolve in both prevalence and content over time, and are easily crowded out by a large, shifting background of everyday content. We introduce DRIFT (Dynamic Recovery of Infrequent Fine-grained Trends), a constrained dynamic topic model that jointly represents time-varying topic prevalence through covariate-driven document-level priors, time-varying topic content through a random-walk prior over topic-word logits, and mild guidance for a domain of interest, without requiring seed guidance to be pre-assigned to individual trends. Two deterministic reparameterizations enforce this guidance by construction rather than by penalty: a seed-mass floor on the content of weak-signal trends, and a downweighting of their prevalence in documents without seed words, so that every topic-proportion and topic-word distribution remains a valid probability simplex throughout inference. We derive a variational EM algorithm that handles the resulting nonconjugate terms with reparameterized Monte Carlo estimation. On a synthetic benchmark built to test recovery of rare, and drifting trends, DRIFT outperforms LDA, DTM, STM, and a constrained matrix-factorization baseline on rare-trend identification, topic-content quality, and temporal recovery, at a modest cost in held-out perplexity relative to the strongest baselines. On a real case study of mental-health discourse in youth YouTube vlogs, DRIFT recovers interpretable, temporally coherent weak-signal trends without ground-truth labels. These results suggest that explicit, feasibility-preserving constraints are an effective way to keep dynamic topic models sensitive to rare but meaningful signals.

1 Introduction

Probabilistic topic models are a popular methodology for quantitative and qualitative analysis of large text collections. Topics models can be used to recover underlying topical themes of the collection and their presence across documents, which can then be used to summarize corpora and to organize and explore document collections at scale. Topic models represent each document as a mixture of latent topics, and each topic as a distribution over words. This representation combines advantages of simple interpretability with sufficient power to extract latent themes, and it has proven useful in many settings. However, current topic models suffer from two major deficiencies, as we outline below: they are either unable to represent well **minority topics**, or are unable to represent **relationship of topics to covariates including time**. **Minority topics are lost by current models.** Most topic models focus on representing majority content of documents, namely the most prevalent themes and their most prominent words. This approach is insufficient when low-prevalence minority themes can contain crucial information. Consider the case of mental health discussion online: mentions of mental health issues are typically low-prevalence and are easily lost among everyday conversation, yet they can yield crucial information of mental health concerns in vulnerable populations such as youth. Seed-guided topic models partly resolve this problem by guiding the model to focus on particular domains of interest using prior words or lexical seeding Andrzejewski et al. (2009); Jagarlamudi et al. (2012). This models works well if the analyst has a rough idea of the domain, but it does not take into account the evolving trends of minor topics over time. Recently, Ebrahimi & Peltonen, 2025 introduced the concept of of constrained matrix factorization for extracting minor topics in presence of major background. Although, their outcomes match our motivation behind this research effort, the algorithms lack any temporal dynamics.

Most topic models are static. Another typical downside of several current topic models is that their representation of themes is static, unable to directly model how content and prevalence of topics should vary with respect to covariates of documents, in particular with respect to time. In many real collections, documents arrive over time. Topics shift. Words that once described a concept fall out of use. New vocabulary enters. A topic active in one period may quiet down and reappear years later under slightly different language. **Representing both minority topics and their association to time and covariates is challenging but crucial.** The ability to represent relationship of topics to covariates including temporal trends is especially crucial when it comes to minority topics. This becomes problematic when the topics/trends of interest are minority patterns. A topic may appear in only a small part of the corpus and still account for a large share of the words in the few documents where it is active. This is especially practical in many applications when dynamic corpora are available. New topics often begin small. They may appear in short bursts and fade away. Their wording may evolve as time goes on. Standard topic models tend to recover the broad and frequent themes first. Weak but meaningful trends are then easy to ignore by standard models. In such cases, it is not enough to model only topic proportions. One also needs a way to preserve weak signals while tracking how topic prevalence and topic content change over time.

Existing topic modeling solutions are not able to address both challenges together, modeling minority topics and temporal trends. In this paper we address this setting and introduce a solution. We build a novel model designed for dynamic minority-topic recovery, which we call DRIFT (Dynamic Recovery of Infrequent Fine-grained Trends). The target topics are weak (low-prevalence) at the corpus level but still important and evolving over time. The model is built to recover such minority topics without requiring prior knowledge of existing minority topics, using only mild overall level seed-word guidance of the minority domain. It incorporates the guidance via well-founded probabilistic constraints on prevalence of minority topics, and prevalence of seed words in their word distributions, without requiring additional loss terms and while explicitly keeping all model representations as feasible and interpretable probability distributions. Moreover, the model does not force topics to remain fixed over time: the model tracks shift of topic prevalence over time, and also allows topic content, meaning the word distributions, to drift. While we focus on temporal trends (modeled by relationship of topics to time covariates) our solution can model relationship of topics to other covariates as well. To infer the model we develop a practical variational expectation maximization (EM) algorithm, where nonconjugate terms from constrained mappings are handled with Monte Carlo estimates. This makes the model workable in settings where exact inference is not available.

We evaluate the performance of our model and its advantage versus comparison models on a synthetic benchmark designed for dynamic minority topic recovery with known ground truth, allowing comprehensive evaluation with several quantitative metrics. In order to quantify the performance of our models a synthetic evaluation framework that is specifically designed for evaluating emergence is proposed. The evaluation framework focuses on rare topics undergoing temporal drift for testing the model's recoverability under constrained yet challenging conditions. The proposed model is then compared against existing models, i.e., Latent Dirichlet Allocation (LDA; Blei et al., 2003), Dynamic Topic Models (DTM; Blei & Lafferty, 2006), Structural Topic Model (STM; Roberts et al., 2016), and ConstraintNMF (Ebrahimi & Peltonen, 2025) on synthetic data. The results show the proposed model yields promising improvements for minority topic identification, topic content representation quality and temporal recovery metrics.

We also demonstrate the performance of our model on a real case study of minority topics, mental health discussion of youth in YouTube vlog discussions. We apply the model to detect mental health discourse in trend detection in this social media domain where mental health arises as an important minority topic. Although our proposed modeling framework is general, mental health is chosen as a domain of interest both for its societal importance and because it has exactly the structure that motivates the work: a broad noisy background, and within it, smaller evolving signals that matter a great deal to the people studying them.

The rest of the paper is organized as follows. Section 2 reviews the related literature. Section 3 states the estimation problem and the main constraints. We then present the DRIFT model and its generative process, its estimation algorithm in Section 4.1, and discuss variational inference and also theoretical properties in detail in Section 4.5. Section 5 reports the experiments through various experiments on synthetic data, real dataset and two derived real data. This includes baselines, evaluation metrics, ablations, sensitivity and robustness checks, and qualitative analyses. Section 6 discusses the main findings briefly and ends with a conclusion.

2 Related Work

2.1 Dynamic Topic Models for Temporal Analysis

Topic models are a common technique for extracting latent topics from texts. The traditional foundation in this research direction is Latent Dirichlet Allocation (LDA, Blei et al., 2003), where each document is modeled as arising from a mixture of topics, and each topic is a mixture of words. However, since LDA is a static model, many extensions have been proposed for dynamic topic models to uncover evolving themes and trends. Early temporal topic models differ in how they represent time. Well-known ones such as include Topics Over Time (TOT, Wang & McCallum, 2006) models topic prevalence as a function of continuous timestamps while keeping topic semantics fixed, whereas Dynamic Topic Models (DTM, Blei & Lafferty, 2006) and continuous-time dynamic topic models (Wang et al., 2008) model the evolution of topic content over time. Other temporal topic models were developed for more task-specific trend analysis, such as term volume or trend structure (Hong et al., 2011; Kawamae, 2011).

Later works extended these ideas toward more scalable,generalized, and neural approaches. There are extensions that generalize the temporal prior beyond the original Wiener process (Jähnichen et al., 2018). Other approaches rely on embedding-based or neural parameterizations for scalability and better representations, including the dynamic embedded topic model and neural dynamic focused topic model(Dieng et al., 2019; Cvejoski et al., 2023a;b). Topic models that have been designed for particular types of temporal settings include multiscale temporal modeling(Nallapati et al., 2007), topics over time modeling with birth and death of topics(Ahmed & Xing, 2010), multivariate or multisource dynamics(Lu et al., 2020; Zhu et al., 2022) and temporal document networks (Zhang & Lauw, 2022).

Dynamic topic models can find topic evolution patterns, but they are normally designed for general temporal changes and gradual change of trends. In many cases, the main concern is not only the general theme changes but also the extraction of low-prevalence topics over time, even though they are not strong in the corpus. Recent work in trend and event discovery has moved in this direction, for example through contextual trend tracking, time-aware document embeddings, and weak-signals (Boutaleb et al., 2024; Jiang et al., 2024; Ma et al., 2024; Nguyen et al., 2016). Recently, there have been discussions about the importance of carefully evaluating comparative dynamic topic models since different models that share the same basic objective of topic evolution over time may still vary significantly in terms of temporal structure and topic quality recovery(Karakkaparambil James et al., 2024; Onah et al., 2025)

Related work combining topic modeling and change detection, both as a two-stage process and as a unified model, can be found in (Holz & Teresniak, 2010; Brüggermann et al., 2016; Wang & Goutte, 2018; Rieger et al., 2022; Zhong & Schweidel, 2020; Lu et al., 2022). The purpose of these techniques is to identify the temporal change in the thematic structure of a corpus. However, they do not explicitly focus on recovering low-prevalence themes and modeling how both their prevalence and content evolve over time.

2.2 Guided and Metadata-Aware Topic Models

A second thread of research focuses on improving topic discovery by using side information. For example, Structural Topic Models (STM) have become a popular choice with covariates and have been used in a wide range of applications (Roberts et al., 2014; 2016; 2019). Other methods have also been developed to incorporate labels and supervision in a more direct fashion, such as in supervised topic models (Mcauliffe & Blei, 2007). Then other methods also incorporate domain knowledge more directly, such as in the context of Dirichlet forest priors, lexical priors, and seeded methods (Andrzejewski et al., 2009; Jagarlamudi et al., 2012). These methods have the advantage of guiding the model toward interpretable and domain-specific topics Ebrahimi & Peltonen (2026b) but differ in how they impose prior knowledge and how they allow for unconstrained topic discovery. However, these above mentioned models are not designed to address the minority-topic setting considered here, where the target themes are low-prevalence at corpus level whereas themes of background content have strong prevalence, and both kinds of themes evolve over time. Constraint-based matrix factorization methods provide another interesting line of work. Recent research on a nonnegative matrix factorization approach with domain constraints (ConstraintNMF; (Ebrahimi & Peltonen, 2025)) has shown effectiveness of incorporating mild constraints and seed-word domain information in recovering minority topics in noisy and imbalanced data, with a further extension to clustering and topic modeling (Ebrahimi & Peltonen, 2026a). The ConstraintNMF work is related to ours as it also deals with low prevalence and minority themes and roughly decoupling them from background patterns. The main difference is that the ConstraintNMF is not designed to incorporate covariates and temporal trend dynamics so that it is static; moreover its model is a matrix factorization model whereas our proposed model is a probabilistic model directly for discrete term-counts of documents.

2.3 Rare, Subtle, and Minority Topic Discovery

In human-generated documents such as social media typically a broad variety of discussion topics are present, with some being much more prevalent than others. When the distribution of topics over the documents is highly skewed, topic models tend to recover the more general topics that appear in majority of documents rather than rare ones (Das et al., 2013; Cvejoski et al., 2023b). A related problem arises when the topics of interest are not only rare in the corpus but also weakly expressed within the documents where they appear. Focused topic models address this setting by proposing a generative model of sparse topic occurrence within the documents (Williamson et al., 2010). Das et al. (2013) use the term "subtle" to define the topics that are both rare and only weakly represented within individual documents, This is particularly relevant to minority topic modeling and weak signal detection, when the issue is that significant but infrequent topics might be hidden under major topics in the corpus (Ma et al., 2024; Ebrahimi & Peltonen, 2025; 2026a). These works motivate our setting, in which the target topics are low prevalence but meaningful trends and dynamically evolving.

Generally speaking, classical topic models used for unsupervised approaches are not capable of detecting those subtle or emerging trends. Thus, these studies motivate our setting where we consider target topics to be low-prevalence and evolving.

2.4 Neural Topic Models and Evaluation

While classical topic models are based on probabilistic graphical models, in recent years neural topic models have emerged with strong ability in theme discovery and also scalability in handling large text corpora. They capture topical features through variational inference, neural encoders, and embedding-based representations. Notable examples include neural variational document models, discrete latent topic models, and autoencoding variational topic models (Miao et al., 2016; 2017; Srivastava & Sutton, 2017). Adversarial and autoregressive neural topic models are other examples of these types of models (Wang et al., 2019; Gupta et al., 2019). Topic models in embedding spaces and dynamic neural topic models have further advanced the flexibility of neural representations, especially for large timestamped data sets (Dieng et al. (2020; 2019); Cvejoski et al. (2023a;b)). Other examples of temporal neural topic models include dynamic neural models that can further be structured, aligned, and chain-free (Miyamoto et al., 2023; Rahimi et al., 2024; Wu et al., 2024a). Other surveys of neural topic models can be found in (Wu et al., 2024b). While such neural models are arguably more flexible than probabilistic topic models, they don't directly address the three requirements that are highly important in our setting: guided recovery of rare themes, separation from dominant background content, and fine-grained temporal modeling of trends. Another related issue is evaluation. Topic coherence and perplexity are widely used, but they do not always reflect success on rare or low-prevalence themes, especially in noisy settings (Newman et al. 2010; Lau et al. 2014; Röder et al. 2015; Ebrahimi & Peltonen 2025). This is especially important in our case where highly domain-relevant topics are rare and changing over time, since the quality of the output from the model needs to be evaluated not just in terms of its ability to fit the text overall, but also in terms of its ability to pick up the domain-relevant minority topics and their trends, which might be easily missed.

Overall, the literature points to a clear gap. Dynamic topic models capture temporal evolution patterns, guided topic models focus on finding domain-specific themes, and minority-topic models focus on low-prevalence patterns. However, the literature does not yet address these settings in a single model for dynamic minority-topic detection. Our work is motivated by this gap. We study trend discovery in a scenario where the topics of interest are not only rare in the corpus and in individual documents, but where the background content is dominant, and evolving over time in both prevalence and content.

3 Problem Statement

We are interested in modeling a collection of documents, such as social media discussion comments, which arise over time and may have also other covariates. We expect the documents to contain not only prominent topics of everyday discussion but also rarely occurring topics ("minority topics") for a domain of interest such as mental health. We do not know beforehand the content of the topics (including the minority topics), the aim is to discover this from data. Our focus is in particular to model well the minority topics and how they vary over time.

As our data, we observe a corpus of documents along with their covariates. We also have available a small set of seed words about the domain of interest—note that we do not assume the seed words to be organized into any pre-existing

topics. We build a probabilistic generative model for the data, based on latent topical trends both for minority topics and for ordinary topics, where the seed words are used to form mild constraints on prevalence and content of the topics.

Formally, let the corpus be $\mathcal{D} = \{(w_d, t_d, x_d)\}_{d=1}^{D}$ where D is the number of documents, w_d is the token sequence of document d, $t_d \in \{0, \ldots, T-1\}$ is its timestamp or time bin where T is the number of bins, and $x_d \in \mathbb{R}^P$ is a P-dimensional vector of observed covariates of the document. In our setting, x_d may include features extracted from time as well as other document-level covariates. The variable t_d is listed separately for ease of notation because it indexes the time-varying topic-word distributions. We also have available mild guidance about the domain of interest overall: a seed set S of words relevant to the domain. Note that the seed set is not specific to any individual topics, only about the overall domain. We do not assume prior knowledge of what specific general or minority themes should exist in the data.

The aim is to model such data as arising out of a set of topics, some of which are general topics and some are rare minority topics relevant to the domain. For all topics, the aim is to estimate their time-varying prevalences as well as their time-varying content (word distributions). The problem combines three challenges. **1)** Minority topics have low prevalence and are easily dominated by background topics. **2)** Their prevalence can change over time in short and irregular episodes rather than smooth long-term trends. **3)** Their content can drift over time even when the underlying topic remains the same.

Existing models usually address only part of this setting. Some model temporal change but not minority-topic recovery, while others use guidance or constraints, but remain static. This setting also changes how the model should be **evaluated**. Standard topic-model metrics such as coherence and perplexity do not fully capture whether weak emerging topics are actually recovered. In our setting, the important question is whether the model can recover minority-topic prevalence, separate it from dominant background content, and track how both prevalence and content change over time. The practical goal is therefore to develop a dynamic topic model and a corresponding optimization algorithm that remain tractable, interpretable, and effective in this regime.

4 Method

We propose a constrained dynamic topic model DRIFT (Dynamic Recovery of Infrequent Fine-grained Trends) for weak-signal themes discovery that takes into account the three challenges in Section 3: we seek to identify topics that 1) have *low base prevalence*, 2) become *episodically active* and 3) may exhibit *content drift*: slow shifts in wording over and between those episodes. The proposed model is designed for weak-signal themes discovery under challenges 1)-3). The model combines covariate-driven prevalence and time-varying content trajectories, addressing challenges 2)-3), with mild and interpretable constraints based on seed-word domain knowledge that let the model focus on minority content, addressing challenge 1). The constraints are incorporated into the probabilistic distributions in the model in a way that lets the model emphasize minority content while preserving simplex geometry. The constraints also benefit challenges 2)-3): they are designed to improve identifiability by achieving **disentanglement of prevalence vs. content**: they prevent shifts of prevalence in the underlying topics from being misrepresented as vocabulary changes only, while still allowing adaptive lexical evolution within topics. This section presents the model definition, objective, and optimization procedure. Detailed derivations are deferred to the appendix.

4.1 Generative Model

For ease of presentation, in the following we first present the generative model and its notation as a whole, then we discuss and motivate each of its parts.

The topic model is shown as a plate model in Figure 1 and corresponds to the following generative procedure. Let D be the number of documents, V the vocabulary size, T the number of time points, K the number of trends/topics, and $\mathcal{K}_{\min} \subset \{1, \ldots, K\}$ the subset of domain-relevant minority/weak-signal trends.

In brief, the model works as follows. **To model topic content**, we introduce time-varying topic-word distributions $\phi_{k,t} \in \Delta^{V-1}$ for each topic k at each time point t, where $\Delta^{V-1} = \{u \in \mathbb{R}^V : u \geq 0, \mathbf{1}^\top u = 1\}$ denotes the $(V-1)$-dimensional probability simplex. The topic content will be allowed to vary across time through the time point specific distributions $\phi_{k,t}$ which will be governed by a random walk prior with parameters σ_0^2 and σ_q^2. The content distributions will further be mapped through a domain constraint-preserving mapping as discussed later. **To model topic prevalences**, we introduce document-level latent topic proportions $\theta_d \in \Delta^{K-1}$ which will be modeled through a logistic-normal latent regression structure driven by the covariates x_d, with parameters $B \in \mathbb{R}^{P \times (K-1)}$ and $\Sigma \in$

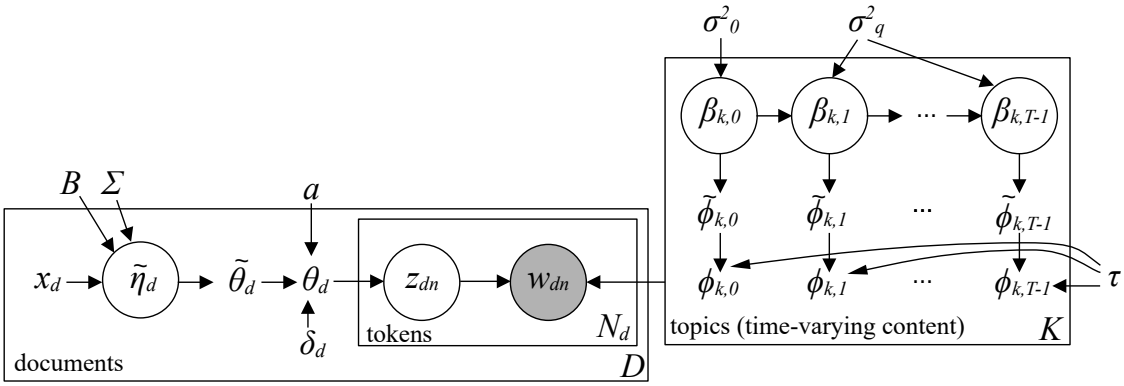

Figure 1: Plate model of the DRIFT topic model. Circles are random variables (shaded background for observed content). Non-circled symbols inside plates are covariates and deterministic mappings; symbols outside plates are hyperparameters.

$\mathbb{R}^{(K-1)\times(K-1)}$, where P is the number of covariates. Since the covariates include features derived from timepoint indices, the structure naturally allows the prevalences to vary over time. The prevalences will further be mapped through a second domain constraint-preserving mapping as discussed later. **To enable the model to focus on minority content**, we introduce two reparameterizations corresponding to soft constraints. The reparameterization for topic prevalences has document-level parameter a, and the reparameterization for topic content has parameter τ. These deterministic maps transform unconstrained distributions into feasible constrained distributions. Observed tokens are then generated from constrained topic proportions and constrained topic-word distributions. This yields a fully specified generative model with explicit dependence on covariates and time. The model has five main parts.

1. **Time-varying topic content trajectories.** For each topic k, draw a latent trajectory $\{\beta_{k,t}\}_{t=0}^{T-1}$ of V-dimensional vectors:
$$\beta_{k,0} \sim \mathcal{N}(0, \Sigma_\beta), \qquad \beta_{k,t} \mid \beta_{k,t-1} \sim \mathcal{N}(\beta_{k,t-1}, Q_\beta) \quad (t = 1, \ldots, T-1)\ , \tag{1}$$
where $\Sigma_\beta = \sigma_0^2 I_V$ and $Q_\beta = \sigma_q^2 I_V$.
2. **Constraint mapping for topic content.** For each topic k at each timepoint t, denote
$$\tilde{\phi}_{k,t} = \mathrm{softmax}(\beta_{k,t}) \in \Delta^{V-1}\ , \tag{2}$$
and compute the constrained topic-word distribution with probability $\phi_{k,t,w}$ for each word w as follows. For non-minority topics $k \notin \mathcal{K}_{\min}$, set
$$\phi_{k,t,w} = \tilde{\phi}_{k,t,w}\ . \tag{3}$$
For minority topics $k \in \mathcal{K}_{\min}$, set
$$\phi_{k,t,w} = \begin{cases} \left(\frac{\tau + (1-\tau)\sum_{v\in S}\tilde{\phi}_{k,t,v}}{\sum_{v\in S}\tilde{\phi}_{k,t,v}}\right)\tilde{\phi}_{k,t,w} & w \in S\ , \\ (1-\tau)\tilde{\phi}_{k,t,w} & w \notin S\ . \end{cases} \tag{4}$$
3. **Covariate-dependent latent variables for topic prevalence.** For each document $d = 1, \ldots, D$, draw:
$$\tilde{\eta}_d \sim \mathcal{N}(B^\top x_d, \Sigma)\ . \tag{5}$$
4. **Constraint mapping for topic prevalence.** For each document d, set
$$\tilde{\theta}_d = \mathrm{softmax}([\tilde{\eta}_d, 0]) \in \Delta^{K-1}\ , \tag{6}$$
and denote $\delta_d = 1$ for documents that will carry minority content and $\delta_d = 0$ for generic documents; δ_d will be inferred from data as detailed in Section 4.5. For minority content carrying documents ($\delta_d = 1$) set
$$\theta_{d,k} = \tilde{\theta}_{d,k} \tag{7}$$
and for generic documents ($\delta_d = 0$) downweight minority topics by factor $a \in [0, 1]$ and renormalize so that
$$\theta_{d,k} = \begin{cases} a\,\tilde{\theta}_{d,k} & k \in \mathcal{K}_{\min}\ , \\ \left(\dfrac{1 - a\sum_{l\in\mathcal{K}_{\min}}\tilde{\theta}_{d,l}}{1 - \sum_{l\in\mathcal{K}_{\min}}\tilde{\theta}_{d,l}}\right)\tilde{\theta}_{d,k} & k \notin \mathcal{K}_{\min}\ . \end{cases} \tag{8}$$

5. **Tokens.** For each document d and each token index $n = 1, \ldots, N_d$ where N_d is the number of tokens in the document, draw the latent topic assignment z_{dn} and the word w_{dn}:

$$z_{dn} \mid \theta_d \sim \text{Cat}(\theta_d) \ , \tag{9}$$

$$w_{dn} \mid z_{dn}, t_d, \{\phi_{k,t}\} \sim \text{Cat}(\phi_{z_{dn}, t_d}) \ . \tag{10}$$

We next detail the five main parts of the model starting from the direct observations and then describing the latent structure and the constraints.

4.2 Token Generation

The core topic model equations (part 5 of the generative model) govern how word tokens are drawn from the underlying topics. As usual in probabilistic topic models, we model this as draws from multinomial (categorical) distributions. For each document $d = 1, \ldots, D$ and each token index $n = 1, \ldots, N_d$, the observed word w_{dn} is modeled arising from a latent topic $z_{dn} \in \{1, \ldots, K\}$. We first draw the latent topic assignment z_{dn} for the token by Equation (9), and then the observed word w_{dn} by Equation (10), where θ_d is the latent topic prevalence distribution in the document and ϕ_{z_{dn}, t_d} is the word distribution of the chosen topic at the time index t_d of the document.

Note that the word distribution ϕ_{z_{dn}, t_d} explicitly depends on time, whereas the topic prevalence distribution θ_d is specific to the document and we will model it depending on time and other covariates of the document. As usual, these categorical distributions are parameterized as vectors of nonnegative probabilities summing to one over the K topics and V words respectively, i.e., they are part of the $K - 1$ and $V - 1$ dimensional probability simplex denoted as $\theta_d \in \Delta^{K-1}$ and $\phi_{z_{dn}, t_d} \in \Delta^{V-1}$. Moreover, these distributions are obtained from mappings that enable the model to focus on domain-relevant minority topics. We describe the mappings next.

4.3 Constraint Design for Focusing on Minority Topics

In order to enable the model to focus on identifying the domain-relevant minority topics under the challenges of low base prevalence, episodic activity and content drift, we use mild constraints on the topic content and topic prevalence, guided by the overall seed word set of the domain. These constraints form **parts 2 and 4 of the generative model, respectively.** The constraints make use of only mild guidance in the form of a *shared seed set* $S \subseteq \{1, \ldots, V\}$ of words that are relevant to the domain. Importantly, the seed set lists words known to be relevant in the minority domain overall, but *we do not require knowledge of prespecified minority topics* with known guidance for each.

Both constraints are achieved through deterministic mappings (reparameterizations) so that the prevalence distributions and topic-content distributions satisfy the constraints and remain valid probability distributions throughout the modeling and optimization, without requiring any additional modeling penalty terms to maintain the constraints.

In detail, **the first constraint (part 2 of the model) is on topic content**: it maintains, for each minority topic and time step, a minimum requirement of probability mass to be placed in total over the seed words. For non-minority topics, no constraint is placed: the probability $\phi_{k,t,w}$ of word w in topic k at time t is simply computed from an underlying latent variable vector $\beta_{k,t}$ by a softmax transformation in equations (2) and (3), so that the individual word probabilities become $\phi_{k,t,w} = \tilde{\phi}_{k,t,w} = \exp(\beta_{k,t,w}) / \sum_{w'} \exp(\beta_{k,t,w'})$. However, for minority topics the intermediate values $\tilde{\phi}_{k,t,w}$ are given to equation (4) which maintains the constraint. In that equation, weights of seed words $w \in S$ are upweighted by a multiplier constructed so that the total probability mass over S becomes at least τ. Correspondingly, weights of non-seed words $w \notin S$ are downweighted by a multiplier $1 - \tau$ so that their total weight is at most $1 - \tau$, and so that all probabilities still sum to 1 as required.

Note that this is a mild constraint that only maintains a minimum total value of probability mass over seed words in minority topics. This has several advantages. The constraint does not restrict non-minority topics at all, allowing them to model majority content as needed. For minority topics, the constraint does not restrict the probability mass on seed words to a fixed value, it only ensures a minimum and allows the mass to grow larger if needed in modeling. Moreover, the constraint does not require probability to be placed on any individual seed words: the constraint does not specify how the probability mass should be distributed among seed words (uniformly or otherwise), it allows the model freedom to allocate which seed words to emphasize in each minority topic. This is an important distinction from conventional seed-word techniques which might assume and enforce pre-known seed word profiles per topic. The constraint further allows the model to allocate probability mass in minority topics to non-seed words too. Thus the constraint is flexible and ensures the model remains *data-adaptive* such that it can discover a data-driven division of

minority topics, differentiating their content based on both seed words and on other co-occurring words. The constraint is simple and easily understandable, it maintains a valid probability distribution $\phi_{k,t} \in \Delta^{V-1}$ for all topics at all times without need of any extra cost function penalty terms, and it is governed in a clear way by one parameter $\tau \in [0, 1)$ representing the minimum seed mass in minority topics.

The **second constraint (part 4 of the model) is on topic prevalence**: it reduces the total prevalence of the minority topics in each document by a document-level factor, so that minority topics are not over-represented unless the data support them. For each document d we start by computing a vector $\tilde{\theta}_d$ of intermediate topic probabilities from the underlying latent variable vector $\tilde{\eta}_d$ by equation (6). Since a softmax operation is invariant to a constant shift of its latent-value arguments, we chose it so that the latent value of the last topic is zero, to reduce the number of free parameters. Thus the intermediate probability of each topic k in the document becomes $\tilde{\theta}_{d,k} = \exp(\tilde{\eta}_{d,k})/(1+\sum_{k'=1}^{K-1}\exp(\tilde{\eta}_{d,k'}))$ for $1 \leq k \leq K-1$ and $\tilde{\theta}_{d,K} = 1/(1+\sum_{k'=1}^{K-1}\exp(\tilde{\eta}_{d,k'}))$. We distinguish documents that will carry minority content, denoted $\delta_d = 1$, from generic documents that will not have notable minority content, denoted $\delta_d = 0$. The δ_d could be drawn from a Bernoulli prior but in inference we will infer it directly from data with a simple seed word based rule in Section 4.5. For documents that will carry minority content ($\delta_d = 1$) no constraint is applied and the final topic probabilities (7) are the same as the intermediate ones. However, for generic documents ($\delta_d = 0$) we constrain the probability of the minority topics by equation (8), where probabilities of the minority topics $k \in \mathcal{K}_{\text{min}}$ are downweighted by a multiplier $a \in [0, 1]$ and probabilities of non-minority topics are renormalized by a corresponding multiplier so that the sum over all topics remains 1 as required.

This is again a mild constraint, which only downweights minority topics for documents identified as generic documents ($\delta_d = 0$). The constraint has several advantages. It does not restrict topic prevalences in documents identified as carrying minority content ($\delta_d = 1$), particularly it does not upweight minority topics or downweight non-minority topics in such documents; this is reasonable since sensitive minority topics may not dominate even those documents where they appear. In generic documents ($\delta_d = 0$), the probability mass of minority topics is not set to a fixed constant, rather the multiplicative downweighting ensures it remains a factor a times the unconstrained minority mass of the document, and hence remains below the maximum a in all generic documents. The downweighting is applied to the overall probability mass of minority topics, hence individual minority topics are still allowed to occur in the generic documents and during inference the model fitting again decides if occurrence probability should be allocated to some minority topics. The constraint is again simple and understandable, maintains a valid probability distribution $\theta_d \in \Delta^{K-1}$ in each document, and is governed by one simple parameter $a \in [0, 1]$ representing the maximum probability mass over minority topics in generic documents. The two constraint mappings discussed above were applied to latent variables $\beta_{k,t}$ and $\tilde{\eta}_d$ governing the topic content and topic prevalences respectively. We discuss their model structure next.

4.4 Time-varying Latent Structure for Topic Content and Topic Prevalence

The model allows topic content to vary over time and topic prevalence to vary depending on document covariates including time-based covariates (**parts 1 and 3 of the generative model**).

The variation of topic content over time is controlled by **latent content trajectories (part 1 of the generative model)**. The latent variables $\beta_{k,t} \in \mathbb{R}^V$ are V-dimensional vectors of topic-word logits for each topic k at each time bin t. We model their evolution through a random walk prior in equation (1). The first latent variable vector $\beta_{k,0}$ is drawn from a multivariate normal distribution, and for each successive time bin a normally distributed innovation is added, i.e., the latent variable vector $\beta_{k,t}$ is drawn from a multivariate normal centered on the previous value $\beta_{k,t-1}$. For simplicity we use spherical normal distributions with variances σ_0^2 for the initial prior and σ_q^2 for the successive steps respectively. This random walk prior ensures the topic content can vary over time while keeping the latent trajectories smooth.

Note that $\beta_{k,t}$ will be identified only up to an additive constant in $\mathbf{1}$, since the later softmax based mapping (equation (2) in **part 2** of the model) is invariant to constant shift so that $\text{softmax}(\beta_{k,t} + c\mathbf{1}) = \text{softmax}(\beta_{k,t})$ for any scalar c; the random walk prior will then give a smooth regularization effect that resolves this ambiguity, as usual in dynamic topic models.

The variation of topic prevalence over documents is controlled by **covariate-dependent latent variables (part 3 of the generative model)**. For each document d, the latent variable vector $\tilde{\eta}_d$ is drawn from a multivariate normal prior distribution in Equation (5) where the mean $B^\top x_d$ is a function of the covariates x_d of the document and the covariance matrix Σ governs variation around the mean. Since the topic prevalence probabilities are further mapped from $\tilde{\eta}_d$ through a logistic transformation (**part 4** of the model as discussed above), overall this yields a logistic-

normal specification for the prevalences. The covariates x_d are defined by the modeler for the respective problem domain according to what can be extracted from the data. In our experiments we focus on temporal variation and extract x_d as a P-dimensional vector of values of temporal nonlinear basis functions computed from the time index t_d of the document, as detailed later in equation (22), thus the mean $B^\top x_d$ of the prior will model latent variation of topic prevalence over time. This model allows flexible temporal variation of all topics; since the x_d contains nonlinear basis functions of time the linear regression $B^\top x_d$ suffices to model nonlinear temporal variation, however if needed it could naturally be replaced by a nonlinear regression from x_d.

Overall, the generative model specification (parts 1-4) allows very flexible temporal variation while incorporating essential constraints that allow the model to focus on minority topics rather than being overfocused on majority content only. Despite the flexibility, all parts of the model remain clearly understandable and are governed by few parameters each, and all constraints are handled by direct mappings not requiring any additional penalties.

Prevalence vs. content disentanglement. A key implication of the specification is that the model *encourages* separation of topic prevalence from the content. Topic prevalence at the document level is modeled using the logistic-normal specification of θ_d, whereas topic-content over time is modeled via $\phi_{k,t}$. The two sets of parameters are separate, and the seed-mass constraint $\sum_{v \in S} \phi_{k,t,v} \geq \tau$ in minority topics serves as a temporal anchor for the contents of the minority topics: it limits the extent to which a change in the prevalence of a minority topic k can be re-expressed as a shift of $\phi_{k,t}$ away from the seed words, because the seed mass is bounded below by τ at every t. This reduces, though it does not by itself eliminate, the prevalence–content trade-off along the seed-word subspace, and we treat it as a modeling rationale rather than a formal identifiability guarantee.

4.5 Inference

We will perform Bayesian inference for the model by variational approximation. We thus formulate estimation of the model as the maximization of a constrained variational objective

$$\max_{q,\, \beta,\, B,\, \Sigma} \mathcal{L}(q, \beta, B, \Sigma), \tag{11}$$

subject to simplex constraints on the document-topic proportions $\{\theta_d\}$ and the topic-word distributions $\{\phi_{k,t}\}$. Here q is a variational distribution over the document-level latent variables $\{\tilde{\eta}_d\}$ and token assignments $\{z_{dn}\}$, while the time-varying topic logits $\{\beta_{k,t}\}$ are treated as parameters to be MAP-estimated under the random-walk prior introduced in Section 4.4; this is the standard treatment in dynamic topic models and avoids the cost of a full variational posterior over long β trajectories. In addition, we impose seed-mass constraints through reparameterization mappings as discussed in Section 4.3 and Equations (2)-(4) and (6)-(8), so that minority topics maintain a minimum probability mass on the shared seed set S and non-minority topics maintain a maximum probability mass in generic documents. For simplicity we infer which documents are generic by a simple filter rule: documents containing any seed word in S are denoted as potentially carrying minority content (i.e., we estimate $\delta_d = 1$) and generic otherwise ($\delta_d = 0$). The parameters B and Σ control topic prevalence, and β controls the time-varying topic-word distributions, and the random-walk variances σ_0^2, σ_q^2 (fixed hyperparameters) control the smoothness of the content trajectories. The main technical difficulty is that the constraint mappings introduce nonconjugate terms into the objective.

Exact posterior inference is intractable because of nonconjugate transformations induced by the constraint mappings. We therefore use a mean-field variational approximation for local latents and MAP estimation for global dynamic content trajectories. The optimization target is the evidence lower bound (ELBO) with a random-walk prior term on dynamic logits β. In the main text we present the objective and final decomposition; complete derivations from the Bayesian posterior (Eq. (27)) to the Kullback-Leibler divergence (Eq. (28)) to the ELBO (Eq. (30)) are in Appendix B.

For brevity, we write $\phi_{k,t}(\beta)$ and $\theta_d(\tilde{\eta}_d)$ for the constrained quantities defined by Eqs. (2)–(4) and (6)–(8) respectively; their dependence on the seed set S, the seed-mass parameter τ, and the prevalence rescaling factor a is left implicit.

Joint distribution. Collecting the parts of the generative model, the joint distribution of the observed words w, the assignments z, and the latent variables $(\tilde{\eta}, \beta)$ given the covariates x and timestamps t factorizes as

$$p(w, z, \tilde{\eta}, \beta \mid x, t) = \left[\prod_{k=1}^{K} p(\beta_{k,0:T-1}) \right] \prod_{d=1}^{D} p(\tilde{\eta}_d \mid x_d) \prod_{n=1}^{N_d} p\big(z_{dn} \mid \theta_d(\tilde{\eta}_d)\big)\, p\big(w_{dn} \mid z_{dn}, \phi_{z_{dn},t_d}(\beta)\big), \tag{12}$$

where $p(\tilde{\eta}_d \mid x_d) = \mathcal{N}(B^\top x_d, \Sigma)$ is the prevalence prior of Eq. (5), $p(\beta_{k,0:T-1})$ is the random-walk prior of Eq. (1), and the dependence of θ_d and $\phi_{k,t}$ on $(\tilde{\eta}, \beta)$ is through the deterministic constraint maps.

Variational family. We approximate the posterior over the local latent variables by a mean-field family that factorizes over documents and tokens,

$$q(\tilde{\eta}, z) = \prod_{d=1}^{D} q_d(\tilde{\eta}_d) \prod_{d=1}^{D} \prod_{n=1}^{N_d} q_{dn}(z_{dn}), \tag{13}$$

with a diagonal Gaussian distribution for each document prevalence variable and a categorical distribution for each token assignment,

$$q_d(\tilde{\eta}_d) = \mathcal{N}\big(m_d, \mathrm{diag}(s_d^2)\big), \qquad q_{dn}(z_{dn}) = \mathrm{Cat}(\gamma_{dn}). \tag{14}$$

Here $m_d \in \mathbb{R}^{K-1}$ and $s_d \in \mathbb{R}_{>0}^{K-1}$ are the variational mean and standard deviation for the latent variable of document d, and $\gamma_{dn} \in \Delta^{K-1}$ are the variational means of the token responsibilities. The content logits β are not given a variational factor; they enter $\mathcal{L}$ as parameters and are MAP-estimated.

Evidence lower bound. For a fixed value of β, the standard variational identity relates the marginal log-likelihood, the ELBO, and the divergence of q from the conditional posterior $p(\tilde{\eta}, z \mid w, x, t, \beta)$:

$$\log p(w \mid x, t, \beta) = \mathcal{L}(q, \beta) + \mathrm{KL}\big(q(\tilde{\eta}, z) \,\big\|\, p(\tilde{\eta}, z \mid w, x, t, \beta)\big). \tag{15}$$

Because the left-hand side $\log p(w \mid x, t, \beta)$ does not depend on q, the sum $\mathcal{L}(q, \beta) + \mathrm{KL}(\cdots)$ is constant in q; since the divergence is nonnegative, maximizing $\mathcal{L}$ over q is equivalent to minimizing the divergence; full Bayes-rule-to-KL-to-ELBO steps are given in Appendix B. Including the random-walk prior on β, the objective we maximize is

$$\mathcal{L}(q, \beta, B, \Sigma) = \mathbb{E}_q\big[\log p(w, z, \tilde{\eta} \mid x, t, \beta)\big] - \mathbb{E}_q\big[\log q(\tilde{\eta}, z)\big] + \sum_{k=1}^{K} \log p(\beta_{k,0:T-1}). \tag{16}$$

ELBO decomposition. Substituting the joint factorization (12) and the mean-field q, the ELBO separates into interpretable terms,

$$\begin{aligned}
\mathcal{L}(q, \beta, B, \Sigma) = {} & \underbrace{\sum_{d=1}^{D} \mathbb{E}_{q_d}\big[\log p(\tilde{\eta}_d \mid x_d)\big]}_{\text{(A) prevalence prior}} + \underbrace{\sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{k=1}^{K} \gamma_{dn}(k)\ \mathbb{E}_{q_d}\big[\log \theta_{d,k}(\tilde{\eta}_d)\big]}_{\text{(B) topic assignment}} \\
& + \underbrace{\sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{k=1}^{K} \gamma_{dn}(k)\ \log \phi_{k,t_d,w_{dn}}(\beta)}_{\text{(C) word likelihood}} + \underbrace{\sum_{d=1}^{D} \sum_{n=1}^{N_d} H\big(\gamma_{dn}\big)}_{\text{(D1) categorical entropy}} \\
& + \underbrace{\sum_{d=1}^{D} H\big(q_d(\tilde{\eta}_d)\big)}_{\text{(D2) Gaussian entropy}} + \underbrace{\sum_{k=1}^{K} \log p(\beta_{k,0:T-1})}_{\text{(E) random-walk prior}}.
\end{aligned} \tag{17}$$

Terms (A), (C), (D1), (D2), and (E) are available in closed form given $(m_d, s_d, \gamma_{dn}, \beta)$. The only intractable term is the expectation $\mathbb{E}_{q_d}[\log \theta_{d,k}(\tilde{\eta}_d)]$ in (B), because $\theta_d(\tilde{\eta}_d)$ is the nonlinear constraint map of a Gaussian variable. We estimate this expectation by reparameterized Monte Carlo: drawing $\epsilon^{(s)} \sim \mathcal{N}(0, I)$ and setting $\tilde{\eta}_d^{(s)} = m_d + s_d \odot \epsilon^{(s)}$ where $\odot$ denotes elementwise multiplication, we have

$$\mathbb{E}_{q_d}\big[\log \theta_{d,k}(\tilde{\eta}_d)\big] \approx \frac{1}{N_{\mathrm{MC}}} \sum_{s=1}^{N_{\mathrm{MC}}} \log \theta_{d,k}\big(\tilde{\eta}_d^{(s)}\big), \tag{18}$$

where s indexes the N_{MC} samples.

Variational EM optimization. We maximize $\mathcal{L}$ by alternating updates of the variational factors and the parameters (variational EM):

$$\textbf{E-step:}\quad q \leftarrow \arg\max_{q} \mathcal{L}(q, \beta, B, \Sigma) \ \ (\beta, B, \Sigma \text{ fixed}), \qquad \textbf{M-step:}\quad (\beta, B, \Sigma) \leftarrow \arg\max_{\beta, B, \Sigma} \mathcal{L}(q, \beta, B, \Sigma) \ \ (q \text{ fixed}). \tag{19}$$

Table 1: Synthetic data set and additional topic-capacity datasets used in the sweep. All sweep datasets keep the same document count, token budget, and time span; only topic capacity and minority/background partition vary. Each sweep data set is named with a code listing the overall topic number and number of minority topics.

Setting	**K**	$K_{\min}$	K_{BG}	**Docs**	**Tokens**	**T**
Synthetic main	16	4	12	1,800	90,272	36
K40_KMH10	40	10	30	1,800	90,272	36
K60_KMH15	60	15	45	1,800	90,272	36
K80_KMH20	80	20	60	1,800	90,272	36
K100_KMH25	100	25	75	1,800	90,272	36

In the E-step, the token responsibilities γ_{dn} have the closed-form update

$$\gamma_{dn}(k) \;\propto\; \exp\Big(\widehat{\mathbb{E}}_{q_d}\big[\log\theta_{d,k}(\tilde{\eta}_d)\big] + \log\phi_{k,t_d,w_{dn}}(\beta)\Big), \qquad \sum_{k=1}^{K}\gamma_{dn}(k) = 1, \tag{20}$$

where $\widehat{\mathbb{E}}_{q_d}[\,\cdot\,]$ is the Monte Carlo estimate on the right-hand side of (18), while the Gaussian parameters (m_d, s_d) are updated by stochastic gradient ascent using the reparameterization trick. In the M-step, the content logits β are updated by gradient ascent on terms (C) and (E), with gradients through the constraint mapping (2)-(4) obtained by automatic differentiation, and the prevalence parameters B by a closed-form ridge-regression working approximation that regresses the variational means $\{m_d\}$ on the covariates (with ridge strength $\lambda = 1$), and Σ is updated as the residual covariance augmented with the average variational variance (Appendix D). After fitting, the document–topic proportions are estimated by Monte Carlo integration over the fitted variational posterior, applying the constrained minority-rescaling map $\theta_d(\cdot)$ of Eqs. (6)-(8) to reparameterized samples:

$$\hat{\theta}_{dk} \approx \frac{1}{N_{\mathrm{MC}}}\sum_{s=1}^{N_{\mathrm{MC}}}\theta_{d,k}\Big(\mathbf{m}_d + \boldsymbol{\epsilon}^{(s)} \odot \mathbf{s}_d\Big), \quad \boldsymbol{\epsilon}^{(s)} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{21}$$

5 Experiments

We evaluate the proposed model DRIFT on both synthetic and real-world data. The synthetic experiments assess recovery of known minority-topic temporal structure under controlled conditions, while the real-data experiments examine temporal and qualitative behaviour in a large YouTube comment corpus. We first describe the synthetic experimental setup, baselines, and evaluation criteria, then report the main comparison results and follow-up analyses. After that, we present the real-data experiments, including analyses on domain-relevant partitions of the data and qualitative analysis of topic evolution.

5.1 Experiments on Synthetic Data

We evaluate the synthetic performance of the proposed model and the comparison methods on a controlled time-stamped corpus with known ground-truth minority-topic/trend trajectories and seed structure. For the main comparison, ablation, sensitivity, robustness, and significance analyses, we use a synthetic corpus containing 1,800 documents and 90,272 tokens over 36 time bins. We additionally examine the performance of our model on four data sets with more topics, which we call a topic-capacity sweep. For the topic-capacity sweep, we consider larger-capacity synthetic settings with varying values of the overall number of topics K, the size $K_{\min}$ of the minority-relevant topic subset, and the size K_{BG} of the non-minority-relevant (background) topic subset. Table 1 lists the settings of the five data sets.

5.1.1 Experimental Setup and Evaluation

We generate a time-stamped document collection with (i) time-varying topic-word distributions, (ii) document-topic proportions driven by covariates via a logistic-normal model, and (iii) seed-word constraints that ensure at least τ probability mass on each minority topic's seed set.

Notation. The synthetic corpus is generated with a vocabulary of size V and $K = K_{\min} + K_{\mathrm{BG}}$ topics, partitioned into $K_{\min}$ minority topics and K_{BG} background topics, so that the minority and background topic index sets are $\mathcal{K}_{\min} =$

$\{1, \ldots, K_{\min}\}$ and $\mathcal{K}_{\text{BG}} = \{K_{\min} + 1, \ldots, K\}$. Documents are distributed over T time bins indexed as $t \in \{1, \ldots, T\}$, with M documents per bin, giving $D = T \cdot M$ documents in total. Each document is indexed by $d \in \{1, \ldots, D\}$ and carries a time index $t_d \in \{1, \ldots, T\}$ together with a covariate vector $x_d \in \mathbb{R}^P$, where P is the covariate dimension. We use the following parameter shapes:

$$B \in \mathbb{R}^{P\times(K-1)}, \quad \Sigma \in \mathbb{R}^{(K-1)\times(K-1)}, \quad \beta_{k,t} \in \mathbb{R}^V, \quad \phi_{k,t} \in \Delta^{V-1}, \quad \theta_d \in \Delta^{K-1}.$$

Covariates. For each synthetic document d we use $P = 5$ time-based covariates computed from a normalized time index of the document $u_d = t_d/(T-1)$, so that

$$x_d = \left[1,\ u_d,\ \sin(2\pi u_d),\ (u_d - c)^2,\ \exp\left(-\frac{1}{2}\left(\frac{u_d - c}{\sigma}\right)^2\right)\right]^\top \qquad c = 0.65,\ \sigma = 0.12\ . \tag{22}$$

Ground-truth minority topics for the synthetic data. In the synthetic data, we create ground-truth minority topics based on the overall seed-word set S of the domain, so that each ground-truth minority topic emphasizes a few of the seed words in its word content; here each minority topic $k \in \mathcal{K}_{\min}$ emphasizes a disjoint subset of $s = 5$ seed words, $\mathcal{S}_k = \{(k-1)s + 1, \ldots, ks\} \subseteq S$. No emphasis is used for the non-minority (background) topics. The ground truth sets $\mathcal{S}_k$ are not provided to learning algorithms, they must learn topics from data given only the overall seed-word set.

The synthetic data is then generated with a procedure somewhat similar to the model in Section 4.1, as follows.

1. **Topic prevalences in documents** are drawn: first latent values are drawn from a logistic-normal prior in Eq. (5), with regression weights B and covariance Σ listed in Table 2. Document type is then drawn with probability $\pi(u_d) = 0.125 + 0.025\sin(2\pi u_d)$ for a minority-content carrying document ($\delta_d = 1$) and $1 - \pi(u_d)$ for a generic document ($\delta_d = 0$), where $u_d = t_d/(T-1)$ is normalized time. Then the topic prevalences θ_d are yielded by the constraint mapping in Eqs. (6)-(8). This mild seasonal pattern keeps the minority-active fraction between roughly 10% and 15% across the corpus.
2. **Topic-word trajectories** are drawn: first latent values $\beta_{k,t}$ are drawn from the Gaussian random walk prior of Eq. (1), with variances $\Sigma_\beta = \sigma_0^2 I_V$ and $Q_\beta = \sigma_q^2 I_V$ set as in Table 2, and probabilities $\tilde{\phi}_{k,t}$ are computed by Eq. (2). For background topics we set $\phi_{k,t} = \tilde{\phi}_{k,t}$. For minority topics $k \in \mathcal{K}_{\min}$ we emphasize their ground truth seed words by a minimum weight τ, so that
$$\phi_{k,t,v} = \begin{cases} \tilde{\phi}_{k,t,v}\left(\frac{\tau + (1-\tau)\sum_{v'\in\mathcal{S}_k}\tilde{\phi}_{k,t,v'}}{\sum_{v'\in\mathcal{S}_k}\tilde{\phi}_{k,t,v'}}\right) & v \in \mathcal{S}_k, \\ \tilde{\phi}_{k,t,v}(1-\tau) & v \notin \mathcal{S}_k. \end{cases}$$
which ensures each minority topic places at least τ total weight on its ground truth seed words $\mathcal{S}_k$.
3. **Document generation.** Document lengths are sampled requiring a minimum length $N_{\min}$, so that $\tilde{N}_d \sim \text{Poisson}(\mu_N)$ and $N_d = \max(N_d, N_{\min})$. Then tokens are generated as in Eqs. (9)-(10), so that $z_{dn} \sim \text{Cat}(\theta_d)$ and $w_{dn} \sim \text{Cat}(\phi_{z_{dn},t_d})$.

5.1.2 Baselines

We compare our intelligently constrained model DRIFT (denoted 'Ours' in the result tables and figures for brevity) against four baselines spanning the relevant design space — static, temporal, structural, and seed-guided: (i) **LDA** (Blei et al., 2003), an established bag-of-words topic model with no temporal or structural constraints; (ii) **DTM** (Blei & Lafferty, 2006), which allows topic–word distributions to evolve over time via a Gaussian random walk but does not incorporate domain knowledge via domain-specific constraints; and (iii) **STM** (Roberts et al., 2016; 2014) which models topic prevalence using document-level covariates, and captures some temporal variation, but does not allow incorporation of domain knowledge via seed-word constraints. As a result, STM tends to recover frequent dominant trends rather than rare or domain-specific minority topics. (iv) **ConstraintNMF** (Ebrahimi & Peltonen, 2025) is a constrained nonnegative matrix factorization approach that uses seed guidance constraints to guide topic discovery; the model is close to our approach, but it does not model temporal dynamics. All baselines use $K = 16$ topics, matching the total number of topics in our model. For our model we set the constraint parameters as $a = 0.03$ and $\tau = 0.2$.

Reproducibility. All primary model-evaluation experiments (main results, ablation, sensitivity, robustness, and efficiency) are repeated over $n_{\text{seed}} = 8$ random seeds (seeds 0–9). For statistical significance testing only, we run an independent larger evaluation with $n_{\text{seed,sig}} = 15$ seeds to increase test power. Robustness perturbations use offset seeds

Table 2: Synthetic generation parameters.

Symbol / Name	Default	Meaning
V	900	Vocabulary size
K_{min}	4	Number of minority topics.
K_{BG}	12	Number of background topics.
K	16	Total topics $K_{\text{min}} + K_{\text{BG}}$.
T	36	Number of time bins (months).
M	50	Documents per time bin; $D = T \cdot M$.
μ_N	50	Mean document length (Poisson).
N_{min}	5	Minimum document length clamp.
P	5	Covariate dimension.
c, σ	0.65, 0.12	Center/spread of bell-shaped covariate.
B	(fixed)	Regression weights for covariate-driven topic prevalence ($P \times (K-1)$).
Σ	$I \cdot 1.0$	Covariance for logistic-normal latent $\tilde{\eta}_d$.
Σ_β	$I \cdot 0.30$	Initial covariance for topic-word logits $\beta_{k,0}$.
Q_β	$I \cdot 1.00$	Innovation covariance for random walk in $\beta_{k,t}$.
τ	0.20	Seed-mass lower bound applied to every minority topic.
$\lvert\mathcal{S}_k\rvert$	5	Seed words per minority topic; disjoint across minority topics.
a	0.03	Minority-topic scaling used in θ_d rescaling.

(seed + 1000 for noise, +2000 for subsampling, +3000 for dropped bins) to ensure independence between perturbation randomness and model initialization.

5.1.3 Metric Computation

We evaluate our model and the comparison models from three main perspectives: minority-topic classification quality, temporal recovery, and predictive performance. All evaluation metrics are computed from the row-normalized documents-topic matrix $\boldsymbol{\theta} \in \mathbb{R}^{D \times K}$. For our model the document-topic proportions are estimated via Monte Carlo integration over the variational posterior by Eqs. (21) and (6)-(8) as described in Section 4.5.

Before computing the minority-topic focused metrics, inferred topics of each model are matched to the four ground-truth minority-topic trajectories using the Hungarian algorithm Kuhn (1955). The matching is based on the Pearson correlation between inferred topic prevalence over time (average in each time bin over its documents) and the four ground-truth minority-topic trajectories (average of the ground-truth topic prevalence in each time bin over its documents) across the $T = 36$ time bins. Once the best-matching topic for each minority topic is found, the same matching is then used for all minority-topic focused evaluation metrics.

For the main comparison, we report six metrics. Together, they describe document-level minority-topic classification, recovery of temporal patterns, recovery of minority-topic distributions, and heldout predictive performance.

Minority-topic classification quality To measure how well the inferred document-level minority-topic assignments agree with the four ground-truth minority-topic categories, we use Purity and Normalized mutual information (NMI) Manning et al. (2008) following Wu et al. (2023); Zhao et al. (2021). We consider the subset of synthetic documents generated as carrying minority content ($\delta_d = 1$), and for each such document the minority topic with highest ground-truth prevalence is taken as the ground truth label. From model fitting the estimated highest-probability topic, among those that were matched to the ground truth topics by the Hungarian algorithm, is taken as the predicted label. Purity and NMI are the computed comparing these predicted versus ground truth labels. These two metrics show whether the model can separate the minority-topic classes clearly at the document level. Higher values indicate better agreement.

Temporal recovery. Temporal recovery is assessed by Trend Correlation and Temporal mean absolute error (Temporal MAE). Trend Correlation is defined as the mean Pearson correlation between the four aligned minority-topic/trend trajectories and the four ground-truth trajectories, so higher values indicate better recovery of the temporal patterns. Temporal MAE measures the average absolute difference between the inferred and ground-truth trajectories, so lower values indicate better recovery.

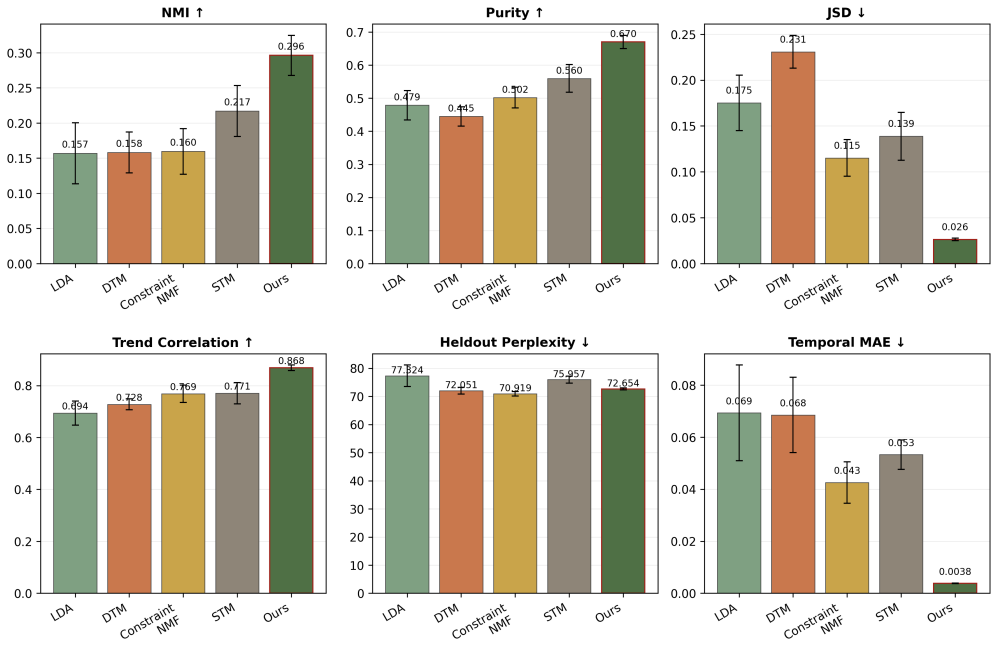

Figure 2: Core MH-focused synthetic results across several random seed runs. Bars show mean performance and error bars show standard deviations.

Distributional recovery. Jensen–Shannon divergence (JSD) is used to measure how closely the inferred distribution over minority topics matches the ground-truth distribution within each time bin. For each time bin, we compute the average ground-truth proportions and inferred topic proportions for the four minority topics (as averages over documents in the bin), normalize them into distributions, and compute their JSD. Lower JSD means the model better recovers the relative composition of minority trends over time.

Heldout perplexity. We compute heldout perplexity (denoted Heldout PPL for brevity in some tables) on documents from the last (20%) of time bins. Topic-word probabilities are estimated from the training portion. For simplicity, instead of fitting document-topic probabilities separately to held-out time bins, we first fit the model to all data and then set the topic-word probabilities in held-out bins to equal the probabilities from the most recent training bin; predictive log-likelihood is then evaluated on the heldout tokens. We consider this shortcut to have negligible impact on perplexity results compared to the large differences between methods. Lower perplexity values mean better predictive performance.

5.1.4 Main Comparison Results

Table 3 and Figure 2 summarize the core minority topic focused results on synthetic data as means and standard deviations over 10 runs with different random seeds. Our model achieves the best performance on five of six reported metrics as shown. Ours reaches NMI = 0.296 ± 0.028, and Purity = 0.670 ± 0.020, outperforming the next best method STM which has NMI = 0.217 ± 0.036 and Purity = 0.560 ± 0.042. Our model also achieves the strongest distributional and temporal recovery. It achieves JSD = 0.026 ± 0.002, considerably lower than the range 0.115–0.231 of the baseline method results, while it also achieves TrendCorr = 0.868 ± 0.010 which exceeds the baseline range of 0.694–0.771. Also the lowest Temporal MAE score at 0.0038 ± 0.0001 belongs to our model, compared with 0.043–0.069 for the competing baselines. Together, these results show that the proposed model recovers minority topic dynamics more accurately and more consistently than the competing methods. For Heldout Perplexity, ConstraintNMF performs best with 70.919 ± 0.781, followed by DTM at 72.051 ± 1.193 and our model at 72.654 ± 0.368. STM and LDA have worse scores at 75.956 ± 1.268 and 77.324 ± 3.795 respectively. Thus, our model remains competitive in terms of time-based heldout prediction, but it is not the best-performing method on this metric. However, the held-out perplexity is just one dimension of performance; our model still performs much better on the other five minority topic recovery metrics.

Table 3: Core MH-focused performance on synthetic data. For JSD, Heldout Perplexity, and Temporal MAE, lower is better (↓); for all other metrics higher is better (↑). The best result in each column is **bold**. For each score stars denote our method differs from the compared method statistically significantly (** denotes p-value < 0.01, * p-value < 0.05).

Model	NMI ↑		Purity ↑		JSD ↓		Trend Correlat. ↑		Heldout Perplexity ↓		Temporal MAE ↓	
LDA	0.157±0.043	**	0.479±0.044	**	0.175±0.030	**	0.694±0.046	**	77.324±3.795	**	0.069±0.018	**
DTM	0.158±0.029	**	0.445±0.029	**	0.231±0.018	**	0.728±0.021	**	72.051±1.193	**	0.068±0.014	**
ConstraintNMF	0.160±0.033	**	0.502±0.031	**	0.115±0.020	**	0.769±0.033	**	**70.918±0.780**	**	0.043±0.008	**
STM	0.217±0.036	*	0.560±0.042	**	0.139±0.026	**	0.771±0.041	**	75.956±1.268	**	0.053±0.006	**
Ours	**0.296±0.028**		**0.670±0.020**		**0.026±0.002**		**0.868±0.010**		72.654±0.367		**0.0038±0.0001**	

5.1.5 Statistical Significance Test (Wilcoxon Test)

In order to verify whether the differences between methods in Table 3 are statistically significant or not, we used paired Wilcoxon signed-rank tests. For all these metrics, our method is compared to each baseline on the seed-matched runs. For those metrics for which lower values are better (JSD, Temporal MAE, and Heldout Perplexity), we changed the sign of the paired differences so that our model shows better performance for larger values.

Our model outperforms all four baselines on the five MH-focused recovery metrics: NMI, Purity, JSD, Trend Correlation, and Temporal MAE. These results are statistically significant in nearly all cases. The majority of the reported p-values are 0.0078, showing a highly consistent significance pattern throughout the comparisons. The only bigger p-value among the five metrics is for NMI against STM, where $p = 0.0156$. Although this value is still statistically significant, it is less so compared to the other values.

Another interesting finding is the size of these mean differences, which is quite large. For instance, our model outperforms all baselines in terms of NMI by up to 0.079 and 0.139. Similarly, for Purity, our model outperforms all baselines by between 0.110 and 0.225. The same applies to temporal metrics, where our model outperforms all baselines on Trend Correlation by 0.098 and 0.174, and on Temporal MAE by 0.039 and 0.066 after direction correction. The same trend holds for JSD, although our model outperforms DTM by the highest margin and ConstraintNMF by the lowest.

For Heldout Perplexity, the pattern is varied. Our model outperforms LDA and STM significantly, and there is no difference between our method and DTM. However, after direction correction for the lower-is-better metric , the results show that our model underperforms ConstraintNMF. The mean differences are +4.670 against LDA ($p = 0.0156$), -0.603 against DTM ($p = 0.4609$), +3.303 against STM ($p = 0.0078$), and -1.735 against ConstraintNMF ($p = 0.0156$).

5.1.6 In-depth Analysis: Per-Topic Temporal Trajectories

While we reported overall correlation between estimated and ground truth minority topic trajectories for our model and the baselines in the core comparison Table 3, here we examine the correlation in more detail.

In the synthetic data we generated four minority topics, here denoted Anxiety, Mental Health, Depression, and Suicide. For each true minority topic, in Table 4 we report two Pearson correlation coefficients. The first correlation coefficient r_t is the Pearson correlation between the ground-truth mean prevalence trajectory across time bins and the corresponding inferred trajectory. The second correlation coefficient r_d is the Pearson correlation between the ground-truth document-level topic weights and the inferred topic weights across documents. The first one evaluates temporal tracking while the second measure evaluates topic identity. As stated before, this difference is critical since there are cases where a model may correlate well with the ground-truth trajectory while not being able to recover the true minority topic.

Figure 3 plots the overall values of the two measures (mean and standard deviation over the topics) for the compared methods; for both measures higher is better. The results clearly demonstrate that our model performs better compared to the rest in terms of consistency across the four minority topics. For these four minority trends (Mental Health, Depression, and Suicide), our model outperforms the other methods with respect to both r_t and r_d.

Table 4: Per-topic comparison of temporal tracking and topic identity recovery, mean ± std across seeds. In each cell, the first line is r_t and the second line is r_d.

Topic	Ours	LDA	DTM	STM	ConstraintNMF
Anxiety	0.76 ± 0.03 0.77 ± 0.01	0.81 ± 0.04 0.13 ± 0.07	0.83 ± 0.02 0.12 ± 0.05	0.86 ± 0.01 0.17 ± 0.02	0.85 ± 0.02 0.16 ± 0.15
Mental Health	0.86 ± 0.01 0.85 ± 0.00	0.76 ± 0.07 0.09 ± 0.08	0.71 ± 0.04 0.05 ± 0.08	0.78 ± 0.04 0.10 ± 0.03	0.71 ± 0.04 0.11 ± 0.22
Depression	0.91 ± 0.01 0.88 ± 0.01	0.53 ± 0.07 0.07 ± 0.18	0.57 ± 0.04 0.07 ± 0.11	0.68 ± 0.11 0.60 ± 0.31	0.73 ± 0.11 0.78 ± 0.06
Suicide	0.94 ± 0.01 0.90 ± 0.01	0.68 ± 0.09 0.30 ± 0.26	0.80 ± 0.07 0.41 ± 0.30	0.76 ± 0.05 0.18 ± 0.19	0.79 ± 0.06 0.15 ± 0.28

Looking at the comparative scores for each individual topic (full results omitted for brevity) supports the good performance of our method. With regards to Anxiety, although some baseline methods perform better in terms of r_t, they lag behind in terms of r_d. For example, for the Anxiety trend, we obtain an r_t of 0.76 and an r_d of 0.77 for our model, while the values obtained by STM and ConstraintNMF are 0.86 and 0.17; 0.85 and 0.16 respectively. The above demonstrates how baselines are able to replicate the trajectory without identifying the true minority topic correctly. The above trend repeats itself when comparing temporal versus identity correlation. Only our model is able to maintain high values in both cases.

Figure 4 shows for our model the estimated topic activity of the best-matching topic from the Hungarian algorithm matching for each ground-truth topic, as well as the corresponding per-topic scores of r_t and r_d. The estimated activities from our model follow the ground truth overall very well.

As an additional qualitative check, we visualize how well the matched minority topics recover both the temporal patterns and the document-level topic identity. Figure 3 compares temporal tracking correlation (r_t) with document-level identity correlation (r_d). Figure 4 shows the normalized temporal trajectories of the four matched minority topics together with the corresponding ground-truth trajectories.

5.1.7 Distributional Divergence Over Time

Figure 5 shows the Jensen–Shannon divergence (JSD) between the probability distribution of ground truth minority topics and the distribution of the estimated minority topics (after Hungarian algorithm matching), for our model and for the comparison models, at each time bin $T = 36$ on a log scale; smaller is better. This gives a more detailed view of how distributional recovery changes over time for each model. Our model keeps JSD in the range $[0.001, 0.08]$ across the full sequence, with most values lying around 0.01–0.02. The minimum value is close to 0.001 near $t = 4$, which shows that the model can recover the target distribution very closely at some time points. By comparison, all baselines remain clearly higher over most of the sequence, with larger variation from one time bin to the next. The baseline trajectories also show different failure patterns. STM (grey diamond dashed) has a pronounced spike around $t = 23$–28, which overlaps with the period where the Suicide topic changes most sharply. DTM (orange triangles) shows higher divergence in the early time bins ($t = 0$–5), while LDA (blue dashed)remains consistently high across the whole sequence. ConstraintNMF stays below the other baselines for much of the time bins, but it is still clearly above our model at nearly all time points. The log-scale view is useful here because it makes the separation between methods easier to see. While the average JSD statistics (mean JSD 0.149–0.290) already favor our model, the per-time-bin plot shows more clearly that this advantage is present throughout the full time horizon rather than only at a few isolated points.

5.1.8 Ablation Study

We perform ablation studies to understand the contribution of each modeling component to the performance of our model while keeping the dataset, optimizer, and evaluation fixed. Each ablated model turns off one component of our model. The compared settings are as follows: **Full**: all components active. **No Seed**: the seed-mass constraint is disabled by setting $\tau \to 0$ (equivalently, $\tau_{\text{scale}} = 0$ in our sensitivity parameterization below). **No Prev-Time**: temporal prevalence covariates are removed, so the prevalence model contains only an intercept. **No Content-Time**:

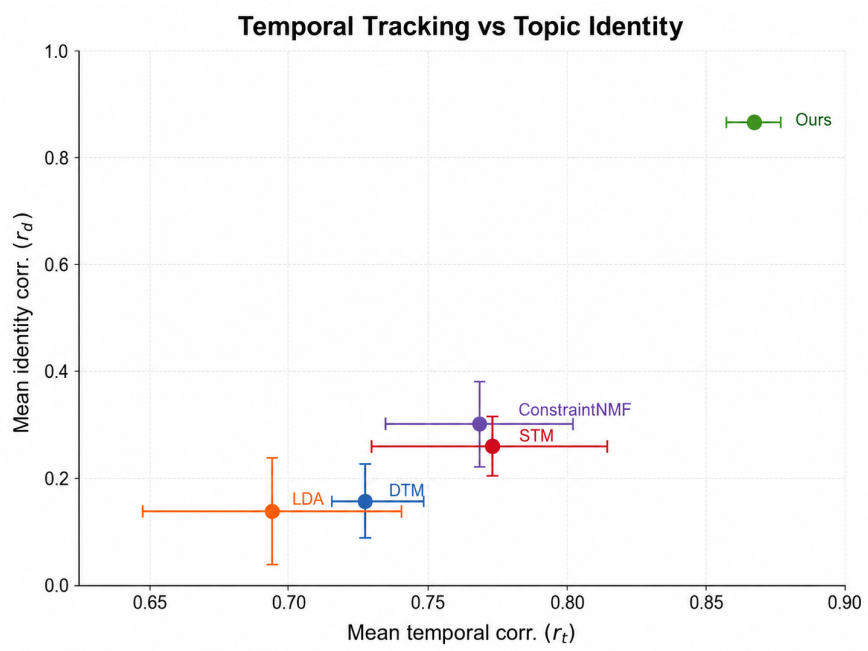

Figure 3: Qualitative comparison of temporal tracking and document-level topic identity for the matched minority topics on synthetic data. The value r_t is the Pearson correlation between the inferred and ground-truth topic prevalence trajectories over time. The value r_d is the Pearson correlation between inferred and ground-truth document-level topic proportions. Points show mean values over the four minority topics and over random seeds, and whiskers show (± 1) standard deviation over the four minority topics and random seeds.

temporal content coupling is removed in topic-word trajectories, and each time slice is regularized independently with a zero-mean Gaussian prior instead of being linked over time. **No** a: minority-topic rescaling is removed ($a \equiv 1$).

Implementation note. For ablation and sensitivity analyses, we parameterize the seed-mass threshold as $\tau_k = \tau_{\text{scale}} \cdot \tau$ for $k \in \mathcal{K}_{\text{min}}$, where $\tau = 0.20$ is the default seed-mass target from Section 5.1.1. Setting $\tau_{\text{scale}} = 0$ disables the seed constraint entirely (the "No Seed" ablation), while $\tau_{\text{scale}} = 1$ recovers the default. We sweep $\tau_{\text{scale}} \in \{0.1, 0.3, 0.5, 0.9\}$ in the sensitivity analysis.

Figure 6 shows that the full model gives the best overall minority-topic trends performance. Removing the seed constraints causes the largest drop, especially for NMI, Purity, Trend Correlation, and JSD. This shows that the seed constraint part of the model is important for finding the target minority trends. Removing the document-level rescaling also reduces performance clearly, which suggests that this component helps the model separate minority content more effectively.

By contrast, removing temporal prevalence covariates or temporal content coupling has a smaller effect. Both ablations remain closer to the full model, although performance still drops compared with the complete setup. This proves that the temporal components help, but their contribution is weaker than the contribution of other factors in this synthetic setting.

5.1.9 Sensitivity and Robustness

We examine the sensitivity of our model by varying one parameter at a time while keeping the rest of the parameters fixed. To this end, there are two hyperparameters Q_β and τ_{scale} in this process as main hyperparameters, plus the number of variational EM iterations during inference n_{iter}. As shown in Table 5, in the case of the number of iterations, gradually all the metrics improve. In the case of Q_β, the effect is small. NMI and Purity remain almost the same across the range,while increasing Q_β slightly improves perplexity but reduces Trend Correlation. Again, the strongest effect is seen in τ_{scale}. At $\tau_{\text{scale}} = 0.1$, the performance of the six metrics is clearly worse. At around 0.3 or above, the results

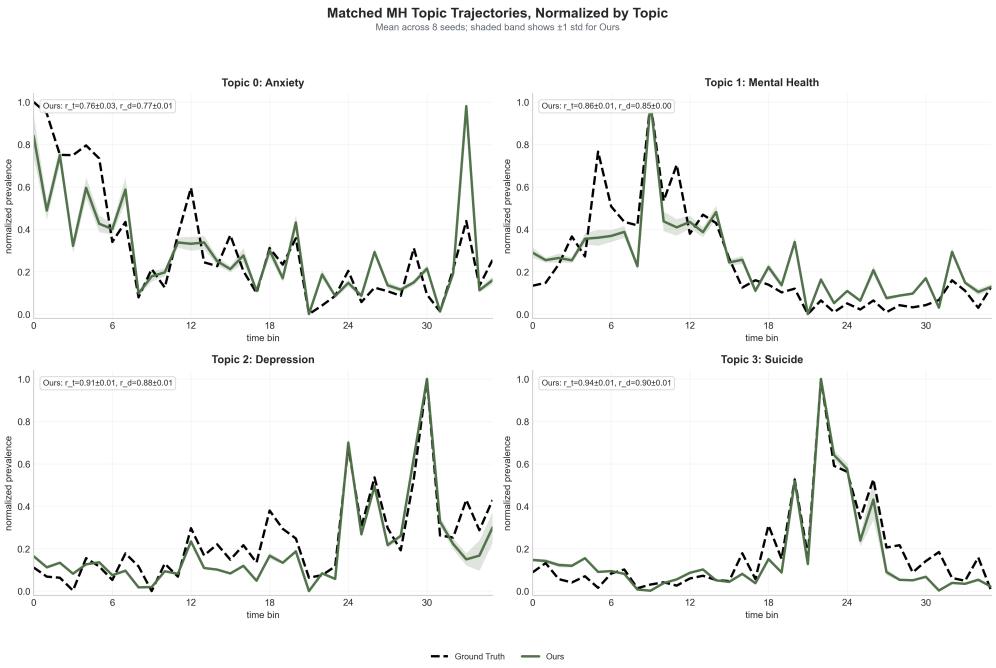

Figure 4: Normalized temporal trajectories of the four matched minority topics on synthetic data. Each panel compares the ground-truth trajectory with the inferred trajectory of the best-matching topic. Solid lines show the mean over random seeds, and shaded areas show (±1) standard deviation over random seeds.

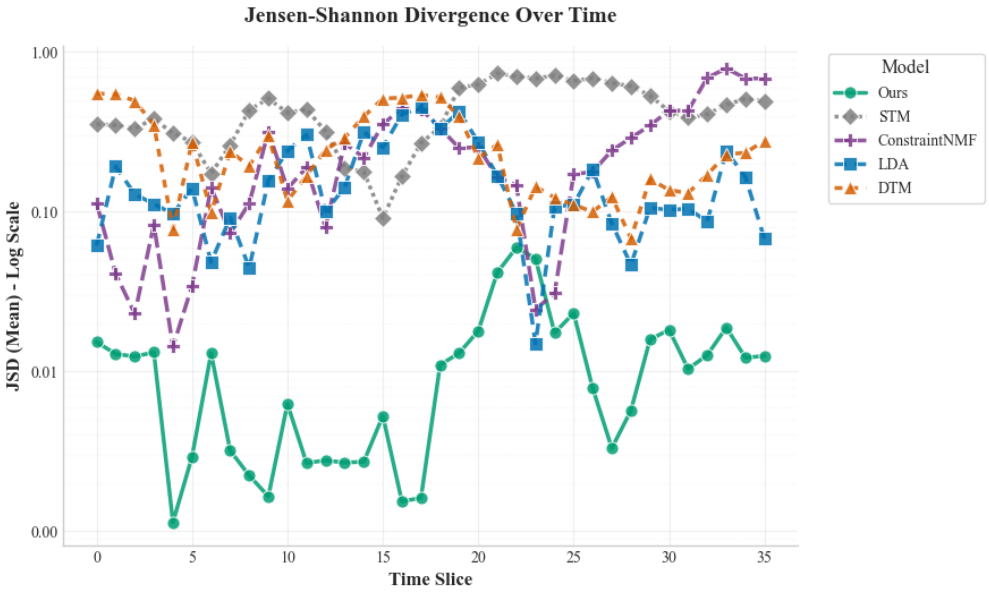

Figure 5: Jensen–Shannon divergence per time bin on a log scale across all 36 time bins.

converge. The best values are between 0.3 and 0.5. It is clear that strong seed constraints play a big role in this case, while the model is not sensitive to changes in Q_β.

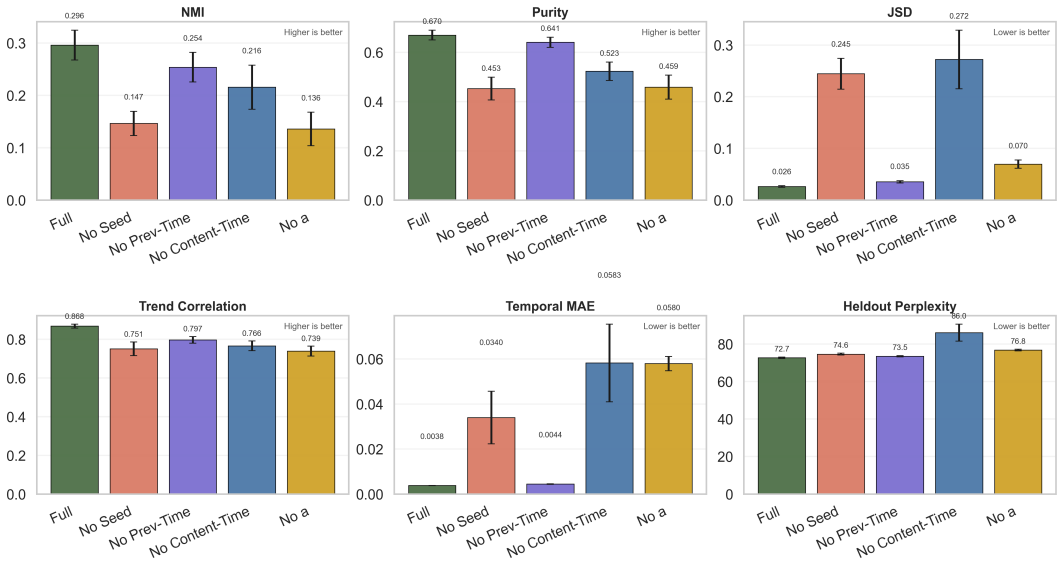

Figure 6: Ablation results across several random seeds. Error bars show standard deviations.

Table 5: Sensitivity analysis, mean ± std across eight random seeds.

Parameter	Value	NMI ↑	Purity ↑	JSD ↓	TrendCorr ↑	Heldout PPL ↓	Temporal MAE ↓
n_{iter}	20	0.2280 ± 0.0254	0.5939 ± 0.0424	0.1361 ± 0.1226	0.8031 ± 0.0168	98.2532 ± 0.2125	0.0116 ± 0.0084
	40	0.2963 ± 0.0284	0.6702 ± 0.0197	0.0264 ± 0.0015	0.8685 ± 0.0103	72.6542 ± 0.3677	0.0038 ± 0.0001
	60	0.2883 ± 0.0197	0.6686 ± 0.0160	0.0409 ± 0.0573	0.8950 ± 0.0117	62.1016 ± 1.1717	0.0047 ± 0.0043
	80	0.3108 ± 0.0154	0.6827 ± 0.0092	0.0188 ± 0.0004	0.9085 ± 0.0029	55.7799 ± 1.2197	0.0029 ± 0.0001
Q_β	0.3	0.3071 ± 0.0332	0.6736 ± 0.0235	0.0523 ± 0.0753	0.8883 ± 0.0172	75.5605 ± 0.5260	0.0059 ± 0.0059
	1.0	0.2963 ± 0.0284	0.6702 ± 0.0197	0.0264 ± 0.0015	0.8685 ± 0.0103	72.6542 ± 0.3677	0.0038 ± 0.0001
	3.0	0.2931 ± 0.0252	0.6680 ± 0.0193	0.0279 ± 0.0015	0.8526 ± 0.0111	71.1809 ± 0.2830	0.0039 ± 0.0001
	10.0	0.2967 ± 0.0329	0.6708 ± 0.0256	0.0287 ± 0.0012	0.8443 ± 0.0115	70.3041 ± 0.2575	0.0039 ± 0.0001
τ_{scale}	0.1	0.2721 ± 0.0373	0.6324 ± 0.0382	0.1321 ± 0.1058	0.8111 ± 0.0429	73.1609 ± 0.4089	0.0140 ± 0.0117
	0.3	0.3066 ± 0.0264	0.6776 ± 0.0206	0.0248 ± 0.0019	0.8671 ± 0.0158	72.5234 ± 0.3150	0.0035 ± 0.0002
	0.5	0.3061 ± 0.0299	0.6782 ± 0.0233	0.0240 ± 0.0016	0.8754 ± 0.0123	72.3816 ± 0.3079	0.0035 ± 0.0001
	0.9	0.2917 ± 0.0298	0.6669 ± 0.0211	0.0256 ± 0.0015	0.8715 ± 0.0104	72.5592 ± 0.3784	0.0037 ± 0.0001

5.1.10 Robustness Analysis

We also test robustness of our model with respect to three other factors: token noise, data reduction, and removing time bins. Token noise means we randomly replace a fraction of the tokens in each document with tokens sampled uniformly from the vocabulary. Data reduction means we reduce the size of the data set by uniform subsampling, keeping a fraction of the documents. For removal of time bins, we uniformly sample a fraction of the time bins to remove and discard all documents belonging to those bins. Together, these experiments evaluate the sensitivity of the model to noisy text, reduced corpus size, and incomplete temporal coverage.

Refer to Table 6 for results. For token noise, all six metrics worsen with increased noise. NMI and Purity decrease continuously, JSD and Temporal MAE worsen, TrendCorr drops from 0.8685 to 0.8341, and Heldout perplexity (Heldout PPL) increases dramatically from 72.65 to 154.21. In the case of Data reduction, temporal recovery and stability change the most. At 25% data, TrendCorr drops to 0.7065, JSD increases to 0.1725, Temporal MAE increases to 0.0347, and the standard deviations become much larger. This means that while the model is still capable of recovering minority topic structure with subsampled data, temporal recovery is much less stable with reduced data. The least impactful factor is Removing time bins. TrendCorr remains steady between 0.844 and 0.873. Heldout PPL, NMI and Purity metrics have moderate changes, but the range increases with more removed time bins. The model is quite robust overall, but the results clearly show that heavy token corruption and data reduction have more impact than moderate irregularity in temporal coverage.

Table 6: Robustness analysis, mean ± std across eight random seeds.

Factor	Level	NMI ↑	Purity ↑	JSD ↓	TrendCorr ↑	Heldout PPL ↓	Temporal MAE ↓
Token noise	0.00	0.2963 ± 0.0284	0.6702 ± 0.0197	0.0264 ± 0.0015	0.8685 ± 0.0103	72.6542 ± 0.3677	0.0038 ± 0.0001
	0.05	0.2783 ± 0.0297	0.6533 ± 0.0234	0.0552 ± 0.0754	0.8579 ± 0.0127	89.1631 ± 0.8002	0.0060 ± 0.0057
	0.10	0.2713 ± 0.0264	0.6442 ± 0.0322	0.1082 ± 0.1089	0.8484 ± 0.0175	108.2023 ± 1.3234	0.0102 ± 0.0084
	0.20	0.2434 ± 0.0313	0.6103 ± 0.0413	0.1752 ± 0.1165	0.8341 ± 0.0261	154.2103 ± 2.0862	0.0149 ± 0.0086
Data fraction	0.25	0.2853 ± 0.0796	0.5818 ± 0.0643	0.1725 ± 0.1222	0.7065 ± 0.0554	81.9204 ± 13.1241	0.0347 ± 0.0353
	0.50	0.2860 ± 0.0337	0.6283 ± 0.0268	0.1698 ± 0.1083	0.7927 ± 0.0680	82.7910 ± 1.0850	0.0179 ± 0.0126
	0.75	0.3130 ± 0.0596	0.6535 ± 0.0370	0.1704 ± 0.1176	0.8441 ± 0.0381	76.8922 ± 0.8101	0.0186 ± 0.0138
	1.00	0.2963 ± 0.0284	0.6702 ± 0.0197	0.0264 ± 0.0015	0.8685 ± 0.0103	72.6542 ± 0.3677	0.0038 ± 0.0001
Dropped bins	0.00	0.2963 ± 0.0284	0.6702 ± 0.0197	0.0264 ± 0.0015	0.8685 ± 0.0103	72.6542 ± 0.3677	0.0038 ± 0.0001
	0.10	0.2730 ± 0.0216	0.6433 ± 0.0284	0.0816 ± 0.1022	0.8727 ± 0.0235	74.3735 ± 3.1216	0.0079 ± 0.0075
	0.20	0.2661 ± 0.0562	0.6306 ± 0.0556	0.1206 ± 0.1010	0.8519 ± 0.0368	73.7145 ± 3.7621	0.0173 ± 0.0156
	0.30	0.3048 ± 0.0758	0.6565 ± 0.0629	0.1001 ± 0.0979	0.8442 ± 0.0362	75.4381 ± 3.3420	0.0136 ± 0.0141

5.1.11 Effect of Trend Capacity

We then examine how the results change as trend capacity K increases. For settings with $K_{\min} > 4$, we basically follow the same synthetic dataset generation process in Section 5.1.1 for each setting, changing the number of the topics while the vocabulary size, time horizon, and number of documents per time bin are kept fixed. To construct varying numbers of topics each with their temporal dynamics, the minority-topic prevalence patterns are constructed by repeating four temporal shapes across minority trends: U-shape, seasonal, linear, and bell-shaped. Small scaling changes are added so that higher-index minority trends remain active. Table 7 reports the results for four settings: $(40, 10, 30)$, $(60, 15, 45)$, $(80, 20, 60)$, and $(100, 25, 75)$ for $(K, K_{\min}, K_{\text{BG}})$. Several patterns are evident.

Larger topic capacity helps along some axes and hurts along others; the pattern across the four metrics families is clear. NMI rises monotonically with capacity, from 0.2092 ± 0.0157 at $K = 40$ to 0.4473 ± 0.0163 at $K = 100$, and Temporal MAE falls in step, from 0.0170 to 0.0106. With more topics available, the model allocates them in ways that distinguish minority sub-themes more finely and recovers their temporal trajectories at higher resolution.

The remaining metrics tell a different story. Purity, JSD, and heldout perplexity (Heldout PPL) are all best at the smallest setting and worsen monotonically as K grows: JSD increases from 0.0757 to 0.1240, and Heldout PPL nearly triples from 231.5 to 644.5. Since there are more topics to model and hence more alternative explanations for observations, accurately modeling probabilities of the individual minority topics becomes harder; on the other hand with increasing $K_{\min}$ less of the total topics are devoted to the majority content. Both behaviors should be expected consequences of the scaling behavior of our model.

The TrendCorr is not monotonic; its peaks occur at $K = 40$ (0.7171) and $K = 60$ (0.7021), while it decreases at $K = 80$ (0.6376), but increases slightly again at $K = 100$ (0.6580). The decrease overall again indicates that with a greater variety of topics modeling the presence, and hence temporal trends, of the individual topics becomes harder. The purity is also non-monotonic, fluctuating between a small range (0.30–0.38) and achieving the peak value at $K = 100$ (value = 0.3795). Unlike the constant growth in NMI, the almost flat purity profile suggests that for large values of K, while the model may represent several variations (sub-topics) of the minority topics, the Hungarian algorithm matching can identify some variation to match each of the ground truth topics well, leading to subtle improvements reflected more strongly in NMI than in purity.

This pattern shows a well-designed trade-off, not an inconsistency. NMI and Temporal MAE give preference to separation and temporal accuracy; the larger value of K is thus preferable. On the other hand, JSD and Heldout PPL correlate positively with distributional accuracy and predictability. Therefore, they are more advantageous when K is small. The value of TrendCorr characterizes trajectory consistency and achieves its maximum value in case of minimum K. Thus, depending on how important the properties above are for the task at hand, K and $K_{\min}$ may differ considerably.

5.2 Experiments on Real Data

5.2.1 Real-Data Source and Preprocessing

In the real-data experiment, we use Finnish YouTube comments as a noisy social media setting where MH-related discussion is expected to appear only occasionally. Most comments are ordinary reactions to the videos or the YouTubers, including support, appreciation, jokes, short replies, and sometimes trolling. The MH-related signal is therefore mixed

Table 7: Effect of trend capacity K settings on Ours performance. Values are mean ± std over 8 random seeds. Bold indicates the best setting for each metric. Across all four configurations the seed input follows the per-topic regime described in Section 5.1.1, and the prevalence covariate basis matches the main comparison.

Metric	K40/MH10	K60/MH15	K80/MH20	K100/MH25
NMI (↑)	0.2092 ± 0.0157	0.2805 ± 0.0130	0.3352 ± 0.0207	**0.4473 ± 0.0163**
Purity (↑)	0.3303 ± 0.0179	0.3196 ± 0.0102	0.2998 ± 0.0220	**0.3795 ± 0.0154**
JSD (↓)	**0.0757 ± 0.0018**	0.0970 ± 0.0024	0.1066 ± 0.0014	0.1240 ± 0.0008
TrendCorr (↑)	**0.7171 ± 0.0111**	0.7021 ± 0.0130	0.6376 ± 0.0117	0.6580 ± 0.0117
Heldout PPL[1] (↓)	**231.47 ± 0.60**	395.51 ± 0.04	542.17 ± 0.05	644.51 ± 0.05
TemporalMAE (↓)	0.0170 ± 0.0002	0.0139 ± 0.0001	0.0116 ± 0.0001	**0.0106 ± 0.0000**

with a much larger amount of general viewer interaction. This experiment tests whether the model can still recover interpretable MH-related minority topics in such data. We evaluate the model on one full real-world corpus together with several derived real-data datasets constructed from that corpus. For the real-data analysis, the domain of interest is mental health (MH). We use "minority topic" and "MH topic" interchangeably throughout this section.

The corpus contains Finnish-language YouTube discussion: comments in the comments sections of YouTube videos from channels of 17 mental health relevant Finnish vloggers recommended by a public health expert; their channels occasionally have mental health content for younger audiences. Mental health is expected to be a minority domain in the comments, with the majority content being generic comment responses typical among followers and viewers of vloggers. In addition to the main corpus we derive several datasets from it by partitioning. The derived datasets correspond to temporal partitions (Pre-COVID and Post-COVID, chosen as a partition due to the impact of the COVID-19 pandemic on online participation) and channel-owner gender partitions (Male, Female, and family, chosen as a partition due to potentially reaching different audiences and hence yielding differing discussion). A seed word list of words relevant to mental health was provided in collaboration with a public health expert, note that this list was an overall list for the domain and does not presuppose any individual topics. Each dataset is modeled independently.

The raw real corpus contains 5,737,253 comments (top-level comments and replies) collected from 17 channels and 10,210 unique videos, covering 2011-02-23 to 2026-01-04. After lemmatization, deduplication, and normalization, this becomes 1,669,052 text rows with 11,135,472 tokens (mean length 6.67, median 4). We take all comments (either direct comments under a vlog or replies to another comment) as documents. In the full corpus, the reply rate (proportion of comments receiving replies) is 16.85%, the mean number of likes per comment is 4.10 (median 0), and 2.92% of documents contain at least one seed term.

To create the training set, we build a vocabulary of 30,000 words from the lemmatized corpus, remove the 100 most frequent non-informative words, remove words that appear in fewer than 5 documents, and then discard comments that become too short after filtering. The resulting training corpus has 835,258 documents and 3,297,848 tokens (mean length 3.95, median 3), divided into 36 time bins.

In results on this sensitive-domain data we report only aggregate corpus statistics; no individual-level user data, comment text, or identifying information is shared.

5.2.2 Temporal Coverage and Binning

To reduce temporal sparsity effects in training, we use two parallel time views: (i) calendar months for descriptive reporting, and (ii) balanced quantile-based time bins for modeling. The balanced bins enforce near-uniform document counts per slice, improving stability of temporal parameter estimation. Figure 7 shows document volume over time under both calendar-month aggregation and balanced modeling bins. The calendar-month view highlights substantial natural variation in activity (including low-volume periods), while the balanced-bin view confirms that the training setup controls sample-size imbalance across time slices. This separation is important: calendar-month counts are used for interpretability and reporting, whereas balanced bins are used for robust temporal topic estimation.

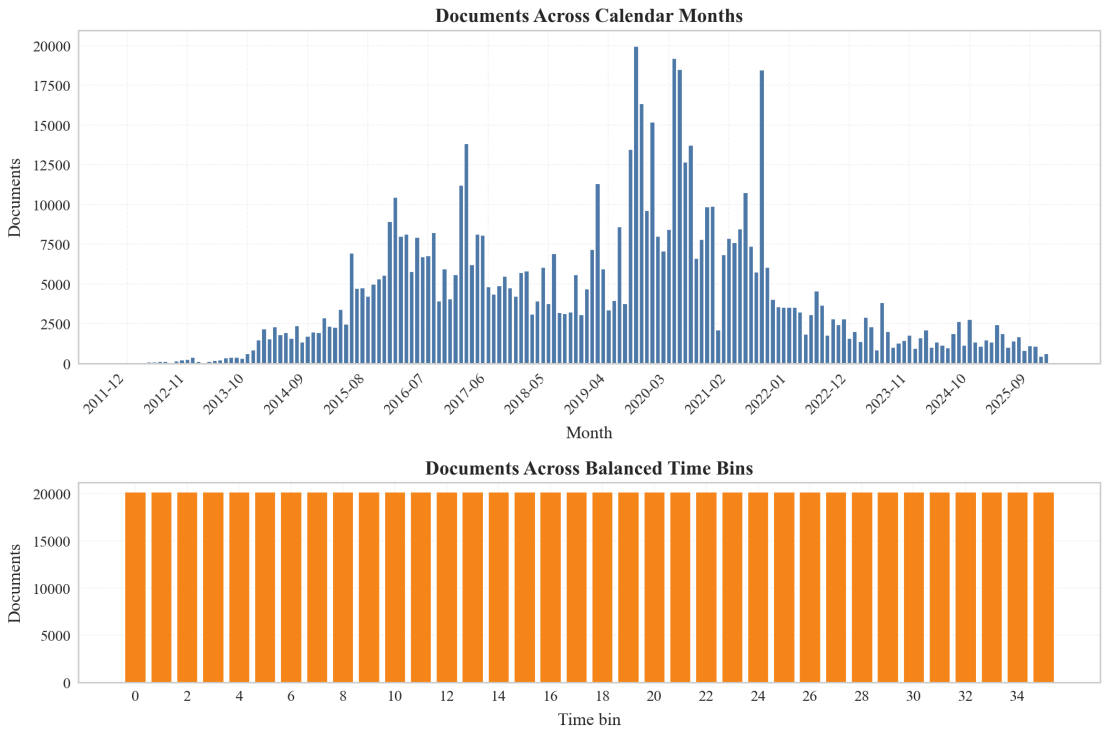

Figure 7: Real data document volume across calendar months (top) and balanced modeling bins (bottom).

5.2.3 Results on the Full Data Set

On the full real-data corpus, we visualize the estimated prevalence of the first five minority topics across the balanced modeling bins. This figure is used as a descriptive view of the discovered topics over time.

Figure 8 shows the estimated prevalence of the five minority topics across the modeling bins. The topics share some broad movement over time, but their trajectories are not identical. Some topics increase more clearly in later bins, while others show smaller or more gradual changes. This gives a qualitative indication that the model captures both a general MH-related signal and differences between the discovered minority topics.

Table 8 shows the word content (top most probable words) for the estimated minority topics and how they evolve over time: word lists are shown at six representative time bins for each topic. The table shows words translated to English; a corresponding table with the original Finnish words is available in the Appendix. The table shows that our model has been able to find topics with meaningful and differing mental health content, and moreover has found a gradual drift of the topic content for each topic. For example Topic 00 consistently contains words related to feelings, dreams, and happiness and joy versus fear, hatred and worry; but the specific emphasis among these varies over time as is natural.

5.2.4 Results on the Derived Datasets

For the real-data analysis, we use the same full corpus under two comparison settings. The first partition is temporal: comments are divided into pre-COVID and post-COVID periods using 16 March 2020 as the cutoff date, corresponding to Finland's official declaration of a state of emergency over COVID-19. As stated in Section 5.2.1, this is a relevant partition for study since COVID-19 had major impact in online participation and it could also have impact on mental health visible in the discussion. The second is a channel-owner gender partition, where comments are grouped according to whether they come from a channel owned by *male*, *female*, and *family* creators. Table 9 summarizes the size of each subset. To keep the modeling setup comparable, the pre-COVID, post-COVID, male, and female models were each sampled to 200,000 comments before preprocessing; vocabulary filtering and minimum-length filtering reduced these to between 105k and 117k documents (Table 9). The family subset used all available comments since the source pool was much smaller, yielding 46k documents after filtering.

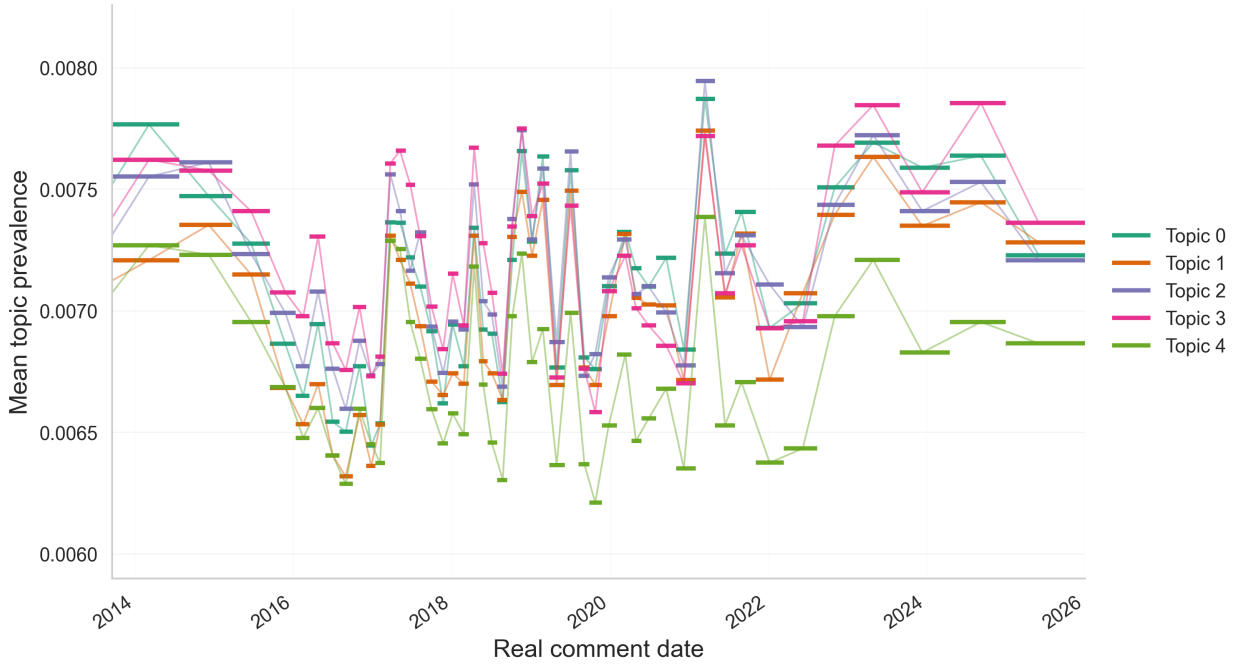

Figure 8: Estimated prevalence of the first five minority topics in the full real-data model across balanced modeling bins. Each line shows one minority topic discovered by the model. Topic indices are specific to this full-corpus model.

We will show topic prevalence plots for the found minority topics in each data division. Figure 9 shows the plots which are discussed below for each division. To complement the yearly plots, Table 10 summarizes selected representative words from the five minority topics in each subset model, along with their English translations. Full lists of topic words over topics, divisions and time points are omitted for brevity. The representative words are used only for qualitative interpretation. They help show which MH-related themes are more visible in each subset, but they are not used as a direct cross-model topic alignment. The topic word profiles found in each subset emphasize different themes. In the pre-COVID subset, top minority-topic words centre on feelings, worry, self-esteem, and bullying. In the post-COVID subset, the most prominent minority-topic words shift toward addiction, sleep, doctors, support, and care. The male subset discovers vocabulary around fear, crying, violence, drugs, and treatment, while the female-channel subset emphasizes fear, appearance, self-esteem, and social experience. The family subset is more health- and care-oriented, with words such as illness, pain, depression, therapy, treatment, and love. Because the subset models are fit independently, these descriptions reflect the most salient themes within each subset model, not a quantitative comparison of MH discourse across subsets.

COVID-Period Datasets Figure 9, panels (d) and (e), shows yearly real-time topic trends for the first five minority topics in the pre-COVID and post-COVID models. Because each subset was modeled separately, the topic numbers only have meaning inside that subset. This means that Topic 0 in one subset is not automatically the same as Topic 0 in another subset. For this reason, we compare the general trend patterns and the main topic words across subsets, rather than matching topics by number. For the COVID split, both subsets show fairly stable MH-topic prevalence across most years, but the post-COVID model is slightly higher in 2023–2025. The sharp endpoint increase in 2026 should not be over-interpreted because it is based on only 13 documents in the post-COVID subset.

Channel-Owner Gender Datasets Figure 9 panels (a), (b), (c) show yearly real-time topic trends for the first five minority topics in the male-owned, female-owned, and family-channel datasets. The three channel-owner groups show different MH-related patterns. As in the COVID-period analysis, each subset is modeled independently, so topic indices are only meaningful within the corresponding subset model. The male-owned and female-owned channel datasets differ in their trends and topic words. The family-channel dataset is smaller and contains more health- and care-related themes.

Table 8: Evolution of estimated minority (mental health) topic content over time in the full real data set. For each topic, a list of the most probable words are listed (translated to English) for six representative time bins; for each bin the bin number and corresponding time period are shown.

Topic	Time bin	Top words
Topic 00	t00 [2011-12–2014-07]	to listen, feeling, dream, illness attack, joy, fear, envious, disappointment
	t06 [2016-05–2016-08]	book, feeling, to feel, have to, to send, to hope, crazy, conversation
	t12 [2017-11–2018-03]	problem, to feel, happy, to take care of, to hate, hatred, to fear, annoying
	t19 [2019-10–2019-11]	feeling, to send, crazy, book, to hope, to fear, to feel, difference/divorce
	t25 [2020-07–2020-10]	700k, crazy, book, to feel, to send, psychopath, variety, hours
	t35 [2024-10–2026-01]	to hope, book, to rest, drug, happy, man, outside, worry
Topic 01	t00 [2011-12–2014-07]	man, annoying, dream, feeling, to fear, illness attack, joy, drug
	t06 [2016-05–2016-08]	man, friend, dream, letter, book, to send, crazy, hour
	t12 [2017-11–2018-03]	to hope, crazy, joy, book, difference/divorce, happy, love, to listen
	t19 [2019-10–2019-11]	to feel, difference/divorce, hours, joy, man, to fear, to listen, book
	t25 [2020-07–2020-10]	feeling, 700k, man, problem, book, to feel, hours, villager
	t35 [2024-10–2026-01]	horror, (amount of) horror, (amount of) horror, friend, hours, problem, fox, have to
Topic 02	t00 [2011-12–2014-07]	crazy, feeling, to fear, dream, joy, illness attack, disappointment, envious
	t06 [2016-05–2016-08]	to hope, problem, book, to listen, to send, disappointment, variety, crazy
	t12 [2017-11–2018-03]	have to, to send, to feel, to hope, to hate, book, difference/divorce, crazy
	t19 [2019-10–2019-11]	have to, to feel, book, man, friend, to send, to hope, advertising income
	t25 [2020-07–2020-10]	700k, have to, cry, to hope, hours, doctor, book, to listen
	t35 [2024-10–2026-01]	fox, joy, have to, to hope, sad, anxiety-inducing, man, cry
Topic 03	t00 [2011-12–2014-07]	to fear, feeling, dream, illness attack, envious, joy, disappointment, difference/divorce
	t06 [2016-05–2016-08]	have to, nightmare, happy, to send, to hope, love, distress, variety
	t12 [2017-11–2018-03]	feeling, horror, to hope, sad, to listen, happy, difference/divorce, I hate
	t19 [2019-10–2019-11]	book, hatred, to hope, horror, stress, worry, sad, variety
	t25 [2020-07–2020-10]	700k, to hope, sad, I hate, to fear, hours, positive, annoying
	t35 [2024-10–2026-01]	bara, bastu, dream, bada, sauna, paralysis, ooooh, som
Topic 04	t00 [2011-12–2014-07]	have to, annoying, dream, feeling, illness attack, joy, to fear, outside
	t06 [2016-05–2016-08]	to send, annoying, trauma, sad, difference/divorce, crazy, horror, hope
	t12 [2017-11–2018-03]	I hate, to send, :-d, man, cry, positive, sad, to listen
	t19 [2019-10–2019-11]	to hope, book, distress, nightmare, happy, sad, dream, to fear
	t25 [2020-07–2020-10]	700k, negative, joy, thought, to fear, happy, doctor, friend
	t35 [2024-10–2026-01]	to hope, crazy, paralysis, feeling, happy, to fear, fox, have to

Table 9: Real-data subsets used in the comparative analysis. Comment rows count both top-level comments and replies. Docs refer to the number of documents retained after sampling, vocabulary filtering, and minimum-length filtering.

Subset	Comment rows	Videos	Channels	Docs
Pre-COVID	3,981,193	6,302	17	108,852
Post-COVID	1,756,060	8,560	17	105,617
Male	4,652,532	6,425	7	105,715
Female	1,003,971	3,091	8	116,216
Family	80,750	689	2	46,488

6 Discussion and Conclusions

We introduced a solution for probabilistic constrained topic modeling in a setting where the target topics/trends are low-prevalence, mixed with dominant background, and evolving over time. Our proposed model DRIFT integrated

Table 10: Representative words from the five minority topics in each subset model. Because the subset models were trained separately, the table is descriptive rather than a direct topic-to-topic comparison.

Subset	Representative MH words	English meanings
Pre-COVID	tunne, pelko, huoli, itsetunto, kiusaaminen, surullinen, kiusattu	feelings, worry, self-esteem, bullying
Post-COVID	addiktio, uni, lääkäri, tuki, hoitaa, kriisi, rakkaus	addiction, sleep, care, support, crisis
Male	hullu, itku, kauhu, lääkäri, huume, väkivalta, kannabis	fear, crying, treatment, drugs, violence
Female	pelko, kiusattu, ulkonäkö, ajatus, itsetunto, sosiaalinen, positiivinen	fear, appearance, self-esteem, social experience
Family	sairaus, kipu, masennus, huoli, terapia, hoito, rakkaus	illness, pain, depression, therapy, care

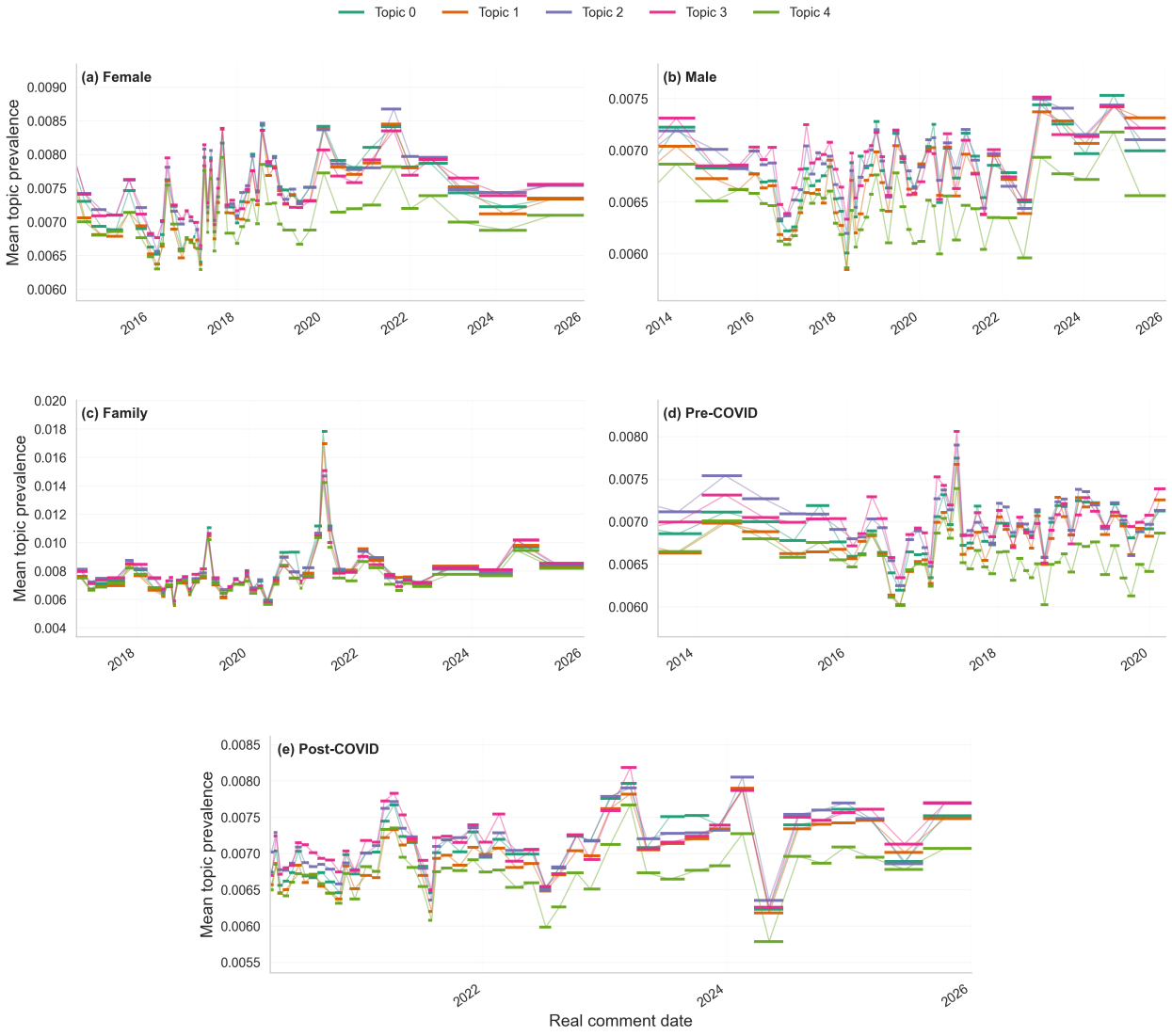

Figure 9: Yearly minority-topic trends across the five derived datasets: (a) female-owned channels, (b) male-owned channels, (c) family channels, (d) pre-COVID, (e) post-COVID. Each line shows one of the five minority topics from the separately fitted subset model; topic indices are subset-specific and are not comparable across panels.

covariate-driven prevalence, evolving topic content, and seed-based constraints into a single probabilistic generative framework. On synthetic data, it performed best on minority topic identification and temporal recovery. Qualitatively, the model consistently captured the desired target topics/trends, which was not the case for the competitive models. Additionally, from the ablation studies, we learned that in the synthetic experiment the seed constraints and minority

topic rescaling within each document were the two most critical parts of our method, while the temporal components provide an extra advantage. From the robustness experiments, our model could maintain relative stability under perturbation to some extent, while token-level noise and aggressive data trimming would result in poor performance. While our model was not the best in heldout perplexity it clearly outperformed the comparison methods on all other metrics.

Also in a case study on real YouTube discussion data with minority mental health content, the model recovers interpretable mental health related topics and meaningful temporal differences across subsets. Future directions of study could include applications in other domains, incorporation of alternative temporal priors and alternative forms of seed constraints, and incorporation into interactive systems for social media analysis focusing on minority themes.

Acknowledgments

(Omitted for anonymous submission.)

References

Amr Ahmed and Eric P. Xing. Timeline: A dynamic hierarchical dirichlet process model for recovering birth/death and evolution of topics in text stream. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, pp. 20–29, 2010.

David Andrzejewski, Xiaojin Zhu, and Mark Craven. Incorporating domain knowledge into topic modeling via dirichlet forest priors. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 25–32. ACM, 2009. doi: 10.1145/1553374.1553378.

David M. Blei and John D. Lafferty. Dynamic topic models. In *Proceedings of the 23rd International Conference on Machine Learning*, pp. 113–120. ACM, 2006. doi: 10.1145/1143844.1143859. URL https://www.cs.columbia.edu/~blei/papers/BleiLafferty2006a.pdf.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003. URL https://www.jmlr.org/papers/v3/blei03a.html.

Allaa Boutaleb, Jerome Picault, and Guillaume Grosjean. BERTrend: Neural topic modeling for emerging trends detection. In *Proceedings of the Workshop on the Future of Event Detection (FuturED)*, pp. 1–17. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.futured-1.1. URL https://aclanthology.org/2024.futured-1.1/.

Daniel Brüggermann, Yannik Hermey, Carsten Orth, Darius Schneider, Stefan Selzer, and Gerasimos Spanakis. Storyline detection and tracking using dynamic Latent Dirichlet Allocation. In Tommaso Caselli, Ben Miller, Marieke van Erp, Piek Vossen, and David Caswell (eds.), *Proceedings of the 2nd Workshop on Computing News Storylines (CNS 2016)*, pp. 9–19, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/W16-5702. URL https://aclanthology.org/W16-5702/.

Kostadin Cvejoski, Ramsés J. Sánchez, and César Ojeda. Neural dynamic focused topic model. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):12719–12727, 2023a. doi: 10.1609/aaai.v37i11.26496.

Kostadin Cvejoski, Ramsés J Sánchez, and César Ojeda. Neural dynamic focused topic model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 12719–12727, 2023b.

Mrinal Das, Suparna Bhattacharya, Chiranjib Bhattacharyya, and Gopinath Kanchi. Subtle topic models and discovering subtly manifested software concerns automatically. In Sanjoy Dasgupta and David McAllester (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 253–261, Atlanta, Georgia, USA, June 2013. PMLR. URL https://proceedings.mlr.press/v28/das13.html.

Adji B. Dieng, Francisco J. R. Ruiz, and David M. Blei. The dynamic embedded topic model. *arXiv preprint arXiv:1907.05545*, 2019. doi: 10.48550/arXiv.1907.05545. URL https://arxiv.org/abs/1907.05545.

Adji Bousso Dieng, Francisco J. R. Ruiz, and David M. Blei. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453, 2020. doi: 10.1162/tacl_a_00325. URL https://aclanthology.org/2020.tacl-1.29/.

Seyedeh Fatemeh Ebrahimi and Jaakko Peltonen. Constrained non-negative matrix factorization for guided topic modeling of minority topics. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 35573–35598. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1802. URL https://aclanthology.org/2025.emnlp-main.1802/.

Seyedeh Fatemeh Ebrahimi and Jaakko Peltonen. Nonnegative matrix factorization for joint clustering and topic modeling with minority topics. In Tadahiro Taniguchi, Chi Sing Andrew Leung, Tadashi Kozuno, Junichiro Yoshimoto, Mufti Mahmud, Maryam Doborjeh, and Kenji Doya (eds.), *Neural Information Processing*, pp. 492–506, Singapore, 2026a. Springer Nature Singapore. ISBN 978-981-95-4367-0.

Seyedeh Fatemeh Ebrahimi and Jaakko Peltonen. Voices between lines: Interpretable labeling of mental health minority topics with seed guidance and llms. In Irena Koprinska, João Mendes-Moreira, and Paula Branco (eds.), *Machine Learning and Principles and Practice of Knowledge Discovery in Databases*, pp. 263–279, Cham, 2026b. Springer Nature Switzerland. ISBN 978-3-032-19105-2.

Pankaj Gupta, Yatin Chaudhary, Florian Buettner, and Hinrich Schuetze. texttovec: Deep contextualized neural autoregressive topic models of language with distributed compositional prior. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=rkgoyn09KQ.

Florian Holz and Sven Teresniak. Towards automatic detection and tracking of topic change. In Alexander Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing*, pp. 327–339, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-12116-6.

Liangjie Hong, Dawei Yin, Jian Guo, and Brian D. Davison. Tracking trends: Incorporating term volume into temporal topic models. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 484–492. ACM, 2011. doi: 10.1145/2020408.2020485.

Jagadeesh Jagarlamudi, Hal Daumé III, and Raghavendra Udupa. Incorporating lexical priors into topic models. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 204–213. Association for Computational Linguistics, 2012. URL https://aclanthology.org/E12-1021/.

Patrick Jähnichen, Florian Wenzel, Marius Kloft, and Stephan Mandt. Scalable generalized dynamic topic models. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 1427–1435, 2018. URL https://proceedings.mlr.press/v84/jahnichen18a.html.

Hang Jiang, Doug Beeferman, Weiquan Mao, and Deb Roy. Topic detection and tracking with time-aware document embeddings. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pp. 16293–16303. ELRA and ICCL, 2024. URL https://aclanthology.org/2024.lrec-main.1416/.

Charu Karakkaparambil James, Mayank Nagda, Nooshin Hajighassemi, Marius Kloft, and Sophie Fellenz. Evaluating dynamic topic models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 160–176. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.11. URL https://aclanthology.org/2024.acl-long.11/.

Noriaki Kawamae. Trend analysis model: Trend consists of temporal words, topics, and timestamps. In *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, pp. 317–326. ACM, 2011. doi: 10.1145/1935826.1935880.

Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2): 83–97, 1955.

Jey Han Lau, David Newman, and Timothy Baldwin. 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*, pp. 530–539. Association for Computational Linguistics, 2014. URL https://aclanthology.org/E14-1056/.

Chien Lu, Jaakko Peltonen, Jyrki Nummenmaa, and Kalervo Järvelin. Probabilistic dynamic non-negative group factor model for multi-source text mining. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management*, pp. 1035–1043. ACM, 2020. doi: 10.1145/3340531.3411956.

Xiaoling Lu, Yuxuan Guo, Jiayi Chen, and Feifei Wang. Topic change point detection using a mixed bayesian model. *Data Mining and Knowledge Discovery*, 36(1):146–173, 2022.

Ming Ma, Jin Mao, and Gang Li. Discovering weak signals of emerging topics with a triple-dimensional framework. *Information Processing & Management*, 61(5):103793, 2024. doi: 10.1016/j.ipm.2024.103793.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. *Introduction to information retrieval*. Cambridge University Press, 2008.

Jon Mcauliffe and David Blei. Supervised topic models. In J. Platt, D. Koller, Y. Singer, and S. Roweis (eds.), *Advances in Neural Information Processing Systems*, volume 20. Curran Associates, Inc., 2007. URL https://proceedings.neurips.cc/paper_files/paper/2007/file/d56b9fc4b0f1be8871f5e1c40c0067e7-Paper.pdf.

Yishu Miao, Lei Yu, and Phil Blunsom. Neural variational inference for text processing. In *Proceedings of the 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 1727–1736. PMLR, 2016. URL https://proceedings.mlr.press/v48/miao16.html.

Yishu Miao, Edward Grefenstette, and Phil Blunsom. Discovering discrete latent topics with neural variational inference. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 2410–2419. PMLR, 2017. URL https://proceedings.mlr.press/v70/miao17a.html.

Nozomu Miyamoto, Masaru Isonuma, Sho Takase, Junichiro Mori, and Ichiro Sakata. Dynamic structured neural topic model with self-attention mechanism. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 5916–5930, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.366. URL https://aclanthology.org/2023.findings-acl.366/.

Ramesh M. Nallapati, Susan Ditmore, John D. Lafferty, and Kin Ung. Multiscale topic tomography. In *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 520–529, 2007.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 100–108, 2010.

Khanh-Ly Nguyen, Byung-Joo Shin, and Seong Joon Yoo. Hot topic detection and technology trend tracking for patents utilizing term frequency and proportional document frequency and semantic information. In *2016 International Conference on Big Data and Smart Computing (BigComp)*, pp. 223–230. IEEE, 2016. doi: 10.1109/BIGCOMP.2016.7425917.

Ngozichukwuka Onah, Nadine Steinmetz, Hani Al-Sayeh, and Kai-Uwe Sattler. Experimental evaluation of dynamic topic modeling algorithms. *arXiv preprint arXiv:2508.00710*, 2025. doi: 10.48550/arXiv.2508.00710. URL https://arxiv.org/abs/2508.00710.

Hamed Rahimi, Hubert Naacke, Camelia Constantin, and Bernd Amann. *ANTM: Aligned Neural Topic Models for Exploring Evolving Topics*, pp. 76–97. Springer Berlin Heidelberg, Berlin, Heidelberg, 2024. ISBN 978-3-662-69603-3. doi: 10.1007/978-3-662-69603-3_3. URL https://doi.org/10.1007/978-3-662-69603-3_3.

Jonas Rieger, Kai-Robin Lange, Jonathan Flossdorf, and Carsten Jentsch. Dynamic change detection in topics based on rolling ldas. In Ricardo Campos, Alípio Mário Jorge, Adam Jatowt, Sumit Bhatia, and Marina Litvak (eds.), *Proceedings of Text2Story - Fifth Workshop on Narrative Extraction From Texts held in conjunction with the 44th European Conference on Information Retrieval (ECIR 2022), Stavanger, Norway, April 10, 2022*, CEUR Workshop Proceedings, pp. 5–13. CEUR-WS.org, 2022. URL https://ceur-ws.org/Vol-3117/paper1.pdf.

Margaret E. Roberts, Brandon M. Stewart, Dustin Tingley, Christopher Lucas, Jetson Leder-Luis, Shana Kushner Gadarian, Bethany Albertson, and David G. Rand. Structural topic models for open-ended survey responses. *American Journal of Political Science*, 58(4):1064–1082, 2014. doi: 10.1111/ajps.12103.

Margaret E. Roberts, Brandon M. Stewart, and Edoardo M. Airoldi. A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, 111(515):988–1003, 2016. doi: 10.1080/01621459.2016.1141684.

Margaret E. Roberts, Brandon M. Stewart, and Dustin Tingley. stm: An R package for structural topic models. *Journal of Statistical Software*, 91(2):1–40, 2019. doi: 10.18637/jss.v091.i02. URL https://www.jstatsoft.org/article/view/v091i02.

Michael Röder, Andreas Both, and Alexander Hinneburg. Exploring the space of topic coherence measures. In *WSDM*, pp. 399–408, 2015.

Akash Srivastava and Charles Sutton. Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*, 2017. doi: 10.48550/arXiv.1703.01488. URL https://arxiv.org/abs/1703.01488.

Chong Wang, David M. Blei, and David Heckerman. Continuous time dynamic topic models. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, pp. 579–586. AUAI Press, 2008. URL https://dblp.org/rec/conf/uai/WangBH08.

Rui Wang, Deyu Zhou, and Yulan He. Atm: Adversarial-neural topic model. *Information Processing & Management*, 56(6):102098, 2019. ISSN 0306-4573. doi: https://doi.org/10.1016/j.ipm.2019.102098. URL https://www.sciencedirect.com/science/article/pii/S0306457319300500.

Xuerui Wang and Andrew McCallum. Topics over time: A non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 424–433. ACM, 2006. doi: 10.1145/1150402.1150450.

Yunli Wang and Cyril Goutte. Real-time change point detection using on-line topic models. In Emily M. Bender, Leon Derczynski, and Pierre Isabelle (eds.), *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 2505–2515, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://aclanthology.org/C18-1212/.

Sinead Williamson, Chong Wang, Katherine A. Heller, and David M. Blei. The IBP compound dirichlet process and its application to focused topic modeling. In *Proceedings of the 27th International Conference on Machine Learning*, pp. 1151–1158, 2010.

Xiaobao Wu, Xinshuai Dong, Thong Thanh Nguyen, and Anh Tuan Luu. Effective neural topic modeling with embedding clustering regularization. In *International Conference on Machine Learning*, pp. 37335–37357. PMLR, 2023.

Xiaobao Wu, Xinshuai Dong, Liangming Pan, Thong Nguyen, and Anh Tuan Luu. Modeling dynamic topics in chain-free fashion by evolution-tracking contrastive learning and unassociated word exclusion. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 3088–3105, Bangkok, Thailand, August 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.183. URL https://aclanthology.org/2024.findings-acl.183/.

Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. A survey on neural topic models: Methods, applications, and challenges. *Artificial Intelligence Review*, 57(2), 2024b. doi: 10.1007/s10462-023-10661-7. URL https://doi.org/10.1007/s10462-023-10661-7.

Delvin Ce Zhang and Hady Lauw. Dynamic topic models for temporal document networks. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 26281–26292. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/zhang22n.html.

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

N. Zhong and D. A. Schweidel. Capturing changes in social media content: A multiple latent changepoint topic model. *Marketing Science*, 39(4):827–846, 2020. doi: https://doi.org/10.1287/mksc.2019.1212.

Yandi Zhu, Xiaoling Lu, Jingya Hong, and Feifei Wang. Joint dynamic topic model for recognition of lead-lag relationship in two text corpora. *Data Mining and Knowledge Discovery*, 36(6):2272–2298, 2022. doi: https://doi.org/10.1007/s10618-022-00873-w.

A Constraint Proofs

For word $w \in \mathcal{S}_k$:

$$\phi_{k,t,w} = \frac{\tilde{\phi}_{k,t,w}}{\sum_{v \in \mathcal{S}_k} \tilde{\phi}_{k,t,v}} \cdot \left(\tau_k + (1 - \tau_k) \sum_{v \in \mathcal{S}_k} \tilde{\phi}_{k,t,v} \right) \tag{23}$$

For word $w \notin \mathcal{S}_k$:

$$\phi_{k,t,w} = (1 - \tau_k) \tilde{\phi}_{k,t,w} \tag{24}$$

Then we have:

$$\begin{aligned} \sum_{w \in \mathcal{S}_k} \phi_{k,t,w} &= \sum_{w \in \mathcal{S}_k} \frac{\tilde{\phi}_{k,t,w}}{\sum_{v \in \mathcal{S}_k} \tilde{\phi}_{k,t,v}} \cdot \left(\tau_k + (1 - \tau_k) \sum_{v \in \mathcal{S}_k} \tilde{\phi}_{k,t,v} \right) \\ &= \frac{\sum_{w \in \mathcal{S}_k} \tilde{\phi}_{k,t,w}}{\sum_{v \in \mathcal{S}_k} \tilde{\phi}_{k,t,v}} \cdot \left(\tau_k + (1 - \tau_k) \sum_{v \in \mathcal{S}_k} \tilde{\phi}_{k,t,v} \right) \\ &= \tau_k + (1 - \tau_k) \sum_{v \in \mathcal{S}_k} \tilde{\phi}_{k,t,v} \end{aligned} \tag{25}$$

And from the construction it is guaranteed that

$$\tau_k \leq \sum_{w \in \mathcal{S}_k} \phi_{k,t,w} \leq 1$$

And the overall sum still satisfies

$$\begin{aligned} \sum_{w} \phi_{k,t,w} = \sum_{w \in \mathcal{S}_k} \phi_{k,t,w} + \sum_{w \notin \mathcal{S}_k} \phi_{k,t,w} &= \left(\tau_k + (1 - \tau_k) \sum_{v \in \mathcal{S}_k} \tilde{\phi}_{k,t,v} \right) + (1 - \tau_k) \sum_{w \notin \mathcal{S}_k} \tilde{\phi}_{k,t,w} \\ &= \tau_k + (1 - \tau_k) \left(\sum_{v \in \mathcal{S}_k} \tilde{\phi}_{k,t,v} \right) + (1 - \tau_k) \sum_{w \notin \mathcal{S}_k} \tilde{\phi}_{k,t,w} \\ &= \tau_k + (1 - \tau_k) \sum_{v} \tilde{\phi}_{k,t,v} = \tau_k + (1 - \tau_k) = 1 \end{aligned} \tag{26}$$

Validity of the prevalence map. We now show that the prevalence constraint map (8) produces a valid distribution.

Proposition 1 (Prevalence simplex validity)**.** Let $\tilde{\theta}_d \in \Delta^{K-1}$, let $\mathcal{K}_{\min} \subsetneq \{1, \ldots, K\}$ be a proper subset so that $c_{d,\min} = \sum_{k \in \mathcal{K}_{\min}} \tilde{\theta}_{d,k} < 1$, and let $a \in [0, 1]$. Then the vector θ_d defined by (8) satisfies $\theta_{d,k} \geq 0$ for all k and $\sum_{k=1}^{K} \theta_{d,k} = 1$; that is, $\theta_d \in \Delta^{K-1}$.

Proof. Nonnegativity. For $k \in \mathcal{K}_{\min}$, $\theta_{d,k} = a\,\tilde{\theta}_{d,k} \geq 0$ since $a \geq 0$ and $\tilde{\theta}_{d,k} \geq 0$. For $k \notin \mathcal{K}_{\min}$, the factor is $(1 - ac_{d,\min})/(1 - c_{d,\min})$. The denominator is positive because $c_{d,\min} < 1$. The numerator is nonnegative because $a \leq 1$ gives $ac_{d,\min} \leq c_{d,\min} < 1$, hence $1 - ac_{d,\min} > 0$. Therefore $\theta_{d,k} \geq 0$.

Sum to one. Split the sum over the two groups,

$$\sum_{k=1}^{K} \theta_{d,k} = a \sum_{k \in \mathcal{K}_{\min}} \tilde{\theta}_{d,k} + \frac{1 - ac_{d,\min}}{1 - c_{d,\min}} \sum_{k \notin \mathcal{K}_{\min}} \tilde{\theta}_{d,k}.$$

By definition $\sum_{k\in\mathcal{K}_{\min}} \tilde{\theta}_{d,k} = c_{d,\min}$ and, since $\tilde{\theta}_d \in \Delta^{K-1}$, $\sum_{k\notin\mathcal{K}_{\min}} \tilde{\theta}_{d,k} = 1 - c_{d,\min}$. Substituting,

$$\sum_{k=1}^{K} \theta_{d,k} = a\, c_{d,\min} + \frac{1 - ac_{d,\min}}{1 - c_{d,\min}} (1 - c_{d,\min}) = a\, c_{d,\min} + \left(1 - ac_{d,\min}\right) = 1. \qquad \square$$

The total mass in generic documents placed on the minority topics after the map is $\sum_{k\in\mathcal{K}_{\min}} \theta_{d,k} = a\, c_{d,\min} \le c_{d,\min}$, so the map reduces (never increases) the minority mass, with equality only when $a = 1$. In the implementation the factor a is additionally capped for each generic document at $a \le (1-\varepsilon)/c_{d,\min}$ for a small $\varepsilon > 0$; since the hypothesis already uses $a \le 1 \le (1-\varepsilon)/c_{d,\min}$ when $c_{d,\min} \le 1-\varepsilon$, this cap only guards against the degenerate case $c_{d,\min} \to 1$ and does not otherwise alter the map.

B Posterior and ELBO Derivations

Using Bayes' rule,(27),

$$\log p(\tilde{\eta}, z \mid w, x, t, \beta) = \log p(w, z, \tilde{\eta} \mid x, t, \beta) - \log p(w \mid x, t, \beta). \tag{27}$$

Plug this into the KL expression (28):

$$\begin{aligned} \mathrm{KL}\left(q \,\|\, p(\,\cdot\, \mid w, x, t, \beta)\right) &= \mathbb{E}_q\Big[\log q(\tilde{\eta}, z) - \log p(w, z, \tilde{\eta} \mid \beta, x, t) + \log p(w \mid \beta, x, t)\Big] \\ &= \mathbb{E}_q\Big[\log q(\tilde{\eta}, z) - \log p(w, z, \tilde{\eta} \mid \beta, x, t)\Big] + \log p(w \mid \beta, x, t). \end{aligned} \tag{28}$$

Note that $\log p(w \mid x, t, \beta)$ does not depend on q, $\tilde{\eta}$, or z, so it is a constant with respect to the optimization over q. Therefore, minimizing the KL divergence is equivalent to minimizing

$$\mathbb{E}_q\left[\log q(\tilde{\eta}, z) - \log p(w, z, \tilde{\eta} \mid \beta, x, t)\right], \tag{29}$$

or, equivalently, to *maximizing* the negative of this quantity.

B.1 Evidence Lower Bound (ELBO)

We define the ELBO as the negative of (29), with B and Σ still held fixed as in (11) and suppressed from the notation below:

$$\mathcal{L}(q;\beta) = \mathbb{E}_q\left[\log p(w, z, \tilde{\eta} \mid x, t, \beta)\right] - \mathbb{E}_q\left[\log q(\tilde{\eta}, z)\right]. \tag{30}$$

From (28) and (30) we have the standard identity

$$\mathrm{KL}\big(q(\tilde{\eta}, z) \,\big\|\, p(\tilde{\eta}, z \mid w, x, t, \beta)\big) = -\mathcal{L}(q;\beta) + \log p(w \mid x, t, \beta). \tag{31}$$

so maximizing $\mathcal{L}(q;\beta)$ is exactly the same as minimizing the KL divergence.

C ELBO Term Expansion

(A) Closed form for $\mathbb{E}_{q_d}[\log p(\tilde{\eta}_d \mid x_d)]$. Let $p(\tilde{\eta}_d \mid x_d) = \mathcal{N}(\mu_d, \Sigma)$ with $\mu_d = \mu(x_d)$ and $q_d(\tilde{\eta}_d) = \mathcal{N}(m_d, \mathrm{diag}(s_d^2))$. Then

$$\begin{aligned} \mathbb{E}_{q_d}[\log p(\tilde{\eta}_d \mid x_d)] &= -\frac{1}{2}\Big((K-1)\log(2\pi) + \log|\Sigma| + \mathbb{E}_{q_d}\left[(\tilde{\eta}_d - \mu_d)^\top \Sigma^{-1} (\tilde{\eta}_d - \mu_d)\right]\Big) \\ &= -\frac{1}{2}\Big((K-1)\log(2\pi) + \log|\Sigma| + (m_d - \mu_d)^\top \Sigma^{-1}(m_d - \mu_d) + \mathrm{tr}\big(\Sigma^{-1}\mathrm{diag}(s_d^2)\big)\Big). \end{aligned} \tag{32}$$

(B) Topic-assignment term Because $q_{dn}(z_{dn}) = \mathrm{Cat}(\gamma_{dn})$ and $p(z_{dn} = k \mid \theta_d) = \theta_{d,k}$,

$$\begin{aligned} \mathbb{E}_q[\log p(z_{dn} \mid \theta_d)] &= \mathbb{E}_{q_d}\, \mathbb{E}_{q_{dn}}[\log p(z_{dn} \mid \theta_d(\tilde{\eta}_d))] \\ &= \mathbb{E}_{q_d}\Big[\sum_{k=1}^{K} \gamma_{dn}(k) \log \theta_{d,k}(\tilde{\eta}_d)\Big] \\ &= \sum_{k=1}^{K} \gamma_{dn}(k)\; \mathbb{E}_{q_d}[\log \theta_{d,k}(\tilde{\eta}_d)]. \end{aligned}$$

The difficulty is that $\theta_d(\tilde{\eta}_d)$ is a nonlinear transformation:

$$\tilde{\eta}_d \mapsto \eta_d = (\tilde{\eta}_d, 0) \mapsto \tilde{\theta}_d = \text{softmax}(\eta_d) \mapsto \theta_d = \text{minority_rescale}(\tilde{\theta}_d; a),$$

so $\mathbb{E}_{q_d}[\log \theta_{d,k}(\tilde{\eta}_d)]$ has no closed form. We therefore approximate it by Monte Carlo using the reparameterization trick:

$$\epsilon^{(s)} \sim \mathcal{N}(0, I), \qquad \tilde{\eta}_d^{(s)} = m_d + s_d \odot \epsilon^{(s)}, \qquad s = 1, \ldots, S, \tag{33}$$

$$\mathbb{E}_{q_d}[\log \theta_{d,k}(\tilde{\eta}_d)] \approx \frac{1}{N_{\text{MC}}} \sum_{s=1}^{N_{\text{MC}}} \log \theta_{d,k}(\tilde{\eta}_d^{(s)}). \tag{34}$$

Only $\mathbb{E}_{q_d}[\log \theta_{d,k}(\tilde{\eta}_d)]$ requires Monte Carlo; all other ELBO terms are closed form given $(m_d, s_d, \gamma, \beta)$.

(C) Expected log word-likelihood (closed form given β). Because $p(w_{dn} \mid z_{dn} = k, t_d, \beta) = \phi_{k,t_d,w_{dn}}(\beta)$ and β is treated as a parameter,

$$\mathbb{E}_q[\log p(w_{dn} \mid z_{dn}, t_d, \beta)] = \sum_{k=1}^{K} \gamma_{dn}(k) \, \log \phi_{k,t_d,w_{dn}}(\beta).$$

This term is closed form once γ_{dn} and β are known.

(D1) Categorical entropy (closed form). For $q_{dn}(z_{dn}) = \text{Cat}(\gamma_{dn})$,

$$H(\gamma_{dn}) = -\mathbb{E}_{q_{dn}}[\log q_{dn}(z_{dn})] = -\sum_{k=1}^{K} \gamma_{dn}(k) \log \gamma_{dn}(k).$$

(D2) Gaussian entropy (closed form). For $q_d(\tilde{\eta}_d) = \mathcal{N}(m_d, \text{diag}(s_d^2))$,

$$H(q_d) = -\mathbb{E}_{q_d}[\log q_d(\tilde{\eta}_d)] = \frac{1}{2} \sum_{i=1}^{K-1} \log\big(2\pi e \, s_{d,i}^2\big).$$

D Variational EM Derivations

D.1 Mathematical preliminaries (derivative identities)

A few basic identities used repeatedly in the E-step and M-step.

(i) Log-derivative rule. For any differentiable positive function $f(x) > 0$,

$$\nabla_x \log f(x) = \frac{1}{f(x)} \nabla_x f(x).$$

(ii) Softmax and its Jacobian. For $\pi = \text{softmax}(\eta) \in \mathbb{R}^K$ with

$$\pi_k = \frac{e^{\eta_k}}{\sum_{j=1}^{K} e^{\eta_j}},$$

the partial derivative is

$$\frac{\partial \pi_k}{\partial \eta_j} = \pi_k(\mathbb{I}[k = j] - \pi_j).$$

Equivalently, the Jacobian matrix is

$$J_{\text{softmax}}(\eta) = \text{diag}(\pi) - \pi\pi^\top.$$

(iii) Reparameterization trick. If $q(\tilde{\eta}_d) = \mathcal{N}(m_d, \text{diag}(s_d^2))$, we can sample as

$$\epsilon \sim \mathcal{N}(0, I), \qquad \tilde{\eta}_d = m_d + s_d \odot \epsilon.$$

Then for any differentiable function f,

$$\nabla_{m_d} \mathbb{E}_q[f(\tilde{\eta}_d)] = \mathbb{E}_\epsilon \left[\nabla_{\tilde{\eta}_d} f(\tilde{\eta}_d)\right],$$

and (elementwise)

$$\nabla_{s_{d,i}} \mathbb{E}_q[f(\tilde{\eta}_d)] = \mathbb{E}_\epsilon \left[\epsilon_i \frac{\partial}{\partial \tilde{\eta}_{d,i}} f(\tilde{\eta}_d) \right].$$

In practice we approximate these expectations using Monte Carlo averages.

M-step (update β; q fixed). In the M-step we maximize the ELBO w.r.t. β while keeping q fixed (i.e., keeping γ fixed). Only two terms depend on β: the expected word-likelihood (C) and the prior $\log p(\beta)$:

$$\mathcal{L}(\beta) = \underbrace{\sum_{d=1}^{D} \sum_{n=1}^{N_d} \sum_{k=1}^{K} \gamma_{dn}(k) \log \phi_{k,t_d,w_{dn}}(\beta)}_{\text{expected log-likelihood}} + \underbrace{\log p(\beta)}_{\text{random-walk prior}}.$$

Step 1: rewrite the likelihood using expected word counts. Define expected counts for each topic k, time slice t, and word v:

$$c_{k,t,v} = \sum_{d:t_d=t} \sum_{n:w_{dn}=v} \gamma_{dn}(k).$$

This counts how many times word v is assigned to topic k (in expectation) at time t. Using $c_{k,t,v}$, the expected log-likelihood becomes

$$\sum_{d,n,k} \gamma_{dn}(k) \log \phi_{k,t_d,w_{dn}} = \sum_{k=1}^{K} \sum_{t=0}^{T-1} \sum_{v=1}^{V} c_{k,t,v} \log \phi_{k,t,v}(\beta).$$

Step 2: gradient of the likelihood term. For a fixed (k, t), define

$$\ell_{k,t}(\beta_{k,t}) = \sum_{v=1}^{V} c_{k,t,v} \log \phi_{k,t,v}(\beta_{k,t}).$$

Then

$$\nabla_{\beta_{k,t}} \ell_{k,t} = \sum_{v=1}^{V} c_{k,t,v} \nabla_{\beta_{k,t}} \log \phi_{k,t,v}.$$

Because $\phi_{k,t}$ is a differentiable transformation of $\beta_{k,t}$ (softmax followed by deterministic seed-rescaling), we can compute this gradient using the chain rule. In implementation we use automatic differentiation to avoid errors in the seed-rescaling Jacobian.

Step 3: gradient of the RW1 prior. Recall the random-walk prior from Section 4.4:

$$\beta_{k,0} \sim \mathcal{N}(0, \Sigma_\beta), \qquad \beta_{k,t} \mid \beta_{k,t-1} \sim \mathcal{N}(\beta_{k,t-1}, Q_\beta), \quad t = 1, \ldots, T-1.$$

For the implementation we use isotropic covariances $\Sigma_\beta = \sigma_0^2 I_V$ and $Q_\beta = \sigma_q^2 I_V$. Ignoring additive constants, the log prior is

$$\log p(\beta) = -\frac{1}{2\sigma_0^2} \sum_k \|\beta_{k,0}\|^2 - \frac{1}{2\sigma_q^2} \sum_k \sum_{t=1}^{T-1} \|\beta_{k,t} - \beta_{k,t-1}\|^2.$$

The gradient with respect to $\beta_{k,t}$ depends on which neighbours contribute. For the initial time point $t = 0$,

$$\nabla_{\beta_{k,0}} \log p(\beta) = -\frac{1}{\sigma_0^2} \beta_{k,0} - \frac{1}{\sigma_q^2} (\beta_{k,0} - \beta_{k,1}).$$

For an interior time $0 < t < T-1$, both the backward and forward differences contribute, giving

$$\nabla_{\beta_{k,t}} \log p(\beta) = -\frac{1}{\sigma_q^2} \big(2\beta_{k,t} - \beta_{k,t-1} - \beta_{k,t+1}\big).$$

For the final time point $t = T-1$, only the backward difference contributes:

$$\nabla_{\beta_{k,T-1}} \log p(\beta) = -\frac{1}{\sigma_q^2} \big(\beta_{k,T-1} - \beta_{k,T-2}\big).$$

Step 4: final gradient ascent update. Combining likelihood and prior:

$$\nabla_{\beta_{k,t}} \mathcal{L}(\beta) = \nabla_{\beta_{k,t}} \ell_{k,t} + \nabla_{\beta_{k,t}} \log p(\beta).$$

We update by gradient ascent:

$$\beta_{k,t} \leftarrow \beta_{k,t} + \rho_\beta \nabla_{\beta_{k,t}} \mathcal{L}(\beta).$$

In practice we compute $\nabla_{\beta_{k,t}} \ell_{k,t}$ using automatic differentiation due to the constrained mapping $\beta \mapsto \phi$.

Because the seed-rescaling map (4) is piecewise-defined and involves global normalization over the seed set, we compute gradients with automatic differentiation rather than deriving the full Jacobian manually.

M-step: updating prevalence parameters (B, Σ). In addition to updating the content trajectories β, we may also update the *prevalence* parameters that define the logistic-normal prior

$$p(\tilde{\eta}_d \mid x_d; B, \Sigma) \;=\; \mathcal{N}\big(\tilde{\eta}_d \mid \mu(x_d), \Sigma\big), \qquad \mu(x_d) = B^\top x_d,$$

while keeping the variational factors $q(\tilde{\eta}, z)$ fixed. Each coefficient $B_{p,k}$ measures how covariate p shifts logit k (relative to the baseline topic with appended zero logit).

In our variational family we use

$$q_d(\tilde{\eta}_d) = \mathcal{N}\big(m_d, \mathrm{diag}(s_d^2)\big).$$

Therefore, the only ELBO term that depends on (B, Σ) is the expected log-likelihood of the priors (term (A)):

$$\mathcal{L}_{\text{prev}}(B, \Sigma) \;=\; \sum_{d=1}^{D} \mathbb{E}_{q_d}\big[\log \mathcal{N}(\tilde{\eta}_d \mid B^\top x_d, \Sigma)\big] \;-\; \frac{\lambda}{2}\|B\|_F^2,$$

where we optionally include an ℓ_2 (ridge) penalty with strength $\lambda \geq 0$.

Step 1: expand the expected Gaussian log-density. We need to evaluate the expectation of the quadratic form inside the Gaussian log-density. Let $q_d(\tilde{\eta}_d) = \mathcal{N}(m_d, S_d)$, where $S_d = \mathrm{diag}(s_d^2)$ is the diagonal variational covariance matrix. Let $\mu_d = B^\top x_d$ be the prior mean.

For each document d, using the standard Gaussian identity,

$$\mathbb{E}_{q_d}\big[\log \mathcal{N}(\tilde{\eta}_d \mid B^\top x_d, \Sigma)\big] = -\frac{1}{2}\Big((K-1)\log(2\pi) + \log|\Sigma| + \mathbb{E}_{q_d}\big[(\tilde{\eta}_d - \mu_d)^\top \Sigma^{-1} (\tilde{\eta}_d - \mu_d)\big]\Big), \tag{35}$$

$$\log \mathcal{N}(\tilde{\eta}_d \mid \mu_d, \Sigma) = \underbrace{-\frac{1}{2}\log|\Sigma| - \frac{K-1}{2}\log(2\pi)}_{C} - \frac{1}{2}(\tilde{\eta}_d - \mu_d)^\top \Sigma^{-1} (\tilde{\eta}_d - \mu_d).$$

Taking the expectation $\mathbb{E}_{q_d}[\cdot]$ yields:

$$\mathbb{E}_{q_d}[\log P(\tilde{\eta}_d)] = C - \frac{1}{2}\underbrace{\mathbb{E}_{q_d}\left[(\tilde{\eta}_d - \mu_d)^\top \Sigma^{-1} (\tilde{\eta}_d - \mu_d)\right]}_{\text{Quadratic Expectation}}. \tag{36}$$

To evaluate the **Quadratic Expectation**, we use the algebraic trick of adding and subtracting the variational mean m_d. Rewrite the deviation as $(\tilde{\eta}_d - \mu_d) = (\tilde{\eta}_d - m_d) + (m_d - \mu_d)$. Substituting this into the quadratic form and expanding gives three terms:

$$\begin{aligned}\mathbb{E}_{q_d}\Big[(\dots)^\top \Sigma^{-1} (\dots)\Big] = {} & \underbrace{(m_d - \mu_d)^\top \Sigma^{-1} (m_d - \mu_d)}_{\text{Bias Term}} + \underbrace{2 \cdot \mathbb{E}_{q_d}[(\tilde{\eta}_d - m_d)]^\top \Sigma^{-1} (m_d - \mu_d)}_{\text{Cross Term (Zero)}} \\ & + \underbrace{\mathbb{E}_{q_d}[(\tilde{\eta}_d - m_d)^\top \Sigma^{-1} (\tilde{\eta}_d - m_d)]}_{\text{Variance Term}}.\end{aligned} \tag{37}$$

The **Bias Term** is constant with respect to the expectation. The **Cross Term** vanishes because the expectation of the centered variable is zero: $\mathbb{E}[\tilde{\eta}_d - m_d] = 0$.

The **Variance Term** is evaluated using the **Trace Trick**. First, note that the quadratic form is a scalar, and a scalar is equal to its own trace ($\mathrm{tr}(x) = x$). Using the cyclic property of the trace ($\mathrm{tr}(ABC) = \mathrm{tr}(BCA)$), we can move the vector to the end:

$$(\tilde{\eta}_d - m_d)^\top \Sigma^{-1} (\tilde{\eta}_d - m_d) \;=\; \mathrm{tr}\left(\Sigma^{-1} (\tilde{\eta}_d - m_d)(\tilde{\eta}_d - m_d)^\top\right).$$

Since the Expectation and Trace are both linear operators, we can swap their order:

$$\mathbb{E}\left[\,\mathrm{tr}\left(\Sigma^{-1}(\tilde{\eta}_d - m_d)(\tilde{\eta}_d - m_d)^\top\right)\right] \;=\; \mathrm{tr}\Big(\Sigma^{-1} \underbrace{\mathbb{E}[(\tilde{\eta}_d - m_d)(\tilde{\eta}_d - m_d)^\top]}_{\text{Covariance } S_d}\Big) \;=\; \mathrm{tr}(\Sigma^{-1} S_d).$$

Substituting these back into Equation (36) and summing over all documents yields the final objective:

$$\mathcal{L}_{\text{prev}}(B, \Sigma) = -\frac{1}{2} \sum_{d=1}^{D} \Big(\log|\Sigma| + (m_d - B^\top x_d)^\top \Sigma^{-1} (m_d - B^\top x_d) + \mathrm{tr}(\Sigma^{-1} S_d) \Big) - \frac{\lambda}{2} \|B\|_F^2 + \text{const}.$$

Step 2: update B (ridge regression working update). A common STM-style working approximation is to update B using only the variational means (and optionally treat the weighting by Σ^{-1} as isotropic for simplicity). Let

$$X \in \mathbb{R}^{D \times P} \text{ with rows } x_d^\top, \qquad M \in \mathbb{R}^{D \times (K-1)} \text{ with rows } m_d^\top, \qquad B \in \mathbb{R}^{P \times (K-1)}.$$

Then $B^\top x_d$ corresponds to the row-vector product $x_d^\top B$, and in matrix form the mean predictions are $XB \in \mathbb{R}^{D \times (K-1)}$. Dropping constants and using the (unweighted) working objective, updating B reduces to the ridge problem

$$B \leftarrow \arg\min_B \; \|M - XB\|_F^2 + \lambda \|B\|_F^2.$$

Taking derivatives,

$$\frac{\partial}{\partial B}\Big(\|M - XB\|_F^2 + \lambda \|B\|_F^2 \Big) = -2X^\top (M - XB) + 2\lambda B,$$

and setting the gradient to zero gives the normal equations

$$(X^\top X + \lambda I_P) B = X^\top M,$$

so the closed-form update is

$$\boxed{B \;=\; (X^\top X + \lambda I_P)^{-1} X^\top M.} \tag{38}$$

Step 3: update Σ (residual covariance with uncertainty). Given the updated coefficients B, we update the global covariance Σ. Unlike the simplified STM approach, our implementation incorporates the variational uncertainty from $q(\tilde{\eta})$. Differentiating the objective (**??**) with respect to Σ^{-1}, we typically solve for the maximum likelihood estimator:

$$\frac{\partial \mathcal{L}}{\partial \Sigma^{-1}} \;=\; \frac{D}{2}\Sigma - \frac{1}{2} \sum_{d=1}^{D} \Big[(m_d - B^\top x_d)(m_d - B^\top x_d)^\top + \mathrm{diag}(s_d^2) \Big] \;=\; 0.$$

Solving for Σ yields the closed-form update:

$$\Sigma \;=\; \underbrace{\frac{1}{D} \sum_{d=1}^{D} (m_d - B^\top x_d)(m_d - B^\top x_d)^\top}_{\text{Empirical Covariance of Residuals}} \;+\; \underbrace{\frac{1}{D} \sum_{d=1}^{D} \mathrm{diag}(s_d^2)}_{\text{Average Variational Variance}} \quad .$$

D.2 Numerical Stability and Implementation Details

In implementation, we enforce feasibility of minority rescaling with clamping, renormalize topic proportions to mitigate floating-point drift, optimize scale parameters in log-space for positivity, and add covariance jitter for stable inversion. These details preserve theoretical constraints in finite-precision training.

In our implementation, we compute the first term using the residuals matrix $R = M - XB$, and the second term using the average of the exponentiated log-scales:

$$\boxed{\Sigma \;=\; \frac{1}{D} R^\top R \;+\; \mathrm{diag}\left(\frac{1}{D} \sum_{d=1}^{D} s_d^2 \right).} \tag{39}$$

We also add a small jitter ϵI for numerical stability.

E-step (update γ and (m_d, s_d); β fixed). Fix β and thus $\phi(\beta)$. The E-step maximizes the ELBO w.r.t. the variational factors $q_{dn}(z_{dn}) = \mathrm{Cat}(\gamma_{dn})$ and $q_d(\tilde{\eta}_d) = \mathcal{N}(m_d, \mathrm{diag}(s_d^2))$.

Step 1: update γ (token responsibilities). For a given document d and token n, we update γ_{dn} by maximizing the ELBO subject to $\sum_k \gamma_{dn}(k) = 1$. Only two ELBO terms depend on γ_{dn}: the topic-assignment term (B) and word-likelihood term (C), plus the categorical entropy (D1). Collecting the relevant part for token (d, n):

$$\mathcal{L}_{dn}(\gamma_{dn}) = \sum_{k=1}^{K} \gamma_{dn}(k) \left(\mathbb{E}_{q_d}[\log \theta_{d,k}(\tilde{\eta}_d)] + \log \phi_{k,t_d,w_{dn}}(\beta) \right) - \sum_{k=1}^{K} \gamma_{dn}(k) \log \gamma_{dn}(k).$$

Using a Lagrange multiplier for the simplex constraint and taking derivatives yields:

$$\gamma_{dn}(k) \propto \exp\left(\widehat{\mathbb{E}}_{q_d}[\log \theta_{d,k}] + \log \phi_{k,t_d,w_{dn}}(\beta) \right),$$

where $\widehat{\mathbb{E}}_{q_d}[\log \theta_{d,k}]$ is approximated by Monte Carlo via (34).

Step 2: objective for (m_d, s_d). For each document d, the ELBO terms that depend on (m_d, s_d) are:

$$\mathcal{L}_d(m_d, s_d) = \underbrace{\mathbb{E}_{q_d}[\log p(\tilde{\eta}_d \mid x_d)]}_{(A)} + \underbrace{\sum_{n=1}^{N_d} \sum_{k=1}^{K} \gamma_{dn}(k)\, \mathbb{E}_{q_d}[\log \theta_{d,k}(\tilde{\eta}_d)]}_{(B)} + \underbrace{H(q_d)}_{(D2)}.$$

The difficult part is (B) because $\theta_d(\tilde{\eta}_d)$ is nonlinear.

Define aggregated expected topic counts for document d:

$$n_{d,k} = \sum_{n=1}^{N_d} \gamma_{dn}(k).$$

Then the topic-assignment term becomes

$$(B)_d = \sum_{k=1}^{K} n_{d,k}\, \mathbb{E}_{q_d}[\log \theta_{d,k}(\tilde{\eta}_d)].$$

Step 3: reparameterized Monte Carlo estimate. Using $\tilde{\eta}_d = m_d + s_d \odot \epsilon$ with $\epsilon \sim \mathcal{N}(0, I)$:

$$(B)_d = \mathbb{E}_\epsilon \left[\sum_{k=1}^{K} n_{d,k} \log \theta_{d,k}(m_d + s_d \odot \epsilon) \right].$$

We approximate the expectation by S Monte Carlo samples $\epsilon^{(1)}, \ldots, \epsilon^{(S)}$.

Step 4: gradients for m_d. Let

$$f(\tilde{\eta}_d) = \sum_{k=1}^{K} n_{d,k} \log \theta_{d,k}(\tilde{\eta}_d).$$

Then by the reparameterization identity,

$$\nabla_{m_d}(B)_d = \mathbb{E}_\epsilon \left[\nabla_{\tilde{\eta}_d} f(\tilde{\eta}_d) \right] \approx \frac{1}{N_{\mathrm{MC}}} \sum_{s=1}^{N_{\mathrm{MC}}} \nabla_{\tilde{\eta}_d} f(\tilde{\eta}_d^{(s)}).$$

The prior term (A) contributes a closed form gradient:

$$\nabla_{m_d}(A)_d = -\Sigma^{-1}(m_d - \mu(x_d)).$$

Therefore the total gradient used for updating m_d is

$$\nabla_{m_d} \mathcal{L}_d \approx \frac{1}{N_{\mathrm{MC}}} \sum_{s=1}^{N_{\mathrm{MC}}} \nabla_{\tilde{\eta}_d} f(\tilde{\eta}_d^{(s)}) - \Sigma^{-1}(m_d - \mu(x_d)).$$

We then perform gradient ascent:

$$m_d \leftarrow m_d + \rho_m \nabla_{m_d} \mathcal{L}_d.$$

Step 5: gradients for s_d (diagonal std). For each coordinate i, again using reparameterization:

$$\frac{\partial}{\partial s_{d,i}}(B)_d = \mathbb{E}_\epsilon \left[\epsilon_i \frac{\partial}{\partial \tilde{\eta}_{d,i}} f(\tilde{\eta}_d) \right] \approx \frac{1}{N_{\mathrm{MC}}} \sum_{s=1}^{N_{\mathrm{MC}}} \epsilon_i^{(s)} \frac{\partial}{\partial \tilde{\eta}_{d,i}} f(\tilde{\eta}_d^{(s)}).$$

The Gaussian prior term contributes:

$$\frac{\partial}{\partial s_{d,i}}(A)_d = -(\Sigma^{-1})_{ii}\, s_{d,i}.$$

The entropy term for q_d is

$$H(q_d) = \frac{1}{2} \sum_{i=1}^{K-1} \log(2\pi e s_{d,i}^2),$$

and its gradient is

$$\frac{\partial}{\partial s_{d,i}} H(q_d) = \frac{1}{s_{d,i}}.$$

Thus the total gradient is

$$\nabla_{s_{d,i}} \mathcal{L}_d \approx \frac{1}{N_{\mathrm{MC}}} \sum_{s=1}^{N_{\mathrm{MC}}} \epsilon_i^{(s)} \frac{\partial}{\partial \tilde{\eta}_{d,i}} f(\tilde{\eta}_d^{(s)}) - (\Sigma^{-1})_{ii}\, s_{d,i} + \frac{1}{s_{d,i}}.$$

In implementation we optimize $\log s_{d,i}$ for positivity:

$$\log s_{d,i} \leftarrow \log s_{d,i} + \rho_s \left(s_{d,i}\, \nabla_{s_{d,i}} \mathcal{L}_d \right).$$

D.3 Synthetic Data Generation Algorithm

Algorithm 1 Constrained Synthetic Data Generation

Require: $V, K_{\min}, K_{\mathrm{BG}}, T, M, P, B, \Sigma, \Sigma_\beta, Q_\beta, \mu_N, N_{\min}, a, \tau$
$K \leftarrow K_{\min} + K_{\mathrm{BG}},\ D \leftarrow T \cdot M$
$\mathcal{K}_{\min} \leftarrow \{0, \ldots, K_{\min} - 1\},\ \mathcal{K}_{\mathrm{BG}} \leftarrow \{K_{\min}, \ldots, K - 1\}$
Construct t_d by repeating each $t \in \{0, \ldots, T-1\}$ exactly M times; compute $u_d = \frac{t_d}{T-1}$
Build covariates $x_d = [1, u_d, \sin(2\pi u_d), (u_d - c)^2, \exp(-\frac{1}{2}((u_d - c)/\sigma)^2)]^\top$ with $c = 0.65,\ \sigma = 0.12$
Build seed sets $\mathcal{S}_k$ for $k \in \mathcal{K}_{\min}$ (size $s = 5$, disjoint) and $r_{k,v}$ from support; set $\tau_k = \tau$ for minority topics else 0
for topic $k = 0$ to $K - 1$ **do**
  Sample $\beta_{k,0} \sim \mathcal{N}(0, \Sigma_\beta)$
  for $t = 1$ to $T - 1$ **do**
    Sample $\beta_{k,t} \sim \mathcal{N}(\beta_{k,t-1}, Q_\beta)$
  end for
  for $t = 0$ to $T - 1$ **do**
    $\tilde{\phi}_{k,t} \leftarrow \mathrm{softmax}(\beta_{k,t})$
    Obtain $\phi_{k,t}$ from $\tilde{\phi}_{k,t}$, τ_k, and $\mathcal{S}_k$ using Eq. (4) (with S replaced by the per-topic $\mathcal{S}_k$)
  end for
end for
for $d = 0$ to $D - 1$ **do**
  $\mu_d \leftarrow B^\top x_d$
  Sample $\tilde{\eta}_d \sim \mathcal{N}(\mu_d, \Sigma)$
  Compute $\tilde{\theta}_d = \mathrm{softmax}([\tilde{\eta}_d, 0])$
  Sample $\delta_d \sim \mathrm{Bernoulli}(\pi(u_d))$**, where** $\pi(u_d) = 0.125 + 0.025 \sin(2\pi u_d)$
  if $\delta_d = 1$ **then**
    If the document is minority content carrying ($\delta_d = 1$) set $\theta_d = \tilde{\theta}_d$.
  else
    If the document is generic ($\delta_d = 0$) obtain θ_d by minority-topic rescaling with factor a using Eq. (8)
  end if
  Sample $N_d \sim \mathrm{Poisson}(\mu_N)$; set $N_d \leftarrow \max(N_d, N_{\min})$
  for $n = 1$ to N_d **do**
    Sample $z_{dn} \sim \mathrm{Cat}(\theta_d)$
    Sample $w_{dn} \sim \mathrm{Cat}(\phi_{z_{dn}, t_d})$
  end for
end for
return documents $\{(w_d, t_d, x_d, \mathrm{metadata}_d)\}_{d=0}^{D-1}$

D.4 Implementation Details

D.4.1 Proposed Model

The proposed model is implemented in Python using NumPy. The core fitting routine is `fit_constrained_stm_vi`, which performs mean-field variational inference with Monte Carlo gradient estimation for the non-conjugate terms.

Topic partition. The total number of topics is $K = 16$, partitioned into $K_{\min} = 4$ dedicated mental-health topics (minority topics) (indices $\{0, 1, 2, 3\}$) and $K_{\mathrm{BG}} = 12$ background topics. The minority-topic indices are fixed prior to inference and are not subject to re-ordering during optimization.

Corpus covariate encoding. For the synthetic benchmark, document-level time covariates are constructed from the normalized time index

$$u_d = \frac{t_d}{T-1}, \qquad t_d \in \{0, \ldots, T-1\},$$

with $T = 36$. Under the full covariate mode, we use

$$x_d = \big[1,\ u_d,\ \sin(2\pi u_d),\ (u_d - 0.65)^2,\ \exp\big(-\tfrac{1}{2}((u_d - 0.65)/0.12)^2\big)\big]^\top.$$

Table 11: Variational inference hyperparameters used in the synthetic benchmark unless stated otherwise. All parameters except n_{iter}, Q_β, and τ_{scale} are held constant.

Parameter	Value	Description
n_{iter}	40 (default)	VI iterations; swept $\in \{20, 40, 60, 80\}$ in sensitivity
$n_{\text{mc}}^{\mathbb{E}\log\theta}$	5	MC samples for $\mathbb{E}[\log\theta]$ gradient
$n_{\text{mc}}^{\nabla m}$	5	MC samples for mean gradient ∇_m
$n_{\text{mc}}^{\nabla s}$	5	MC samples for log-variance gradient ∇_s
η_m	0.05	Step size for variational mean m
η_s	0.05	Step size for variational log-variance s
η_β	0.05	Step size for topic–word matrix β
Q_β	1.0 (default)	Topic–word prior scale; swept $\in \{0.3, 1, 3, 10\}$
Σ_β	0.3	Topic–word prior SD in log space
τ_{scale}	1.0 (default)	Seed constraint scale; swept $\in \{0.1, 0.3, 0.5, 0.9\}$

Seed constraints. Seed constraints are encoded by a topic-level threshold vector $\tau_{\text{scale}} \in [0, 1)^K$, and a topic-word seed-support matrix $r \in \{0, 1\}^{K \times V}$. In the synthetic benchmark, only the minority topics have nonzero τ_k. We scale these thresholds by a scalar $\tau_{\text{scale}} \in [0, 1)$:

$$\tilde{\tau} = \text{clip}(\tau_{\text{scale}} \cdot \tau,\ 0.0,\ 0.999)\,. \tag{40}$$

The default value is $\tau_{\text{scale}} = 0.2$. Setting $\tau_{\text{scale}} = 0$ disables seed constraints entirely (the *No Seed* ablation condition).

Topic–word prior. The topic–word trajectories are regularized by two scalar variance parameters. The first, $\Sigma_\beta = 0.3$, controls the spread of the initial topic-word logits at the first time point. The second, $Q_\beta = 1.0$, controls how much the topic-word logits are allowed to change from one time step to the next in the random-walk prior. In the implementation, both are applied equally to all vocabulary dimensions, that is, as isotropic covariance matrices $\Sigma_\beta I$ and $Q_\beta I$. The vocabulary size is $V = 900$,[2] indexed as $\{\text{word}_0, \ldots, \text{word}_{V-1}\}$.

Variational inference configuration. The VI optimiser is configured via `VIConfig` with the fixed settings listed in Table 11.

E Finnish-language Topic Content Evolution in the Full Data Set

Table 12 shows the evolution of the five minority topics found in the full data set, listing the top words of each topic at selected representative time points. The table corresponds to Table 8 in the main paper but shows the original Finnish-language words for the topics.

[2]The maximum word index observed in the experimental outputs is word_{866}; the vocabulary is $\{\text{word}_0, \ldots, \text{word}_{V-1}\}$ with $V = 900$.

Table 12: Evolution of estimated minority (mental health) topic content over time in the full real data set. For each topic, a list of the most probable words are listed (in original Finnish language) for six representative time bins; for each bin the bin number and corresponding time period are shown.

Topic	**Time bin**	**Top words**
Topic 00	t00 [2011-12–2014-07]	kuunnella, tunne, uni, kohtaus, ilo, pelä, kateellinen, pettymys
	t06 [2016-05–2016-08]	kirja, tunne, tuntea, pakko, lähettää, toivoa, hullu, keskustelu
	t12 [2017-11–2018-03]	ongelma, tuntea, onnellinen, hoitaa, vihata, viha, pelätä, ärsyttävä
	t19 [2019-10–2019-11]	tunne, lähettää, hullu, kirja, toivoa, pelätä, tuntea, ero
	t25 [2020-07–2020-10]	700k, hullu, kirja, tuntea, lähettää, psykopaatti, kirjo, tunni
	t35 [2024-10–2026-01]	toivoa, kirja, levätä, huume, iloinen, man, ulkona, huoli
Topic 01	t00 [2011-12–2014-07]	man, ärsyttävä, uni, tunne, pelä, kohtaus, ilo, huume
	t06 [2016-05–2016-08]	man, ystävä, uni, kirjain, kirja, lähettää, hullu, tuntii
	t12 [2017-11–2018-03]	toivoa, hullu, ilo, kirja, ero, iloinen, rakkaus, kuunnella
	t19 [2019-10–2019-11]	tuntea, ero, tunni, ilo, man, pelätä, kuunnella, kirja
	t25 [2020-07–2020-10]	tunne, 700k, man, ongelma, kirja, tuntea, tunni, kyläläinen
	t35 [2024-10–2026-01]	kauhu, kauhuä, kauhua, ystävä, tunni, ongelma, kettunen, pakko
Topic 02	t00 [2011-12–2014-07]	hullu, tunne, pelä, uni, ilo, kohtaus, pettymys, kateellinen
	t06 [2016-05–2016-08]	toivoa, ongelma, kirja, kuunnella, lähettää, pettymys, kirjo, hullu
	t12 [2017-11–2018-03]	pakko, lähettää, tuntea, toivoa, vihata, kirja, ero, hullu
	t19 [2019-10–2019-11]	pakko, tuntea, kirja, man, ystävä, lähettää, toivoa, mainostulo
	t25 [2020-07–2020-10]	700k, pakko, itku, toivoa, tunni, lääkäri, kirja, kuunnella
	t35 [2024-10–2026-01]	kettunen, ilo, pakko, toivoa, surullinen, ahdistava, man, itku
Topic 03	t00 [2011-12–2014-07]	pelä, tunne, uni, kohtaus, kateellinen, ilo, pettymys, ero
	t06 [2016-05–2016-08]	pakko, painajainen, onnellinen, lähettää, toivoa, rakkaus, hätä, kirjo
	t12 [2017-11–2018-03]	tunne, kauhu, toivoa, surullinen, kuunnella, iloinen, ero, vihaan
	t19 [2019-10–2019-11]	kirja, viha, toivoa, kauhu, stressi, huoli, surullinen, kirjo
	t25 [2020-07–2020-10]	700k, toivoa, surullinen, vihaan, pelä, tunni, positiivinen, ärsyttävä
	t35 [2024-10–2026-01]	bara, bastu, uni, bada, sauna, halvaus, ooooh, som
Topic 04	t00 [2011-12–2014-07]	pakko, ärsyttävä, uni, tunne, kohtaus, ilo, pelä, ulko
	t06 [2016-05–2016-08]	lähettää, ärsyttävä, trauma, surullinen, ero, hullu, kauhu, toivo
	t12 [2017-11–2018-03]	vihaan, lähettää, :-d, man, itku, positiivinen, surullinen, kuunnella
	t19 [2019-10–2019-11]	toivoa, kirja, hätä, painajainen, onnellinen, surullinen, uni, pelätä
	t25 [2020-07–2020-10]	700k, negatiivinen, ilo, ajatus, pelätä, iloinen, lääkäri, ystävä
	t35 [2024-10–2026-01]	toivoa, hullu, halvauknen, tunne, onnellinen, pelä, kettunen, pakko