Constrained Non-negative Matrix Factorization for Guided Topic Modeling of Minority Topics

Seyedeh Fatemeh Ebrahimi and Jaakko Peltonen[†]
Faculty of Information Technology and Communication Sciences
Tampere University
{seyedeh.ebrahimi, jaakko.peltonen}@tuni.fi

Abstract

Topic models often fail to capture low-prevalence, domain-critical themes—so-called *minority topics*—such as mental health themes in online comments. While some existing methods can incorporate domain knowledge such as expected topical content, methods allowing guidance may require overly detailed expected topics, hindering the discovery of topic divisions and variation. We propose a topic modeling solution via a specially constrained NMF. We incorporate a seed word list characterizing minority content of interest, but we do not require experts to pre-specify their division across minority topics. Through prevalence constraints on minority topics and seed word content across topics, we learn distinct data-driven minority topics as well as majority topics. The constrained NMF is fitted via Karush-Kuhn-Tucker (KKT) conditions with multiplicative updates. We outperform several baselines on synthetic data in terms of topic purity, normalized mutual information, and also evaluate topic quality using Jensen-Shannon divergence (JSD). We conduct a case study on YouTube vlog comments, analyzing viewer discussion of mental health content; our model successfully identifies and reveals this domain-relevant minority content.

1 Introduction

A central problem in many data analysis domains is identifying and extracting both dominant and minority themes from extensive corpora (Jagarlamudi et al., 2012). Topic modeling is a well-established task proposed to discover latent themes from a collection of texts based on word occurrence (Srivastava and Sutton, 2017; Egger and Yu, 2022; Wu et al., 2024a), so that each topic represents a theme by grouping together related words.

Topic modeling remains an important approach in modern text analytics. Unlike black-box classifiers or embedding-based clustering methods, topic models provide interpretable groupings of words and documents. This makes them especially useful in domains like social science, public health, or digital humanities, where understanding what the model has found matters as much as performance metrics. Moreover, topic modeling yields quantitative output: an overall prevalence for each topic, a prevalence for each word within the topic, and prevalences for each topic in each document. Hence topic modeling allows quantifiable exploration and tracking of topical trends. This gives topic modeling an advantage of transparency and trustworthiness over, e.g., large language model (LLM)-based queries whose outputs are not so easily quantifiable and verifiable in terms of their relation to the data. Our goal is to support such interpretability while enabling the discovery of *minority topics* that are often overlooked.

Minority topics are generally defined as themes that have low prevalence both corpus-wide and within individual documents, and we specifically focus on *domain-relevant minority content*. An example is mental health discussion in YouTube comments: it is rare compared to other discussion, and often mentioned only briefly within a comment. Minority topics are crucial for understanding niche but meaningful content—especially in domains like mental health—but tend to be overlooked by state-of-the-art (SOTA) models.

Firstly, conventional methods adopt probabilistic approaches such as Latent Dirichlet Allocation (LDA; Blei et al. 2003), and others such as non-negative matrix factorization (NMF; Lee and Seung 2000) which has been extended to several scenarios, as well as clustering-based techniques (Chen et al., 2019). Neural and LLM-based topic models (e.g., Top2Vec (Angelov, 2020), BERTopic (Grootendorst, 2022), FASTopic (Wu et al., 2024b), (Bianchi

[†]Equal contributions.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 35562–35587
November 4-9, 2025

et al., 2021a,b)), and embedding-clustering approaches (Sia et al., 2020) propose contextual representations, but still struggle to identify minority themes due to dominance of frequent topics.

Secondly, most topic models also lack the flexibility to incorporate domain knowledge, such as expert expectations about content. Methods such as Anchored Correlation Explanation (Gallagher et al., 2016) that do allow expert guidance typically require detailed specifications of expected topics (Steeg and Galstyan, 2014). An expert may not be able to predefine them, and relying on them can limit discovery of variations within topics. Other models also employ guided (Vendrow et al., 2021) or semi-supervised approaches (Lee et al., 2010; Jia et al., 2020; Lindstrom et al., 2022) by incorporating prior knowledge to guide the model toward finding desired topics in various fashions (Li et al., 2022). However, such guided topic models often struggle to detect such low-prevalence themes, and many require strong assumptions, hyperparameter tuning, or rigid supervision, limiting their ability to generalize to subtle or unexpected variations.

Our solution. To address the issues we propose a novel topic model using a specially constrained NMF. Our method integrates soft prevalence constraints and a unified seed word list, without requiring topic-specific supervision. We set inequality constraints on topic distributions in documents and word distributions within topics. The model is optimized to minimize a generalized Kullback-Leibler divergence reconstruction error under the constraints, using KKT conditions (Lange, 2013; Ghojogh et al., 2021), yielding multiplicative updates. This lets us distinguish data-driven topics in a nuanced way, ensuring minority themes are well represented in addition to majority ones.

Note that our solution is a topic model, which does more than just flag (label) content as relevant to the minority domain — it breaks that domain into meaningful subtopics, which is essential for deeper analysis. Our model does this in a well-grounded probabilistic manner, so that interpretable and quantitative exploration of topics becomes possible even for hard-to-capture minority topics.

Crucially, our model aims to discover minority topics without requiring them to be dominant in the corpus and without requiring seed words to be prominent within such topics. We do not enforce presence of minority content in all topics, nor do we maximize prevalence of minority content or prominence of seed words, as such approaches could distort their modeling. Rather, by mild constraints on prevalence of minority topics and distribution of seed word content, we let the model learn distinct, data-driven minority and majority topics.

Unlike models that force seed alignment or rely heavily on neural decoding (Lin et al., 2023), and models that require exact specified guidance and prior domain knowledge, our approach not only enhances representation of minority themes but also enables flexible topic discovery without imposing rigid prior structures on the data. This can be crucial for analyzing data where subtle variations in themes are key to understanding the domain.

Our key contributions are: 1. We target underrepresented topics: instead of enforcing guidance on all topics, we identify a subset under guidance constraints while leaving others unconstrained for flexibility. **2.** We incorporate domain knowledge without overspecification: our guidance does not require preexisting knowledge of topic divisions. **3.** We apply soft prevalence constraints to avoid overfitting to seed words, enabling balanced topic emergence. **4.** Our model is grounded in constrained NMF, and optimized via KKT optimization with multiplicative updates. **5.** Experiments show improvements over several baselines on synthetic data and extract high-quality domain-relevant topics in a case study on real-world mental health data.

Next, Section 2 discusses the related work. Sections 3, and 4 detail our method and its optimization. Section 5 details comparison experiments and a mental health case study. Section 5.3 gives results and A.12 findings; Section 7 concludes.

2 Related Work

We review baseline models and their variants (Zhao et al., 2021a), existing NMF and semi-supervised NMF models (Carbonetto et al., 2021; Lindstrom et al., 2022), and other supervised models.

Among probabilistic topic models LDA (Blei et al., 2003) has been widely used (Chen et al., 2019). LDA, LSA, and PLSA (Hofmann, 1999; Albalawi et al., 2020) are probabilistic models using Bayesian graphical structures with topics as latent variables. The methods prioritize discovering the most common patterns over documents as latent themes, but may struggle to represent less frequent trends (Das and Jain, 2024). LDA models documents as bags-of-words (BOW) with word counts drawn from topic-specific word distributions (Blei et al., 2003). This can overemphasize

frequent terms, and reliance on Gibbs sampling degrades performance on short texts with sparse co-occurrence (Chen et al., 2019). In contrast, NMF factorizes any non-negative matrix (e.g., TF-IDF) into interpretable components, and has proven effective in unsupervised clustering tasks (Chen et al., 2019; Obadimu et al., 2019; Carbonetto et al., 2021). We use NMF as a foundation of our method due to its flexibility but our constrained formulation is extendable to LDA-style count models too.

Due to its simplicity and effectiveness, NMF has become influential in data mining (Zhang, 2012). While NMF can outperform LDA, yielding higher-quality topics on short-text datasets, traditional NMF shows limited effectiveness in discovering expected topics and often overlooks crucial minority content in document collections (Chen et al., 2019; Egger and Yu, 2022) and other downstream tasks (Vendrow et al., 2021). Unsupervised NMF approaches may learn meaningless or biased topics and often suffer from redundancy particularly when the data set is biased toward a set of features (Li et al., 2022; Chang et al., 2009; Jagarlamudi et al., 2012; Vendrow et al., 2021). To address the limitations, researchers have used slight supervision (Lee et al., 2010), such as incorporating class label knowledge in semi-supervised approaches (Jia et al., 2020) for downstream tasks (Jia et al., 2021). Another study (Haddock et al., 2020) used maximum likelihood estimators under specific uncertainty distributions with multiplicative updates, showing flexibility across supervised tasks.

SeededLDA (Jagarlamudi et al., 2012) associates each topic with a seed set and biases topic assignment in documents containing matching seed words. KeyATM (Eshima et al., 2023) extends this by supporting topics without seeds and improving empirical robustness through selective seed specification and term weighting. Anchored CorEx (Gallagher et al., 2016) takes an information-theoretic approach, anchoring seed words to specific topics. GuidedNMF (Vendrow et al., 2021) incorporates weighted seed supervision to guide topic formation and support tasks like classification (Li et al., 2022). A recent semi-supervised variant (Lindstrom et al., 2022) further combines prior knowledge with label information for improved latent topic discovery. However, such models often overfit to seed content, limiting generalization. Zhang et al. (2023) also proposed a seed-guided method based on contextual pattern alignment, but this can reinforce predefined structure and constrain topic diversity.

HGTM (Das and Jain, 2024) models rare topics with multiple topic-prevalence distributions, but needs topic-specific seed initialization. STM (Das et al., 2013) uses hierarchical Bayesian modeling with stick-breaking processes to capture low-frequency themes. Top2Vec (Angelov, 2020), FASTopic (Wu et al., 2024b), and BERTopic (Grootendorst, 2022) have gained attention. However, they lack ability to identify hidden topical patterns in a corpus (Srivastava and Sutton, 2017) when prevalences of topics or themes are imbalanced. STM is a fully unsupervised Bayesian model which is unable to make use of seed word domain knowledge; in contrast, our method uses seed word supervision with explicit constraints—encouraging minority topics to include (but not rely exclusively on) seed words, while ensuring that documents without such cues are not misattributed to minority themes. HGTM requires strong supervision: it needs a specific seed word set for each rare event or topic to find. In contrast, our method only needs an overall seed word list for the domain and does not require known supervision of seed words for each individual topic; we use the overall list through flexible constraints, and we find the individual minority topics organically through fitting the model to the data, hence our model is better suited for domains where strong existing knowledge of each minority topic is not available.

Recent neural topic models improve interpretability by incorporating supervision or structural signals. SeededNTM (Lin et al., 2023) uses multi-level seed word guidance at both word and document levels. NeuroMax (Pham et al., 2024) aligns topics with PLM-based embeddings through mutual information and optimal transport. Anchor-based models build unsupervised hierarchies by clustering seed anchors into interpretable trees (Liu et al., 2024). Prior work has also explored covariate-based (Eisenstein et al., 2011; Card et al., 2018) and taxonomy-driven topic modeling (Lee et al., 2022a,b), but such methods rely on structured metadata or external hierarchies. In contrast, our approach requires only a single seed list and mild constraints, enabling flexible recovery of minority topics from imbalanced data.

We focus on minority topics specific to our domain (e.g., mental health), rather than all rare content. This makes our task more challenging and sets our work apart from prior approaches (Das and Jain, 2024; Das et al., 2013; Zhang et al., 2023; Eshima et al., 2023; Haddock et al., 2020), etc.

Another strand of work extends NMF to federated learning. (Si et al., 2022) proposed FedNMF and its variant FedNMF+MI, which enable collaborative topic modeling without sharing raw data. The latter incorporates a mutual information regularizer between local representations and topic weights to improve coherence when data are non-i.i.d. across clients. These approaches are well-suited for privacy-sensitive and decentralized applications. In contrast, our method is designed for centralized corpora and modifies the factorization itself through seed-guided constraints, with the goal of uncovering low-prevalence minority topics.

3 METHOD

Classical NMF (Wang and Zhang, 2013; Egger and Yu, 2022; Wu et al., 2024a) creates a low-rank approximation of a non-negative valued matrix V representing a corpus. It is approximately decomposed as $V \approx WH$ (Carbonetto et al., 2021). The matrices W and H are fitted to minimize a reconstruction error objective function, here a generalized Kullback-Leibler (KL) divergence (Joyce, 2011):

$$D_{KL}(V \parallel WH) = \sum_{i,j} \left(V_{ij} \log \frac{V_{ij}}{(WH)_{ij}} - V_{ij} + (WH)_{ij} \right)$$

Alternative forms of the penalty function are available, such as the Frobenius norm. Optimization iterates multiplicative update rules similar to those by Lee and Seung (2000); Lin (2007). This process ensures W and H are adjusted to provide an optimal low-rank approximation of V, while maintaining non-negativity of the factors. The classical NMF update rules for W and H are:

$$W_{ik} \leftarrow W_{ik} \frac{\sum_{j'} \frac{H_{kj'} V_{ij'}}{(WH)_{ij'}}}{\sum_{j'} H_{kj'}}, \quad H_{kj} \leftarrow H_{kj} \frac{\sum_{i'} \frac{W_{i'k} V_{i'j}}{(WH)_{i'j}}}{\sum_{i'} W_{i'k}}$$

3.1 Our Constrained NMF

Let $V \in \mathbb{R}^{M \times N}$ be a document-term frequency matrix, where each row i corresponds to a document and each column j represents the frequency of a word across all documents. Let $W \in \mathbb{R}^{M \times K}$ be the document-topic distribution matrix, where K is the number of topics, and $H \in \mathbb{R}^{K \times N}$ is the topic-word distribution matrix. In the context of topic modeling, W shows the distribution of topics across the documents in the corpus, and H captures the significance of terms across the topics.

Our model sets constraints on W and H to target minority themes while reducing noise. Constraints on W ensure documents without any seed words do not have high prevalence of minority topics, while constraints on H align topics with domain-specific seed words, anchoring them to the content of interest. This dual technique allows effective minority topic detection without overfitting or emphasizing noise. Our new NMF method minimizes the generalized KL divergence with constraints on the low-rank matrices W and H, given a user-defined seed word list. The constraints on W and H are detailed below; they can be written as two sets of inequality constraints, g_1 and g_2, applied to the two factor matrices. We show the constraints satisfy KKT conditions (Lange, 2013; Ghojogh et al., 2021). Based on the conditions, we derive an optimization algorithm to find W and H minimizing the cost under the constraints. This yields a new multiplicative update rule differing from standard NMF.

In our case study we use mental health discussion as a minority domain of interest, within YouTube comment data where the majority of content is not mental health related. We denote minority topics as 'mental health topics' for concreteness, but the method is general. The user can set the number K_{MH} of mental health topics (minority) to be modeled with guidance. The other $K - K_{MH}$ topics model other topical content: majority topics and minority content not interesting to the expert. K_{MH} can be set as a desired level of detail in minority content but could also be chosen by typical model selection criteria. Guidance is provided as a set of seed words: a subset of terms known to be of interest in the minority domain [1]. We do not require the list to be comprehensive, and we do not require known divisions of words to predefined topics. We also do not require a known prior of minority content prevalences, or seed word prevalences. Thus the guidance is easy for experts to provide, and modeling will discover divisions and prevalences of minority topics in a data driven way.

The constraints g_1 (Constraints on W) ensure that in documents where none of the known seedwords occur, prevalence of each minority topic should be at most an upper bound value. The constraints g_2 (Constraints on H) ensure that in each minority topic, at least some of the known-to-be-relevant seed words should have sufficient prevalence in the topic-to-word distribution, so that the total prevalence of the seed words is at least a lower

[1] The seed word list provided in the GitHub repository consists of a single collection of mental health-related terms, without any predefined division into specific topics.

bound value. Both constraints are designed to make maximal use of the seed words, and separate minority topics from noise, while also allowing the model to discover how minority content is divided across topics, and without needing separate seed words for each topic to be discovered.

Constraint on W. Given the list of seed word indices SI, we identify the subset of documents not having any seed words, $I_0 = \{i : \sum_{j \in SI} V_{ij} = 0\}$. For documents $i \in I_0$, we limit prevalence of each minority topic (in our case mental health topic) to at most a maximum W_{max} which can be set by the user or by model selection strategies. This yields per-element inequality constraints $g_{1,ik}$,

$$g_{1,ik}(W) = W_{ik} - W_{\text{max}} \leq 0 \quad \forall k \in S_{MH},\ \forall i \in I_0 \quad (1)$$

where for convenience we denote minority topics as the first K_{MH} topics $S_{MH} = \{1, \ldots, K_{MH}\}$. This mild constraint states documents without seed words should not have *strong* prevalence of minority topics; prevalence up to the user-set maximum is allowed. We do not set a converse constraint: documents having seed words are *not* constrained to have high prevalence of minority topics.

Constraint on H. We define mild constraints $g_{2,k}$ on seed word content in minority topics:

$$g_{2,k}(H) = \theta_{\text{min}} - \frac{\sum_{j' \in SI} H_{kj'}}{\sum_{j'=1}^{N} H_{kj'}} \leq 0, \quad \forall k \in S_{MH} \quad (2)$$

The user can set θ_{min} to control the focus on seed words in minority topics. We only use the mild overall constraint, and do not constrain prevalence of specific seed words per topic: which seed words may become prevalent in each minority topic is found by model fitting. We do not require a converse constraint, i.e., other topics are not required to avoid seed words.

3.2 Lagrangian Formulation

The Lagrangian (Leech, 1965), $L(W, H, \lambda, \mu)$, integrates the objective and the constraints. The objective is to minimize KL divergence $D_{KL}(V \parallel WH)$ measuring how different the document-word matrix V is from the product of W (document-topic distribution) and H (topic-word distribution). Penalty terms are added to ensure constraints on W and H are met. The Lagrangian L is:

$$L(W, H, \lambda, \mu) = D_{KL}(V \parallel WH) + \lambda \cdot g_1(W) + \mu \cdot g_2(H) \quad (3)$$

where $g_1(W)$ and $g_2(H)$ are sums over sets of constraints, and λ and μ are Lagrange multipliers penalizing constraint violations. The $g_1(W)$ sums constraints $g_{1,ik}(W)$ over indices i, k, corresponding to elements of W, while $g_2(H)$ sums $g_{2,k}(H)$ over indices k, corresponding to rows of H.

3.3 Karush-Kuhn-Tucker (KKT) Conditions

An optimal solution must satisfy KKT conditions (Lange, 2013; Ghojogh et al., 2021) as follows.

Stationarity. To establish the stationarity condition in our optimization problem, we set gradients of the Lagrangian L to zero with respect to the W and H. With respect to the KL divergence and inequality constraints $g_1(W)$ and $g_2(H)$, the requirements guarantee we are at a crucial point where changes in the cost are balanced.

$$\frac{\partial L}{\partial W_{ik}} = \frac{\partial D_{KL}(V \parallel WH)}{\partial W_{ik}} + \lambda_{ik} \frac{\partial g_1(W)}{\partial W_{ik}} = 0 \quad (4)$$

$$\frac{\partial L}{\partial H_{kj}} = \frac{\partial D_{KL}(V \parallel WH)}{\partial H_{kj}} + \mu_k \frac{\partial g_2(H)}{\partial H_{kj}} = 0 \quad (5)$$

Primal Feasibility. For W, this means the sum of specific elements (related to minority topics, like mental health) is capped by a maximum value W_{max}. For H, we ensure the proportion of seed words in each topic is at least a user-defined minimum θ_{min}, as given in Eq. (1), and Eq. (2). This ensures the solution is within reasonable bounds and remains meaningful.

Dual Feasibility. The Lagrange multipliers, λ and μ, must be non-negative. They denote strength of the penalty when a constraint is violated. If a constraint isn't violated, the multiplier can be zero. If the constraint is violated, the penalty pushes the solution back within the desired limits.

$$\lambda_{ik} \geq 0 \ , \ \mu_k \geq 0 \quad (6)$$

Complementary Slackness. If a constraint is already satisfied, there's no need for a penalty. For instance, if elements in W for documents without seed words are below W_{max}, the multiplier λ will be zero. Similarly, if the proportion of seed words in mental health topics exceeds θ_{min}, the multiplier μ will be zero. The product of each multiplier and its corresponding constraint must be zero:

$$\lambda_{ik} \left(W_{ik} - W_{\text{max}} \right) = 0 \ , \ \mu_k \left(\theta_{\text{min}} - \frac{\sum_{j' \in SI} H_{kj'}}{\sum_{j'=1}^{N} H_{kj'}} \right) = 0 \quad (7)$$

To meet and solve the KKT criteria we derive the gradients of above equations in Appendix A.1- A.8.

4 Optimization

We derive multiplicative update rules for the factorization matrices W and H, to optimize the objective function under our constraints. Our Constrained NMF updates W and H iteratively. Convergence properties of multiplicative update rules for unconstrained NMF have been studied (Gonzalez and Zhang, 2005; Berry et al., 2007) including by Lin (2007) for Euclidean distance and Finesso and Spreij (2006) for generalized KL divergence. Convergence of multiplicative rules for NMF with inequality constraints remains open as the lifting approach of Finesso and Spreij (2006) is not immediately applicable. Our updates derived using KKT conditions (Lange, 2013; Ghojogh et al., 2021) guarantee constraints are satisfied. In experiments our algorithm consistently demonstrated (Appendix A.13, Figure 3) to i) be nonincreasing with respect to the loss, and ii) converge to a stable point. Convergence theorems are left to future work.

We apply the KKT conditions (Lange, 2013; Ghojogh et al., 2021) to handle the constraints such that optimal solutions satisfy non-negativity of the matrices and our domain constraints. The multiplicative update rules allow efficient updates while retaining non-negativity of W and H. The rules arise from gradients of the Lagrangian, setting them to zero and solving for W and H iteratively. We derive update rules not only for elements W_{ik} and H_{kj} of W and H but also for Lagrange multipliers λ_{ik} and μ_k that control satisfaction of the corresponding constraints (see Eqs. 12 & 13). To optimize the objective under our constraints, we update W_{ik}, H_{kj}, λ_{ik}, and μ_k by the rules.

4.1 The Multiplicative Update Rule for W_{ik}

We set the Lagrangian gradient $\frac{\partial L}{\partial W_{ik}}$ to zero to satisfy the KKT condition (Lange, 2013; Ghojogh et al., 2021). Adding appendix Eq. (14) and Eq. (16), we derive the gradient:

$$\frac{\partial L}{\partial W_{ik}} = \sum_{j'} H_{kj'} \left(1 - \frac{V_{ij'}}{(WH)_{ij'}}\right) + \lambda_{ik}\delta_{i\in I_0,\ k\in S_{MH}} \tag{8}$$

where the δ function is 1 if $k \in S_{MH}$ and $i \in I_0$, and 0 otherwise.

Final Multiplicative Update Rule for W_{ik}: To maintain non-negativity of W_{ik}, we solve $\frac{\partial L}{\partial W_{ik}} = 0$ which yields that the ratio of the positive and negative terms in the gradient must be 1. This yields the multiplicative update for W_{ik}:

$$W_{ik} \leftarrow W_{ik} \cdot \frac{\sum_{j'} H_{kj'} \frac{V_{ij'}}{(WH)_{ij'}}}{\sum_{j'} H_{kj'} + \lambda_{ik}\delta_{i\in I_0,\ k\in S_{MH}}} \tag{9}$$

If the constraint $g_{1,ik}(W)$ is active for some $i \in I_0,\ k \in S_{MH}$, the Lagrange multiplier λ_{ik} adjusts the update to keep the constraint satisfied. If the constraint is not active, then $\lambda_{ik} = 0$. The updating process is presented in Algorithm 1.

4.2 The Multiplicative Update Rule for H_{kj}

The update rule for H_{kj} is derived from setting the gradient $\frac{\partial L}{\partial H_{kj}}$ of the Lagrangian to zero. By appendix Eq. (15) and Eq. (17) the gradient becomes

$$\frac{\partial L}{\partial H_{kj}} = \sum_{i'} W_{i'k} \left(1 - \frac{V_{i'j}}{(WH)_{i'j}}\right) + \mu_k \cdot \delta_{k\in S_{MH}} \left(\frac{\text{Num}_k}{(\text{Den}_k)^2} - \delta_{j\in SI}\frac{1}{\text{Den}_k}\right) \tag{10}$$

where $\text{Den}_k = \sum_{j'} H_{kj'}$, $\text{Num}_k = \sum_{j'\in SI} H_{kj'}$. Setting the above to zero, in the appendix we derive two multiplicative update rules, version 1 is provided below as the **Final Multiplicative Update Rule for H_{kj}**:

$$H_{kj} \leftarrow H_{kj} \cdot \frac{\sum_{i'} W_{i'k} \frac{V_{i'j}}{(WH)_{i'j}}}{\sum_{i'} W_{i'k} + \mu_k \cdot \delta_{k\in S_{MH}} \left(\frac{\text{Num}_k}{(\text{Den}_k)^2} - \frac{\delta_{j\in SI}}{\text{Den}_k}\right)} \tag{11}$$

See the updating process in Algorithm 2.

4.3 Update Rules for Lagrange Multipliers

Lagrange multipliers λ and μ are updated based on whether the constraints are active or not. The λ_{ik} and μ_k are updated using gradient ascent to ensure that they enforce the constraints on W and H.

Update Rule for λ_{ik}.
Given our constraint on W given in Eq. (1):
If $g_{1,ik}(W) < 0$: the constraint is inactive, we set $\lambda_{ik} = 0$.
If $g_{1,ik}(W) \geq 0$: the constraint is active, we update λ_{ik} using the following rule.

For active constraints, λ_{ik} is updated iteratively. One possible method a gradient ascent update:

$$\lambda_{ik}^{\text{new}} = \max\left(0, \lambda_{ik}^{\text{old}} + \eta \cdot g_{1,ik}(W)\right) \tag{12}$$

where η is our learning rate, controlling how aggressively λ_{ik} is updated. The max function ensures λ_{ik} stays non-negative, as required by our dual feasibility.

Update Rule for μ_k.

Recalling Eq. (2), given our constraint on H:

If $g_{2,k}(H) < 0$: The constraint is inactive, we set $\mu_k = 0$.

If $g_{2,k}(H) \geq 0$: The constraint is active, we update μ_k using the following rule.

For active constraints, μ_k is updated similarly to λ_{ik}:

$$\mu_k^{\text{new}} = \max\left(0, \mu_k^{\text{old}} + \eta \cdot g_{2,k}(H)\right) \tag{13}$$

During the optimization, the constraints are not satisfied at every iteration. Each Lagrange multiplier (λ_{ik} and μ_k) becomes active (nonzero) if its corresponding constraint is violated. The algorithm thus adjusts W and H iteratively, striving to balance satisfaction of the constraints with minimizing the primary objective function. Eq. (9) to Eq. (13) provide easy-to-implement updating rules. Until the objective value of Eq. (3) remains unchanged, we iteratively modify W and H. Algorithm 3 gives an outline of this approach. The time complexity of our model is detailed in the Appendix A.10. We demonstrate the strength of our method in the next section by comparing its performance on synthetic data to SOTA models and carrying out a case study.

Algorithm 1 Multiplicative Update Rule for the W_{ik}

Input: $V \in \mathbb{R}^{M\times N}$, $W \in \mathbb{R}^{M\times K}$, $H \in \mathbb{R}^{K\times N}$, λ_{ik} (Lagrange multipliers for $g_{1,ik}(W)$), $W_{\max}$ (upper bound constraint)
Output: Updated W_{ik}
repeat
  for each $i \in [1, M]$ and $k \in [1, K]$ **do**
    numerator $\leftarrow \sum_{j'} H_{kj'} \cdot \left(\frac{V_{ij'}}{(WH)_{ij'}}\right)$
    denominator $\leftarrow \sum_{j'} H_{kj'}$
    if $i \in I_0$ and $k \in S_{MH}$ **then**
      Add λ_{ik} to denominator
    end if
    $W_{ik} \leftarrow W_{ik} \cdot \frac{\text{numerator}}{\text{denominator}}$
  end for
until convergence

5 Experiments

We evaluate our model on synthetic and real-world Finnish YouTube data. We benchmark against a wide range of conventional and neural topic models, including NMF (Lee and Seung, 2000), LDA (Blei et al., 2003), ProdLDA (Srivastava and Sutton, 2017), Top2Vec (Angelov, 2020), BERTopic (Grootendorst, 2022), FASTopic (Wu et al., 2024b), KeyATM (Eshima et al., 2023), GuidedNMF (Vendrow et al., 2021), GuidedLDA and SeededLDA (Jagarlamudi et al., 2012), and Corex (Gallagher et al., 2016). We used publicly available implementations (see Appendix A.10), with shared preprocessing, seed word list, and a uniform topic count [2]. The code of our model is available. [3]

Algorithm 2 Multiplicative Update Rule for the H_{kj}

Input: $V \in \mathbb{R}^{M\times N}$, $W \in \mathbb{R}^{M\times K}$, $H \in \mathbb{R}^{K\times N}$, μ_k, seed indices, S_{MH}, $\theta_{\min}$
Output: Updated H_{kj}
for each $k \in [1, K]$ and $j \in [1, N]$ **do**
  numerator $\leftarrow \sum_{i'} W_{i'k} \cdot \left(\frac{V_{i'j}}{(WH)_{i'j}}\right)$
  denominator $\leftarrow \sum_{i'} W_{i'k}$
  if $k \in S_{MH}$ **then**
    Compute constraint term:
    constraint $\leftarrow \mu_k \cdot \left(\frac{\text{Num}_k}{(\text{Den}_k)^2} - \delta_{j\in SI} \cdot \frac{1}{\text{Den}_k}\right)$
    denominator $\leftarrow$ denominator + constraint
  end if
  $H_{kj} \leftarrow H_{kj} \cdot \frac{\text{numerator}}{\text{denominator}}$
end for

Algorithm 3 Optimization Procedure for W and H

Input: V, initial W, H, constraints $g_1(W)$, $g_2(H)$
Output: Optimized W, H
Initialize λ_{ik}, μ_k (Lagrange multipliers)
repeat
  Update W_{ik} via Alg. 1
  Update H_{kj} via Alg. 2
  Check $g_1(W) \leq 0$ and $g_2(H) \leq 0$
  Update λ_{ik} (see Eq. 1):
  if $g_{1,ik}(W) < 0$ **then**
    $\lambda_{ik} \leftarrow 0$ {Inactive constraint}
  else
    Update λ_{ik} via Eq. 12
  end if
  Update μ_k (see Eq. 2):
  if $g_{2,k}(H) < 0$ **then**
    $\mu_k \leftarrow 0$ {Inactive constraint}
  else
    Update μ_k via Eq. 13
  end if
until convergence or stopping condition is met
Return : optimized W, H

5.1 Datasets

Real Dataset. We selected 20 Finnish YouTubers recommended by a public health expert, focusing on mental health content for younger audiences. We scraped comments and metadata from their videos up to March 15, 2024. This yielded roughly 5.5 million Finnish-language comments. While most comments cover various topics, mental health discussions are a minority. Table 1 summarizes statistics of both datasets.

[2] We used default hyperparameters unless otherwise noted, tuning only when required for baseline stability or fairness.

[3] https://github.com/seyedeh-mona-ebrahimi/Constrained-NMF-for-Minority-Topics

Statistic	Real	Synthetic
Documents (raw)	5,578,289	500
Documents (filtered)	2,979,969	500
Avg. words/doc	7.04	10
Vocabulary size	885,945	2,800

Table 1: Dataset statistics after preprocessing.

Synthetic Dataset. To build a synthetic dataset, topic-related words were injected into randomly selected sentences to simulate mental health discussions. A synthetic ground truth of 18 mental health topics was defined, each with related Finnish words. Topics like *Suicide* and *Anxiety* had terms such as *"itsemurha"* (suicide) and *"ahdistus"* (anxiety), while others like *Mental Health* and *Social Isolation* included words like *"yksinäisyys"* (loneliness) and *"trauma."* To generate the data, we randomly sampled 500 documents. Each had a 10% chance to receive mental health content: if selected, one synthetic topic was randomly assigned, and four related words were injected. The remaining documents were left unchanged and labeled "-1" to indicate no mental health topic. This created a realistic mix of general and topic-specific content.

5.2 Evaluation Metrics

We evaluate clustering quality by Purity and normalized mutual information (NMI) (Manning et al., 2008) following (Wu et al., 2023; Zhao et al., 2021b). We further introduce a customized purity function that focuses on true minority topics ignoring background (non-mental-health) labels. Standard Purity reflects majority content, which is not our goal. To address this, we compute a focused purity score measuring only accuracy of minority topics. In the synthetic dataset, each document $d = 1, \ldots, M$ is either injected with minority-related (mental health) words from one ground-truth minority topic (labeled $y_d = 1, \ldots, Y$) or (labeled $y_d = -1$) for background. The model assigns each document d to its most probable predicted topic t_d. For each predicted topic k, we count how many documents are assigned; $\text{count}(k) = \sum_{d=1}^{M} \delta(t_d = k)$, where $\delta(\cdot)$ is the indicator function. Similarly, how many of those belong to a ground-truth minority label y; $\text{count}(y, k) = \sum_{d=1}^{M} \delta(y_d = y,\, t_d = k)$. The dominant valid label for each topic k is defined as $y^k = \arg\max_{y>-1} \text{count}(y, k)$, excluding the background. The purity score is computed as:

$$\text{Purity} = \frac{\sum_{k=1}^{K} \text{count}(y^k, k)}{\sum_{k=1}^{K} \delta(\text{count}(y^k, k) > 0) \cdot \text{count}(k)}$$

To assess topic quality beyond clustering, we compute JSD between injected ground-truth topic-word distributions and the model's learned topics. For each ground-truth topic, we report the minimum JSD to its closest learned topic:

$$\min_k \text{JSD}(P_{\text{true}}, P^{(k)}_{\text{model}})$$

JSD is more reliable for evaluating low-prevalence content than standard coherence and perplexity metrics, as it compares to ground-truth minority content. In detail, unsupervised coherence and perplexity scores would fail to reflect whether models can recover minority topics: coherence and perplexity both have the problem that they are dominated by the majority content. Coherence tends to reward dominant, frequent themes (coherence is highest for frequent topics which represent the majority content), and thus it does not favor presenting minority topics. Therefore, coherence would be biased in favor of models that focus on the majority content. Similarly, perplexity evaluates the surprise of the entire document content, and for most documents the content is dominated by the majority topics, thus perplexity is biased in favor of models that focus on the majority content. For these reasons coherence and perplexity are not well suited to capture performance on our task; thus we use JSD to evaluate topic quality.

5.3 Results

On the synthetic set the models were tasked to find 20 topics (7 minority and 13 majority). Conventional NMF, LDA, Top2Vec, FASTopic, ProdLDA were run without seed word guidance. Corex, GuidedLDA, SeededLDA, BERTopic, KeyATM, and GuidedNMF were run with the same seed words as our model's setting. The quality of the models was then evaluated by the metrics discussed above.

Figure 1 reports NMI and Purity scores across the models on the synthetic dataset and Appendix C Figure 9 shows a corresopnding scatterplot of Purity and NMI. [4] Our model achieves the highest scores on both metrics and outperforms all the baselines, including both unsupervised and seed-guided approaches. Furthermore, our approach guided the model to discover latent themes within the topics, rather than forcing seed words into the content, unlike models such as Corex, GuidedLDA, and

[4] We do not report NMI, Purity, or JSD for the real dataset, as it lacks ground-truth topic labels, but we present the discovered topics for interpretability.

SeededLDA. Notably, an inspection of the results revealed that their strategy predominantly incorporated seed words directly into the discovered topics, resulting in poor interpretability and performance. We also performed additional analyses on using various numbers of topics (see Appendix C, Figures 7 and 8, again showing good performance of our model.)

Figure 2 and Table 2 report the topic quality of the models according to JSD. Here as well, our method yielded best performance (smallest JSDs) in discovering minority themes. Thus we achieve both high topic quality (small JSD) and high clustering ability (high purity & NMI) outperforming others.

The detailed topics discovered by our model are shown in Appendix A.12, Table 7.

Model	Purity	NMI
NMF	0.0969	0.1846
LDA	0.0500	0.1131
Corex	0.0677	0.1697
GuidedLDA	0.0635	0.1721
SeededLDA	0.0923	0.1660
KeyATM	0.0878	0.1791
GuidedNMF	0.0723	0.2247
Top2Vec	0.0184	0.1055
FASTopic	0.0762	0.1617
BERTopic	0.0381	0.0914
ProdLDA	0.0577	0.1104
Ours	**0.1765**	**0.2453**

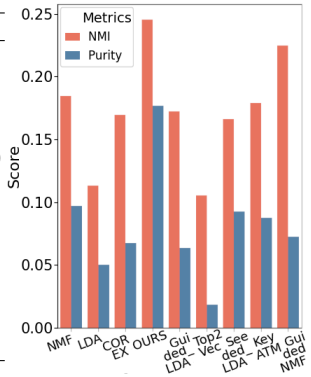

Figure 1: Comparison of NMI and Purity Scores across Baselines on synthetic dataset (20 topics, 7 mental health topics, 500 samples). Left: Result table, the best is in **bold**. Right: results as a bar graph.

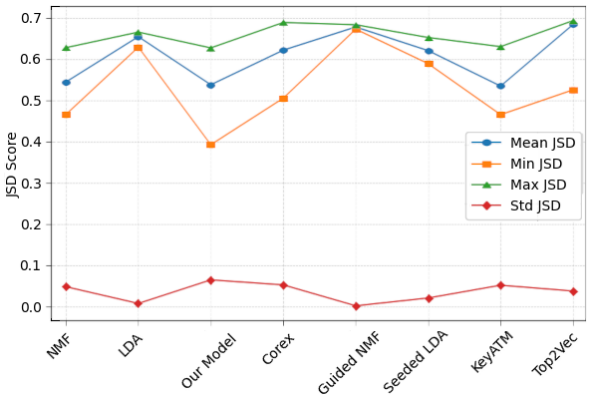

Figure 2: Topic Quality using JSD Score

5.4 Ablation Study

We ran an ablation study on the synthetic dataset to assess impact of modeling choices. We varied the number of total topics (30, 50, 80), number of minority topics (10, 15, 20), and hyperparameters $W_{\max}$ and $\theta_{\min}$. We assessed their effect on KL divergence, NMI, and purity. Our model is robust and outperformed baselines on all settings. Detailed results are in Appendix B in Figures 5, 6 and corresponding Tables 8, 9.

Model	Mean_JSD	Min_JSD	Max_JSD	Std_JSD
NMF	0.54451	0.46581	0.62761	0.04880
LDA	0.65429	0.62900	0.66529	0.00827
Ours	0.53783	0.39268	0.62705	0.06538
Corex	0.62160	0.50528	0.68862	0.05325
Guided NMF	0.67771	0.67248	0.68317	0.00262
Seeded LDA	0.62048	0.58895	0.65224	0.02149
KeyATM	0.53462	0.46541	0.63016	0.05245
Top2Vec	0.68383	0.52559	0.69314	0.03837

Table 2: JSD statistics across models.

6 Discovered Topics

We present example outputs from our Constrained NMF. Topics discovered in the real data include highly meaningful mental health concerns such as *How mental health differs from outward appearance* (Topic 0, top words crazy, appearance, medicine), *How mental health problems may exacerbate around holidays* (Topic 1; expectation, christmas, mental health problem), *Sadness and suicide* (Topic 2; sad, suicide, human), and *Support and ADHD* (Topic 3; to support, adhd, crisis). Full lists in Finnish with English translations for clarity are provided in the Appendix: discovered topics on the synthetic dataset in Appendix A.12, Table 7, and Tables 3–6 for our real-world YouTube dataset.

7 Conclusions and Future Work

We introduced a constrained NMF model for discovering domain-specific minority topics without explicit topic-level supervision. We use soft prevalence constraints and a single seed word list to guide discovery of distinct, data-driven minority themes. In experiments on synthetic and real-world YouTube comment data, our method outperforms strong baselines in clustering and topic quality, successfully modeling low-prevalence mental health discussions. This shows potential of constrained matrix factorization to identify patterns in noisy, imbalanced corpora. Future work includes expanding to new domains and exploring integration with neural or contextualized topic models.

Limitations

Evaluating low-frequency topics remains challenging, as standard metrics often fail to reflect minority theme quality. While we use synthetic ground truth, real-world datasets lack annotated minority themes, and benchmarks in domains like mental health are scarce. To estimate the quality of a topic, we compute the JSD in discovered and ground-truth topic distributions in synthetic settings. Although, NMI, purity, JSD and similar automated measures are informative, these methods risk being misaligned or biased when it comes to the value of meaning (Hoyle et al., 2021). More reliable measures of semantic value include human assessment, which can enhance evaluation of topic coherence and relevance in so-called minority content. Future work should address this gap.

Acknowledgement

This work was supported by Academy of Finland decision 348523.

References

Rania Albalawi, Tet Hin Yeap, and Morad Benyoucef. 2020. Using topic modeling methods for short-text data: A comparative analysis. *Frontiers in Artificial Intelligence*, 3.

Dimo Angelov. 2020. Top2vec: Distributed representations of topics. *ArXiv*, abs/2008.09470.

Michael W. Berry, Murray Browne, Amy N. Langville, V. Paul Pauca, and Robert J. Plemmons. 2007. Algorithms and applications for approximate nonnegative matrix factorization. *Computational Statistics & Data Analysis*, 52(1):155–173.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021a. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 759–766, Online. Association for Computational Linguistics.

Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021b. Cross-lingual contextualized topic models with zero-shot learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1676–1683, Online. Association for Computational Linguistics.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Peter Carbonetto, Abhishek Sarkar, Zihao Wang, and Matthew Stephens. 2021. Non-negative matrix factorization algorithms greatly improve topic model fits. *arXiv preprint arXiv:2105.13440*.

Dallas Card, Chenhao Tan, and Noah A. Smith. 2018. Neural models for documents with metadata. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2031–2040, Melbourne, Australia. Association for Computational Linguistics.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan Boyd-graber, and David Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

Yong Chen, Hui Zhang, Rui Liu, Zhiwen Ye, and Jianying Lin. 2019. Experimental explorations on short text topic mining between LDA and NMF based schemes. *Knowledge-Based Systems*, 163:1–13.

Mrinal Das, Suparna Bhattacharya, Chiranjib Bhattacharyya, and Gopinath Kanchi. 2013. Subtle topic models and discovering subtly manifested software concerns automatically. In *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 253–261, Atlanta, Georgia, USA. PMLR.

Mrinal Das and Gaurav Jain. 2024. Human guided multi-proportions topic model for rare event detection without using labels. In *Intelligent Systems and Applications*, pages 368–385, Cham. Springer Nature Switzerland.

Roman Egger and Joanne Yu. 2022. A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. *Frontiers in Sociology*, 7.

Jacob Eisenstein, Amr Ahmed, and Eric P Xing. 2011. Sparse additive generative models of text. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 1041–1048.

Shusei Eshima, Kosuke Imai, and Tomoya Sasaki. 2023. Keyword-assisted topic models. *American Journal of Political Science*, 68.

Lorenzo Finesso and Peter Spreij. 2006. Nonnegative matrix factorization and I-divergence alternating minimization. *Linear Algebra and its Applications*, 416(2–3):270–287.

Ryan J. Gallagher, Kyle Reing, David C. Kale, and Greg Ver Steeg. 2016. Anchored correlation explanation: Topic modeling with minimal domain knowledge. *Transactions of the Association for Computational Linguistics*, 5:529–542.

Benyamin Ghojogh, Ali Ghodsi, Fakhri Karray, and Mark Crowley. 2021. KKT conditions, first-order

and second-order optimization, and distributed optimization: tutorial and survey. *arXiv preprint arXiv:2110.01858*.

Edward F Gonzalez and Yin Zhang. 2005. Accelerating the lee-seung algorithm for non-negative matrix factorization. *Dept. Comput. & Appl. Math., Rice Univ., Houston, TX, Tech. Rep. TR-05-02*, pages 1–13.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Jamie Haddock, Lara Kassab, Sixian Li, Alona Kryshchenko, Rachel Grotheer, Elena Sizikova, Chuntian Wang, Thomas Merkh, R. W. M. A. Madushani, Miju Ahn, Deanna Needell, and Kathryn Leonard. 2020. Semi-supervised NMF models for topic modeling in learning tasks. *ArXiv*, abs/2010.07956.

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, page 50–57, New York, NY, USA. Association for Computing Machinery.

Alexander Hoyle, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken? the incoherence of coherence. *Advances in neural information processing systems*, 34:2018–2033.

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

Yuheng Jia, Sam Kwong, Junhui Hou, and Wenhui Wu. 2020. Semi-supervised non-negative matrix factorization with dissimilarity and similarity regularization. *IEEE Transactions on Neural Networks and Learning Systems*, 31(7):2510–2521.

Yuheng Jia, Hui Liu, Junhui Hou, and Sam Kwong. 2021. Semisupervised adaptive symmetric nonnegative matrix factorization. *IEEE Transactions on Cybernetics*, 51(5):2550–2562.

James M. Joyce. 2011. Kullback-Leibler divergence. In *International Encyclopedia of Statistical Science*, Berlin, Heidelberg. Springer.

Kenneth Lange. 2013. *Karush-Kuhn-Tucker Theory*, pages 107–135. Springer New York, New York, NY.

Daniel Lee and H Sebastian Seung. 2000. Algorithms for non-negative matrix factorization. *Advances in neural information processing systems*, 13.

Dongha Lee, Jiaming Shen, SeongKu Kang, Susik Yoon, Jiawei Han, and Hwanjo Yu. 2022a. Taxocom: Topic taxonomy completion with hierarchical discovery of novel topic clusters. In *Proceedings of the ACM Web Conference 2022*, pages 2819–2829.

Dongha Lee, Jiaming Shen, Seonghyeon Lee, Susik Yoon, Hwanjo Yu, and Jiawei Han. 2022b. Topic taxonomy expansion via hierarchy-aware topic phrase generation. *arXiv preprint arXiv:2211.01981*.

Hyekyoung Lee, Jiho Yoo, and Seungjin Choi. 2010. Semi-supervised nonnegative matrix factorization. *IEEE Signal Processing Letters*, 17:4–7.

J. W. Leech. 1965. *The Lagrangian Formulation*, pages 17–25. Springer Netherlands, Dordrecht.

Pengyu Li, Christine Tseng, Yaxuan Zheng, Joyce A. Chew, Longxiu Huang, Benjamin Jarman, and Deanna Needell. 2022. Guided semi-supervised nonnegative matrix factorization. *Algorithms*, 15(5).

Chih-Jen Lin. 2007. On the convergence of multiplicative update algorithms for nonnegative matrix factorization. *IEEE Transactions on Neural Networks*, 18(6):1589–1596.

Yang Lin, Xin Gao, Xu Chu, Yasha Wang, Junfeng Zhao, and Chao Chen. 2023. Enhancing neural topic model with multi-level supervisions from seed words. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13361–13377.

Michael R. Lindstrom, Xiaofu Ding, Feng Liu, Anand Somayajula, and Deanna Needell. 2022. Continuous semi-supervised nonnegative matrix factorization. *Algorithms*, 16:187.

Jiyuan Liu, Hegang Chen, Chunjiang Zhu, and Yanghui Rao. 2024. Unsupervised hierarchical topic modeling via anchor word clustering and path guidance. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7505–7517.

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

Adewale Obadimu, Esther Mead, and Nitin Agarwal. 2019. Identifying latent toxic features on YouTube using non-negative matrix factorization. In *The Ninth International Conference on Social Media Technologies, Communication, and Informatics, IEEE*.

Duy-Tung Pham, Thien Trang Nguyen Vu, Tung Nguyen, Linh Van Ngo, Duc Anh Nguyen, and Thien Huu Nguyen. 2024. NeuroMax: Enhancing neural topic modeling via maximizing mutual information and group topic regularization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7758–7772, Miami, Florida, USA. Association for Computational Linguistics.

Gerard Salton and Christopher Buckley. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, 24(5):513–523.

Shijing Si, Jianzong Wang, Ruiyi Zhang, Qinliang Su, and Jing Xiao. 2022. Federated non-negative matrix factorization for short texts topic modeling with mutual information. In *2022 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE.

Suzanna Sia, Ayush Dalmia, and Sabrina J. Mielke. 2020. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1728–1736, Online. Association for Computational Linguistics.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *International Conference on Learning Representations*.

Greg Ver Steeg and A. G. Galstyan. 2014. Discovering structure in high-dimensional data through correlation explanation. In *Neural Information Processing Systems*.

Joshua Vendrow, Jamie Haddock, Elizaveta Rebrova, and Deanna Needell. 2021. On a guided nonnegative matrix factorization. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3265–32369.

Yu-Xiong Wang and Yu-Jin Zhang. 2013. Nonnegative matrix factorization: A comprehensive review. *IEEE Transactions on Knowledge and Data Engineering*, 25(6):1336–1353.

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

Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. 2024a. A survey on neural topic models: Methods, applications, and challenges. *ArXiv*, abs/2401.15351.

Xiaobao Wu, Thong Thanh Nguyen, Delvin Ce Zhang, William Yang Wang, and Anh Tuan Luu. 2024b. FASTopic: Pretrained transformer is a fast, adaptive, stable, and transferable topic model. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Yu Zhang, Yunyi Zhang, Martin Michalski, Yucheng Jiang, Yu Meng, and Jiawei Han. 2023. Effective seed-guided topic discovery by integrating multiple types of contexts. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining*, WSDM '23, page 429–437, New York, NY, USA. Association for Computing Machinery.

Zhong-Yuan Zhang. 2012. Nonnegative matrix factorization: models, algorithms and applications. *Data Mining: Foundations and Intelligent Paradigms: Volume 2: Statistical, Bayesian, Time Series and other Theoretical Aspects*, pages 99–134.

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

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

A Appendix

This Appendix contains several derivations and details complementing the main paper, as follows.

We provide derivatives (gradients) of our generalized Kullback-Leibler divergence loss function in Sections A.1 and A.2, derivatives of the constraint functions in Sections A.3 and A.4, and derivatives of the full Lagrangian in Sections A.5 and A.6. The derivatives are used to solve multiplicative update rules satisfying the Karush-Kuhn-Tucker conditions; we provide the details in Sections A.7 and A.8. Next, Sections A.9, A.10, and A.11 include additional details, such as details concerning the training datasets, hyperparameter choices, the computer system utilized, time complexity, and training details. Section A.12 provides additional detailed experimental results for the synthetic and real data experiments, showing topics discovered by our proposed method in terms of their top words. Section A.13 provides error bars from multiple runs. Lastly, Appendix B provides an ablation study on the synthetic dataset, evaluating varying hyperparameter settings on different metrics, and Appendix C provides additional analysis on how varying the number of topics counts, minority supervision impacts NMI and purity, along with an evaluation of topic quality using JSD across several baselines.

A.1 KL Divergence Gradient with Respect to W_{ik}

As stated in Section 3 of the main paper, the generalized Kullback-Leibler (KL) divergence between V and WH is

$$D_{KL}(V \parallel WH) = \sum_{i',j'} \Big(V_{i'j'} \log \frac{V_{i'j'}}{(WH)_{i'j'}} - V_{i'j'} + (WH)_{i'j'} \Big)$$

where we use i' and j' as sum indices for clarity, to distinguish them from indices of specific elements whose derivatives we will compute.

The generalized KL divergence is a special case of a Bregman divergence. While the classical KL divergence compares two probability distributions whose elements are probabilities that sum to 1, the generalized KL divergence applies between two sets of nonnegative numbers whose elements do not need to sum to 1: here the sets are the elements of V and the corresponding elements of WH.

To find the derivative $\frac{\partial D_{KL}(V\|WH)}{\partial W_{ik}}$ with respect to an element W_{ik}, we will differentiate each term in D_{KL}.

First, recall:

$$(WH)_{i'j'} = \sum_{k'} W_{i'k'} H_{k'j'} \ .$$

The derivative of $(WH)_{i'j'}$ with respect to W_{ik} is then

$$\frac{\partial}{\partial W_{ik}}(WH)_{i'j'} = \frac{\partial}{\partial W_{ik}} \sum_{k'} W_{i'k'} H_{k'j'} = \delta_{i'i} H_{kj'}$$

where $\delta_{i'i} = 1$ if $i' = i$ and zero otherwise.

The first term inside the sum in the KL divergence is $V_{i'j'} \log \frac{V_{i'j'}}{(WH)_{i'j'}}$. Its derivative with respect to W_{ik} is

$$\frac{\partial}{\partial W_{ik}} \left(V_{i'j'} \log \left(\frac{V_{i'j'}}{(WH)_{i'j'}} \right) \right) = - \left(\frac{V_{i'j'}}{(WH)_{i'j'}} \right) \left(\frac{\partial}{\partial W_{ik}} (WH)_{i'j'} \right) = - \left(\frac{V_{i'j'}}{(WH)_{i'j'}} \right) (\delta_{i'i} H_{kj'})$$

Summing this over i' and j', we get the derivative of the first part of the sum:

$$\frac{\partial}{\partial W_{ik}} \sum_{i'j'} \left(V_{i'j'} \log \left[\frac{V_{i'j'}}{(WH)_{i'j'}} \right] \right) = \sum_{i'j'} - \left(\frac{V_{i'j'}}{(WH)_{i'j'}} \right) (\delta_{i'i} H_{kj'}) = \sum_{j'} - \left(\frac{V_{ij'}}{(WH)_{ij'}} \right) H_{kj'} = - \sum_{j'} \frac{V_{ij'} H_{kj'}}{(WH)_{ij'}}$$

The second term inside the sum in the KL divergence is $-V_{i'j'}$ which is constant with respect to W_{ik} hence its derivative is zero.

The third term is $(WH)_{i'j'}$ itself. Its derivative with respect to W_{ik} was already derived above to be

$$\frac{\partial (WH)_{i'j'}}{\partial W_{ik}} = \delta_{i'i} H_{kj'} \ .$$

Summing this over i' and j', we get the derivative of the second part of the sum:

$$\frac{\partial}{\partial W_{ik}} \sum_{i'j'} (WH)_{i'j'} = \sum_{i'j'} \frac{\partial}{\partial W_{ik}} (WH)_{i'j'} = \sum_{i'j'} \delta_{i'i} H_{kj'} = \sum_{j'} H_{kj'} \ .$$

Combining the Derivatives: The sum of the two parts above yields the full derivative of $D_{KL}(V \parallel WH)$ with respect to W_{ik} as Eq. (14):

$$\frac{\partial}{\partial W_{ik}} D_{KL}(V \parallel WH) = \sum_{j'} H_{kj'} - \sum_{j'} \frac{V_{ij'} H_{kj'}}{(WH)_{ij'}} = \sum_{j'} H_{kj'} \left(1 - \frac{V_{ij'}}{(WH)_{ij'}} \right) \ . \quad (14)$$

A.2 KL Divergence Gradient with Respect to H_{kj}

Similarly, we differentiate the KL divergence with respect to H_{kj}. Firstly, the derivative of $(WH)_{i'j'}$ with respect to H_{kj} is

$$\frac{\partial}{\partial H_{kj}}(WH)_{i'j'} = \frac{\partial}{\partial H_{kj}} \sum_{k'} W_{i'k'} H_{k'j'} = \delta_{j'j} W_{i'k}$$

where $\delta_{j'j} = 1$ if $j' = j$ and zero otherwise. The first term inside the sum of the D_{KL} is $V_{i'j'} \log \frac{V_{i'j'}}{(WH)_{i'j'}}$ and its derivative becomes

$$\frac{\partial}{\partial H_{kj}} \left(V_{i'j'} \log \left(\frac{V_{i'j'}}{(WH)_{i'j'}} \right) \right) = - \frac{V_{i'j'}}{(WH)_{i'j'}} \left(\frac{\partial}{\partial H_{kj}} (WH)_{i'j'} \right) = - \frac{V_{i'j'}}{(WH)_{i'j'}} (\delta_{j'j} W_{i'k})$$

Summing over i and j, we get the derivative of the first part of the D_{KL} sum:

$$\frac{\partial}{\partial H_{kj}} \sum_{i'j'} \left(V_{i'j'} \log \left(\frac{V_{i'j'}}{(WH)_{i'j'}} \right) \right) = \sum_{i'j'} \left(- \frac{V_{i'j'}}{(WH)_{i'j'}} \right) \delta_{j'j} W_{i'k} = - \sum_{i'} \frac{V_{i'j} W_{i'k}}{(WH)_{i'j}} \ .$$

The second term inside the sum in the KL divergence is $-V_{i'j'}$ is constant with respect to H_{kj} hence its derivative is zero.

The derivative of the third part of the D_{KL} sum becomes

$$\frac{\partial}{\partial H_{kj}} \sum_{i'j'} (WH)_{i'j'} = \sum_{i'j'} (\delta_{j'j} W_{i'k}) = \sum_{i'} W_{i'k} \ .$$

Combining the derivatives: Adding the parts together, the derivative of D_{KL} with respect to H_{kj} is

$$\frac{\partial}{\partial H_{kj}} D_{KL}(V \parallel WH) = \sum_{i'} W_{i'k} - \sum_{i'} \frac{V_{i'j} W_{i'k}}{(WH)_{i'j}} = \sum_{i'} W_{i'k} \left(1 - \frac{V_{i'j}}{(WH)_{i'j}}\right) . \quad (15)$$

A.3 Gradient of the constraints $g_1(W)$

The constraints $g_1(W)$ are given for each element of W by

$$g_{1,ik}(W) = W_{ik} - W_{\max} \leq 0 \quad \forall i \in I_0, \forall k \in S_{MH} .$$

The constraints $g_1(W)$ ensure that each weight W_{ik} individually does not exceed the maximum value $W_{\max}$.

Derivative of the constraints $g_1(W)$ with respect to W_{ik}. It is easy to see each individual constraint affects only one element W_{ik} of W, hence the constraint has a nonzero derivative only with respect to that element. Thus the sum of the constraints $g_1(W)$ has a nonzero derivative only if W_{ik} is one of the constrained elements. The derivative is

$$\frac{\partial g_1(W)}{\partial W_{ik}} = \frac{\partial}{\partial W_{ik}} \sum_{i' \in I_0, k' \in S_{MH}} (W_{i'k'} - W_{\max}) = \begin{cases} 1 & \text{if } i \in I_0, k \in S_{MH} \\ 0 & \text{otherwise} \end{cases} \quad (16)$$

where S_{MH} is the subset of mental health topics, and the derivative indicates that the constraint is active when $i \in I_0$ and $k \in S_{MH}$, and inactive otherwise.

A.4 Gradient of the constraints $g_2(H)$

The constraints $g_2(H)$ are given for each element of H by

$$g_{2,k}(H) = \theta_{\min} - \frac{\sum_{j' \in SI} H_{kj'}}{\sum_{j'=1}^{N} H_{kj'}} \leq 0, \ \forall k \in S_{MH}$$

Derivative of the constraints $g_2(H)$ with respect to H_{kj}. It is easy to see that each individual constraint $g_{2,k}(H)$ concerns all elements H_{kj} in a particular row k of H hence the constraint has a nonzero derivative only with respect to elements in that row.

We first define helper notations Num and Denom which are also used in the main paper to state the Final Multiplicative Update Rule for H_{kj}. The detailed definitions were left in the appendix due to space limitations, and we provide them here.

Step 1: We define the Components first

Let Num_k be the numerator:

$$\text{Num}_k = \sum_{j' \in SI} H_{kj'}$$

Let Den_k be the denominator:

$$\text{Den}_k = \sum_{j'=1}^{N} H_{kj'}$$

Thus, we can rewrite the constraint as:

$$g_{2,k}(H) = \theta_{\min} - \frac{\text{Num}_k}{\text{Den}_k}$$

Step 2: The Quotient Rule

To find the derivative of $g_{2,k}(H)$ with respect to H_{kj}, we use the quotient rule:

$$\frac{\partial g_{2,k}(H)}{\partial H_{kj}} = -\frac{\frac{\partial(\text{Num}_k)}{\partial H_{kj}} \cdot \text{Den}_k - \text{Num}_k \cdot \frac{\partial(\text{Den}_k)}{\partial H_{kj}}}{(\text{Den}_k)^2}$$

Partial derivative of Num_k with respect to H_{kj}:

$$\frac{\partial(\text{Num}_k)}{\partial H_{kj}} = \begin{cases} 1 & \text{if } k \in S_{MH}, j \in SI \\ 0 & \text{otherwise} \end{cases}$$

Partial derivative of Den_k with respect to H_{kj}:

$$\frac{\partial(\text{Den}_k)}{\partial H_{kj}} = \begin{cases} 1 & \text{if } k \in S_{MH} \\ 0 & \text{otherwise} \end{cases}$$

Step 3: Substituting Back into the Quotient Rule

If $j \in SI, k \in S_{MH}$:

$$\frac{\partial g_{2,k}(H)}{\partial H_{kj}} = -\frac{1 \cdot \text{Den}_k - \text{Num}_k \cdot 1}{(\text{Den}_k)^2} = -\frac{\text{Den}_k - \text{Num}_k}{(\text{Den}_k)^2}$$

If $j \notin SI, k \in S_{MH}$:

$$\frac{\partial g_{2,k}(H)}{\partial H_{kl}} = -\frac{0 \cdot \text{Den}_k - \text{Num}_k \cdot 1}{(\text{Den}_k)^2} = \frac{\text{Num}_k}{(\text{Den}_k)^2}$$

Thus, the final derivative is:

$$\frac{\partial g_{2,k}(H)}{\partial H_{kj}} = \begin{cases} -\frac{\text{Den}_k - \text{Num}_k}{(\text{Den}_k)^2} & \text{if } k \in S_{MH}, \ j \in SI \\ \frac{\text{Num}_k}{(\text{Den}_k)^2} & \text{if } k \in S_{MH}, \ j \notin SI \\ 0 & \text{otherwise} \end{cases} \quad (17)$$

A.5 Lagrangian Derivative with Respect to W_{ik}

The Lagrangian is:

$$L(W,H,\lambda,\mu) = D_{KL}(V \parallel WH) + \sum_{i',k'} \lambda_{i'k'} g_{1,i'k'}(W) + \sum_{k'} \mu_{k'} g_{2,k'}(H)$$

Taking the derivative with respect to W_{ik}, and noting that the only constraint affected by W_{ik} is $g_{1,ik}(W)$:

$$\frac{\partial L}{\partial W_{ik}} = \frac{\partial D_{KL}(V \parallel WH)}{\partial W_{ik}} + \lambda_{ik} \frac{\partial g_{1,ik}(W)}{\partial W_{ik}}$$

Inserting the results from Eqs. (14) and (16), we get:

$$\frac{\partial L}{\partial W_{ik}} = \sum_{j'} H_{kj'} \left(1 - \frac{V_{ij'}}{(WH)_{ij'}}\right) + \lambda_{ik} \times \begin{cases} 1 & \text{if } i \in I_0, k \in S_{MH} \\ 0 & \text{otherwise} \end{cases} \quad (18)$$

A.6 Lagrangian Derivative with Respect to H_{kj}

Similarly, the derivative with respect to H_{kj} is:

$$\frac{\partial L}{\partial H_{kj}} = \frac{\partial D_{KL}(V \parallel WH)}{\partial H_{kj}} + \mu_k \frac{\partial g_{2,k}(H)}{\partial H_{kj}}$$

Inserting the derivatives from Eqs. (15) and (17), we can write:

$$\frac{\partial L}{\partial H_{kj}} = \sum_{i'} \left(W_{i'k} - \frac{V_{i'j} W_{i'k}}{(WH)_{i'j}}\right) + \mu_k \times \begin{cases} -\frac{\text{Den}_k - \text{Num}_k}{(\text{Den}_k)^2} & \text{if } k \in S_{MH}, j \in SI \\ \frac{\text{Num}_k}{(\text{Den}_k)^2} & \text{if } k \in S_{MH}, j \notin SI \\ 0 & \text{otherwise} \end{cases} \quad (19)$$

A.7 Multiplicative Update Rule for W_{ik}

To optimize the objective under our constraints, we will update H_{kj} and W_{ik} according to the gradients we derived. In this section, we first derive the multiplicative update rule for W_{ik}.

We set the gradient to zero to satisfy the KKT condition:

$$\frac{\partial L}{\partial W_{ik}} = 0 \Rightarrow \sum_{j'} H_{kj'} \left(1 - \frac{V_{ij'}}{(WH)_{ij'}}\right) + \lambda_{ik} \frac{\partial g_{1,ik}(W)}{\partial W_{ik}} = 0 \,.$$

Substituting the derivative of $g_{1,ik}(W)$ we get:

$$\sum_{j'} H_{kj'} \left(1 - \frac{V_{ij'}}{(WH)_{ij'}}\right) = - \lambda_{ik} \cdot \begin{cases} 1 & \text{if } i \in I_0,\ k \in S_{MH} \\ 0 & \text{otherwise} \end{cases}$$

We now rearrange the terms:

$$\sum_{j'} H_{kj'} - \sum_{j'} H_{kj'} \frac{V_{ij'}}{(WH)_{ij'}} = - \lambda_{ik} \cdot \begin{cases} 1 & \text{if } i \in I_0,\ k \in S_{MH} \\ 0 & \text{otherwise} \end{cases}$$

This can be further rearranged as

$$\sum_{j'} H_{kj'} \frac{V_{ij'}}{(WH)_{ij'}} = \sum_{j'} H_{kj'} + \lambda_{ik} \delta_{i \in I_0,\ k \in S_{MH}}$$

where $\delta_{i \in I_0,\ k \in S_{MH}} = 1$ if $i \in I_0$ and $k \in S_{MH}$ and zero otherwise. The above then yields

$$\frac{\sum_{j'} H_{kj'} \frac{V_{ij'}}{(WH)_{ij'}}}{\sum_{j'} H_{kj'} + \lambda_{ik} \delta_{i \in I_0,\ k \in S_{MH}}} = 1 \,.$$

The multiplicative update rule for W_{ik} can thus be written as:

$$W_{ik} \leftarrow W_{ik} \cdot \frac{\sum_{j'} H_{kj'} \frac{V_{ij'}}{(WH)_{ij'}}}{\sum_{j'} H_{kj'} + \lambda_{ik} \delta_{i \in I_0,\ k \in S_{MH}}}$$

In the multiplier on the right-hand side, the numerator contains the terms that had a negative sign in the gradient of the Lagrangian, i.e., the terms that would have a positive sign when moving in the opposite direction of the gradient to minimize the Lagrangian. Similarly, the terms in the denominator are the ones that had a positive sign in the gradient, i.e., the terms that would have a negative sign when moving in the opposite direction of the gradient.

Therefore the numerator $\sum_{j'} H_{kj'} \frac{V_{ij'}}{(WH)_{ij'}}$ corresponds to the positive part of the gradient during optimization, which pulls W_{ik} toward reducing the reconstruction error. The denominator $\sum_{j'} H_{kj'} + \lambda_{ik} \delta_{i \in I_0,\ k \in S_{MH}}$ includes the regularization term that controls the magnitude of W_{ik}.

It is easy to see that the update rule maintains the nonnegativity of W_{ik} as all terms that multiply W_{ik} on the right-hand side of the rule are nonnegative.

A.8 Multiplicative Update Rule for H_{kj}

The update rule for H_{kj} is derived from setting the gradient of the Lagrangian with respect to H_{kj} to zero. Given the expression

$$\frac{\partial L}{\partial H_{kj}} = \sum_{i'} W_{i'k}\left(1 - \frac{V_{i'j}}{(WH)_{i'j}}\right) + \mu_k \frac{\partial g_{2,k}(H)}{\partial H_{kj}} = 0$$

and inserting the gradient of the constraint, this becomes

$$\frac{\partial L}{\partial H_{kj}} = \sum_{i'} W_{i'k}\left(1 - \frac{V_{i'j}}{(WH)_{i'j}}\right) + \left(\mu_k \cdot \begin{cases} -\frac{\text{Den}_k - \text{Num}_k}{(\text{Den}_k)^2} & \text{if } k \in S_{MH},\ j \in SI, \\ \frac{\text{Num}_k}{(\text{Den}_k)^2} & \text{if } k \in S_{MH},\ j \notin SI \\ 0 & \text{otherwise} \end{cases}\right) = 0$$

which can be simplified as

$$\frac{\partial L}{\partial H_{kj}} = \sum_{i'} W_{i'k}\left(1 - \frac{V_{i'j}}{(WH)_{i'j}}\right) + \mu_k \cdot \delta_{k\in S_{MH}}\left(\frac{\text{Num}_k}{(\text{Den}_k)^2} - \delta_{j\in SI}\frac{1}{\text{Den}_k}\right) = 0\,. \quad (20)$$

We next derive two versions of the update rule.

Version 1. Starting from Eq. (20) we can rearrange the terms as

$$\sum_{i'} W_{i'k}\frac{V_{i'j}}{(WH)_{i'j}} = \sum_{i'} W_{i'k} + \mu_k \cdot \delta_{k\in S_{MH}}\left(\frac{\text{Num}_k}{(\text{Den}_k)^2} - \delta_{j\in SI}\frac{1}{\text{Den}_k}\right)$$

which then yields

$$\frac{\sum_{i'} W_{i'k}\frac{V_{i'j}}{(WH)_{i'j}}}{\sum_{i'} W_{i'k} + \mu_k \cdot \delta_{k\in S_{MH}}\left(\frac{\text{Num}_k}{(\text{Den}_k)^2} - \delta_{j\in SI}\frac{1}{\text{Den}_k}\right)}\,.$$

From this, we get the update rule for H_{kj}:

$$H_{kj} \leftarrow H_{kj} \cdot \frac{\sum_{i'} W_{i'k}\frac{V_{i'j}}{(WH)_{i'j}}}{\sum_{i'} W_{i'k} + \mu_k \cdot \delta_{k\in S_{MH}}\left(\frac{\text{Num}_k}{(\text{Den}_k)^2} - \delta_{j\in SI}\frac{1}{\text{Den}_k}\right)} \quad (21)$$

The update rule in Eq. (21) above is appealing due to its symmetrical form to the update rule of W_{ik}. In the multiplier on the right-hand side the numerator is always nonnegative and the denominator is nonnegative if $\sum_{i'} W_{i'k}$ is larger than the term with μ_k. In our experiments, the denominator has consistently remained nonnegative and hence the updates maintain nonnegativity of H. Thus we use this update rule due to its appealing symmetry. However, it is also possible to use an alternate update rule with guaranteed nonnegativity and we derive it below.

Version 2. Starting from Eq. (20) we can rearrange the terms as

$$\sum_{i'} W_{i'k}\frac{V_{i'j}}{(WH)_{i'j}} + \mu_k \cdot \delta_{k\in S_{MH}}\delta_{j\in SI}\frac{1}{\text{Den}_k} = \sum_{i'} W_{i'k} + \mu_k \cdot \delta_{k\in S_{MH}}\frac{\text{Num}_k}{(\text{Den}_k)^2}$$

This then yields

$$\frac{\sum_{i'} W_{i'k}\frac{V_{i'j}}{(WH)_{i'j}} + \mu_k \cdot \delta_{k\in S_{MH}}\delta_{j\in SI}\frac{1}{\text{Den}_k}}{\sum_{i'} W_{i'k} + \mu_k \cdot \delta_{k\in S_{MH}}\frac{\text{Num}_k}{(\text{Den}_k)^2}} = 1\,.$$

The multiplicative update for H_{kj} then becomes

$$H_{kj} \leftarrow H_{kj} \cdot \frac{\sum_{i'} W_{i'k}\frac{V_{i'j}}{(WH)_{i'j}} + \mu_k \cdot \delta_{k\in S_{MH}}\delta_{j\in SI}\frac{1}{\text{Den}_k}}{\sum_{i'} W_{i'k} + \mu_k \cdot \delta_{k\in S_{MH}}\frac{\text{Num}_k}{(\text{Den}_k)^2}} \quad (22)$$

The update rule in Eq. (22) always maintains the nonnegativity of H_{kj} as all terms in the multiplier are nonnegative. Therefore this rule can be used as an alternative to the first update rule (Eq. (21)) if any situation arises where Eq. (21) would not maintain nonnegativity, e.g. if the multiplier μ_k were set to a very large value. In practice, we have found Eq. (21) to work well across all our experiments but we provide this alternate update rule for completeness.

A.9 Details of the Datasets

We discuss the formation of our case study data set and synthetic data set below.

Real data set. Our real data, discussed in Section 5.1 of the main paper, concerns YouTube discussion of viewers for vlogs of Finnish YouTubers and relates to an ongoing study of peer mental health support. YouTube is a massive video sharing platform where many YouTubers have risen to prominence as influencers. The focus of interest in this data domain is the viewers' discussion of mental health aspects, rather than all discussion, hence the domain is a suitable case study for topic modeling of minority topics. To form the data set, we began by selecting 20 pre-identified Finnish YouTubers from our public health experts, who focus primarily on vlogs related to topics such as mental

health issues, targeting a younger audience. Our data is anonymized and is not made public, and reporting of modeling results is only on an aggregate level without personally identifying information. The topics found by our proposed method from this real data are discussed later in this Appendix in Section A.12.

Synthetic data set. We also formed a synthetic data set where we started from the real data set and injected known mental health content as ground truth topical contents, as described below. This data set was used to compare the performance of several methods in modeling the ground truth topical content in Sections 5.2 and 6 of the main paper.

To build our synthetic dataset, we injected topic-related words into randomly selected sentences to simulate mental health discussions. We began by defining a synthetic ground truth as 18 mental health topics to be injected, each with a collection of related Finnish words. This synthetic ground truth was designed to be realistic for the mental health case.

To generate the synthetic data we first chose 500 documents (comment sentences under YouTube vlogs) from our dataset at random. Each document then had a probability of 10 percent to be injected with additional words relating to mental health. When choosing a document, we selected one of the synthetic mental health topics randomly and inserted four of its words in random spots into the document. As a result, some documents have been injected with a relevant mental-health topic and the label of the injected topic is recorded for each document. Note that mental health content was injected into a minority of documents and forms a minority of words in those documents. The majority of the 500 sentences were left unmodified and were given a ground-truth label "-1" indicating they do not contain injected mental health topics. This method allowed us to create a realistic synthetic dataset by combining regular documents with topic-specific terms.

Since we know the ground-truth injected mental health topic (if any) for each of the 500 documents, we can use them for performance evaluation. We use clustering quality measures for this purpose: given a topic modeling result by one of the compared methods, we assign each document to its estimated majority topic as a cluster label. The estimated cluster labels are then evaluated against the ground truth labels described above, by the well-known clustering quality measures normalized mutual information (NMI) and by the Purity score.

Note that for the Purity score in Section 5.2 (which estimates what proportion of all documents have the same ground-truth label as the majority label of their assigned cluster), since we are only interested in mental health related documents we pick the majority class of each cluster from its mental health injected documents, and ignore any clusters having no mental health injected documents. This rewards clusterings where non-mental health documents are placed into separate clusters from mental health injected ones.

A.10 Details of the hyper-parameter tuning, computing systems, and time complexity

For all the compared SOTA baselines, we sought the same number of topics and ensured fair comparison with equal hyperparameter configurations. For NMF, we used default settings: $\alpha_W = 0.0$, $\alpha_H =$ 'same', and $l1\_ratio = 0.0$. We tested both solvers and observed no substantial differences in performance. For LDA, we used the default priors: `doc_topic_prior` $= 1/$`n_components` and `topic_word_prior` $= 1/$`n_components`. We found that varying these priors did not significantly affect model performance. For Top2Vec, we followed the authors' recommended setting, as these were optimized for best performance. SeededLDA was evaluated with three parameter configurations: the default settings suggested by its authors and two additional random samples around the defaults. For the Anchored CoreX method, the key hyperparameter was anchor strength, which we tested across three reasonable values, selecting the best-performing value 4 for comparison. Our method similarly used three configurations, ensuring no method had an unfair advantage. We used publicly available baseline codes; including NMF [5], LDA [6], Top2Vec [7], KeyATM [8], GuidedNMF [9], SeededLDA [10], Corex [11], GuidedLDA [12], ProdLDA [13],

[5] https://scikit-learn.org/dev/modules/generated/sklearn.decomposition.NMF.html

[6] https://scikit-learn.org/1.5/modules/generated/sklearn.decomposition.LatentDirichletAllocation.html

[7] https://github.com/ddangelov/Top2Vec

[8] https://keyatm.github.io/keyATM/

[9] https://github.com/jvendrow/GuidedNMF

[10] https://github.com/bsou/cl2_project/tree/master/SeededLDA

[11] https://github.com/gregversteeg/corex_topic

[12] https://github.com/vi3k6i5/GuidedLDA

[13] https://pyro.ai/examples/prodlda.html

BERTopic [14], and FASTopic [15].

For our experiments, we used two alternative computing systems. We processed 150k data points using a local workstation with 32GB of RAM and an NVIDIA RTX 2000 Ada GPU. We used a High-Performance Computing (HPC) on the web interface to analyze the entire dataset, which has 4.5 million comments.

Regarding time complexity, since the core of our constrained NMF technique is matrix operations, time complexity increases linearly with the number of features (rows of matrix W) and data points (columns of matrix V). In particular, $O(T \cdot MNK)$ is the time complexity per iteration, where M is the number of rows in V, K is the number of latent features, and N is the number of data points. The computational cost grows with dataset size due to the matrix multiplications needed, especially when N increases. Furthermore, the total computation time may be affected by the growth of T, the number of iterations needed for convergence.

A.11 Training Details

Pre-processing the data involved eliminating symbols and stopwords from the NLTK[16] stopwords list and a custom Finnish stopwords list[17]. The Voikko library[18] for spellchecking and stemming was then used to reduce inflected words to their simplest versions which is required particularly for Finnish language. We used Term Frequency-Inverse Document Frequency (TF-IDF) (Salton and Buckley, 1988) transformation with `max_df= 0.9` and `min_df= 0.2` for the real dataset, but no `max_df` or `min_df` values were used for the synthetic dataset. In our model, the parameter θ_{min}= 0.4, and one-third of the total topics were determined as mental health-related. The model demonstrated rapid improvement, dropping significantly in the first 20 iterations, followed by minor changes. The learning rate $\eta = 0.001$ was used to update the Lagrange multipliers λ and μ. We found $W_{\text{max}} = 1 \times 10^{-9}$ to be optimal across trials. Our code is available[19].

[14] https://maartengr.github.io/BERTopic/index.html

[15] https://github.com/BobXWu/FASTopic

[16] https://www.nltk.org/

[17] https://github.com/stopwords-iso/stopwords-fi

[18] https://github.com/voikko/corevoikko

[19] https://github.com/seyedeh-mona-ebrahimi/Constrained-NMF-for-Minority-Topics

A.12 Topics Discovered by the Proposed Method

We display the themes found by our Constrained NMF model applied to both synthetic and real-world datasets respectively in Tables 7 and Tables 3–6. These tables demonstrate how well the methodology can yield insightful and superior mental health-related conversation topics from adolescent peer support discussions. Tables 3–6 in 4 parts, demonstrates the topics discovered in the real dataset, while Table 7 displays those discovered in our synthetic dataset. Each table contains the top ten terms for each topic in both Finnish and English translation.

Topics Discovered in the Real Dataset. Table 3 lists first 15 mental health-related topics, with topics 15-20 in the table and the rest in Tables 4–6 being non-mental health topics. For each topic, the words are listed twice: first as the original Finnish words and then as English translations. The topics are interpretable, when analyzed together with their top documents, as is common in topic modeling.

Topic 1 features odotus (expectation), joulu (Christmas) and mielenterveysongelma (mental health problem) as top words, this is because its top documents feature discussion of expectation, such as expectation of Christmas, but also comments such as "expectation versus reality", as well as low assessments of the future such as seeing mental health problems, intoxicants, and grudges being issues. Similarly, Topic 5 is on relationships in families, and words like perhe (family), isä (father), äiti (mother), and lapsi (child) appear prominently. Topic 6 focuses on mental health treatment and support, including terminology such as hoito (therapy), tuki (support), and ahdistus (anxiety). Topic 4 focuses on trauma and psychosis, using terms such as trauma (trauma) and psyko (psycho). Topic 7 delves into psychosis, trust, and emotional issues which reflects trust issues in relationships and the impact of psychosis on mental well-being: the topic features top documents having discussion of trust and critical attitude towards various sources of information (internet, celebrities, research, politicians, cults), and betrayal of trust by frauds, and even trusting computer security; on the other hand there was also a comment suggesting one would rather bear pain at home rather than going to a hospital, connecting pain to issues of trust. The remaining subjects cover a variety of mental health issues, including self-harm (Topic

0), schizophrenia (Topic 9), and interactions with healthcare providers (Topic 14), demonstrating the model's ability to represent varied topics related to mental health.

Topics Discovered in the Synthetic Dataset. Table 7 highlights the topics observed in the synthetic dataset, which help to validate the efficiency of our model. Again, for each topic, the words are listed twice: first as the original Finnish words and then as English translations. We specified the total number of topics as well as the mental health issues. Table 7 has 7 mental health issues (1-7), and the remaining topics from 8 to 20 are non-mental health. For example, Topic 1 includes words like väkivalta (violence), psykoosi (psychosis), and skitsofrenia (schizophrenia), whereas Topic 2 covers topics about masennus (depression), hoito (treatment), and vertaistuki (peer support). Topic 4 addresses broader mental health issues, including persoonallisuushäiriö (personality disorder) and emotional struggles. Our model's consistent understanding of family, trauma, treatment, and support tendencies across both datasets highlights its effectiveness in discovering high-quality mental health topics. Note that in this synthetic data the mental health content was injected from synthetic original ground truth mental health topics. The topics found by the model correspond well to the original ground truth topics. Note that the quality of the model was also quantitatively shown by its best clustering performance out of all compared models in Section 6 of the main paper.

A.13 Error Bars of KL Divergence from Multiple Model Runs

In our research, the model was run using 10 different random seeds to ensure robustness and reduce biases in performance evaluation. In each iteration, the elements of the matrices V and H were initialized as absolute values of normally-distributed zero-mean random numbers with standard deviation 0.01. For each iteration, we determined the mean KL divergence over initializations, which reflects the average divergence between the learned and target distributions and therefore serves as an important performance parameter. To measure the variability of KL divergence data, we calculated the standard deviation, which represents the degree of dispersion between values obtained from different trials. Figure 3 shows the mean KL divergence (curve) and 10 times the standard deviation (error bars). It shows that the objective function stabilizes at a stationary point after decreasing monotonically, as shown by the mean curve. In line with the theoretical constraints of multiplicative update rules in non-convex optimization, we point out that this does not always imply convergence to a local minimum.

B Ablation Studies

We provide the ablation study referenced in the main text. These results highlight how the model responds to changes in key hyperparameters across various settings. As shown in Figure 5 and the corresponding Table 8 and in Figure 6 and the corresponding Table 9, the model consistently performs well across different settings in terms of KL divergence, NMI, and purity.

C Additional Analysis on Varying Topic Counts and Topic Quality (JSD)

We further examined how changes in the number of total topics and minority topics affect evaluation metrics. As shown in Figures 7 and 8 and in the corresponding Table 10, our model consistently outperforms others across all configurations. Notably, it maintains strong performance even as the number of topics increases, which often challenges other models. We report results from synthetic dataset experiments using 30, 50, and 80 topics. In each setting, we designated 10, 15, and 20 topics, respectively, as minority (mental health-related) topics, with the remaining 20, 35, and 60 topics representing majority (non-mental health) themes. While most baselines show a drop in NMI and Purity, our approach increases, especially with larger sets of minority topics (e.g., K=80, MH=20).

Additionally, to assess topic coherence, we computed the JSD between topic distributions. Lower JSD indicates more distinct and well-formed topics. In Figure 4 (also shown in the main paper as Figure 2 and corresponding Table 2), curves are shown for the mean, standard deviation, minimum, and maximum of JSD over the ground-truth topics. Results show our model achieves high quality (small JS divergence); our model, NMF and KeyATM are the best three models outperforming others and are comparable to each other. Hence our model both attains high topic quality (small JS divergence) and outperforms all models in clustering ability (high purity & NMI).

Lastly, Figure 9 shows a scatterplot of Purity and NMI for different models, corresponding to

Topic	Top 10 Topic Words
Topic 0	hullu, ulkonäkö, lääke, lehti, ihminen, yrittää, itsemurha, mieli, keskustelu, mies crazy, appearance, medicine, magazine, human, try, suicide, mind, conversation, man
Topic 1	odotus, joulu, mielenterveysongelma, ihminen, pettymys, tunnistaa, pelko, autismi, yksinäisyys expectation, christmas, mental health problem, human, disappointment, recognize, fear, autism, loneliness
Topic 2	surullinen, itsemurha, ihminen, kärsimys, kuolla, metsä, jutella, johtaa, saada, keskustelu sad, suicide, human, suffering, die, forest, chat, lead, get to, conversation
Topic 3	tukea, adhd, kriisi, onnellisuus, itsemurha, suomi, julkisuus, tieto, addiktio, anna to support, adhd, crisis, happiness, suicide, finland, publicity, knowledge, addiction, to give
Topic 4	trauma, psyko, ihminen, empatia, keskustelu, väkivalta, pitää, saada, yhteiskunta, tsemppiä trauma, psycho, human, empathy, conversation, violence, have to, get to, society, rooting for you
Topic 5	perhe, isä, äiti, lapsi, vanha, koti, yhteisö, vuotiaana, lestadiolainen, tappaa family, father, mother, child, old, home, community, years old, laestadian, to kill
Topic 6	hoito, tuki, ahdistus, vihapuhe, mieli, henkilö, kipu, jakaa, saada, oire treatment, support, anxiety, hate speech, mind, person, pain, share, get to, symptom
Topic 7	luottaa, kipu, psykoosi, hallusinaatio, ongelma, mielenterveys, paha, tuntea, tosi, keskustelu trust, pain, psychosis, hallucination, problem, mental health, bad, feel, real, conversation
Topic 8	itku, kannabis, potilas, saada, arvo, tukea, ilo, mieli, tasa, israel crying, cannabis, patient, get to, value, to support, joy, mind, even, israel
Topic 9	pelko, neuvo, jutella, masennus, skitsofrenia, lapsi, ihminen, keskustelu, tunne, saada fear, advice, to chat, depression, schizophrenia, child, human, conversation, feeling, get to
Topic 10	huume, poliisi, suomi, jengi, väkivalta, rikollinen, käyttäjä, saada, itsetunto, nykyään drug, police, finland, gang, violence, criminal, user, get to, self confidence, nowadays
Topic 11	viha, keskustelu, mielenterveys, motivaatio, tukea, itsemurha, väkivalta, ihminen, aihe, pelko hate, conversation, mental health, motivation, to support, suicide, violence, human, topic, fear
Topic 12	tunne, ilo, huume, raamattu, rakkaus, ihminen, jumala, stressi, kestää, rauha feeling, joy, drug, bible, love, human, god, stress, bear it, peace
Topic 13	terapia, keskustelu, vakava, ihminen, tuomita, väite, tuntea, pelastua, puhua, paine therapy, conversation, serious, human, sentence, claim, feel, be saved, speak, pressure
Topic 14	lääkäri, ärsyttävä, masennus, paniikki, häiriö, rutto, syy, jaksaa, autismi, saada doctor, annoying, depression, panic, disorder, plague, reason, bear to, autism, get to
Topic 15	pelottava, tehä, mulle, vois, vittu, jeesus, vesi, israel, uskaltaa, uida scary, to do, to me, one could, fuck, jesus, water, israel, dare, to swim
Topic 16	tietää, ottaa, lukea, kysymys, odottaa, tosta, vastata, totuus, pelottaa, mr know, take, read, question, wait, from there, answer, truth, be scared, mr
Topic 17	mun, kaveri, vaa, lähteä, pelata, mieli, seuraavaks, selittää, poi, paa my, friend, just, leave, play, mind, next, explain, away, to put
Topic 18	vaarallinen, mäkipelto, kuulla, määrä, homma, nuori, meinata, minecraft, sota, tsemppiä dangerous, youtuber's name, hear, amount, matter, young, to mean, minecraft, war, rooting for you
Topic 19	kiva, venäjä, sä, salaliittoteoria, metsä, mainita, sota, mainos, jakso, sua nice, russia, you, conspiracy theory, forest, mention, war, ad, episode, of you
Topic 20	kohta, kommentti, katto, hieno, maa, tosi, like, paljo, olo, planeetta soon, comment, ceiling, nice, earth, really, near, much, feeling, planet

Table 3: Topics discovered by our model in the real dataset, part 1, topics 0-20.

Figure 1 in the main paper.

Topic	Top Topic Words
Topic 21	mahtava, asu, seuraava, kaupunki, tubettaja, ehdoton, elokuva, supo, helsinki, mielenkiintonen, pitänyt, huippu, vaatia, polttaa great, outfit, next, city, youtuber, absolute, movie, security and intelligence service, helsinki, interesting, should have, top, demand, burn
Topic 22	kanava, nähdä, tykätä, kiinnostava, alkaa, alku, selvä, raha, toivottava, kuullu, löysä, tilaus, artolauri channel, see, like, interesting, begin, beginning, clear, money, hopeful, heard, loose, order, artolauri (personname)
Topic 23	tarina, historia, sana, taitaa, löytyä, laki, tärkeä, käynyt, loistava, asua, legenda, netti, tapahtunut, pitää, rakentaa, tarkoitus, sattua, mieli story, history, word, be likely, find, law, imporant, happened, brilliant, live, legend, internet, happened, must, build, purpose, hurt, mind
Topic 24	saada, kuu, tossa, huomata, ostaa, lähde, jatko, seurata, liikkua, kuula, selkeä, oikeus, tapahtuma, pyytää, peru, selvitä get, month, there, notice, buy, source, continuation, follow, move, ball, clear, justice, event, request, cancel, survive
Topic 25	katsoa, kiinnostaa, elämä, vähä, jaksaa, pari, äiti, uus, supervoima, ulos, suku, elää, sopimus, vapaa see, interest, life, little, to bear, couple, mother, new, superpower, out, family, live, contract, free
Topic 26	fakta, onni, ihmeellinen, outo, kissa, kuulostaa, koira, väärä, rakastaa, ase, musiikki, joulu, mite, asukas, onnettomuus, syntyä fact, happiness, wonderful, strange, cat, sound like, dog, wrong, love, weapon, music, christmas, how, tenant, accident, be born
Topic 27	osata, miettiä, kuolema, ruotsi, sitte, naapuri, mysteeri, siisti, kylmä, hahmo, sanottu, kestää, luokka, laulu know how to, think, death, sweden, then, neighbor, mystery, clean, cold, character, said, to bear, class, song
Topic 28	aihe, mielenkiintoinen, tehty, petteri, kans, lopettaa, väärin, tehny, mieli, mikkonen, halloween, katsonut, tuleva, mahtava, saanu topic, interesting, done, peter, also, stop, wrong, done, mind, mikkonen (surname), halloween, watched, future, great, gotten
Topic 29	ihminen, top, lapsi, eläin, ymmärtää, jäädä, vaikuttaa, tyhmä, syödä, tappaa, luonto, koe, käyttö, tuntea, lukenut, tarvita human, top, child, animal, understand, stay, influence, stupid, eat, kill, nature, experiment, use, feel, read, need
Topic 30	työ, koulu, oikeesti, henkilö, musta, vika, jenna, käsi, talo, vitsi, tieten, numero, kauhu, milkyllänaa, kallio, puhe, haluinen work, school, really, person, black, flaw, jenna, hand, house, joke, of course, number, horror, milkyllänaa, rock, speech, wanting

Table 4: Topics discovered by our model in the real dataset, part 2, topics 21-30.

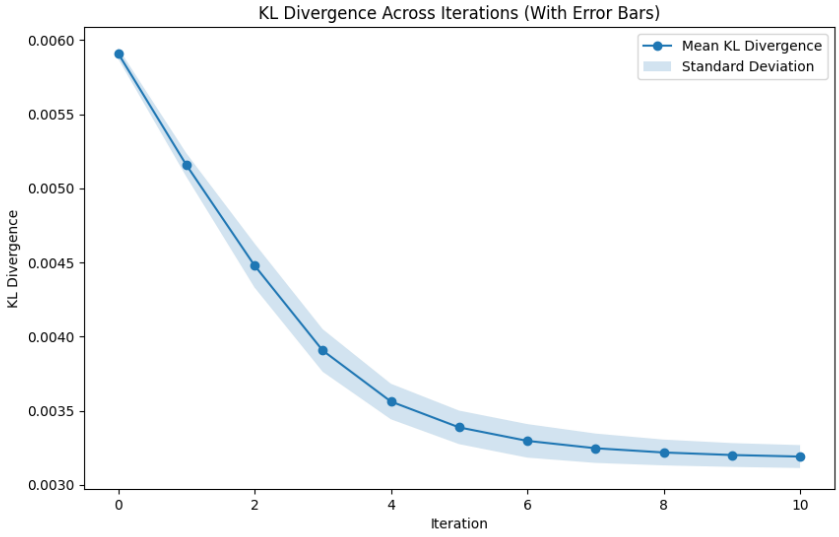

Figure 3: KL Divergence Across Iterations with Error Bars.

Topic	Top Topic Words
Topic 31	maailma, suomi, peli, tieto, löytää, rikas, alue, gta, keksi, kuoltu, ykkönen, poistaa, korona, rakennus, auttaa, ikä, rokote, synty world, finland, game, knowledge, find, rich, area, gta, cookie, died, number one, remove, corona, building, help, age, vaccine, birth
Topic 32	muistaa, paikka, ajatella, haluta, pystyä, käärme, jättää, onks, päättää, nähdä, todistaa, valo, kuunnella, raamattu, ero, mahdoton remember, place, think, want, be able to, snake, leave, is it, decide, see, prove, light, listen, bible, difference, impossible
Topic 33	paha, mielenkiintoinen, vastaus, avaruus, katsoma, eurooppa, suositella, paikko, maa, samanlainen, vanha, suomi, vankila, välttää, etsiä, sydän, todiste evil, interesting, answer, space, looking, europe, recommend, replacement, country, similar, old, finland, prison, avoid, seek, heart, proof
Topic 34	tuoda, vanha, yö, jännä, silmä, kuuluu, uni, nään, isä, herätä, normaali, tie, ihmetellä, aamu, kommentoida, veli bring, old, night, exciting, eye, belong, dream, i see, father, wake, normal, road, wonder, morning, comment, brother
Topic 35	juttu, presidentti, iso, veikata, suomi, japani, setti, sopia, turvallinen, mäki, isojalka, miljardi, murha story, president, big, wager, finland, japan, set, fit, safe, hill, bigfoot, billion, murder
Topic 36	ääni, nauraa, nähny, mieli, puuttua, tapaus, ihmeellinen, lahko, kuollut, mahdollinen, tulo, kiinni, nähä, rikollinen, alue, sotilas, esine sound, laugh, seen, mind, lack, event, wondrous, cult, dead, possible, product, caught, see, criminal, area, soldier, object
Topic 37	nimi, kyl, loppu, mielenkiintoisa, pitkä, katto, näkyä, tuttu, poika, loppua, tutkia, matka, nähnyt, leffa, huone, lemppari name, yes, end, interesting, long, roof, see, familiar, son, end, investigate, trip, seen, movie, room, favorite
Topic 38	luulla, laittaa, sisältö, tullu, kattonu, päin, upea, pistää, valtio, dinosaurus, teko, raja, asiallinen, tylsä think, to put, contents, has come, seen, toward, gorgeous, to stick, government, dinosaur, action, limit, reasonable, boring
Topic 39	uskoa, jumala, puhua, liittyä, ikinä, ihminen, uskonto, syy, paska, tulevaisuus, kirjoittaa, kieli, millon, ihmiskunta, tuhota, kanta, kokea, raamattu believe, god, speak, join, ever, human, religion, reason, shit, future, write, language, when, humanity, destroy, stance, experience, bible
Topic 40	jengi, pää, viikko, iso, voima, unohtaa, virhe, onnee, sais, jalka, porukka, sahein, fiksu, kuulemma, tienny, keksintö, päässyt, hinta gang, head, week, big, power, forget, mistake, congratulations, could get, foot, crowd, neat, smart, as i heard, known, invention, gotten, price

Table 5: Topics discovered by our model in the real dataset, part 3, topics 31-40.

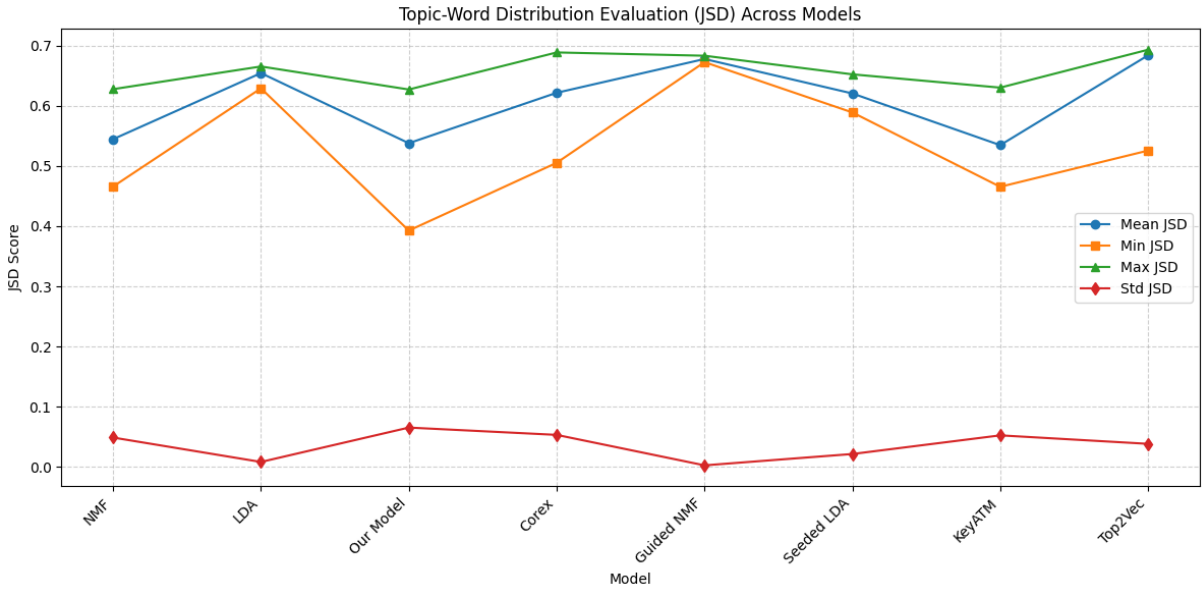

Figure 4: Topic Quality using JSD Score

Topic	Top Topic Words
Topic 41	suomalainen, mies, kertoa, kova, nainen, hauska, kirja, teoria, kuultu, helvetti, tapa, tarkka, kappale, mieli - liittyvä, linna, pitää, finntop finnish, man, tell, hard, woman, funny, book, theory, heard, hell, manner, precise, piece, mind, related, castle, keep, finntop
Topic 42	kattoo, kerto, kuolla, salainen, itsenäisyyspäivä, käyny, saatana, onneks, ohjelma, rauha, toimia, puoli, kiehtova, vaimo, saada, jätkä see, tell, die, secret, independence day, visited, satan, luckily, program, peace, act, half, fascinatinng, wife, get, dude
Topic 43	pakko, oppia, ai, elää, aivo, toivoa, ajatus, turha, herra, muuttua, myöhä, itä, kauhea, ansainnut, nostaa, laskea, kasvi need, learn, oh, live, brain, hope, thought, futile, sir, change, late, east, horrible, deserved, raise, lower, plant
Topic 44	idea, luoja, ongelma, luu, poliisi, pallo, amerikka, värinen, salaliitto, laita, huomio, yllättää, tekemä idea, creator, problem, bone, police, ball, america, colored, conspiracy, put, attention, surprise, made by
Topic 45	käyttää, laiva, riittää, uutinen, arvo, lisätä, mielenkiintosia, laulaa, ruoka, viisas, vaunu, upota, pohja use, ship, suffice, news, value, add, interesting, sing, food, wise, wagon, sink, bottom
Topic 46	mainittu, sarja, parka, arto, tavata, lauri, tori, käsitellä, ajaa, armeija, anna, oikeen, ihme, alotukset, aiheinen, positiivinen, merkitys mentioned, series, poor, arthur, meet, larry, market, deal with, drive, army, give, real, miracle, beginnings, topical, positive, meaning
Topic 47	varma, kuva, näyttää, kokemus, mieli, jakaa, väittää, ilta, vidi, katsella, äijä, uskomaton, haamu, räjäyttää - kone sure, picture, show, experiance, mind, share, claim, evening, vid, watch, dude, incredible, ghost, explore, machine
Topic 48	mennä, jatkaa, tilata, auto, malli, väli, kallis, ihme, super, meri, maksa, väri, salaisuus, niinkuin go, continue, order, car, model, distance, expensive, miracle, super, ocean, pay, color, secret, as in
Topic 49	biisi, tilaaja, päästä, tehnyt, käydä, helppo, tiennyt, sairas, pelkkä, pitäs, vahva, tutkimus, paikata, ansaita, tajuta, kunta song, customer, make it, done, visit, easy, known, sick, mere, should, strong, research, compensate, earn, realize, municipality

Table 6: Topics discovered by our model in the real dataset, part 4, topics 41-49.

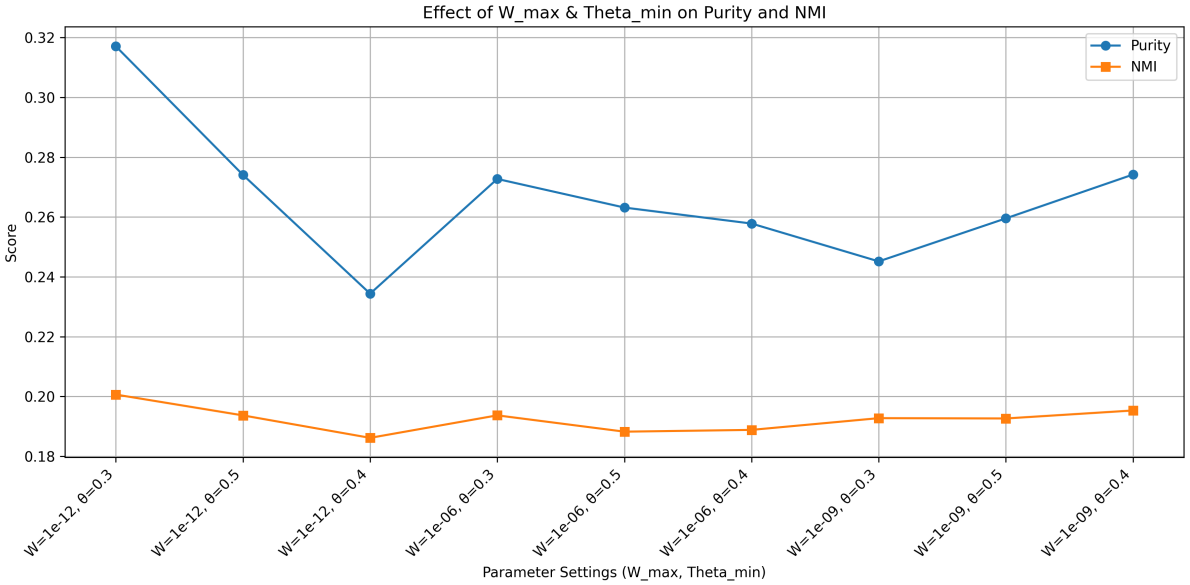

Figure 5: Effect of $W_{\max}$ and $\theta_{\min}$ on NMI and purity scores.

Topic	Top 10 Topic Words
Topic 1	väkivalta, kokemus, psykoosi, itsemurha, oire, syy, ihminen, suru, pöllö, skitsofrenia violence, experience, psychosis, suicide, symptom, cause, human, sorrow, cuckoo, schizophrenia
Topic 2	kannabis, masennus, rentoutuminen, jumala, tuomita, hoito, hoitaja, kuunnella, jeesus, vertaistuki cannabis, depression, relaxation, god, sentence, treatment, nurse, listen, jesus, peer support
Topic 3	keskustelu, viha, tukea, ongelma, tunne, lapsi, ihminen, mulle, skitsofrenia, adhd conversation, hate, to support, problem, feeling, child, human, to me, schizophrenia, adhd
Topic 4	mielenterveys, sairaala, sota, ihminen, tukea, hoito, tunne, kokea, persoonallisuushäiriö, uskoa mental health, hospital, war, human, to support, treatment, feeling, to experience, personality disorder, to believe
Topic 5	uskoa, kokea, perhe, erottaa, sun, parisuhde, huolestua, alotuksista, pitäiskö, tukea to believe, to experience, family, to separate, your, relationship, to get worried, of beginnings, should one, to support
Topic 6	neuvonta, perhe, kuunnella, psykoterapia, kannabis, hoito, skitsofreenikko, lauri, arto, hitto counseling, family, listen, psychotherapy, cannabis, treatment, schizophrenic, larry, artie, damn
Topic 7	vihapuhe, ihminen, tuki, sairaala, päihde, kannabis, vertaistuki, totuus, suhun, voldemortilta hate speech, human, support, hospital, intoxicant, cannabis, peer support, truth, to you, from voldemort
Topic 8	mielenkiintoinen, katsottava, oikeesti, aihe, vitsi, osata, pitäis, päästä, sul, toiminta interesting, watchable, for real, topic, joke, be able, should, get to, with you, action
Topic 9	käydä, ny, hienostaa, venäjä, koulu, lauantai, selittää, ratkaisu, kpl, ku visit, now, make fancy, russia, school, saturday, explain, solution, amount, when
Topic 10	kirja, paa, kiva, suomi, sääli, sisältö, seura, nii, puolustusvoima, eläin book, put, nice, finland, pity, content, company, yeah, defence forces, animal
Topic 11	hieno, para, oot, tunnettu, maailma, onni, sit, kanava, suomi, maa great, poor, you are, known, world, happiness, then, channel, finland, country
Topic 12	malli, jatkaa, kuuluu, kieli, ranska, raha, nato, pitää, pelkkä, ruth model, continue, belong, language, france, money, nato, should, mere, ruth
Topic 13	mun, aihe, kova, tuoda, fakta, ois, mulla, kanava, muistaa, tilata mine, topic, hard, bring, fact, would be, i have, channel, remember, order
Topic 14	jes, homma, mun, pelottava, seuraava, sarja, fiilis, tieto, kuu, kanada yes, matter, mine, scary, next, series, feeling, knowledge, month, canada
Topic 15	sun, suomi, ruumis, mukava, chicago, oho, mies, kiehtovii, uudestaan, ladata your, finland, body, nice, chicago, wow, man, fascinating, again, load
Topic 16	tarina, lista, star, pyrkiä, poi, tieto, suosikki, ruusu, pystyä, suomi story, list, star, attempt, poi, knowledge, favorite, rose, be able to, finland
Topic 17	nato, uskoa, aloitus, viikinki, helvetti, tosi, suomi, liittymä, mua, nokia nato, believe, beginning, viking, hell, real, finland, cell phone plan, me, nokia
Topic 18	oot, sun, jee, mahtava, seuraavaks, lempi, videoo, tykätä, mona, katto you're, your, yeah, great, next, favorite, video, like, mona, ceiling
Topic 19	tapahtua, mieli, kiinnostava, selvä, illuusio, ihminen, päästä, mi, area, seleeni happen, mind, interesting, clear, illusion, human, to get to, mi, area, selenium
Topic 20	ihminen, epäillä, tarvita, oo, ukraina, elää, nähny, hari, kalifornia, kandassa human, doubt, need, oo, ukraine, live, seen, hari, california, in canada

Table 7: Topics discovered by our model in the synthetic dataset.

$(W_{\max}, \theta_{\min})$	Purity	NMI
1×10^{-12}, 0.3	0.317	0.201
1×10^{-12}, 0.5	0.274	0.194
1×10^{-12}, 0.4	0.234	0.186
1×10^{-06}, 0.3	0.273	0.194
1×10^{-06}, 0.5	0.263	0.188
1×10^{-06}, 0.4	0.258	0.189
1×10^{-09}, 0.3	0.245	0.193
1×10^{-09}, 0.5	0.260	0.193
1×10^{-09}, 0.4	0.274	0.195

Table 8: Purity and NMI for parameter settings.

$W_{\max}$	θ=0.3	θ=0.4	θ=0.5
1×10^{-12}	0.001797	0.001796	0.001799
1×10^{-06}	0.001790	0.001794	0.001783
1×10^{-09}	0.001793	0.001799	0.001802

Table 9: Final KL divergence loss for different $W_{\max}$ and $\theta_{\min}$.

	K=30, MH=10		K=50, MH=15		K=80, MH=20	
Model	Purity	NMI	Purity	NMI	Purity	NMI
NMF	0.147	0.189	0.174	0.202	0.181	0.178
LDA	0.072	0.127	0.099	0.134	0.143	0.141
COREX	0.054	0.189	0.064	0.187	0.083	0.173
Ours	0.172	0.207	0.190	0.250	0.292	0.190
Top2Vec	0.016	0.069	0.016	0.037	0.022	0.072
SeededLDA	0.165	0.174	0.182	0.171	0.238	0.158
GuidedLDA	0.081	0.171	0.170	0.174	0.190	0.132
GuidedNMF	0.061	0.119	0.102	0.135	0.142	0.143
KeyATM	0.140	0.185	0.184	0.191	0.195	0.164

Table 10: Purity and NMI scores across different numbers of topics (K) and mental health topics (MH).

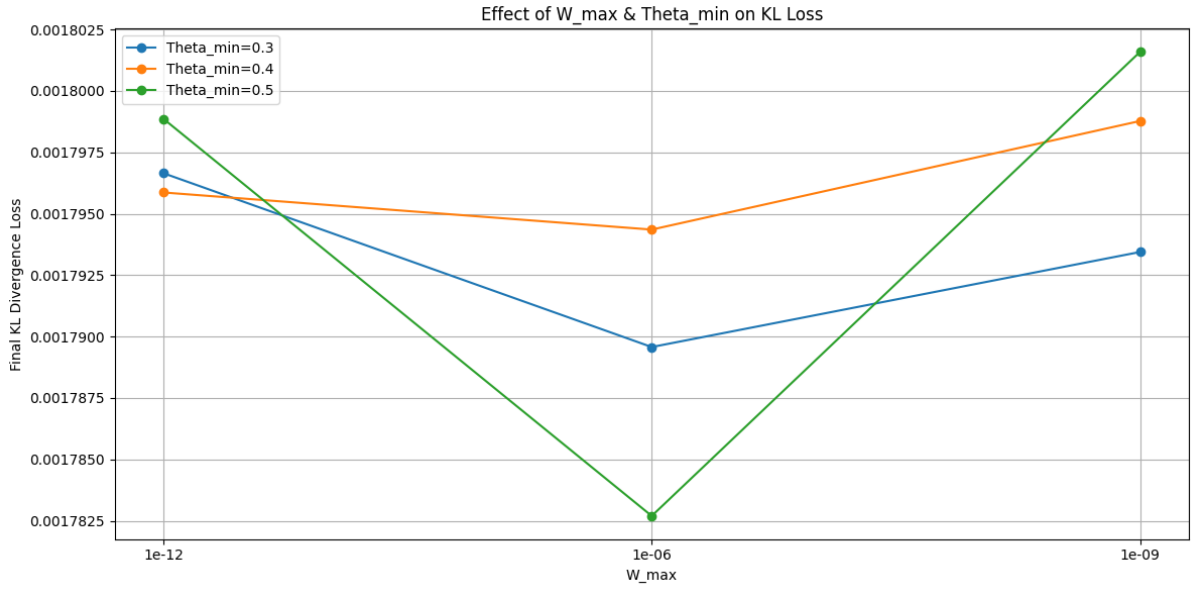

Figure 6: Effect of $W_{\max}$ and $\theta_{\min}$ on KL divergence.

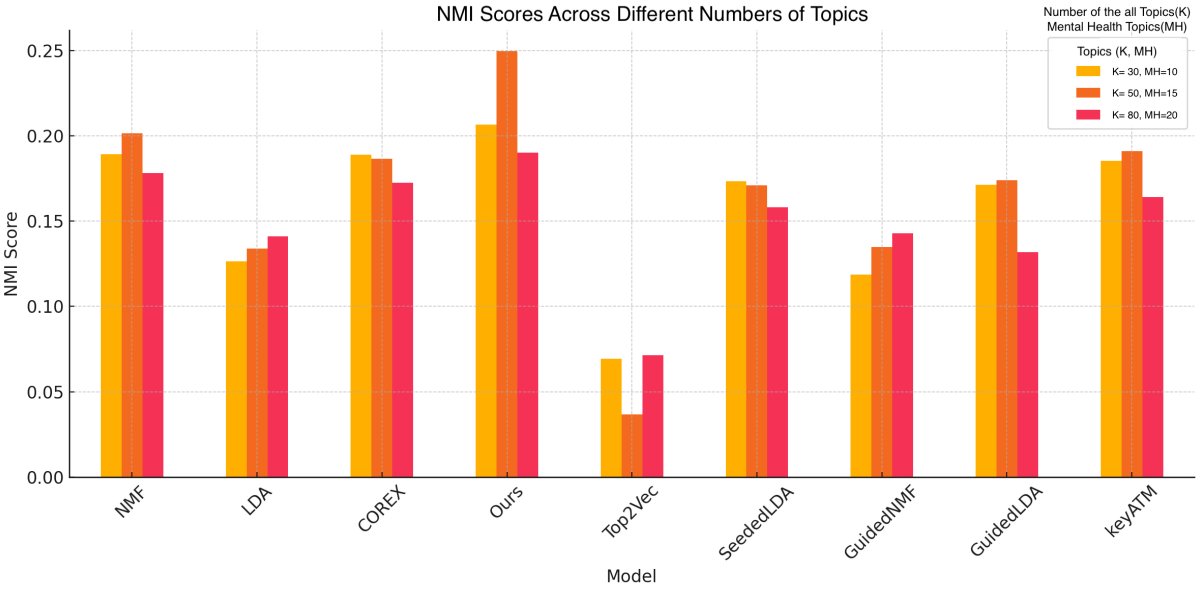

Figure 7: NMI scores for topic counts 30, 50, and 80.

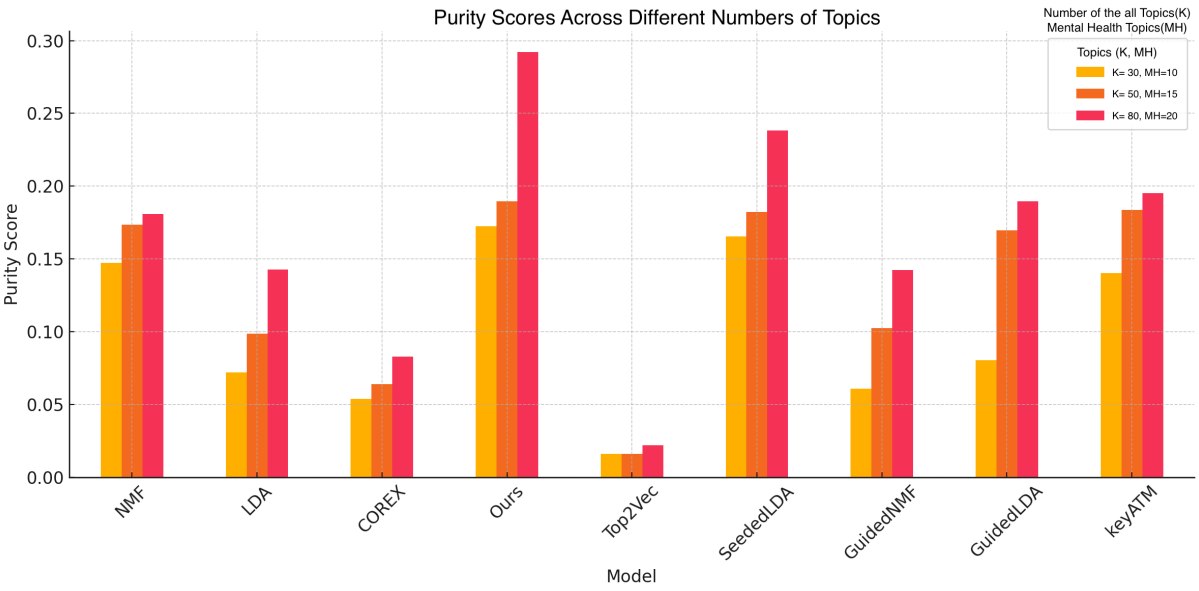

Figure 8: Purity scores for topic counts 30, 50, and 80.

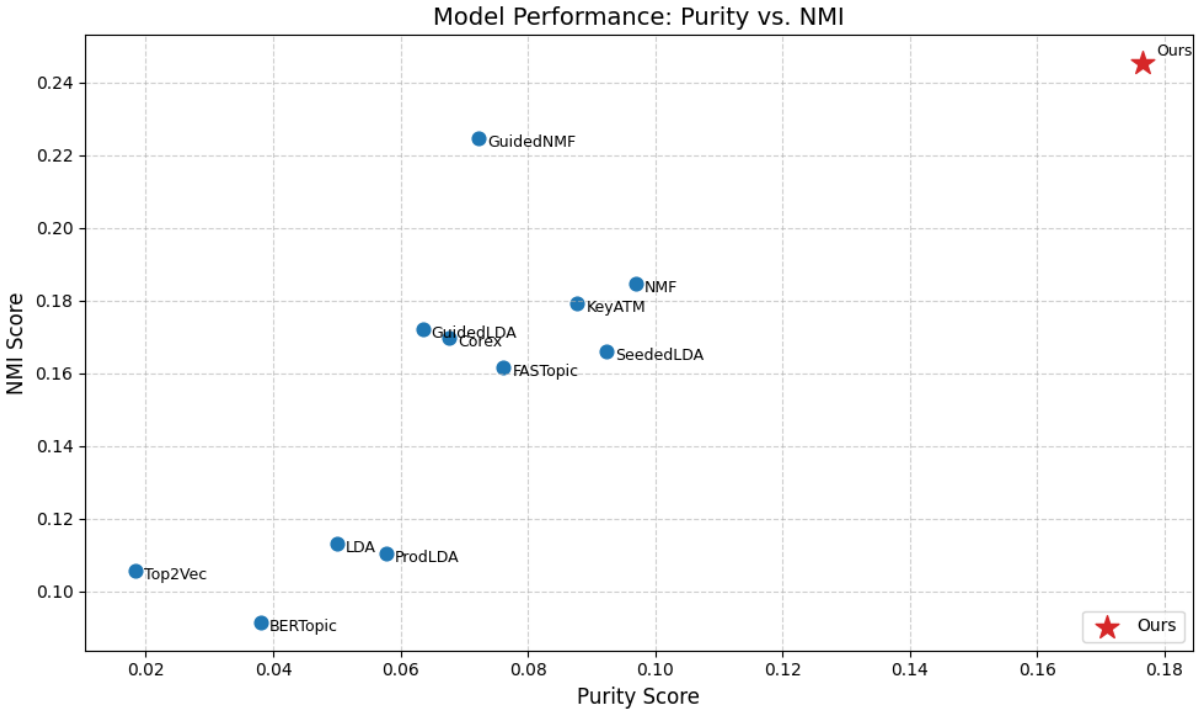

Figure 9: Scatter plot of Purity vs. NMI scores for different models.