Voices Between Lines: Interpretable Labeling of Mental Health Minority Topics with Seed Guidance and LLMs

Seyedeh Fatemeh Ebrahimi[0009−0001−4909−6117] and Jaakko Peltonen[0000−0003−3485−8585]

Faculty of Information Technology and Communication Sciences,
Tampere University, Finland
{seyedeh.ebrahimi, jaakko.peltonen}@tuni.fi
https://www.tuni.fi

Abstract. We present a pipeline that automatically assigns concise, human-readable labels to under-represented mental-health topics discovered by a seed-guided nonnegative matrix factorization model. Given each topic's word distribution, we rank documents via Jensen–Shannon divergence, extract anchored n-gram candidates, score them on informativeness, phraseness, and seed overlap, and finally ask a large language model (LLM) to choose and justify a label. The approach amplifies minority voices while remaining fully automatic and language-agnostic. We demonstrate its efficacy on Finnish-language online discussion of YouTuber vlogs containing mental health minority themes. In two human evaluations of label quality, our model attains high expert scores and outperforms a baseline approach.

Keywords: Automatic topic labeling · Jensen-Shannon divergence · LLM · minority topic · interpretability · mental health discourse.

1 Introduction

Topic modeling is a powerful tool for uncovering latent themes within extensive text corpora [23,27,7,21,26,19]. However, traditional unsupervised topic models often prioritize dominant trends in data sets, making it challenging to detect minority topics like mental health discourse in large-scale social media data [5]. This can hinder the understanding of critical yet minority discussions, such as discussions related to mental health which due to their sensitive nature, are likely to appear as an undercurrent rather than a prominent theme in online discourse. Our prior work addressed this by proposing a constrained nonnegative matrix factorization (NMF) model designed to extract minority topics via seed word supervision-based constraints on topic prevalence and content [5].

Even experienced practitioners acknowledge that topic modeling is far from *"push-button"*. Key choices—such as the number of topics, stopwords, lemmatizers, and other hyperparameters are often unclear to non-experts. Thankfully,

several metrics and tools for model choices are available for practitioners. Yet, even after successfully modeling topics, interpreting them can remain a laborious task, especially when the topics are nuanced minority topics. In this work, we focus on improving topic interpretability.

Although topic models are widely used, their output often remains difficult to interpret semantically, especially for users who are non-experts in the topic modeling methods or in the data domain [14,10,20]. Topic models model the observed text data through latent spaces parameterized by topics. They represent each topic mathematically as a distribution of probabilities or set of weights over words, and experts typically read through the top word lists and possibly example documents from topics, in order to assign human interpretation to each topic. This task remains laborious and challenging, since top words alone rarely provide sufficient context to understand what a topic truly represents or how it differs from others [20]. The topics produced by these models frequently misalign with human interpretations, resulting in vague, generic, and incoherent word groupings [4,3,10]. Moreover, while topic models often generate distributions that are statistically coherent, models with better statistical performance may produce less semantically coherent topics, making human understanding difficult [4]. Indeed, assessing topic interpretability is challenging: recent studies show that coherence and diversity metrics do not always reflect how interpretable a topic is to humans, as users may misinterpret topic words, miss key concepts, or impose unintended meaning [20,4,10].

This mismatch between probabilistic word distributions and human expectations highlights the need for interpretation methods that yield textual, coherent, and context-aware topic labels [20]. Traditional topic interpretation methods rely on top words as primitive labels [15,16,2] or require manual annotation [1], both of which are limited in clarity, consistency, and scalability [10,4]. LLMs like ChatGPT have recently been explored as tools for topic interpretation. Preliminary findings show that while their outputs can occasionally surpass domain expert labels in clarity, they also exhibit inconsistencies and need careful prompting [20]. In this work, we present a post-hoc labeling framework that improves topic interpretability, particularly for minority themes. Unlike prior systems that rely purely on surface-level keyword heuristics or manual effort, we create an end-to-end pipeline combining statistical document-topic alignment with seed-guided candidate label ranking and LLM-based explanation generation.

To identify documents that best represent each topic, we use Jensen-Shannon divergence (JSD) [13], a symmetric measure of similarity between two probability distributions. We use JSD to effectively capture the alignment between word distributions of topics and of candidate representative documents. This is especially important for minority topics, where signals may be diffuse or subtle. Moreover, as minority topics may represent only part of the content in the documents, we use seed word guidance to extract relevant phrase candidates and score them along several metrics. We then use a carefully prompted LLM to propose labels for the topic based on top extracted phrases. We validate our framework using a real-world case study on Finnish-language YouTube comments, focusing on

mental health discourse. The dataset, consisting of over 5.5 million comments across 19 influencers [17], provides a rich but highly imbalanced testbed where mental health discussion is a minority as is realistic. Results show our labeling pipeline adds semantic clarity, linguistic alignment, and practical interpretability to the discovered mental health topics.

Contributions. We introduce an end-to-end labeling pipeline that turns topic word lists produced by a seed-guided NMF model into short, readable titles that non-technical public health professionals can act on. In detail:

- **(i)** We introduce a three-stage automatic labeling pipeline for seed-guided NMF topics, combining distributional document matching, n-gram phrase scoring, and LLM-based selection and refinement.
- **(ii)** We propose a novel use of Jensen-Shannon divergence to select on-theme documents even for sparse minority topics.
- **(iii)** We design a candidate scoring function that integrates informativeness, phraseness, and seed-relevance, improving relevance and label quality.
- **(iv)** We demonstrate that lightweight LLMs can generate coherent, justified labels in Finnish.
- **(v)** Our outputs serve both interpretability and downstream use cases, paving the way for large-scale monitoring of mental-health discourse without costly manual annotation.

2 Background

Minority-Aware Topic Models. Traditional probabilistic models such as CTM and Correlated LDA [1] capture nuanced dependencies but remain fully unsupervised and thus dominated by majority themes. Seed-guided extensions inject weak supervision to surface under-represented phenomena. Examples include SeededLDA [8], GuidedLDA[1], and Guided NMF variants [24,11]. Joint formulations of clustering and topic modeling have also been explored, for example through a recent NMF-based approach that integrates both tasks simultaneously and shows strong performance on minority themes and clusters [6]. Our work builds on the constrained NMF introduced by Ebrahimi & Peltonen [5] which uses an overall domain-relevant seed word list to set mild constraints, requiring minority topics overall to have a minimum prevalence in documents with seed words, and non-minority topics not to have strong prevalence of seed words. The model does not require known seeds per topic; it finds the variety of minority and majority topics through model fitting. The above seed-guided approaches and our approach in this paper can all be seen as semi-supervised/guided matrix factorization, where weak domain signals steer parts of the factorization while preserving NMF's transparency and scalability.

[1] `https://guidedlda.readthedocs.io/`

2.1 Automatic Topic Labeling

To make topic models usable in real-world applications, human-interpretable labels must be assigned to each topic. Early approaches simply show the top-N words [2], a cognitively demanding practice prone to subjectivity. However, Chang et al. [4] showed by large-scale user studies that this assumption does not always hold: topics that score high in likelihood may be harder for humans to interpret. Their findings highlight the need for evaluation metrics and labeling approaches grounded in human judgment rather than purely statistical fit.

Early probabilistic and Wikipedia-based labellers. Mei et al. [14] first cast topic naming as an optimisation problem that minimises the Kullback-Leibler (KL) divergence between a candidate label and the topic–word distribution while maximising mutual information with the corpus. Lau et al. [9] extended this idea by mining Wikipedia article titles and their sub-phrases, then ranking these candidates with supervised and unsupervised association measures. Although both methods yield concise labels, they presuppose that the most descriptive phrase is present verbatim in an external knowledge base—an assumption that rarely holds for minority or non-English topics such as Finnish mental-health discourse.

LLM assisted interpretation. Very recently, Rijcken *et al.* [20] explored ChatGPT as a zero-shot topic explainer, asking the model to summarise topics that were manually assigned by a domain expert. Their findings suggest LLMs can be helpful but also highlight the need for reliable prompts and grounding in evidence. Meanwhile, a growing body of research uses LLMs not only for interpretability but also for evaluating topic model quality. Stammbach et al. [22] and Yang et al. [28] demonstrate that LLM-based evaluations correlate more strongly with human judgments than traditional metrics like coherence or perplexity, while also guiding decisions such as the number of topics. Lieb et al. [12] further show LLM-generated data augmentation can produce more interpretable, targeted topic models for domain-specific applications. Our framework builds on this emerging literature by (i) *automatically* generating a compact, seed-aware candidate set before invoking the LLM, and (ii) grounding the LLM prompt in distributionally matched document snippets as context, which reduces hallucination and keeps the focus on minority signals, particularly in low-resource and sensitive domains like mental health discussions.

Positioning. Compared with the prior studies, our pipeline unifies distributional document retrieval, seed-guided scoring, and evidence-aware LLM prompting, addressing the challenges of minority-topic sensitivity and label interpretability. Our use case further demonstrates the advantage of the pipeline in a low-resource language setting (Finnish) where manual annotation is costly or impractical.

3 Problem Formulation

Let $\mathbf{V} \in \mathbb{R}^{n \times d}$ denote a document-term matrix, where n is the number of documents and d the size of the vocabulary. A topic model trained on $\mathbf{V}$ outputs two

matrices: a topic-word matrix $\mathbf{H} \in \mathbb{R}^{k \times d}$, where each row $\mathbf{H}_t$ represents the word distribution for topic $t \in \{1, \ldots, k\}$, and a document-topic matrix $\mathbf{W} \in \mathbb{R}^{n \times k}$, where each row reflects a document's mixture of topics. The goal of topic labeling is to give each topic t a concise, descriptive label $\ell_t \in \mathcal{L}$, where $\mathcal{L}$ is the space of possible textual labels. The labels should summarize the semantic content of each topic, capture domain-relevant and potentially low-frequency expressions, and be interpretable and useful for non-expert users in applications such as content moderation or public health monitoring. To generate these labels, we consider not only the topic-word distribution $\mathbf{H}_t$, but also top documents associated with each topic. Formally, we define a labeling function:

$$f : (\mathbf{H}_t, \mathcal{D}_t) \mapsto \ell_t,$$

where $\mathcal{D}_t \subseteq \mathcal{D}$ is a set of documents most representative of topic t. These are selected based on distributional similarity between $\mathbf{H}_t$ and document representations derived from $\mathbf{V}$, as detailed in the next section. This formulation reflects that document context, not just top words, is crucial for producing accurate and interpretable labels, especially for underrepresented or noisy themes.

4 Method

We propose a multi-stage pipeline to automatically generate concise and meaningful topic labels from constrained NMF [5] outputs on Finnish social media. Our method combines distributional document ranking, anchored phrase extraction, seed-guided scoring, and grounding LLM-based final stage interpretability, addressing challenges posed by low-frequency, domain-specific topics. The algorithm underlying our pipeline is presented in Algorithm 1. Each stage of the post-hoc label generation pipeline is detailed in the following subsections.

4.1 Document Ranking via Jensen-Shannon Divergence (JSD)

For each topic t, represented by a topic-word distribution $\mathbf{H}_t \in \mathbb{R}^d$ with word probabilities that sum to 1, we aim to find representative documents whose word distributions align closely with the topic. Let $\mathbf{v}_i$ be the TF-IDF vector of document d_i, normalized to sum to one. We compute the JSD between $\mathbf{v}_i$ and $\mathbf{H}_t$:

$$\mathrm{JSD}(\mathbf{v}_i, \mathbf{H}_t) = \frac{1}{2} D_{\mathrm{KL}}\left(\mathbf{v}_i \,\|\, \mathbf{m}\right) + \frac{1}{2} D_{\mathrm{KL}}\left(\mathbf{H}_t \,\|\, \mathbf{m}\right), \tag{1}$$

where $\mathbf{m} = \frac{1}{2}(\mathbf{v}_i + \mathbf{H}_t)$ and D_{KL} denotes KL-divergence. JSD is symmetric and bounded, making it robust for measuring distributional similarity between sparse vectors. We rank documents by increasing JSD and select the top m documents per topic to serve as evidence for label generation.

4.2 Anchored Candidate Label Extraction

While using top documents to help label a topic seems attractive, we must avoid a potential pitfall: for minority topics even the top documents may not be fully about the minority topic, thus we should not use all parts of the top documents for labeling. To solve this, we extract label candidates by a topic and seed word informed approach as follows.

For each topic t, we vectorize the set D_t of the m top-ranked documents using TF–IDF and extract all n-grams of lengths 1 to 3 as phrases. We retain only those candidate phrases that contain both (i) at least one topic keyword from H_t and (ii) at least one seed word from the domain-specific lexicon $\mathcal{S}$. This dual anchoring mechanism filters out noisy or off-topic phrases and encourages candidates that reflect minority-specific lexical signals. The retained phrases are then ranked by their cumulative TF–IDF mass across the documents D_t, and the top k phrases define the initial candidate label set $\mathcal{C}_t = c_1, c_2, \ldots, c_k$.

4.3 Seed-Guided Label Scoring and Ranking

Each candidate $c \in \mathcal{C}_t$ is scored based on three complementary criteria:

1. **Informativeness** $f_{\text{inf}}(c)$: the total relative frequency in D_t of the candidate's tokens which are in the topic's top word list H_t.
2. **Phraseness** $f_{\text{phr}}(c)$: a length-based reward that favors multi-word expressions, computed as $\min\{|c|, 3\}$.
3. **Seed Overlap** $f_{\text{seed}}(c)$: the number of overlapping words between the candidate and the domain-specific seed lexicon $\mathcal{S}$.

The overall score of the candidate is then computed as

$$s(c) = f_{\text{inf}}(c) + \lambda_1 \cdot f_{\text{phr}}(c) + \lambda_2 \cdot f_{\text{seed}}(c) \tag{2}$$

where λ_1 and λ_2 are tunable hyperparameters controlling the weights of phraseness and seed signal, respectively. This scoring function promotes labels that are lexically salient and domain-aware with respect to both the topics and the seed guidance. The top k candidates by score are retained as the refined candidate set $\mathcal{L}'_t$ used for final label selection.

4.4 LLM-Based Label Explainability and Justification

To select the most coherent and descriptive label $l_t \in \mathcal{L}'_t$, we query a causal language model using a structured Finnish prompt. Figure 1 shows an example of the prompt and its English translation. The prompt includes:

- the top keywords for topic t from H_t (labelled as **Avainsanat** in Finnish, **Keywords** in the English translation),
- a few representative document snippets from D_t (**Keskustelukatkelmat** in Finnish, **Conversation Fragments** in English),

– the top candidate labels $\mathcal{L}'_t$ (**Ehdotetut Etiketit** in Finnish, **Suggested Labels** in English).

The prompt is formulated in natural Finnish, instructing the model to propose a short and concise label summarizing the discussion theme, with an explicit instruction to provide both a label and a justification, but not to simply repeat the keywords or candidates.

Algorithm 1: Post-hoc Label Generation using Document Ranking and LLM Explanation

Input: Document corpus $\mathcal{D} = \{d_i\}_{i=1}^{n}$, stopword list $\mathcal{Z}$, seed lexicon $\mathcal{S}$,
topic model (W, H) with K topics, LLM M,
hyperparameters: m (top docs), k (top labels), $\boldsymbol{\lambda} = (\lambda_1, \lambda_2)$
Output: Label set $\mathcal{L} = \{l_1, \ldots, l_K\}$

1. Text Preprocessing and Representation
Lemmatize d_i, remove stopwords in $\mathcal{Z}$.
Construct TF–IDF matrix $V \in \mathbb{R}_{\geq 0}^{n \times |\mathcal{V}|}$ over $\mathcal{D}$.
2. Document Ranking via Jensen–Shannon Divergence
for $t \leftarrow 1$ **to** K **do**
  Let $h_t \in \mathbb{R}^{|\mathcal{V}|}$ be the topic-word distribution for topic t.
  Compute $\mathrm{JS}(v_i, h_t)$ for all documents v_i (rows of V):
    $\mathrm{JS}(v_i, h_t) = \frac{1}{2} D_{\mathrm{KL}}(v_i \parallel m) + \frac{1}{2} D_{\mathrm{KL}}(h_t \parallel m)$,
    where $m = \frac{1}{2}(v_i + h_t)$.
  Select top m documents with lowest divergence:
  $\mathcal{D}_t \leftarrow$ Top-m docs ranked by $\mathrm{JS}(v_i, h_t)$.
3. Anchored Candidate Extraction
From $\mathcal{D}_t$, extract all 1–3-grams from TF–IDF ranked n-grams.
Retain phrases containing **both** topic keywords ($\mathrm{supp}(h_t)$) and at least one seed word in $\mathcal{S}$.
Let $\mathcal{C}_t$ be the resulting candidate set.
4. Seed-Guided Scoring of Candidates
for *each* $c \in \mathcal{C}_t$ **do**
  $I(c) \leftarrow$ normalized frequency of tokens from c among tokens in $\mathcal{D}_t$ appearing in $\mathrm{supp}(h_t)$;
  $P(c) \leftarrow \min(\mathrm{length}(c), 3)$
  $O(c) \leftarrow |\mathrm{tokens}(c) \cap \mathcal{S}|$;
  $s(c) \leftarrow I(c) + \lambda_1 \cdot P(c) + \lambda_2 \cdot O(c)$
Let $\widehat{\mathcal{C}}_t$ be the top k candidates by score $s(c)$.
5. LLM-Based Label Justification and Selection
Construct prompt $\pi_t = \big(h_t, \text{top-3 snippets from } \mathcal{D}_t, \widehat{\mathcal{C}}_t\big)$
Query LLM by Prompting: $l_t \leftarrow M(\pi_t)$ and extract response: `ETIKETTI & PERUSTELU`.
return $\mathcal{L} = \{l_1, l_2, \ldots, l_K\}$

The model's reply is parsed to extract the final label after the token `ETIKETTI:`. This stage enhances the fluency, abstraction, and contextual interpretability of labels—particularly valuable for morphologically complex languages like Finnish language.

Model selection and inference setup. For LLM-assisted selection and interpretability, we initially tested the TurkuNLP GPT-3 Finnish foundation models (3B–13B)[2], but found them unsuitable for direct instruction-following without fine-tuning. To reduce complexity and ensure consistency, we instead use the Ahma-3B-Instruct and Ahma-7B-Instruct models from Finnish-NLP[3], both based on the LLaMA v1 architecture and fine-tuned to follow Finnish-language

[2] `https://huggingface.co/TurkuNLP/gpt3-finnish-3B`
[3] `https://huggingface.co/Finnish-NLP/Ahma-3B-Instruct`

Alla on keskustelun avainsanoja, katkelmia ja koneellisesti ehdotettuja etikettejä. **AVAINSANAT:** hullu, itsemurha, psykiatri **KESKUSTELUKATKELMAT:** – "... psykiatri sanoi että ..." – "... tunsin oloni ahdistuneeksi ..." **EHDOTETUT ETIKETIT:** – mielenterveyden kriisi – itsetuhoiset ajatukset – ahdistuksen hallinta Nämä ehdotukset voivat auttaa sinua ymmärtämään keskustelun teemaa, mutta sinun ei tarvitse valita niistä suoraan. Perustele valintasi viittaamalla sekä keskustelukatkelmiin että ehdotuksiin, jos niistä on hyötyä. Anna lyhyt ja ytimekäs etiketti, joka tiivistää keskustelun pääaiheen. Vastaa täsmälleen seuraavassa muodossa: `ETIKETTI:` "kirjoita tähän lyhyt etiketti" `PERUSTELU:` "perustele etiketti keskustelun sisällön perusteella" Älä toista yllä olevia avainsanoja tai ehdokaslistaa sellaisenaan.	*Below are keywords and conversation fragments and machine-suggested labels.* **KEYWORDS:** crazy, suicide, psychiatrist **CONVERSATION FRAGMENTS:** – "... the psychiatrist said that ..." – "... I felt anxious ..." **SUGGESTED LABELS:** – mental health crisis – suicidal thoughts – anxiety management These suggestions can help you understand the theme of the conversation, but you do not have to choose any of them directly. Justify your choice by referring both to conversation fragments and suggestions, if they are useful. Give a short and succinct label that summarizes the main topic of the conversation. Answer exactly in the following form: `LABEL:` "write here a short label" `JUSTIFICATION:` "justify the label based on the content of the conversation" Do not repeat the above keywords or candidate list as is.

Fig. 1. Our LLM prompt example for extracting the final label.

chat instructions. These models were accessed via Hugging Face[4] using the official system prompts and template-based input formatting. Inference was run locally on an Apple M4 Pro (24GB RAM), using decoding parameters `temperature=0.9`, `top_p=0.85`, and a max length of 300 tokens. No additional fine-tuning was applied.

5 Experiments

5.1 Data

We evaluate on a real-world dataset of Finnish-language YouTube comments, gathered from the audience comments sections of vlogs of 19 Finnish YouTubers. The dataset consists of approximately 5.5 million comments. Following public health experts [17], we identified these 19 YouTubers from channels addressing

[4] `https://huggingface.co/`

youth mental-health themes, such as depression, anxiety, trauma, crisis helpline, and social stigma. Texts were lemmatised with `libvoikko` and stop-words removed[5]. The corpus was encoded as TF–IDF for both topic model training and downstream scoring. Topic modeling was conducted using a constrained non-negative matrix factorization model, guided by by a domain seed lexicon for mental-health themes[5]. This ensured the learned topics were semantically focused on underrepresented themes. The model was run with the topic counts yielding highest normalized mutual information in experiments of [5]: $K = 50$ topics of which 15 are minority topics; constraint strength hyperparameters in the model were the same as in [5]. For each topic t, we computed JSD given by Eq.1 between its word distribution and the normalized TF-IDF vector; the lowest-JSD documents are used as evidence for candidate label generation.

Baseline. We compare to a keywords-only LLM baseline, where the model is prompted with the top-N words of each topic and no documents or seed lexicon. The model's first returned label is used as the prediction; prompt template and decoding parameters are provided in our repository.

5.2 Evaluation

In what follows, we first analyze the model behavior in terms of the contribution of different components to the overall candidate scoring, and the model's confidence in choosing the best candidates based on the overall scores. Finally, we carry out a human evaluation of the topic labels by a public health expert and show our model attains high scores by the expert. We also do a second human evaluation by a non-expert, showing our model outperforms the baseline.

Component Contribution Analysis. To better understand the relative impact of each scoring component, we visualized the distribution of informativeness, phraseness, and seed overlap across all candidate labels in Figure 2. We observe that Phraseness tends to dominate the overall score, as most candidates fall near the upper bound of its weight range (capped by the maximum n-gram length). In contrast, Informativeness shows a wide spread, indicating variability in how well candidate labels capture topic-specific vocabulary. SeedOverlap values cluster tightly, with most labels sharing a similar number of seed word matches. This distribution confirms that while our scoring function balances all three components, Phraseness acts as a strong default prior; however, Informativeness and SeedOverlap introduce meaningful variation across label quality.

Model Confidence Estimation. Given a topic t, let $\mathcal{L}_t = \{\ell_1, \ell_2, \ldots, \ell_n\}$ be the set of n candidate labels. For each label ℓ_i, we compute the composite score $s(c)$ given by Eq. 2 for each candidate $c \in \mathcal{C}_t$. Following prior work that employs softmax-based confidence estimation for label selection [18,25], we define

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

Table 1. Finnish topic words discovered by Constrained NMF (CNMF) model in the real data set, and their English translations.

Topic #	Top Words
0	**Finnish:** itsemurha, julkisuus, jutella, autismi, saada, väkivalta, väsynyt, lestadiolainen, pitää, hullu **English:** suicide, publicity, chat, autism, receive, violence, tired, Laestadian, like, crazy
1	**Finnish:** sairaus, keskustelu, diagnoosi, häiriö, oire, paniikkikohtaus, persoonallisuushäiriö, persoonallisuus, ahdistuneisuus, tuki **English:** illness, discussion, diagnosis, disorder, symptom, panic attack, personality disorder, personality, anxiety, support
2	**Finnish:** hullu, selittämätön, kiusata, mielenterveys, ongelma, vihapuhe, adhd, neuvo, itsetunto, perhe **English:** crazy, unexplained, to bully, mental health, problem, hate speech, ADHD, advice, self-esteem, family
3	**Finnish:** surullinen, jumala, usko, jeesus, raamattu, henkiä, seurakunta, kuolla, elämä, yhteisö **English:** sad, god, faith, jesus, bible, spirits, congregation, die, life, community
4	**Finnish:** perhe, usko, suru, lapsi, käydä, keskustelu, mieli, saada, adhd, pahoinvointi **English:** family, faith, grief, child, visit, discussion, mind, receive, ADHD, nausea
5	**Finnish:** terapia, kriisi, nukahtaa, paniikki, saada, ahdistus, ajatus, mieli, tuki, kriisipuhelin **English:** therapy, crisis, fall asleep, panic, receive, anxiety, thought, mind, support, crisis hotline
6	**Finnish:** tuki, ukraina, venäjä, saada, perhe, liika, suomia, sota, psykoterapia, kertoa **English:** support, Ukraine, Russia, receive, family, too much, criticize, war, psychotherapy, tell
7	**Finnish:** huume, psyko, hoito, alkoholi, väkivalta, päihde, käyttäjä, käyttö, laillinen, suomia **English:** drug, psycho, treatment, alcohol, violence, substance, user, use, legal, criticize
8	**Finnish:** lääkäri, motivaatio, piirre, paniikkihäiriö, mielenterveysongelma, masennus, lääke, saada, skitsofreenikko, perhe **English:** doctor, motivation, trait, panic disorder, mental health issue, depression, medication, receive, schizophrenic, family
9	**Finnish:** kannabis, pelko, ahdistus, lääke, aiheuttaa, paha, kipu, väkivalta, luottaa, alkoholi **English:** cannabis, fear, anxiety, medication, cause, bad, pain, violence, trust, alcohol
10	**Finnish:** ärsyttävä, odotus, persoonallisuushäiriö, saada, laillistaa, jumala, billion, evoluutio, väkivalta, ajatella **English:** annoying, waiting, personality disorder, receive, legalize, god, billion, evolution, violence, think
11	**Finnish:** kuunnella, podcast, ääni, äänikirja, unettomuus, rauhallinen, nukkua, ilta, puhe, keskittyä **English:** listen, podcast, sound, audiobook, insomnia, calm, sleep, evening, speech, focus
12	**Finnish:** tauti, ilo, pelastua, elää, stressi, elämä, itsemurha, saada, usko, alkaa **English:** disease, joy, be saved, live, stress, life, suicide, receive, faith, begin
13	**Finnish:** tunne, trauma, hallusinaatio, usko, saada, ahdistava, päänsärky, herkkä, tuttu, ajatus **English:** feeling, trauma, hallucination, faith, receive, distressing, headache, sensitive, familiar, thought
14	**Finnish:** viha, masennus, jumala, skitsofrenia, raamattu, tuki, saada, lääkitys, puhua, auttaa **English:** anger, depression, god, schizophrenia, bible, support, receive, medication, talk, help

a posterior confidence score using a softmax-based normalization:

$$P(c \mid t) = \frac{\exp(s(c))}{\sum_{c' \in \mathcal{C}_t} \exp(s(c'))}. \tag{3}$$

This yields a probability distribution over label candidates, allowing us to interpret $P(c \mid t)$ as the model's relative confidence in selecting c as the best label for

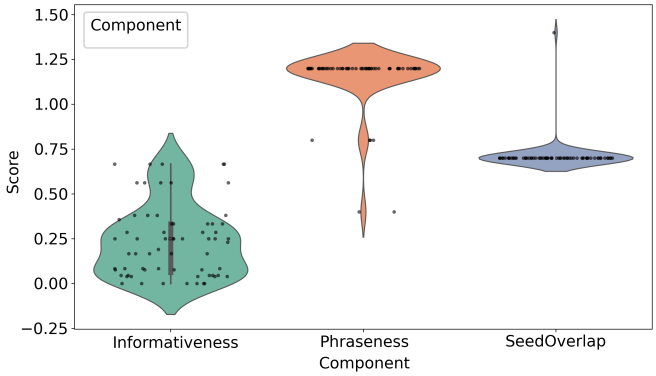

Fig. 2. Distribution of label score components (Informativeness, Phraseness, and SeedOverlap) across all candidate labels. Each violin shows the value distribution of that component, with internal scatter indicating individual label candidates. The plot illustrates which components vary the most and which dominate the scoring dynamics.

topic t. The confidence $P(c \mid t)$, Eq. 3, reflects both internal scoring dynamics and how strongly the top candidate stands out from its peers. Figure 3 shows selected examples spanning both high and moderate confidence levels. High-confidence topics (e.g., *kriisipuhelin*) indicate a clear preference by the scoring function, whereas lower-confidence examples (e.g., *kohdistunut viha tarkoitti*) reveal more ambiguity or competition among similar candidates. These confidence scores serve two key purposes: (i) they enhance interpretability by quantifying label certainty, and (ii) they help identify representative or borderline topics for expert annotation and error analysis.

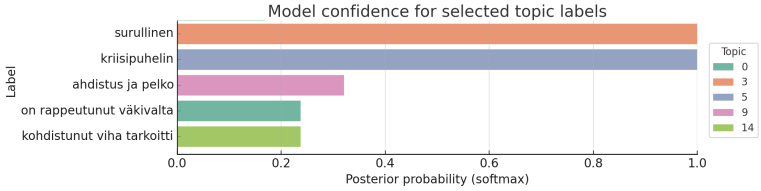

Fig. 3. Model confidence scores (posterior probabilities given by Eq.3) for selected topic labels. Higher values indicate stronger certainty in the selected label among candidates.

Expert Human Judgement of Label Quality. To assess the interpretability and domain relevance of the generated labels, we conducted a human evaluation of our method's output by a public health expert using domain-specific criteria. The experiment was done in Finnish using original-language text, allowing evaluation of the effectiveness of the labeling pipeline in morphologically rich, low-resource language context. In the evaluation, the public health expert was presented with the top keywords for a topic, a sample of representative comments (documents) associated with the topic, and the final topic label from our method. The expert scored each label using a 5-point Likert scale (1=strong dis-

agree, 4=weak disagree, 3=neutral, 4=weak agree, 5=strong agree) in response to four interpretability-focused statements:

(i) The label represents a mental health concept.
(ii) The label reflects the content in the topic word list.
(iii) The label captures the content in the representative sample documents.
(iv) The label is useful for analyzing mental health in this dataset.

For each topic, the expert gave one Likert scale score representing the expert's agreement with the set of four statements overall, and a marking which ones of statements (i)-(iv) were considered agreeable for the topic. This scoring strategy reflects the preferred working mode of the expert, we report results accordingly.

In Table 2 we present the detailed labeling results of our labeling pipeline across the 15 minority topics learned by the topic model. As shown in Table 1, each topic is represented by its top-10 keywords in Finnish and English translation, and Table 2 shows the model's final labels for the topics in Finnish and English as well as the public health expert's overall Likert scale rating of each topic and which of the statements (i)-(iv) the expert agreed with.

Table 2. Final labels from our model and scores by a public health expert. Rating: overall Likert scale agreement rating. Agreed Statements: which of the four statements the expert agreed with.

Topic	Final Label (FI / EN)	Rating	Agreed Statements
0	**itsemurha** *suicide*	1	i, ii
1	**persoonallisuushäiriöiden vaikutukset ja diagnostiikka** *Effects and diagnosis of personality disorders*	5	i, ii, iii, iv
2	**hulluus, poikkeavuus ja luova mieli** *madness, deviance, and creative mind*	4	i
3	**surullinen tunne helluntailaisuudessa** *sad feeling within Pentecostalism*	5	ii, iii
4	**perhe: muutokset ja perherakenteet kristillisessä uskossa** *family: changes and family structures in Christian faith*	2	ii, iii partly
5	**kriisipuhelin ja sen rooli ahdistuksen hallinnassa** *crisis hotline and its role in anxiety management*	5	i, ii, iii, iv
6	**saada Ukrainan tukea** *get support for Ukraine*	4	ii, iii
7	**huumeen vaikutus käyttäytymiseen** *effects of drugs on behavior*	5	i, ii, iii, iv
8	**paniikkihäiriö ja sen hallinta** *panic disorder and its management*	5	i, ii, iii, iv
9	**Kannabis ja ahdistus/pelko** *cannabis and anxiety/fear*	5	i, ii, iii, iv
10	**Ärsytyksen ilmaisu** *expressing irritation*	2	-
11	**Podcast-kuuntelu rauhoittumiseen** *listening to podcasts for relaxation*	5	i, ii, iii, iv
12	**Kokemuspohjainen helluntailainen teologia** *experience-based Pentecostal theology*	3	ii partly, iii
13	**Tunnekokemusten moninaisuus** *diversity of emotional experiences*	5	i, ii, iii, iv
14	**Vihan ja sen seurausten tutkiminen** *exploring anger and its consequences*	5	i, ii, iii, iv

Our model received a very good evaluation from the expert. Most topics received high scores: of the 15 topics, nine topics (topics 1, 3, 5, 7, 8, 9, 11, 13, and 14) received the best score of 5 (strong agreement), and two topics (topics 2 and 6)

Table 3. Comparison of our pipeline (**Ours**) to the baseline (**Base**). For each topic, the label by each model is rated by Likert-scale agreement for statements (i)-(iv) from the main text. In brief: (i) the label reflects mental health, (ii) matches topic word list, (iii) matches sample documents, (iv) is useful for mental health analysis. For each model we also report the average of ratings (i)-(iv). For each comparison better score is bolded.

	Statement i		**Statement ii**		**Statement iii**		**Statement iv**		**Average**	
Topic	**Base**	**Ours**	**Base**	**Ours**	**Base**	**Ours**	**Base**	**Ours**	**Base**	**Ours**
0	4.0	**5.0**	**5.0**	4.0	2.0	**4.0**	3.0	3.0	3.50	**4.0**
1	4.0	**5.0**	4.0	**5.0**	3.0	**4.0**	2.0	**5.0**	3.25	**4.75**
2	2.0	**3.0**	2.0	**3.0**	2.0	**4.0**	2.0	**4.0**	2.0	**3.50**
3	2.0	**3.0**	4.0	4.0	4.0	4.0	3.0	**5.0**	3.25	**4.0**
4	3.0	3.0	2.0	**4.0**	1.0	**4.0**	1.0	**4.0**	1.75	**3.75**
5	4.0	**5.0**	4.0	**5.0**	1.0	**4.0**	1.0	**4.0**	2.50	**4.50**
6	1.0	**2.0**	2.0	**3.0**	2.0	**4.0**	1.0	**5.0**	1.50	**3.50**
7	3.0	**4.0**	4.0	4.0	2.0	**4.0**	1.0	**5.0**	2.50	**4.25**
8	4.0	**5.0**	3.0	**4.0**	2.0	**4.0**	2.0	**5.0**	2.75	**4.50**
9	1.0	**5.0**	1.0	**4.0**	2.0	**4.0**	1.0	**3.0**	1.25	**4.0**
10	1.0	**3.0**	1.0	**3.0**	2.0	**4.0**	2.0	**3.0**	1.50	**3.25**
11	4.0	**5.0**	4.0	4.0	2.0	**4.0**	1.0	**4.0**	2.75	**4.25**
12	**3.0**	2.0	**2.0**	1.0	3.0	**4.0**	2.0	**3.0**	2.50	**2.50**
13	3.0	**5.0**	3.0	**4.0**	2.0	**5.0**	1.0	**5.0**	2.25	**4.75**
14	5.0	5.0	3.0	3.0	2.0	**5.0**	2.0	**5.0**	3.0	**4.50**

received a score of 4 (weak agreement). One topic (topic 12) received a score of 3 (neutral). On the other hand, two topics (topics 4 and 10) received a score of 2 (weak disagreement), and one (topic 0) received a score of 1 (strong disagreement). The result shows strong performance for our method in a challenging scenario of minority topics in a low-resource setting, while also highlighting the benefit of a final human curation for domain relevance. Notably, such human curation is still expected to be far faster than naming topics by manual inspection. Overall, this shows clear benefit of our labeling pipeline for analyzing minority topics.

Human Comparison to the Baseline Model. We also performed a second rating task comparing our pipeline to the baseline method by a non-expert human annotator. We compare to the baseline label generation strategy: an instruction-tuned Finnish LLM that generates topic labels from topic keywords alone, not using documents or seed words; model details, prompt and resulting labels are in GitHub (see *Reproducibility*). The same topics from Table 1 were used, and the labels of our pipeline in Table 2 were compared to the labels by the baseline. For each topic, the non-expert human annotator rated the labels of both methods by Likert scale agreement with respect to each of the four statements (i)-(iv).

The results in Table 3 show we strongly outperform the baseline. The 15 topics and 4 statements yield 60 topic-statement pairs: our model achieves better Likert scores for 49 pairs, equal for 8, and worse for only 3. Thus, our pipeline improves results for almost all cases & criteria. The two human evaluations (expert&non-expert) show our pipeline clearly helps minority topic labeling.

Reproducibility. All our code, prompts, and seed words list are released at `https://github.com/seyedeh-mona-ebrahimi/Voices-Between-Lines`.

6 Discussion and Future Work

This study shows post-hoc topic labeling, guided by seed relevance and distribution aware matching, can greatly improve the interpretability of minority topics in noisy real-world corpora. Notably, focusing the selection of documents and candidate phrases on those matching the minority topic content and seed guidance is crucial to avoid losing the minority theme among unrelated majority content. Our results also highlight the importance of blending statistical methods with LLM understanding. The candidate label scoring balances term frequency and domain relevance, while the LLM acts as a semantic filter, enhancing readability and contextual appropriateness in Finnish language. Our automatic labeling framework is not just an interpretability tool: it creates a bridge to supervised downstream tasks. With topic-wise labels grounded in both data and expert concepts, one may use them to train classifiers for comment-level mental health detection, construct interpretable dashboards for social support analysis, or track evolution of themes across creators and time. Future research could explore the framework across languages and domains.

7 Generalizable Insights about Responsible Application of Machine Learning in Healthcare

Interpretability for Trustworthy Machine Learning in Health. Topic models can discover valuable themes in mental-health representations, but without clear labels they remain black boxes. Our pipeline generates transparent, evidence-grounded topic labels that enhance usability accessible to domain experts, public health officials, and other non-technical stakeholders. This is especially important in healthcare contexts, where interpretability is a prerequisite for trust and accountability to expect practitioners to act on findings. In effect, the framework operationalises equity: it gives minority voices a proportional presence in algorithmic summaries without requiring costly manual annotation.

Quantified Uncertainty for Safe Decision-Making. Responsible deployment needs models to acknowledge what they do not know. We use a softmax-derived confidence score to every candidate label, yielding a posterior probability. High confidence invites swift adoption; low confidence can act as an automatic "second-opinion" trigger, flagging topics for expert review before policy decisions or clinical interventions are made. This lightweight uncertainty estimate is more informative than a single point prediction, balancing rigour and usability.

Transparency, Explainability, and Language Adaptation. Each stage of the pipeline–document retrieval, candidate extraction, seed-aware scoring, and LLM label generation–yields intermediate outputs that can be inspected, critiqued, and replicated. This interpretability and explainability trail is vital to align

with emerging standards of accountability in health care AI. Also important is linguistic adaptability: by using Finnish in-language prompting, the system performs reliably in a morphologically rich, low-resource setting without resorting to translation. These design choices show a broader principle: transparent, modular pipelines for trustworthy ML tools in health.

Ethical and Privacy Considerations, and Limitations. LLM inference was run locally with Finnish instruction-tuned models; no raw YouTube comments were sent to external services. We release only de-identified snippets, aggregated labels, and code. Limitations include dependence on seed-lexicon coverage.

Acknowledgments. This study was supported by by Academy of Finland decision 348523. We thank T. Nygård and E. Valkonen for their help in annotation.

Disclosure of Interests. The authors declare no competing interests relevant to the content of this article.

References

1. Blei, D.M., Lafferty, J.D.: Correlated topic models. In: Advances in Neural Information Processing Systems. vol. 18, p. 147–154. MIT Press (2005)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research **3**(Jan), 993–1022 (2003)
3. Boyd-Graber, J., Mimno, D., Newman, D.: Care and Feeding of Topic Models: Problems, Diagnostics, and Improvements. CRC Handbooks of Modern Statistical Methods, CRC Press, Boca Raton, Florida (2014)
4. Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., Blei, D.: Reading tea leaves: How humans interpret topic models. Advances in neural information processing systems **22** (2009)
5. Ebrahimi, S.F., Peltonen, J.: Constrained non-negative matrix factorization for guided topic modeling of minority topics. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP) (2025), in press
6. Ebrahimi, S.F., Peltonen, J.: Nonnegative matrix factorization for joint clustering and topic modeling with minority topics. In: Proceedings of the 32nd International Conference on Neural Information Processing (ICONIP) (2025), in press
7. Egger, R., Yu, J.: A topic modeling comparison between LDA, NMF, Top2Vec, and BERTopic to demystify Twitter posts. Frontiers in Sociology **7** (2022)
8. Jagarlamudi, J., Daumé III, H., Udupa, R.: Incorporating lexical priors into topic models. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics. pp. 204–213 (2012)
9. Lau, J.H., Grieser, K., Newman, D., Baldwin, T.: Automatic labelling of topic models. In: Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 1536–1545. Association for Computational Linguistics, Portland, Oregon, USA (Jun 2011)
10. Lee, T.Y., Smith, A., Seppi, K., Elmqvist, N., Boyd-Graber, J., Findlater, L.: The human touch: How non-expert users perceive, interpret, and fix topic models. International Journal of Human-Computer Studies **105**, 28–42 (2017)
11. Li, P., Tseng, C., Zheng, Y., Chew, J.A., Huang, L., Jarman, B., Needell, D.: Guided semi-supervised non-negative matrix factorization. Algorithms **15**(5) (2022)

12. Lieb, A., Arora, M., Mustafaraj, E.: Creating targeted, interpretable topic models with llm-generated text augmentation. arXiv preprint arXiv:2504.17445 (2025)
13. Lin, J.: Divergence measures based on the shannon entropy. IEEE Transactions on Information Theory **37**(1), 145–151 (1991)
14. Mei, Q., Shen, X., Zhai, C.: Automatic labeling of multinomial topic models. In: Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 490–499 (2007)
15. Mei, Q., Zhai, C.: Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In: Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. pp. 198–207 (2005)
16. Mei, Q., Zhai, C.: A mixture model for contextual text mining. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 649–655 (2006)
17. Nygård, T., Lindfors, P.: Promoting youth well-being: a qualitative study of finnish youtubers' mental health content. Health Promotion International **40**(3), daaf074 (06 2025)
18. Pearce, T., Brintrup, A., Zhu, J.: Understanding softmax confidence and uncertainty. arXiv preprint arXiv:2106.04972 (2021)
19. Pham, D.T., Nguyen Vu, T.T., Nguyen, T., Ngo, L.V., Nguyen, D.A., Nguyen, T.H.: NeuroMax: Enhancing neural topic modeling via maximizing mutual information and group topic regularization. In: Findings of the Association for Computational Linguistics: EMNLP 2024. pp. 7758–7772. Association for Computational Linguistics, Miami, Florida, USA (Nov 2024)
20. Rijcken, E., Scheepers, F., Zervanou, K., Spruit, M., Mosteiro, P., Kaymak, U.: Towards interpreting topic models with chatgpt. In: The 20th World Congress of the International Fuzzy Systems Association (2023)
21. Srivastava, A., Sutton, C.: Autoencoding variational inference for topic models. In: International Conference on Learning Representations. pp. 2223–2234. Curran Associates (2017)
22. Stammbach, D., Zouhar, V., Hoyle, A., Sachan, M., Ash, E.: Revisiting automated topic model evaluation with large language models. arXiv preprint arXiv:2305.12152 (2023)
23. Vendrow, J., Haddock, J., Rebrova, E., Needell, D.: On a guided nonnegative matrix factorization. In: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3265–32369 (2021)
24. Vendrow, J., Haddock, J., Rebrova, E., Needell, D.: On a guided nonnegative matrix factorization. In: ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 3265–32369 (2021)
25. Vishwakarma, H., Chen, Y., Tay, S.J., Namburi, S.S.S., Sala, F., Korlakai Vinayak, R.: Pearls from pebbles: Improved confidence functions for auto-labeling. Advances in Neural Information Processing Systems **37**, 15983–16015 (2024)
26. Wu, X., Nguyen, T., Luu, A.T.: A survey on neural topic models: Methods, applications, and challenges. ArXiv **abs/2401.15351** (2024)
27. Wu, X., Nguyen, T., Zhang, D., Wang, W.Y., Luu, A.T.: Fastopic: Pretrained transformer is a fast, adaptive, stable, and transferable topic model. Advances in Neural Information Processing Systems **37**, 84447–84481 (2024)
28. Yang, X., Zhao, H., Phung, D., Buntine, W., Du, L.: Llm reading tea leaves: Automatically evaluating topic models with large language models. Transactions of the Association for Computational Linguistics **13**, 357–375 (2025)