Tethering Broken Themes: Aligning Neural Topic Models with Labels and Authors

Mayank Nagda and **Phil Ostheimer** and **Sophie Fellenz**
RPTU Kaiserslautern-Landau, Germany
surname@cs.uni-kl.de

Abstract

Topic models are a popular approach for extracting semantic information from large document collections. However, recent studies suggest that the topics generated by these models often do not align well with human intentions. Although metadata such as labels and authorship information are available, it has not yet been effectively incorporated into neural topic models. To address this gap, we introduce FANToM, a novel method to align neural topic models with both labels and authorship information. FANToM allows for the inclusion of this metadata when available, producing interpretable topics and author distributions for each topic. Our approach demonstrates greater expressiveness than conventional topic models by learning the alignment between labels, topics, and authors. Experimental results show that FANToM improves existing models in terms of both topic quality and alignment. Additionally, it identifies author interests and similarities.

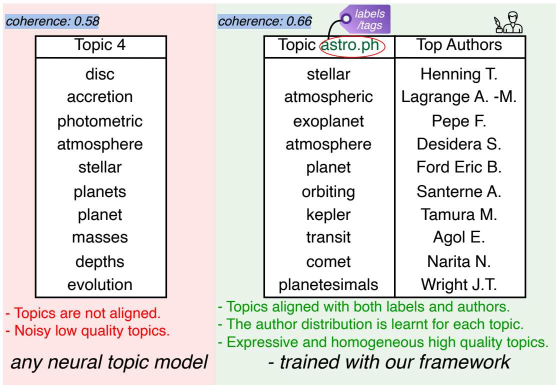

Figure 1: FANToM in action: A comparison of semantically closest topics learned by DVAE (left) and DVAE trained with FANToM (right) for alignment. Notably, FANToM not only accurately aligns the learned topic with the label (astrophysics) and authors but also improves the quality of the learned topic.

1 Introduction

Topic models are a family of generative models that help discover sets of words (called *topics*) describing the semantics of a large document collection (Blei et al., 2003). Topic models find applications in various fields, including healthcare (Rajendra Prasad et al., 2021), political science (Grimmer and Stewart, 2013; Karakkaparambil James et al., 2024), psycholinguistics (Monteiro et al., 2024), bioinformatics (Liu et al., 2016; Burkhardt et al.), among others. These models can be categorized into statistical models, such as latent Dirichlet allocation (LDA) (Blei et al., 2003), or neural topic models (NTMs). NTMs are based on generative models such as variational autoencoders (VAEs) (Miao et al., 2016; Wu et al., 2023; Nagda and Fellenz, 2024) or, more recently, Large Language Models (LLM) (Bianchi et al., 2020). NTMs have been shown to learn topics with improved quality compared to statistical topic models (Hoyle et al., 2021).

However, recent studies reveal that current NTMs often fail to align well with human intentions and labeling (Zhao et al., 2017; Doogan and Buntine, 2021; Hoyle et al., 2022). For instance, in a scenario where the available labels are "Windows" and "MacOS," a misaligned model might merge them into a single "OS" topic, despite the intent to distinguish between these two labels. In other cases, where a general "OS" topic would suffice, the model might inappropriately split topics into specific operating systems.

Beyond labels, authorship information is also crucial for aligning topic models. Typically, an author focuses on a limited range of topics, and understanding these interests is fundamental for NLP and information retrieval tasks that involve large document collections (Rosen-Zvi et al., 2004). Modeling author interests enables us to answer key questions about document content, such as which subjects an author covers, which authors

Findings of the Association for Computational Linguistics: NAACL 2025, pages 740–760
April 29 - May 4, 2025

have similar writing styles, and which authors work on comparable topics (Tang et al., 2022; Li et al., 2015). Although statistical models have been used to model author interests and link topics to authors (Rosen-Zvi et al., 2004), incorporating authorship information into NTMs remains a challenge.

In this paper, we present **FANToM**, a novel **F**ramework for **A**ligning **N**eural **To**pic **M**odels, which incorporates metadata such as labels and authorship information in existing NTMs. As shown in Figure 1, FANToM aligns the learned topics with both the labels and the authors, enhancing the interpretability of the model. This approach not only establishes a connection between latent topics, labels, and authors, but also helps identify author interests based on the topics.

In summary, our contributions are as follows.

- We introduce FANToM, a framework that aligns latent topics in NTMs with document labels and authors (Section 4.2). [1]

- We demonstrate through experiments (Section 5) that FANToM not only effectively aligns NTMs with labels and authors but also improves the quality of the learned topics, outperforming existing models.

- Our extensive experiments (Section 6) show how FANToM facilitates learning a shared embedding space for authors, words, and topics, and provides insight into author interests and similarities.

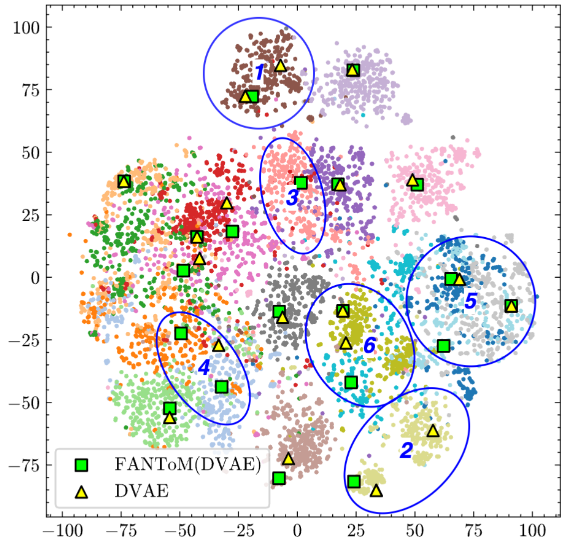

Figure 2: t-SNE projection of topic embeddings from the DVAE model (triangles) and its FANToM variant (squares), alongside document embeddings from the 20NG dataset, color-coded by labels. Ideally, topic embeddings should be positioned near the centroid of their corresponding document clusters. The circled regions highlight discrepancies where DVAE either overrepresents or underrepresents certain topics, while FANToM achieves a more balanced and accurate alignment with document labels, reinforcing its effectiveness in topic representation.

2 Are Neural Topic Models Misaligned?

Recently, the alignment of discovered topics with human-determined labels has come under scrutiny (Zhao et al., 2017; Doogan and Buntine, 2021; Hoyle et al., 2022). Despite decades of application in various domains, existing topic models struggle to align their generated topics with the intentions and expectations of human users. This discrepancy is particularly concerning, given that the primary objective of topic models is to uncover meaningful, interpretable patterns from text data. For instance, Hoyle et al. (2022) highlighted issues of low topic purity and stability by utilizing human-assigned labels. Their findings indicate that numerous models fail to capture the nuanced distinctions that users make between different topics.

To further investigate this issue, we analyze the document and topic embeddings generated by the baseline model, Dirichlet-VAE (DVAE) (Burkhardt and Kramer, 2019a), and its FANToM variant. We obtain the document and topic embeddings by using SBERT (Reimers and Gurevych, 2019) which allocates the embeddings based on the content of the text. Ideally, topic embeddings should be positioned near the centroids of document embeddings, signaling high cluster purity and a strong correlation between topics and their respective documents. In Figure 2, using t-SNE (Van der Maaten and Hinton, 2008) projection we visualize these embeddings and pinpoint six problematic regions that reveal significant concerns regarding topic representation.

In regions 1, 2, and 6, we observe that the DVAE model tends to overrepresent certain clusters, leading to the generation of multiple topics for the same document label. This phenomenon not only diminishes topic diversity but also clouds the interpretability of the model's outputs. For example, when multiple topics are assigned to a single label, it becomes challenging for users to discern

[1] Code: https://github.com/mayanknagda/fantom

the unique contributions of each topic, ultimately undermining the model's utility.

Conversely, in regions 3, 4, and 5, the DVAE model demonstrates an underrepresentation of other clusters, resulting in the complete omission of some niche topics. This imbalance is detrimental as it indicates that the model is not adequately capturing the breadth of the data it is trained on. Such gaps in representation can lead to significant blind spots in the insights generated by the model, leaving users without a comprehensive understanding of the underlying themes present in the corpus.

These observations lead to our main hypothesis. We hypothesize that the misalignment observed in neural topic models, is largely a consequence of the unconstrained nature of the latent topic distribution. In the absence of constraints, the model is free to allocate probability mass as it sees fit, which can result in overrepresentation of certain labels while simultaneously leading to the underrepresentation of others. This inherent instability can manifest itself in two problematic ways: the emergence of redundant topics that offer little new information and the failure to recognize and represent niche topics that are critical for a holistic understanding of the data.

By integrating metadata such as document labels and authorship, our proposed model, FANToM, imposes necessary constraints on the topic distribution. These constraints serve to enhance both the stability and alignment of the model, thereby improving its overall performance.

3 Related Work

In this section, we present related work on topic models and its alignment.

3.1 Topic Models

Topic models based on Latent Dirichlet Allocation (LDA) (Blei et al., 2003) were originally trained using variational inference or Gibbs sampling (Griffiths et al., 2004). Miao et al. (2016) introduced neural topic models (NTMs) based on Variational Autoencoders (VAEs) with a Gaussian prior on latent topic variables, which were later extended with a Dirichlet prior (Srivastava and Sutton, 2017; Burkhardt and Kramer, 2019a). Other NTMs based on Generative Adversarial Networks (GANs) (Wang et al., 2020) and word embeddings (ETMs) (Dieng et al., 2020; Wu et al., 2023) have also been proposed. NTMs generally outperform statistical models in terms of topic quality (Srivastava and Sutton, 2017) due to their more flexible generative distributions. Furthermore, NTMs are compatible with advances in deep learning, such as Large Language Models (LLMs) (Bianchi et al., 2020) and word embeddings (Dieng et al., 2020; Wu et al., 2023). Recent LLM-based topic models cluster document embeddings from LLMs using simple clustering methods, such as k-means (Grootendorst, 2022) or Gaussian mixture models (Sia et al., 2020). However, some researchers do not consider these clustering methods as topic models because they do not produce document-topic distributions (Wu et al., 2023). An example of an LLM-based topic model is CTM (Bianchi et al., 2020), which uses a VAE for clustering document embeddings. Recent state-of-the-art ETMs, such as ECRTM (Wu et al., 2023), cluster word embeddings with topic embeddings as centers using soft assignment. Our framework allows for the first time the alignment of all variants of NTMs based on VAEs, and we compare it to these models in our experiments.

3.2 Alignment of Topic Models

Topic model alignment involves ensuring that latent topics are aligned with metadata such as labels or authorship information (Rahimi et al., 2023; Chuang et al., 2013; Abels et al., 2021). This is typically achieved through supervision using metadata (Ramage et al., 2009; Rosen-Zvi et al., 2004). Topic model supervision can be divided into methods to improve interpretability with metadata and methods with the goal of improving classification (Burkhardt et al., 2018; Burkhardt and Kramer, 2019b,c). However, with advancements in LLMs, the use of topic models for classification has become less relevant. We focus on supervising models to improve alignment and interpretability.

Alignment of Statistical Topic Models. Labeled LDA learns one topic per label (Ramage et al., 2009), while the Author-Topic Model (ATM) learns topic distributions for each author (Steyvers et al., 2004; Rosen-Zvi et al., 2004), enabling the computation of author similarities. Variants include the Author-Conference-Topic (ACT) model (Tang et al., 2008) and the Author-Recipient-Topic (ART) model (McCallum et al., 2005), which adds recipient-based analysis for emails. We use Labeled LDA and ATM as baselines in our experiments.

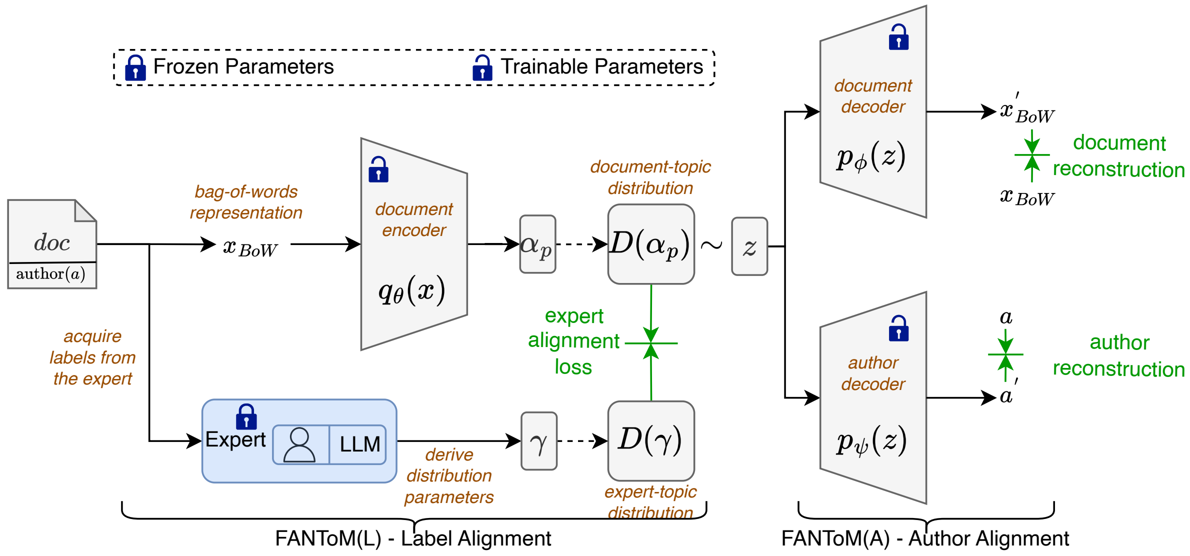

Figure 3: Illustration of FANToM: The framework aligns labels and authorship information with topics. It incorporates expert-assigned labels to establish a prior distribution parameterized by γ, which is then aligned with the posterior. For authorship, a separate decoder is used to learn the multinomial distribution over authors, ensuring a structured representation of author-topic relationships. Overall, FANToM ensures a structured and interpretable alignment between topics, labels, and authors.

Alignment of Neural Topic Models. Neural topic models (NTMs) have seen increasing development (Miao et al., 2016; Srivastava and Sutton, 2017; Burkhardt and Kramer, 2019a); however, the number of supervised NTMs remains limited. SCHOLAR (Card et al., 2018) uses metadata as input labels and constructs a classifier network from the latent vector to predict labels, generating topics relevant for classification. Rahimi et al. (2023) proposed an aligned neural topic model for dynamic, evolving topics, which contrasts with our focus on static topics. TAM (Wang and Yang, 2020) trains an RNN classifier jointly with an NTM to predict labels from word sequences. Other models emphasize classification over alignment (Bai et al., 2018; Korshunova et al., 2019). Generally, NTMs do not enforce topic-label alignment, and aligned variants of supervised NTMs have not been thoroughly explored. Studies reveal that current NTMs often fail to align well with human-defined labels (Zhao et al., 2017; Doogan and Buntine, 2021; Hoyle et al., 2022). In our experiments, we compare FANToM to supervised NTMs such as SCHOLAR and TAM.

4 Methodology

We begin by discussing the background in Sec. 4.1 and present our proposed method in Sec. 4.2.

4.1 Background

VAE-based topic models use an encoder-decoder architecture. Let $\{x_i\}_N$ represent the observed input documents in the Bag-of-Words (BoW) format, where $x_i \in \mathbb{N}^V$ and V is the vocabulary size. The encoder, parameterized by θ, maps the input to a latent vector z, while the decoder, parameterized by ϕ, reconstructs the documents (Burkhardt and Kramer, 2019a; Srivastava and Sutton, 2017).

The objective is to learn the parameters θ and ϕ by minimizing the β-VAE loss:

$$\begin{aligned}\mathcal{L}_{\text{vae}}(\theta, \phi; x) = &-\mathbb{E}_{q_\theta(z|x)}\left[\log p_\phi(x \mid z)\right] \\ &+ \beta D_{KL}\left[q_\theta(z \mid x) \| p_\alpha(z)\right],\end{aligned} \tag{1}$$

where the first term is the reconstruction loss, and the second is the Kullback–Leibler (KL) divergence (Kullback and Leibler, 1951), which acts as a regularizer. β balances these terms (Higgins et al., 2017). The prior $p_\alpha(z)$ is typically a uniform Dirichlet distribution with parameters $\alpha \ll 1$ (Burkhardt and Kramer, 2019a). The approximate posterior, $q(z|x)$, is modeled by a Dirichlet distribution with parameters α_p, derived from the encoder output. The latent vector $z_i \in \mathbb{R}^K$ denotes the document-topic distribution, where K is the number of topics. $\phi \in \mathbb{R}^{K \times V}$ represents the normalized topic-word distributions. This approach is fully unsupervised and does not incorporate labels or authors by default.

4.2 FANToM: Aligning Neural Topic Models

We now introduce our proposed method, as illustrated in Figure 3. We first explain the integration of labels and authorship information, subsequently merging these elements into a unified framework.

FANToM(L): Aligning Topics with Labels. To align topics with labels, we supervise the document-topic distribution using experts[2], ensuring that each document's topics are restricted to its assigned labels. We achieve this by deriving an expert-aligned Dirichlet prior $p_\gamma(z)$ with parameters γ, ensuring that the posterior $q_\theta(z|x)$ aligns with this prior through an expert alignment loss.

Let Λ represent the set of possible labels and K the total number of topics. We define a global topic-label vector $L \in \{\Lambda \cup \{\text{no-label}\}\}^K$, which assigns a label to each topic k. The "no-label" token is used when no label is assigned (e.g., in semi-supervised settings). L is derived from a mapping $f : \{k\}_{k=1}^{K} \rightarrow \{\Lambda \cup \{\text{no-label}\}\}$. For each document d, let λ^{d} be the set of labels assigned to the document. We then derive a multi-hot vector $\mathbb{I}^{\text{d}}$ as: $(\mathbb{I}^{\text{d}}_k = 1 \text{ if } L_k \in \lambda^{\text{d}}, \text{ else } 0)_{k=1}^{K}$. The k^{th} element of $\mathbb{I}^{\text{d}}$ is 1 if L_k is a label for the document. We derive the parameters γ for the expert-aligned Dirichlet prior $p_\gamma(z)$ as $\gamma = \alpha \cdot \mathbb{I}^{\text{d}}$, where $\alpha = (\alpha_1, \ldots, \alpha_K)$ represents the base Dirichlet distribution over topics.

For example, if $L = (1, 1, 2, 2, 3)$ and a document d has labels $\lambda^{\text{d}} = \{2\}$, then $\mathbb{I}^{\text{d}} = (0, 0, 1, 1, 0)$, and $\gamma = (0, 0, \alpha_3, \alpha_4, 0)$, ensuring that the document's topics are restricted to its assigned labels. We illustrate this example extensively in Appendix B.1. This approach ensures that the model focuses on relevant topics, improving interpretability and topic-label alignment.

FANToM(A): Parameterizing the Topic-Author Distribution. The existing VAE-based frameworks do not directly incorporate authorship information. We address this by using a separate decoder to learn a multinomial topic-author distribution based on authorship. Let A represent the author vocabulary, where each document d is associated with one or more authors, represented by a multi-hot vector $a \in \{0, 1\}^{|A|}$. The author-decoder, parameterized by ψ, reconstructs the authors from the latent topics. The learned parameter $\psi \in \mathbb{R}^{K \times |A|}$ represents the topic-author distributions.

Training Objective: Given the expert-aligned prior $p_\gamma(z)$, the observed authors $\{a_i\}_N$ where $a_i \in \{0, 1\}^{|A|}$, and the author likelihood p_ψ, the FANToM training objective is defined as:

$$\begin{aligned} \mathcal{L}_{\text{F}}(\theta, \phi, \psi; x) = & -\mathbb{E}_{q_\theta(z|x)}\left[\log p_\phi(x \mid z)\right] \\ & -\mathbb{E}_{q_\theta(z|x)}\left[\log p_\psi(a \mid z)\right] \\ & +\beta D_{\text{KL}}\left[q_\theta(z \mid x) \| p_\gamma(z)\right], \end{aligned} \tag{2}$$

where the first term represents the document reconstruction error, the second term accounts for the author reconstruction, and the third term is the expert-alignment loss. The document reconstruction identifies latent topics. The author reconstruction term ensures that the authors associated with a document are constrained by the topics, promoting alignment between authors and topics. The expert-alignment KL loss incorporates the expert-aligned prior $p_\gamma(z)$ instead of $p_\alpha(z)$, enforcing alignment between latent topics and labels. FANToM follows a training regime similar to existing VAE-based NTMs. More details on the training process and the algorithm are provided in Appendix B.2.

5 Experiments

In this section, we first describe the datasets, comparison models, and evaluation metrics used in our experiments. We then demonstrate FANToM's alignment capabilities in Section 5.4 and Section 5.5, followed by a benchmark comparison against baselines in Section 5.6.

5.1 Datasets and Preprocessing

We use four well-known datasets in our experiments. The 20 Newsgroups (20NG) dataset contains around 18,000 newsgroup posts categorized into 20 labels (Lang, 1995). The AG News (AGN) corpus has over a million news articles from 2,000+ sources, divided into four groups (Zhang et al., 2015). The DBpedia-14 (DB-14) dataset consists of 14 distinct classes selected from DBpedia 2014 (Zhang et al., 2015). Lastly, we use the arXiv dataset (arxiv) (arXiv.org submitters, 2023), which includes six labels and authors with at least ten papers for evaluating author models.

For tokenization, we utilize SpaCy (Honnibal and Montani, 2017). We remove stop words, punctuation, and words that appear in fewer than 30 documents or in more than 85% of the documents.

[2]An expert can be a human labeler or an external source like a large language model (LLM). In our experiments, we use both types of experts.

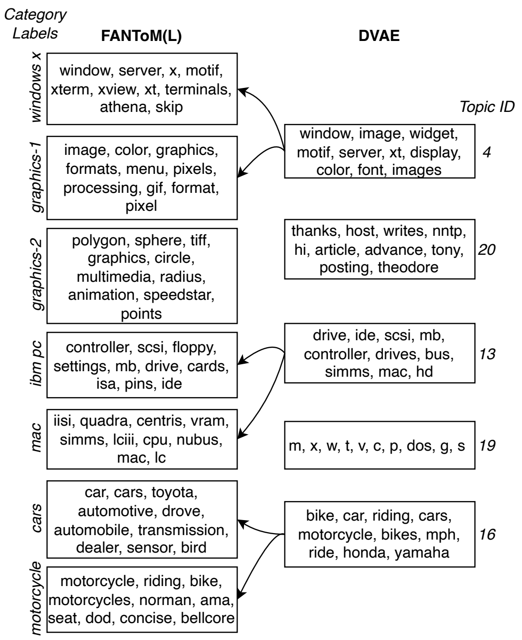

Figure 4: Comparison of topic alignment between FANToM(L) and DVAE (baseline) on the 20NG dataset. The semantically closest topics are linked (right to left). FANToM(L) cleanly separates topics based on labels, while DVAE lacks this distinction. FANToM(L) generates esoteric topics closely aligned with labels and learns multiple topics within the *graphics* label.

Additional details about the datasets and preprocessing are provided in Appendix C.

5.2 Models

Baselines: We compare FANToM against several supervised and unsupervised topic models. As supervised baselines, we use SCHOLAR (Card et al., 2018), TAM (Wang and Yang, 2020), and Labeled LDA (L-LDA) (Ramage et al., 2009). For author modeling, we include the Author-Topic Model (ATM) (Rosen-Zvi et al., 2004).[3] For unsupervised NTMs, we include Dirichlet-VAE (DVAE) (Burkhardt and Kramer, 2019a), Embedded Topic Model (ETM) (Dieng et al., 2020), and ECRTM (Wu et al., 2023), which extends ETM. Additionally, we use Contextualized Topic Models (CTM) (Bianchi et al., 2020), with document embeddings from SBERT (Reimers and Gurevych, 2019).

FANToM: Our proposed framework can be integrated into existing NTMs. We present FANToM variants for all NTM baselines, including SCHOLAR, DVAE, ETM, ECRTM, and CTM. To maintain consistency, all neural baselines and FANToM use a Dirichlet prior. See Appendix A for more details.

[3]We use implementations from (Fenstermacher and Schneider, 2021) and (Řehůřek and Sojka, 2010) for L-LDA and ATM, respectively.

5.3 Evaluation Measures

In addition to qualitative analysis, we evaluate model performance quantitatively using the standard Topic Quality (TQ) measure, which is defined as the product of Topic Coherence (TC) and Topic Diversity (TD) (Dieng et al., 2020). TQ reflects both interpretability and diversity of topics. TC, calculated via the C_V coherence score (Röder et al., 2015), measures the co-occurrence of top words within a topic using a reference corpus. We use the WikiText-103 dataset (Merity et al., 2017) as our reference corpus, consisting of approximately 2 million Wikipedia articles. TD assesses the proportion of unique words across all topics, with scores close to zero indicating redundancy and scores near one reflecting high diversity. Additionally, we perform experiments on document clustering to assess the topic alignment using Purity and NMI, following (Wu et al., 2023; Hoyle et al., 2022). Purity assesses topic homogeneity by assigning each topic the most frequently co-occurring label, while NMI evaluates the mutual information between true labels and predicted topics, thereby measuring alignment in terms of precision and recall. Moreover, across all experiments, we conduct a two-tailed t-test to assess significance.

5.4 Alignment of Topics with Labels

To demonstrate the problem of misaligned topics, we compare the alignment of FANToM(L) with the DVAE baseline on the 20NG dataset (Figure 4). DVAE tends to generate general, noisy topics, prioritizing common words and leading to information loss. For example, in Topic ID 4, DVAE merges the graphics and windows topics, adding noisy words like "font" that do not align with the label. In contrast, FANToM(L) separates topics more effectively, guided by labels. For instance, DVAE fails to generate a distinct *mac* topic, while FANToM(L) successfully separates it. FANToM(L) produces more expressive, homogeneous topics closely aligned with the labels.

To capture multiple topics within a single label, we assigned two topic indices to the *graphics* label using the topic-label vector L, resulting in two distinct topics (*graphics-1* and *graphics-2*). As seen in Figure 4, FANToM(L) distinguishes between topics related to digital graphics software (*graphics-1*)

	Models	20NG			AGN			DB-14			arxiv		
		TQ	Purity	NMI	TQ	Purity	NMI	TQ	Purity	NMI	TQ	Purity	NMI
baseline	L-LDA	0.192	0.629	**0.681**	0.320	0.815	0.747	0.518	0.894	0.876	0.246	0.875	0.806
	SCHOLAR	0.391	0.341	0.375	0.372	0.617	0.581	0.601	0.811	0.809	0.343	0.796	0.779
	TAM	0.377	0.328	0.361	0.385	0.603	0.570	0.620	0.785	0.832	0.355	0.770	0.791
	DVAE	0.354	0.336	0.325	0.393	0.743	0.691	0.628‡	0.829	0.810	0.339	0.862	0.836
	ETM	0.398	0.491	0.482	0.310	0.795	0.774	0.603	0.902	0.896	0.264	0.884	0.875
	CTM	0.412	0.390	0.374	0.371	0.763	0.749	0.603	0.858	0.841	0.324	0.814	0.799
	ECRTM	0.434	0.628	0.617	0.391	0.826	0.809	0.640	0.894	0.881	0.315	0.886	0.862
FANToM(ours)	SCHOLAR	0.403‡	0.649‡	0.601‡	0.397‡	0.841‡	0.819‡	0.608	0.915‡	0.887‡	**0.375**‡	0.919‡	0.893‡
	DVAE	0.424‡	0.629‡	0.611‡	**0.409**‡	0.837‡	0.821‡	0.611	0.898‡	0.874‡	0.336	0.910‡	0.904‡
	ETM	0.399	0.670‡	0.641‡	0.391‡	0.854‡	0.830‡	0.630‡	0.924	0.905	0.310‡	0.920‡	0.915‡
	CTM	0.436‡	0.651‡	0.617‡	0.395‡	0.847‡	0.828‡	0.622‡	0.916‡	0.908‡	0.342‡	0.895‡	0.873‡
	ECRTM	**0.441**	**0.710**‡	0.681‡	0.402	**0.918**‡	**0.873**‡	**0.651**‡	**0.937**‡	**0.917**‡	0.332‡	**0.941**‡	**0.925**‡

Table 1: Comparison of the proposed FANToM(L) (bottom) against the respective non-aligned baselines and L-LDA (top) using Topic Quality, Purity, and NMI measures. The best results across datasets are highlighted in bold, and significantly better results (p-value < 0.05) between baselines and FANToM(L) are marked with ‡. Our FANToM approach significantly outperforms the existing baselines.

and geometry (*graphics-2*). These distinctions are weak or absent in DVAE.

To quantify alignment, we calculate purity and NMI (Wu et al., 2023) to measure how well topics in the latent space correspond to assigned labels. As shown in Table 1, FANToM(L) significantly outperforms baselines in producing more aligned document-topic distributions, as indicated by the higher purity and NMI scores.

5.5 Alignment of Topics with Authors

Authorship information can further refine topic alignment. As shown in Table 2, FANToM(A) outperforms ATM, the baseline author-topic model. Additionally, FANToM, which aligns both labels and authors, achieves higher scores than FANToM(L) and FANToM(A) individually. For this experiment, we use the arXiv dataset, the only one in our benchmarks with author information. Figure 1 illustrates topic-author alignment, with additional examples provided in Appendix F.2.

5.6 Benchmarking FANToM(L)

To measure the quality of generated topics, we benchmark the proposed label alignment models against state-of-the-art NTMs in Table 1. Since FANToM can be integrated with any existing VAE-based topic model variant, we incorporate FANToM with all the baselines in the benchmarking. Labels from the datasets are used to train our models, and to ensure consistency, the total number of topics learned across all models is set to match the number of labels. Our results are averaged over five independent runs. Contrary to our expectation of a trade-off between alignment and topic quality, incorporating labels does not harm the quality of the topics produced; on the contrary, it even improves the quality significantly. Individual coherence and diversity measures are listed in Appendix F.1 for further analysis.

We also conduct benchmarking experiments in a semi-supervised scenario where we do not have labels for all the topics in our corpus. To address this, we assign the "no-label" designation to all documents lacking labels. In our experiment, we randomly remove labels from 50% of the documents across all datasets and learn 50 and 200 topics, resulting in some topics being labeled while the rest are labeled as "no-label." FANToM consistently outperforms the baseline models, as shown in Table 4 (Appendix F.1).

6 Discussion

Our experiments confirm the achieved alignment qualitatively and quantitatively. We do not observe any trade-off between alignment and topic quality; on the contrary, we observe improved quality compared to the SOTA baselines. Crucially, we also show that the statistical topic model Labeled LDA performs worse than all neural topic models, even though Labeled LDA is trained with the Gibbs sampling algorithm, which is less prone to local minima during training than variational inference.

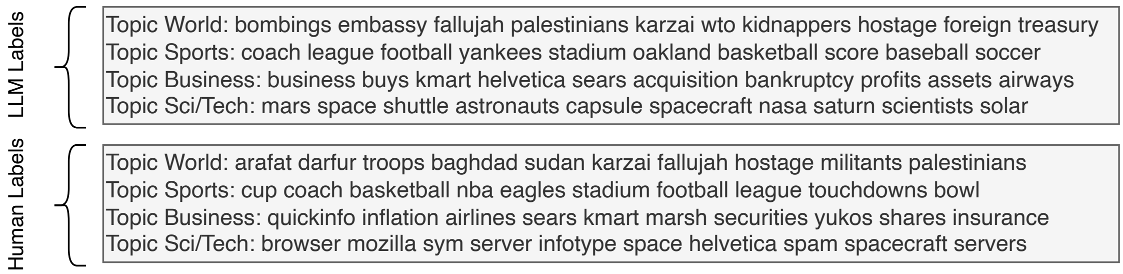

Figure 5: Comparison between the topic words estimated using human labels (bottom) and LLM labels (top) in the ag news corpus, using FANToM(L) for topic estimation. Both align well with the corresponding labels.

Models	TQ	Purity	NMI
ATM	0.287	0.730	0.759
DVAE	0.339	0.862	0.836
FANToM(L)	0.336	0.910	0.904
FANToM(A)	0.354	0.889	0.861
FANToM	**0.362**	**0.951**	**0.948**

Table 2: Comparison of FANToM (L), FANToM (A), and FANToM against non-aligned baseline (DVAE) and aligned statistical baseline (ATM). Best results are highlighted in bold. Both FANToM and FANToM (A) significantly outperform ATM on all aspects.

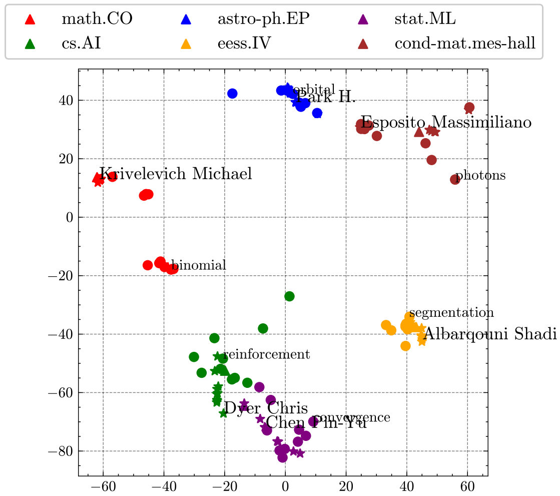

Figure 6: The shared embedding space between author ($\star$), word (‡), and topic ($\blacktriangle$) embeddings shows authors and words in close proximity to their respective topics.

This demonstrates that aligning neural topic models is a crucial step towards further improving their performance and practical relevance. We now discuss some applications of FANToM's alignment.

Learning an Embedding Space Between Topics and Authors: We are able to extract informative word, topic, and author embeddings by using FANToM with embedding (ETM) decoders for both topics and authors. Figure 6 shows the TSNE (Van der Maaten and Hinton, 2008) projection of the embeddings in a shared embedding space. We assign the learned label to each topic, allowing us to easily identify topics. The figure demonstrates that words and authors belonging to a particular topic are close to each other, while each topic cluster remains distinct from others. The learned embeddings further open avenues for exploration and analysis of associated authors and topics.

Using LLMs as Experts Labeling datasets can be challenging, but Large Language Models (LLMs) offer an alternative for assigning labels to unlabeled documents. We explore this approach using BART (Lewis et al., 2019) for zero-shot classification, leveraging a pretrained model (Wolf et al., 2019) fine-tuned on Natural Language Inference (NLI) (Williams et al., 2018) following (Yin et al., 2019). This method frames classification as an NLI task, where a sequence serves as the premise and each potential label is tested via a hypothesis (e.g., "This text is about sports"). By computing entailment and contradiction probabilities, we derive label assignments. Applying this to our dataset and comparing with human labels, we achieve 92% accuracy, validating LLM effectiveness in document categorization. Additionally, using LLM-generated labels for the AG News corpus (AGN) with FANToM(L) yields a mean topic quality of 0.503, outperforming the 0.409 obtained with human labels. Figure 5 illustrates the topic words estimated from both sources, showing strong alignment with their corresponding labels.

Author Interests and Similarity: In addition to aligned topics, we also learn authors' interests as FANToM maps each author to a topic distribution and vice versa. We first validate this comparing FANToM's recommended labels for each author to ground truth labels. We find that FANToM accurately associates 91.6% of authors with their respective labels.

To evaluate author interests, we train the proposed FANToM(A) model on the "stats.ML" label from the (arXiv.org submitters, 2023) corpus, spanning the years 2010 to 2020. Each author is associated with different topics (interests), as illustrated in Figure 7, where the line thickness corresponds to the weight of the association. We specifically select two prominent researchers in the field, namely *Blei David M.* and *Goodfellow Ian*, and identify their top two matching topics. The figure illustrates the correct alignment between these researchers and their respective research interests. Furthermore, the author-topic vectors can also be used to compute the similarity between authors. One such similarity matrix is depicted in Figure 9 in Appendix F.

Topic Models and LLMs: In recent years, the combination of topic modeling with LLMs has emerged as a key area of research. Recent efforts have explored combining topic modeling with LLMs, such as TopicGPT (Pham et al., 2024), which uses prompt engineering to address topic modeling and does not provide a framework for alignment. However, the direct prompting method incurs high computational costs for processing prompts of an individual document. FANToM, on the other hand, employs a VAE for topic modeling and uses a smaller, task-specific LLM for labeling, offering alignment, reduced computational costs, greater flexibility and efficiency. We discuss this aspect in detail in the Appendix G.

7 Conclusion

In this work, we propose FANToM, a novel neural architecture for aligning Neural Topic Models (NTMs) with expert-assigned labels and authorship information. Our results show that (i) the alignment effectively captures the corresponding associations among topics, labels, and authors, and (ii) FANToM outperforms existing state-of-the-art models in terms of topic quality. This underscores the importance of aligning topic models and paves the way for potential downstream applications. We investigate learning a shared embedding space between topics, authors, and words, enabling us to identify author interests and compute the similarity between and within authors and topics. As incorporating prior knowledge into machine learning has recently seen significant success (Nagda et al., 2024b,a; Specht et al., 2024; Vollmer et al., 2024; Jirasek et al., 2023; Manduchi et al., 2024), promising lines of future research include the possibility of combining prior knowledge with the modeling of joint topic, label, and author hierarchies.

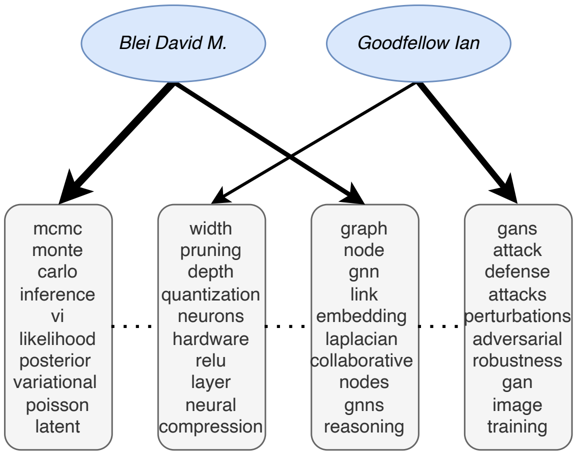

Figure 7: The top two topics of *Blei David M.* and *Goodfellow Ian*, based on cosine similarity, accurately represent the respective research interests of the authors, aligning well with their contributions in probabilistic modeling and generative adversarial networks, respectively.

Limitations

Our approach may face challenges when an author suddenly writes about a completely unrelated topic. However, such cases are relatively rare, and when there is even partial thematic overlap, we believe FANToM remains robust. Addressing extreme cases of topic divergence presents an interesting direction for future research. Although FANToM is a flexible framework, it necessitates the use of labels and authors for alignment, which involves collecting metadata. When expert labels are not readily available, using even a smaller LLM as an alternative can result in higher computational costs and resource demands. Additionally, LLMs face limitations due to context length constraints. In our experiments, truncation was required, but it is not an optimal solution. Future research could explore methods to handle full documents within these length limits, such as processing documents in chunks, selecting representative segments, or summarizing the content. Another challenge with LLMs is multilinguality; for some languages, finding a small-scale LLM for label assignment may be impractical and increase reliance on human experts for label assignment.

Ethical Statement

In this study, we incorporated authorship information into topic models while ensuring ethical considerations. To acquire the dataset, we obtained it from the openly available (arXiv.org submitters, 2023), which is licensed under the *CC0 1.0 Universal (CC0 1.0) Public Domain Dedication*, allowing us unrestricted use of the data. We specifically utilized the abstracts and author names of openly available papers. It is important to note that we have mentioned author names as they were presented to us by our model, without making any alterations. We did not have access to any sensitive or restricted information.

Acknowledgments

The authors acknowledge support by the Carl-Zeiss Foundation, DFG awards BU 4042/2-1, FE 2282/6-1, and FE 2282/1-2, as well as the BMBF award 01|S2407A.

References

Patrick Abels, Zahra Ahmadi, Sophie Burkhardt, Benjamin Schiller, Iryna Gurevych, and Stefan Kramer. 2021. Focusing knowledge-based graph argument mining via topic modeling. *arXiv preprint arXiv:2102.02086*.

arXiv.org submitters. 2023. arxiv dataset.

Haoli Bai, Zhuangbin Chen, Michael R Lyu, Irwin King, and Zenglin Xu. 2018. Neural relational topic models for scientific article analysis. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 27–36.

Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2020. Cross-lingual contextualized topic models with zero-shot learning. *arXiv preprint arXiv:2004.07737*.

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

Sophie Burkhardt and Stefan Kramer. 2019a. Decoupling sparsity and smoothness in the dirichlet variational autoencoder topic model. *J. Mach. Learn. Res.*, 20(131):1–27.

Sophie Burkhardt and Stefan Kramer. 2019b. Multi-label classification using stacked hierarchical dirichlet processes with reduced sampling complexity. *Knowledge and Information Systems*, pages 1–23.

Sophie Burkhardt and Stefan Kramer. 2019c. A survey of multi-label topic models. *ACM SIGKDD Explorations Newsletter*, 21(2):61–79.

Sophie Burkhardt, Julia Siekiera, Josua Glodde, Miguel A. Andrade-Navarro, and Stefan Kramer. *Towards identifying drug side effects from social media using active learning and crowd sourcing*, pages 319–330.

Sophie Burkhardt, Julia Siekiera, and Stefan Kramer. 2018. Semi-supervised bayesian active learning for text classification. In *Bayesian Deep Learning Workshop at NeurIPS*.

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.

Jason Chuang, Sonal Gupta, Christopher Manning, and Jeffrey Heer. 2013. Topic model diagnostics: Assessing domain relevance via topical alignment. In *International conference on machine learning*, pages 612–620. PMLR.

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

Caitlin Doogan and Wray Buntine. 2021. Topic model or topic twaddle? re-evaluating semantic interpretability measures. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3824–3848, Online. Association for Computational Linguistics.

Douglas Fenstermacher and Jonathan Schneider. 2021. bab2min/tomotopy. *Zenodo*.

Thomas Griffiths, Mark Steyvers, David Blei, and Joshua Tenenbaum. 2004. Integrating topics and syntax. *Advances in neural information processing systems*, 17.

Justin Grimmer and Brandon M Stewart. 2013. Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political analysis*, 21(3):267–297.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint*. ArXiv:2203.05794 [cs].

Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Matthew Honnibal and Ines Montani. 2017. spaCy: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

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.

Alexander Hoyle, Pranav Goel, Rupak Sarkar, and Philip Resnik. 2022. Are neural topic models broken? *Preprint*, arXiv:2210.16162.

Fabian Jirasek, Sophie Fellenz, Robert Bamler, Michael Bortz, Marius Kloft, Stephan Mandt, and Hans Hasse. 2023. Making thermodynamic models of mixtures predictive by machine learning: matrix completion of pair interactions. In *ECML/PKDD, Workshop on Neuro-Explicit AI and Expert-informed Machine Learning for Engineering and Physical Sciences*.

Charu Karakkaparambil James, Mayank Nagda, Nooshin Haji Ghassemi, Marius Kloft, and Sophie Fellenz. 2024. Evaluating dynamic topic models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 160–176, Bangkok, Thailand. Association for Computational Linguistics.

Iryna Korshunova, Hanchen Xiong, Mateusz Fedoryszak, and Lucas Theis. 2019. Discriminative topic modeling with logistic lda. *Advances in neural information processing systems*, 32.

Solomon Kullback and Richard A Leibler. 1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In Armand Prieditis and Stuart Russell, editors, *Machine Learning Proceedings 1995*, pages 331–339. Morgan Kaufmann, San Francisco (CA).

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chunshan Li, William K Cheung, Yunming Ye, Xiaofeng Zhang, Dianhui Chu, and Xin Li. 2015. The author-topic-community model for author interest profiling and community discovery. *Knowledge and Information Systems*, 44:359–383.

Lin Liu, Lin Tang, Wen Dong, Shaowen Yao, and Wei Zhou. 2016. An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus*, 5(1):1–22.

Laura Manduchi, Kushagra Pandey, Robert Bamler, Ryan Cotterell, Sina Däubener, Sophie Fellenz, Asja Fischer, Thomas Gärtner, Matthias Kirchler, Marius Kloft, et al. 2024. On the challenges and opportunities in generative ai. *arXiv preprint arXiv:2403.00025*.

Andrew McCallum, Andrés Corrada-Emmanuel, and Xuerui Wang. 2005. Topic and role discovery in social networks. In *Proceedings of the 19th International Joint Conference on Artificial Intelligence*, IJCAI'05, page 786–791, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. In *International Conference on Learning Representations*.

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

Marcio Monteiro, Charu Karakkaparambil James, Marius Kloft, and Sophie Fellenz. 2024. Characterizing text datasets with psycholinguistic features. In *Findings of the Association for Computational Linguistics: EMNLP 2024*. Association for Computational Linguistics.

Mayank Nagda and Sophie Fellenz. 2024. Putting back the stops: Integrating syntax with neural topic models. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 6424–6432. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Mayank Nagda, Phil Ostheimer, Thomas Specht, Frank Rhein, Fabian Jirasek, Marius Kloft, and Sophie Fellenz. 2024a. Pits: Physics-informed transformers for predicting chemical phenomena. In *ECML/PKDD, Workshop on Machine Learning for Chemistry and Chemical Engineering*.

Mayank Nagda, Phil Ostheimer, Thomas Specht, Frank Rhein, Fabian Jirasek, Marius Kloft, and Sophie Fellenz. 2024b. Setpinns: Set-based physics-informed neural networks. *arXiv preprint arXiv:2409.20206*.

Phil Ostheimer, Mayank Nagda, Marius Kloft, and Sophie Fellenz. 2024. Text style transfer evaluation using large language models. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-Coling)*.

Phil Ostheimer, Mayank Kumar Nagda, Marius Kloft, and Sophie Fellenz. 2023. A call for standardization and validation of text style transfer evaluation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10791–10815, Toronto, Canada. Association for Computational Linguistics.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Chau Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024. TopicGPT: A prompt-based topic modeling framework. In *Proceedings of* *the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2956–2984, Mexico City, Mexico. Association for Computational Linguistics.

Hamed Rahimi, Hubert Naacke, Camelia Constantin, and Bernd Amann. 2023. Antm: An aligned neural topic model for exploring evolving topics. *arXiv preprint arXiv:2302.01501*.

K Rajendra Prasad, Moulana Mohammed, and RM Noorullah. 2021. Visual topic models for healthcare data clustering. *Evolutionary Intelligence*, 14(2):545–562.

Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D. Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256, Singapore. Association for Computational Linguistics.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408.

Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, page 487–494, Arlington, Virginia, USA. AUAI Press.

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.

Thomas Specht, Mayank Nagda, Sophie Fellenz, Stephan Mandt, Hans Hasse, and Fabian Jirasek. 2024. Hanna: Hard-constraint neural network for consistent activity coefficient prediction. *Preprint*, arXiv:2407.18011.

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

Mark Steyvers, Padhraic Smyth, Michal Rosen-Zvi, and Thomas Griffiths. 2004. Probabilistic author-topic models for information discovery. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 306–315.

Jie Tang, Ruoming Jin, and Jing Zhang. 2008. A topic modeling approach and its integration into the random walk framework for academic search. In *2008 Eighth IEEE International Conference on Data Mining*, pages 1055–1060. IEEE.

Xudong Tang, Chao Dong, and Wei Zhang. 2022. Contrastive author-aware text clustering. *Pattern Recognition*, 130:108787.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Luisa Vollmer, Sophie Fellenz, Fabian Jirasek, Heike Leitte, and Hans Hasse. 2024. Knowtd - an actionable knowledge representation system for thermodynamics. *Journal of Chemical Information and Modeling*, 64(15):5878–5887. PMID: 39042488.

Rui Wang, Xuemeng Hu, Deyu Zhou, Yulan He, Yuxuan Xiong, Chenchen Ye, and Haiyang Xu. 2020. Neural topic modeling with bidirectional adversarial training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 340–350, Online. Association for Computational Linguistics.

Xinyi Wang and Yi Yang. 2020. Neural topic model with attention for supervised learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1147–1156. PMLR.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

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.

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. *arXiv preprint arXiv:1909.00161*.

Xiang Zhang, Junbo Jake Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, December 7-12, 2015, Montreal, Quebec, Canada*, pages 649–657.

He Zhao, Lan Du, Wray Buntine, and Gang Liu. 2017. Metalda: A topic model that efficiently incorporates meta information. In *2017 IEEE International Conference on Data Mining (ICDM)*, pages 635–644. IEEE.

A Modeling Details

Let x be the BoW representation of a document. Let V and K be vocabulary size and no of topics respectively. The common encoder used across different models is:

Encoder:

$$\begin{aligned} X_1 &= \text{ReLU}(L^1 \cdot x + L^1_{\text{bias}}) \\ X_2 &= \text{dropout}(X_1, p_{\text{keep}}) \\ X_3 &= \text{BatchNorm}(L^2 \cdot X_2 + L^2_{\text{bias}}) \\ \text{enc\_out} &= \text{Softplus}(X_3), \end{aligned}$$

A common linear decoder with Dirichlet prior is given as:

Decoder:

$$\begin{aligned} X_4 &= \text{BatchNorm}(L^3 \cdot z + L^3_{\text{bias}}) \\ \text{recon} &= \text{LogSoftmax}(X_4) \end{aligned}$$

Certain models (such as ETM) use an embedding decoder. α and δ represent topic and word embeddings, respectively:

Decoder ETM:

$$\begin{aligned} X_4 &= z \cdot \eta \cdot \delta \\ \text{recon} &= \text{LogSoftmax}(\text{BatchNorm}(X_4)) \end{aligned}$$

For FANToM we have two decoders for text and authors of the document.

Decoder FANToM:

$$\begin{aligned} X_4 &= \text{BatchNorm}(L^3 \cdot z + L^3_{\text{bias}}) \\ X_5 &= \text{BatchNorm}(L^4 \cdot z + L^4_{bias}) \\ \text{recon\_doc} &= \text{LogSoftmax}(X_4) \\ \text{recon\_author} &= \text{LogSoftmax}(X_5) \end{aligned}$$

The hyperparameters used in the experiments are given in Table 3.

Every model, including its FANToM variant, is constructed as follows:

SCHOLAR: SCHOLAR utilizes the common Encoder with the linear Decoder. The softmax of the output of the Encoder, denoted as softmax(enc_out), is used as input to a separate classifier, which predicts document labels.

DVAE: DVAE employs the common Encoder with the linear Decoder.

ETM: ETM utilizes the common Encoder with an embedding decoder (Decoder ETM). The word embeddings are initialized with GloVe embeddings.

CTM: CTM employs the common Encoder with the linear Decoder. The distinction lies in the input to the encoder, which consists of contextualized embeddings rather than Bag of Words (BoW). The contextualized embeddings are sourced from the SBERT model.

ECRTM: ECRTM employs the common Encoder with the ETM Decoder. The distinction lies in the optimization function, where we use implementation provided by the authors[4].

All of the aforementioned models are also constructed with their corresponding FANToM variants. In these variants, only the priors are manipulated in the objective function (as mentioned in described in Section 4.2 of the main paper), and a separate author decoder is introduced to learn the topic-author multinomial distribution. The core architecture remains unchanged in all cases.

This also confirms the adaptability of our architecture to seamlessly incorporate any VAE-based topic model. Furthermore, it demonstrates the ability to align the learned topics with expert-assigned labels, and to leverage author information for establishing a meaningful correspondence between topics, authors, and labels.

B FANToM

B.1 Illustrative example to construct topic-label vector

Consider a document d in a corpus with five topics ($K = 5$). We define a topic-label linking vector $L_k = (\text{sport}, \text{cars}, \text{weather}, \text{no-label}, \text{no-label})$, where each element links a topic index to a corresponding label. If the label for document d is "sport", we define a label indicator vector $\mathbb{I}^{\text{d}} = (1, 0, 0, 0, 0)$, indicating that only the first topic is

[4]https://github.com/BobXWu/ECRTM

Table 3: Hyperparameter settings for the experiments.

Hyperparameter	Value
batch size	128
α	0.02
β	2
Learning Rate	0.001
Max Epochs	100
p_{keep}	0.25
L^1	$\mathbb{R}^{\text{vocab\_size} \times 512}$
L^2	$\mathbb{R}^{512 \times \text{total\_topics}}$
L^3	$\mathbb{R}^{\text{total\_topics} \times \text{vocab\_size}}$
L^4	$\mathbb{R}^{\text{total\_topics} \times \text{author\_size}}$
δ	$\mathbb{R}^{\text{total\_words} \times 300}$
η	$\mathbb{R}^{\text{total\_words} \times 300}$
word embeddings	(Pennington et al., 2014, GloVe)
train:val:test	70:15:15

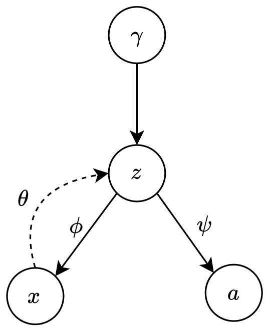

Figure 8: The graphical model of our framework is also shown, where solid lines represent the generative distribution p, and dotted lines represent the variational distribution q.

active. If no labels are provided, the indices corresponding to "no-label" are activated, resulting in $\mathbb{I}^{\mathrm{d}} = (0, 0, 0, 1, 1)$. For cases where multiple labels are assigned, such as both "sport" and "cars", the indicator vector becomes $\mathbb{I}^{\mathrm{d}} = (1, 1, 0, 0, 0)$.

Using this label indicator vector, the modified prior parameter is defined as $\gamma = \alpha \cdot \mathbb{I}^{\mathrm{d}}$, where α is a scaling factor. This allows for flexible topic assignments based on the available labels. For example, if L_k = (sport, sport, cars, cars, weather, weather, no-label, no-label) , multiple topics may be linked to a single label. This allows the model to learn multiple focused topics.

B.2 Training FANToM

The training algorithm for FANToM, outlined in Algorithm 1, begins by taking as input a document set D, an expert E, a prior parameter α, and topic-label assignments L_k. The process starts with the expert providing labels l for the documents, which are then used to create a multi-hot vector $\mathbb{I}^{\mathrm{d}}$ that reflects the association between labels and topics.

The algorithm processes batches $\mathcal{B}$ of documents from the set D. For each batch, the bag-of-words (BoW) representation x and an author representation a are extracted. An encoder network, parameterized by θ, computes the posterior parameters α_p from the BoW representation. Simultaneously, a prior distribution is constructed using the expert-provided labels, with parameters given by $\gamma = \alpha \cdot \mathbb{I}^{\mathrm{d}}$.

A variable z is then sampled from a Dirichlet distribution Dirichlet(α_p), representing the document-topic distribution. This sampled z is input into two decoder networks: one reconstructs the document as x', while the other reconstructs the author representation as a'.

The model parameters are updated by minimizing the objective function defined in Eq. 2. This iterative process of batch processing, parameter updating, and reconstruction continues until the algorithm converges, ensuring the model effectively learns both document and author representations aligned with the provided topic labels.

Throughout the training, the interaction between the posterior parameters derived from the encoder, the prior informed by expert labels, and the outputs from the decoders continually refines the model's understanding of topic, label, and author associations. This optimization process ensures that FAN-

Algorithm 1 Training FANToM

Input: Documents D, Expert E, Prior parameter α, Topic-label assignments L_k
$\lambda \leftarrow$ Get labels from expert E for documents D
$\mathbb{I}^{\mathrm{d}} \leftarrow$ Create multi-hot vector based on λ^{d} and L_k for $d \in D$
Initialize model parameters (θ, ϕ, ψ)
while not converged **do**
  for batch $\mathcal{B}$ in D **do**
    Extract BoW x, authors a, and multi-hot vector $\mathbb{I}^{\mathrm{d}}$ from $\mathcal{B}$
    $\gamma \leftarrow \alpha \cdot \mathbb{I}^{\mathrm{d}}$ (modify prior parameter)
    $\alpha_p \leftarrow$ Encoder$(\theta; x)$ (encode words)
    Sample $z \sim$ Dirichlet(α_p)
    $x' \leftarrow$ Decoder$(\phi; z)$ (reconstruct document)
    $a' \leftarrow$ Decoder$(\psi; z)$ (reconstruct author)
    Compute gradients $\nabla_{(\theta,\phi,\psi)} L_{\mathrm{F}}(\theta, \phi, \psi; x, \gamma)$ (Eq. 2)
    Update parameters (θ, ϕ, ψ)
  end for
end while

ToM captures the underlying topic structures in the document set D, aligning them with expert-provided labels and authors.

C Datasets and Preprocessing

For dataset preparation, we use SpaCy (Honnibal and Montani, 2017) for data tokenization. Additionally, we eliminate common stop words, punctuations, as well as high and low-frequency words. High and low-frequency words are determined by excluding words that appear in over 85% of the documents or in fewer than 30 documents. This standardization is applied across all models.

D Using LLMs as Expert

Not all datasets come with accompanying labels and labeling every corpus can be a challenging task. However, an alternative approach involves using Large Language Models (LLMs) (Ostheimer et al., 2023, 2024) to assign labels to unlabeled documents. In our study, we explore this possibility by using the capabilities of BART (Lewis et al., 2019) for zero-shot text classification. To achieve this, we use a pretrained model from (Wolf et al., 2019), which undergoes fine-tuning on Natural Language Inference (NLI) (Williams et al., 2018) following the methodology described by (Yin et al., 2019).

This method operates by treating the sequence to be classified as the NLI premise and constructing a hypothesis for each potential label. For example, when determining if a sequence pertains to the "sport" label, we can formulate a hypothesis such as "This text is about sports." By converting the probabilities for entailment and contradiction, we obtain the label probabilities.

We apply this approach to our selected dataset and supply prospective labels from the label set. By comparing the results obtained through this method with the available human labels, we achieve an accuracy of 92%. This validation further supports the practicality and effectiveness of using LLMs as experts for document categorization. Furthermore, we use the LLM labels of the ag news corpus (AGN) and use FANToM(L) to estimate topics. The results demonstrate a mean topic quality of 0.503, surpassing the 0.409 achieved through human labeling. Figure 5 provides an illustration of the topic words estimated by both human and LLM labels. Notably, both sets of topics align well with the corresponding labels.

E Runtime

In this section, we present the runtime of the proposed models compared to existing models. All experiments were conducted on NVIDIA V100 GPUs. The runtimes are presented in Table 6. The results demonstrate that the proposed methods have minimal impact on the overall runtime. It is important to note that only the model training time is considered in the runtime analysis, while data preparation time is excluded. The incorporation of labels and authors introduces a data preparation threshold, but it remains constant and does not scale with the model's runtime.

F Additional Evaluation

Apart from the main paper's evaluation, we provide supplementary assessments. Figure 10 showcases extended qualitative analysis by revealing all the topics acquired by Labeled LDA, existing NTMs, and our proposed FANToM on the 20ng dataset. This illustration highlights the concordance with labels and also showcases the esoteric topics uncovered by our method. It further underscores the suboptimal quality of topics obtained through statistical models.

F.1 Additional Quantitative Evaluation

Tables 9 and 10 delve into topic coherence and diversity in comparison to baselines, thus demonstrating the superior quality of topics learned by our model.

We also validate alignment quantitatively by conducting text classification using FANToM(L). We use the encoder to extract posterior probabilities for the test set during inference. We evaluate each model based on Top-3 and Top-5 accuracy and F1 scores. As demonstrated in Tables 7 and 8, FANToM(L) yields good classification and F1 scores which suggests alignment with the expert-assigned labels.

We also conduct benchmarking experiments in a semi-supervised scenario where not all topics in the corpus have labels. To manage this, we designate the "no-label" category for documents lacking labels. In our experiment, we randomly remove labels from 50% of the documents across all datasets and train models to learn 50 and 200 topics. This setup results in a mixture of labeled and "no-label" topics, reflecting real-world situations where not all documents may have known labels. Analyzing the results, shown in Table 4, FANToM consistently outperforms the baseline models, demonstrating its robustness even in the absence of complete label information. Specifically, the FANToM variants show higher topic coherence (TC) and topic quality (TQ) across datasets. Despite the removal of labels, FANToM maintains superior peformance and consistently achieves better overall scores, solidifying its advantage in semi-supervised settings.

F.2 Topic-Author Alignment

Our framework allows us to link topics with authors using two distinct approaches: FANToM, which uses labels, and FANToM(A), which does not. In Table 5, we present results that demonstrates that our method associates topics with authors and labels from the arxiv dataset. The authors associated with the topics belong to their respective field of research. Including labels allows us to validate this association by checking the respective topic authors within their labels. Further we also list results from FANToM(A) on stats.ML subset from the arxiv dataset in Figure 11.

G Discussion on Topic Models and LLMs

In recent years, combining topic modeling with large language models (LLMs) has emerged as a key area of research, aiming to harness the strengths of both techniques.

Traditionally, topic modeling has employed methods like Latent Dirichlet Allocation (LDA) and its variants to uncover hidden topics in a corpus based on word co-occurrence patterns. However, these methods often face challenges with interpretability and coherence.

The introduction of Variational Autoencoders (VAEs) brought a new level of flexibility to topic modeling by incorporating advancements in natural language processing (NLP), such as word embeddings. Later, with the advent of transformer models like BERT, contextualized embeddings were integrated into topic modeling approaches, such as in Contextualized Topic Models (CTM).

The rise of LLMs, known for their ability to generate coherent and contextually relevant text, has opened up new possibilities for enhancing topic modeling. One notable development is the use of prompt engineering, exemplified by recent approaches like TopicGPT (Pham et al., 2024). This method involves prompting an LLM to generate topics by processing each document individually. While this approach can produce high-quality topics, it incurs substantial computational costs due to the need to process prompts for every document, which can be particularly demanding with large corpora and complex LLMs. Additionally, prompting methods often lack a systematic framework for incorporating information such as authors and labels into the topic modeling process.

FANToM offers an alternative approach by integrating LLMs and topic modeling in a different way. It continues to use a VAE framework, known for its efficiency in topic modeling, while employing a smaller, task-specific LLM for labeling tasks. The VAE handles the core topic modeling process, effectively capturing the latent structure of the data and providing compact and interpretable topic representations. For labeling, FANToM utilizes a smaller, specialized LLM rather than a large-scale general-purpose model, which reduces computational demands while maintaining high accuracy. By aligning the outputs of the smaller LLM with the topics generated by the VAE, FANToM achieves a more coherent and adaptable solution that can be tailored to various datasets and applications.

Overall, FANToM opens up new possibilities for leveraging large-scale LLMs to enhance traditional topic modeling frameworks, leading to improved topic generation and alignment.

	Models	20NG			AGN			DB-14			arxiv		
		TC	TD	TQ	TC	TD	TQ	TC	TD	TQ	TC	TD	TQ
50 Topics													
baseline	SCHOLAR	0.410	0.57	0.234	0.372	0.44	0.164	0.629	0.68	0.428	0.364	0.46	0.168
	DVAE	0.354	0.59	0.209	0.393	0.46	0.181	0.658	0.66	0.434	0.360	0.48	0.173
	ETM	0.418	0.56	0.234	0.326	0.43	0.140	0.630	0.67	0.422	0.281	0.45	0.126
	CTM	0.412	0.58	0.239	0.371	0.44	0.163	0.632	0.69	0.436	0.348	0.47	0.164
	ECRTM	0.465	0.60	0.279	0.389	0.45	0.175	0.670	0.71	0.475	0.390	0.49	0.191
FANToM (ours)	SCHOLAR	0.415	0.62	0.257	0.397	0.48	0.191	0.608	0.72	0.438	0.375	0.50	0.188
	DVAE	0.424	0.61	0.258	0.409	0.50	0.205	0.611	0.71	0.434	0.336	0.51	0.171
	ETM	0.410	0.60	0.246	0.391	0.49	0.191	0.639	0.70	0.447	0.310	0.49	0.151
	CTM	0.436	0.64	0.279	0.395	0.47	0.186	0.622	0.69	0.429	0.342	0.52	0.178
	ECRTM	0.440	0.65	0.286	0.400	0.50	0.200	0.645	0.73	0.471	0.395	0.53	0.209
200 Topics													
baseline	SCHOLAR	0.362	0.47	0.170	0.354	0.38	0.134	0.581	0.57	0.331	0.342	0.41	0.140
	DVAE	0.376	0.49	0.184	0.374	0.39	0.146	0.602	0.58	0.349	0.335	0.43	0.144
	ETM	0.368	0.48	0.177	0.361	0.37	0.134	0.590	0.59	0.348	0.328	0.40	0.131
	CTM	0.384	0.50	0.192	0.368	0.40	0.147	0.595	0.60	0.357	0.340	0.42	0.143
	ECRTM	0.397	0.52	0.206	0.370	0.41	0.151	0.612	0.61	0.373	0.355	0.44	0.156
FANToM (ours)	SCHOLAR	0.380	0.54	0.205	0.362	0.40	0.145	0.615	0.62	0.381	0.355	0.46	0.163
	DVAE	0.387	0.53	0.205	0.381	0.42	0.160	0.617	0.61	0.376	0.360	0.48	0.173
	ETM	0.379	0.52	0.197	0.385	0.41	0.158	0.620	0.60	0.372	0.365	0.45	0.164
	CTM	0.399	0.55	0.220	0.380	0.39	0.148	0.624	0.63	0.393	0.369	0.49	0.181
	ECRTM	0.404	0.56	0.226	0.386	0.43	0.166	0.630	0.64	0.403	0.371	0.50	0.185

Table 4: Topic Coherence (TC), Topic Diversity (TD), and Topic Quality (TQ) for baseline and FANToM models across datasets for 50 and 200 topics in semi-supervised setting. In general, FANToM outperforms baseline models across both 50 and 200 topics.

Index/Label	Aspect	Topics/Authors
2	*topic*	graphs, chromatic, graph, bipartite, isomorphic
	authors	Král'_D, Seymour_P, Alon_N, Wood_DR, Koolen_JH
4	*topic*	exoplanet, orbiting, planet, kepler, comet
	authors	Ford Eric B., Agol E., Wright J.T., Pepe F., Henning T.
math.CO	*topic*	conjectured, automorphism, matroids, poset, extremal
	authors	Seymour_P, Král'_D, Rautenbach_D, Klavžar_S, Li_X
astro-ph.EP	*topic*	stellar, atmospheric, exoplanet, atmosphere, planet, orbiting
	authors	Henning T., Lagrange A. -M, Pepe F., Desidera S., Ford Eric B.

Table 5: Associations between Topics and Authors using FANToM(A) (top) and FANToM (bottom) approaches. FANToM builds meaningful correspondence between labels, topics, and authors.

Datasets	*ours (FANToM)*		*baselines*	
	ETM	DVAE	ETM	DVAE
20NG	119.29 ± 1.34	119.36 ± 1.46	115.84 ± 1.76	119.46 ± 1.36
AGN	157.78 ± 2.02	147.48 ± 2.47	148.05 ± 3.06	143.25 ± 3.54
DB-14	155.27 ± 4.03	150.84 ± 2.05	147.16 ± 2.99	144.16 ± 2.79
arxiv	620.48 ± 4.05	605.06 ± 5.88	602.99 ± 4.59	595.23 ± 6.82
arxiv*	601.23 ± 1.28	590.71 ± 8.53	-	-
arxiv**	656.46 ± 0.69	601.85 ± 9.88	-	-

Table 6: This table presents a comparison of the runtime for the proposed FANToM model against baseline models. The upper part of the table displays the runtime values for FANToM(L) when only labels are incorporated, while the bottom part shows the runtime values for FANToM(A) (*) and FANToM (**) respectively. All runtimes are measured in seconds. The results indicate that incorporating the proposed methods does not have a significant impact on the runtime of the models.

Datasets	FANToM(L)	
	%MacroF1	%MicroF1
20NG	0.523 ± 0.02	0.541 ± 0.01
AGN	0.659 ± 0.03	0.655 ± 0.03
DB-14	0.511 ± 0.00	0.532 ± 0.01
arxiv	0.707 ± 0.03	0.725 ± 0.04

Table 7: F1 scores of the FANToM(L) model validates the alignment of topics with labels.

Datasets	FANToM(L)	
	Top-3	Top-5
20NG	0.696 ± 0.02	0.798 ± 0.01
AGN	0.892 ± 0.02	1.000 ± 0.00
DB-14	0.726 ± 0.01	0.832 ± 0.02
arxiv	0.852 ± 0.02	0.918 ± 0.01

Table 8: Top-3 and Top-5 accuracy scores of the FANToM(L) model validates the alignment of topics with labels.

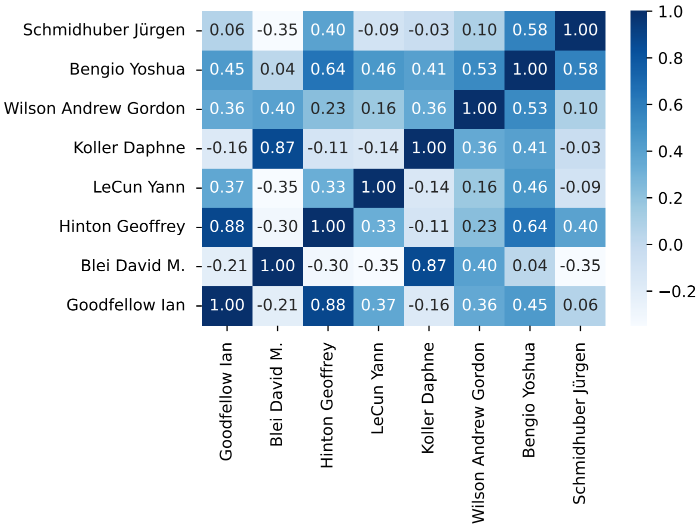

Figure 9: Cosine Similarity Matrix illustrating the relationships between prominent ML authors, determined by analyzing their associated topic vectors through the FANToM(A) model. High similarity scores indicate shared research areas among authors.

FANToM-C
(ours)

Topic alt.atheism: anecdote blew beaverton beauchaine evolution consequences khan argument keith exist
Topic comp.graphics: image removal cad pixel routines images edges sgi modules graphics
Topic comp.os.ms-windows.misc: windows mouse cica m download swap instruction characters microsoft p
Topic comp.sys.ibm.pc.hardware: scsi drives settings dma bus sec floppy port gateway card
Topic comp.sys.mac.hardware: apple mac monitor useragent nuntius dartmouth quadra simm lciii xxmessage
Topic comp.windows.x: window lib string null sunos xview specify manager client der
Topic misc.forsale: sale shipping forsale sell make condition baud brand purdue buyer
Topic rec.autos: car cars dealer vehicles toyota mileage driving concepts brakes bird
Topic rec.motorcycles: ama helmet riding dogs cage bmw ride jonathan bike max
Topic rec.sport.baseball: braves hr mets giants morris outs tigers fischer expos alomar
Topic rec.sport.hockey: playoff coach goals edmonton ice leafs patrick gerald pens devils
Topic sci.crypt: chip keys dorothy encryption tap transmitted cellular authorized security xor
Topic sci.electronics: volt cooling circuit watts voltage amplifier ic amp rf components
Topic sci.med: surrender skepticism literature food skin severe causing banks treatments russell
Topic sci.space: zoology landing solar prize nasa elements jet exploration baalke launch
Topic soc.religion.christian: rome languages blessed easter catholic condemned fallen christianity judged jesus
Topic talk.politics.guns: rifle shooting fired firearms kratz safety weapons elevator guns accidents
Topic talk.politics.mideast: proceeded territory kurds extermination suffered davidsson lebanese argic massacres cambridge
Topic talk.politics.misc: income mutual republicans clinton democracy economy debt taxes commitment spending
Topic talk.religion.misc: existent writings describes weiss harvey sword interpret hanging jesus followers

LabeledLDA

Topic alt.atheism: god people writes article think posting atheists like know atheism
Topic comp.graphics: image graphics university posting file host nntp files use software
Topic comp.os.ms-windows.misc: windows m w file university use t dos files writes
Topic comp.sys.ibm.pc.hardware: drive scsi card ide controller bus disk university drives use
Topic comp.sys.mac.hardware: mac apple posting university nntp host writes drive article know
Topic comp.windows.x: x window file use server program motif widget host posting
Topic misc.forsale: sale new university posting host nntp distribution dos offer mail
Topic rec.autos: car article writes cars like posting nntp host university good
Topic rec.motorcycles: bike article writes dod like posting host nntp know ride
Topic rec.sport.baseball: year writes article team game baseball good university posting think
Topic rec.sport.hockey: team hockey game play university writes nhl season article year
Topic sci.crypt: key encryption clipper chip government use people writes keys security
Topic sci.electronics: use writes host article posting like university nntp power know
Topic sci.med: article writes people msg university science like know gordon use
Topic sci.space: space writes nasa article launch posting like moon orbit earth
Topic soc.religion.christian: god people jesus think church know believe christians writes like
Topic talk.politics.guns: gun people writes article guns like think right firearms posting
Topic talk.politics.mideast: people israel turkish armenian israeli armenians jews said writes article
Topic talk.politics.misc: people writes article q think president government know new mr
Topic talk.religion.misc: god people jesus writes article christian know think bible good

Neural Topic Models

Topic 1: sale printer shipping forsale manuals speakers obo dos covington athens
Topic 2: batf dividian people apartment koresh roby fbi atf children survivors
Topic 3: god belief moral atheist morality bible evidence atheists objective faith
Topic 4: players baseball team game season braves league player sox pitching
Topic 5: nntp host freenet thanks posting article cleveland advance writes sale
Topic 6: space launch orbit shuttle spacecraft jet nasa solar mission satellite
Topic 7: turkish armenian armenians serdar turks armenia argic sahak turkey appressian
Topic 8: q president health mr clinton jobs administration ms people package
Topic 9: encryption key clipper escrow keys chip enforcement crypto privacy secure
Topic 10: israeli israel lebanese arab jews islamic peace islam homosexual arabs
Topic 11: msg shameful banks skepticism food surrender chastity gordon intellect superstition
Topic 12: hockey nhl team season playoff devils leafs cup game teams
Topic 13: scsi ide drive mb card bus ram bios motherboard drives
Topic 14: jesus god church bible christ sin christians lord heaven christian
Topic 15: m x w t v c p dos g s
Topic 16: x bike helmet entry riding zoology key spencer dog biker
Topic 17: car honda cars bike oil tires mph rear engine radar
Topic 18: window image widget motif server xt display colormap font images
Topic 19: mouse water cooling voltage amp drive ground wire wiring circuit
Topic 20: gun guns firearms weapons amendment people militia arms crime laws

Figure 10: This figure compares FANToM(L) (top), Labeled LDA (middle), and Neural Topic Models (bottom) on the 20NG dataset. FANToM(L) exhibits aligned and esoteric topics consistent with human intentions, while Labeled LDA shows repetitive generic words in many topics (e.g., "host," "writes," "articles"), leading to low diversity. Neural Topic Models do not have association with provided labels and fail to produce certain topics completely, unlike FANToM(L) (e.g., "mac", "graphics", "electronics").

Topic 1: mcmc monte carlo inference vi likelihood posterior variational poisson latent
Top Authors: Hensman James ,Ruiz Francisco J. R. ,Broderick Tamara ,Blei David M. ,Domke Justin

Topic 2: lasso causal clustering union penalized covariates demonstrate idea coverage gene
Top Authors: Chamroukhi Faicel ,McNicholas Paul D. ,Kakde Deovrat ,Allen Genevera I. ,Shpitser Ilya

Topic 3: reinforcement rl reward policy meta planning imitation reasoning skills environments
Top Authors: Heess Nicolas ,Levine Sergey ,Abbeel Pieter ,Pathak Deepak ,Peng Jian

Topic 4: rank recovery entries matrix minimax nuclear recovering estimator semidefinite matrices
Top Authors: Zhou Shuheng ,Caramanis Constantine ,Singh Shashank ,Sanghavi Sujay ,Zhang Anru

Topic 5: regret bandit horizon ucb reward policy mdps bandits arms arm
Top Authors: Langford John ,Perchet Vianney ,Wen Zheng ,Pirotta Matteo ,Agarwal Alekh

Topic 6: sgd proximal nonsmooth nonconvex gd nesterov momentum decentralized newton convexity
Top Authors: Richtárik Peter ,Ozdaglar Asuman ,Scheinberg Katya ,Sun Tao ,Lin Qihang

Topic 7: gans attack defense attacks perturbations adversarial robustness gan image training
Top Authors: Goodfellow Ian ,Schiele Bernt ,Zhang Han ,Carlini Nicholas ,Zhang Huan

Topic 8: ehr speech lstm patient clinical temporal medical cnn sequences music
Top Authors: Hojo Nobukatsu ,Xiao Cao ,Cho Kyunghyun ,Sarrafzadeh Majid ,Sarker Iqbal H.

Topic 9: graph node gnn link embedding laplacian collaborative nodes gnns reasoning
Top Authors: Benson Austin R. ,Yang Jian ,Hamilton William L. ,Zhang Chengqi ,Tang Jiliang

Topic 10: width pruning depth quantization neurons hardware relu layer neural compression
Top Authors: LeCun Yann ,Huan Jun ,Hoffer Elad ,Ganguli Surya ,Sohl-Dickstein Jascha

Figure 11: Cosine Similarity Matrix illustrating the relationships between prominent ML authors, determined by analyzing their associated topic vectors through the FANToM(A) model. High similarity scores indicate shared research areas among authors, revealing clusters of researchers with overlapping contributions and thematic focus in machine learning. It can provide insights into academic collaborations, interdisciplinary influences, and the evolution of research trends within the field.

	Models	20NG	AGN	DB-14	arxiv
baseline	L-LDA	0.198 ± 0.00	0.336 ± 0.00	0.551 ± 0.00	0.269 ± 0.02
	SCHOLAR	0.410 ± 0.01	0.372 ± 0.06	$0.629 \pm 0.02\bullet$	0.364 ± 0.03
	DVAE	0.354 ± 0.02	0.393 ± 0.07	$\mathbf{0.658 \pm 0.01}\bullet$	$0.360 \pm 0.03\bullet$
	ETM	$0.418 \pm 0.03\bullet$	0.326 ± 0.02	0.630 ± 0.01	0.281 ± 0.01
	CTM	0.412 ± 0.01	0.371 ± 0.01	$0.632 \pm 0.01\bullet$	$0.348 \pm 0.02\bullet$
ours	FANToM-SCHOLAR	$0.415 \pm 0.02\bullet$	$0.397 \pm 0.04\bullet$	0.608 ± 0.02	$\mathbf{0.375 \pm 0.01}\bullet$
	FANToM-DVAE	$0.424 \pm 0.03\bullet$	$\mathbf{0.409 \pm 0.05}\bullet$	0.611 ± 0.01	0.336 ± 0.02
	FANToM-ETM	0.410 ± 0.01	$0.391 \pm 0.02\bullet$	$0.639 \pm 0.02\bullet$	$0.310 \pm 0.02\bullet$
	FANToM-CTM	$\mathbf{0.436 \pm 0.01}\bullet$	$0.395 \pm 0.02\bullet$	0.622 ± 0.02	0.342 ± 0.01

Table 9: Comparison of the proposed FANToM(L) (bottom) against the baselines (top) using Topic Coherence metrics. The best results across datasets are highlighted in bold, and best results across corresponding models between baselines and FANToM(L) are marked with •. The proposed model generally outperforms the existing baselines.

	Models	20NG	AGN	DB-14	arxiv
baseline	L-LDA	0.97 ± 0.00	0.95 ± 0.00	0.93 ± 0.00	0.91 ± 0.01
	SCHOLAR	0.95 ± 0.01	1.00 ± 0.00	0.95 ± 0.01	0.94 ± 0.01
	DVAE	1.00 ± 0.00	1.00 ± 0.00	0.95 ± 0.01	0.94 ± 0.01
	ETM	0.95 ± 0.01	0.95 ± 0.00	0.96 ± 0.01	0.93 ± 0.02
	CTM	1.00 ± 0.00	1.00 ± 0.00	0.94 ± 0.01	0.93 ± 0.02
ours	FANToM-SCHOLAR	0.97 ± 0.01	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	FANToM-DVAE	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00
	FANToM-ETM	0.97 ± 0.01	1.00 ± 0.00	0.99 ± 0.00	1.00 ± 0.00
	FANToM-CTM	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00	1.00 ± 0.00

Table 10: Comparison of the proposed FANToM(L) (bottom) against the baselines (top) using Topic Diversity metrics. The proposed model produces topics with higher diversity.