CobwebTM: Probabilistic Concept Formation for Lifelong and Hierarchical Topic Modeling

Karthik Singaravadivelan*, **Anant Gupta***, **Zekun Wang, Christopher J. MacLellan**
College of Computing
Georgia Institute of Technology
Atlanta, GA 30332 USA
{ksingara3, agupta886, zwang910}@gatech.edu

Abstract

Topic modeling seeks to uncover latent semantic structure in text corpora with minimal supervision. Neural approaches achieve strong performance but require extensive tuning and struggle with lifelong learning due to catastrophic forgetting and fixed capacity, while classical probabilistic models lack flexibility and adaptability to streaming data. We introduce CobwebTM, a low-parameter lifelong hierarchical topic model based on incremental probabilistic concept formation. By adapting the Cobweb algorithm to continuous document embeddings, CobwebTM constructs semantic hierarchies online, enabling unsupervised topic discovery, dynamic topic creation, and hierarchical organization without predefining the number of topics. Across diverse datasets, CobwebTM achieves strong topic coherence, stable topics over time, and high-quality hierarchies, demonstrating that incremental symbolic concept formation combined with pretrained representations is an efficient approach to topic modeling.

1 Introduction

Topic modeling seeks to uncover latent semantic structure in large document collections by grouping text into coherent topics. It is a fundamental tool for document organization, corpus exploration, and information retrieval, particularly in settings where labeled data is unavailable. As modern text corpora grow in scale, diversity, and temporal span, effective topic modeling increasingly requires methods that support unsupervised topic discovery, adapt to streaming data, and represent topics at multiple levels of abstraction.

Early work in topic modeling was dominated by probabilistic generative models, most notably Latent Dirichlet Allocation (LDA) (Blei et al., 2003b). While influential, LDA requires the number of topics to be specified in advance, assumes independence between topics, and relies on bag-of-words representations that ignore semantic similarity between words. These assumptions limit its ability to model imbalanced, correlated, or evolving topics, making it poorly suited for lifelong or streaming settings.

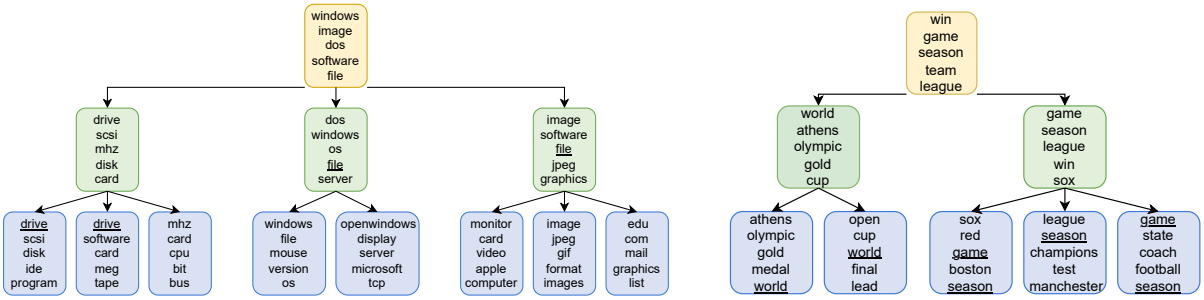
Recent advances in representation learning have led to neural topic models that leverage dense document embeddings (Zheng et al., 2013; Wu et al., 2024a). These approaches often achieve improved topic coherence and richer semantic representations, but at the cost of increased complexity. Neural topic models are typically highly parameterized, sensitive to hyperparameter choices, and trained in batch settings that assume access to the full corpus. Consequently, they struggle in lifelong learning scenarios where data arrives incrementally and topic structure must evolve over time. Moreover, neural architectures are prone to catastrophic forgetting, causing previously learned topics to degrade as new data is introduced.

Lifelong topic modeling addresses these challenges by updating topics incrementally as new documents arrive. Methods such as Online LDA (Hoffman et al., 2010) and neural lifelong topic models mitigate some scalability issues but retain key limitations, including fixed topic capacity, limited topic restructuring, and reliance on corpus-specific training. More recent embedding-based pipelines replace static clustering with incremental clustering algorithms, yet these methods remain sensitive to parameter choices and typically lack principled mechanisms for organizing topics at multiple levels of abstraction.

In practice, however, topic structure is inherently hierarchical: broad themes naturally decompose into progressively finer subtopics. Capturing such hierarchical organization improves interpretability and allows models to represent semantic relation-

*Equal Contribution
Code available at https://github.com/Teachable-AI-Lab/cobweb-language-embedding

Findings of the Association for Computational Linguistics: ACL 2026, pages 35140–35155
July 2-7, 2026 ©2026 Association for Computational Linguistics



(a) Subtree from 20 Newsgroups dataset

(b) Subtree from AG news dataset

Figure 1: A visualization of three levels of the hierarchy induced by CobwebTM. For each node, we display the top five representative words extracting using the c-tf-idf procedure described in Section 3.2.1. Words that appear in multiple nodes at the same level are underlined to highlight shared semantic content across sibling topics.

ships between topics rather than treating them as independent clusters. Consequently, hierarchical topic models have been widely explored in both probabilistic and neural frameworks (Blei et al., 2003a; Koltcov et al., 2021). These approaches aim to learn topic trees that capture varying levels of abstraction within a corpus.

Despite their promise, many hierarchical topic models rely on fixed-depth latent structures or require batch training over the full corpus, limiting their applicability in dynamic or streaming environments. In many modern systems, hierarchy is therefore imposed post hoc after flat topic discovery, rather than learned incrementally as the data evolves. This disconnect between lifelong learning and hierarchical structure motivates the need for topic modeling approaches that can simultaneously support incremental updates and flexible hierarchical organization.

In this work, we revisit incremental concept formation as an alternative paradigm for topic modeling. We introduce CobwebTM, a lifelong hierarchical topic modeling framework based on the Cobweb algorithm (Fisher, 1987) for probabilistic concept formation. By adapting Cobweb to operate over continuous document embeddings, CobwebTM incrementally constructs a semantic hierarchy as documents arrive, enabling unsupervised topic discovery without predefining the number of topics.

Our contributions are threefold: (1) we introduce CobwebTM, an incremental hierarchical topic modeling framework for unsupervised topic discovery over streaming text; (2) we show that probabilistic concept formation in embedding space provides a simple yet effective mechanism for lifelong topic modeling without catastrophic forgetting or fixed topic capacity; and (3) through extensive empirical evaluation, we demonstrate that CobwebTM matches or outperforms recent neural and clustering-based methods in both topic quality and hierarchical structure.

2 Related Works

2.1 Lifelong Topic Modeling

Online LDA (Hoffman et al., 2010) is the most widely used lifelong topic model, updating global topics via mini-batch variational inference. However, it inherits LDA's bag-of-words assumption, requires a predefined number of topics, and lacks mechanisms for restructuring topics as new data arrives.

Most neural topic models are trained in batch settings and struggle with sequential updates without retraining (Wu et al., 2024b). They are also prone to catastrophic forgetting (Luo et al., 2025). Mitigation techniques such as replay or elastic weight consolidation (Gupta et al., 2020) reduce forgetting but still rely on fixed latent dimensions.

Embedding-based pipelines instead perform topic discovery through clustering over neural representations. BERTopic (Grootendorst, 2022), for example, combines transformer embeddings and clustering. Lifelong variants replace static clustering with incremental methods such as DBStream (Bär et al., 2014) or MiniBatch KMeans (Sculley, 2010), though these approaches typically assume flat clustering and remain sensitive to parameter choices. Recent approaches such as TopicGPT (Pham et al., 2024) and FASTopic (Wu et al., 2024c) improve topic quality through LLM-based generation or embedding-level semantic modeling, but are either computationally expensive at scale or do not support hierarchical

and incremental topic discovery.

2.2 Hierarchical Topic Modeling

Hierarchical topic models organize topics across levels of abstraction. Early Bayesian approaches such as hLDA (Blei et al., 2003a) and related models (Mimno et al., 2007; Perotte et al., 2011) learn topic trees through generative processes. More recent methods construct hierarchies over embedding-based topic representations. Examples include CluHTM (Viegas et al., 2020), HyHTM (Shahid et al., 2023), and hierarchical variants of BERTopic (Grootendorst, 2022), which typically derive hierarchies through clustering or linkage procedures applied after flat topic discovery. Neural hierarchical topic models further learn structured latent representations using VAEs (Kingma and Welling, 2013), including tree-based (Isonuma et al., 2020), fixed-depth (Duan et al., 2021), and geometrically regularized models (Wu et al., 2024d; Lu et al., 2024). However, these models are generally trained in batch settings and impose structural constraints that limit their flexibility in lifelong or streaming scenarios.

2.3 Incremental Concept Formation

Humans organize knowledge hierarchically using prototypes and graded category membership (Rosch and Mervis, 1975). Incremental clustering methods formalize this process by building taxonomies whose internal nodes summarize concept-level statistics.

Cobweb (Fisher, 1987) incrementally constructs a probabilistic taxonomy through conceptual clustering, dynamically creating and restructuring nodes to maximize category utility. Recent work has extended Cobweb to neural settings and demonstrated robustness in vision and language tasks (MacLellan et al., 2022; MacLellan and Thakur, 2021; Wang et al., 2025; Barari et al., 2024b,a; Lian et al., 2025).

Unlike probabilistic topic models such as LDA, which directly learn $P(\text{word}|\text{topic})$ and $P(\text{topic}|\text{document})$ through Dirichlet priors, our approach derives these quantities through clustering in embedding space. Continuous Cobweb incrementally partitions transformer document embeddings into a hierarchical mixture of clusters, estimating document–topic associations via category utility. Topic–word distributions are computed post hoc using class-based TF–IDF over the documents assigned to each node.

3 Methodology

We propose CobwebTM, a topic modeling framework that incrementally organizes document embeddings into a dynamic semantic hierarchy. Unlike batch clustering methods such as k-Means or HDBSCAN, CobwebTM supports continual updates without retraining through a two-step neuro-symbolic process.

First, we perform document–topic inference directly in the latent space of pretrained transformer embeddings. Assuming the embedding space reflects an underlying mixture of topics, we apply the continuous Cobweb algorithm to incrementally partition the space, assigning each document to a node that maximizes category utility. This procedure produces a hierarchical clustering that implicitly defines the document–topic distribution.

Second, we derive topic–word representations from the resulting hierarchy. Each node represents a topic defined by the documents in its subtree. Treating nodes as classes, we compute word–topic distributions using c-TF-IDF, producing interpretable topic descriptors from the highest-ranked words.

3.1 Probabilistic Concept Formation

At the core of our approach is a variant of Cobweb adapted for continuous-valued attributes (Barari et al., 2024b). Each concept node c maintains a D-dimensional multivariate Gaussian with diagonal covariance,

$$p(x \mid c) = \mathcal{N}\big(x; , \mu_c, , \mathrm{diag}(\sigma_c^2)\big),$$

where $\mu_c \in \mathbb{R}^D$ is the node mean and $\sigma_c^2 \in \mathbb{R}^D$ is the variance vector. These statistics are updated incrementally as new documents are incorporated.

Cobweb constructs a hierarchy of concepts online. Given a new document embedding x, the algorithm performs a top-down search over the tree guided by *Category Utility* (CU) (Gluck and Corter, 1985; Corter and Gluck, 1992). Following Barari et al. (2024b), we adopt an information-theoretic formulation that measures the expected reduction in feature uncertainty obtained by knowing the child concept.

Let a parent node c_p have children $\mathcal{C}(c_p)$, each with count N_c. The empirical probability of concept c under the parent is

$$P(c \mid c_p) = \frac{N_c}{\sum_{c' \in \mathcal{C}(c_p)} N_{c'}} = \frac{N_c}{N_{c_p}}. \qquad (1)$$

We measure node uncertainty using the differential entropy of the Gaussian:

$$U(c) = \frac{1}{2} \sum_{d=1}^{D} \log\left(2\pi e, \sigma_{c,d}^2\right). \quad (2)$$

The category utility of a parent node is then

$$\mathrm{CU}(c_p) = \sum_{c \in \mathcal{C}(c_p)} P(c \mid c_p) \left[U(c_p) - U(c)\right]. \quad (3)$$

Maximizing CU favors partitions that reduce feature uncertainty while maintaining sufficient support, balancing intra-cluster similarity and inter-cluster separation. For continuous attributes, this corresponds to maximizing variance reduction induced by the partition, allowing Cobweb to determine the depth and breadth of the hierarchy without specifying the number of topics K.

At each node, Cobweb evaluates four operators to determine how x should be incorporated into the hierarchy: (1) insert x into the best-matching existing child and update its Gaussian parameters; (2) create a new singleton child node for x; (3) merge the two best-matching children and assign x to the merged node; or (4) split the best-matching child, promoting its children to the current level.

3.2 Topic Extraction

3.2.1 Hierarchical Topic Extraction

The Cobweb tree forms a multi-level topic hierarchy: the root represents the entire corpus, intermediate nodes correspond to increasingly specific topics, and leaves represent individual documents or fine-grained micro-topics.

To extract interpretable topics, we treat each node as a candidate topic and compute class-based TF–IDF (c-TF-IDF) scores following Grootendorst (2022). For a node C, all documents in its subtree are concatenated into a single document. The importance of word w in topic C is then computed as

$$W_{C,w} = tf_{C,w} \times \log\left(1 + \frac{A}{f_w}\right), \quad (4)$$

where $tf_{C,w}$ is the frequency of word w within topic C, f_w is the frequency of w across all topics, and A is the average number of words per topic. The highest-scoring words serve as descriptive keywords for the topic.

3.2.2 Dynamic Flat Topic Extraction

While the hierarchical structure provides rich semantic organization, many incremental topic modeling baselines operate on a flat set of topics. To enable direct comparison and support applications requiring fixed topic sets, we extract a dynamic flat partition from the hierarchy.

Because Cobweb is incremental, documents may appear at varying depths depending on their semantic specificity. To obtain a coherent flat topic set, we identify a cut through the tree that balances topic coverage and granularity. Specifically, we traverse the hierarchy top-down and select nodes such that (1) the number of selected nodes does not exceed a user-defined `max_clusters`, and (2) the ratio of leaf nodes to total nodes in the layer does not exceed a `leaf_ratio` threshold.

This procedure filters out shallow outliers near the root while grouping deeper semantically similar documents into stable clusters. As new documents arrive and the hierarchy evolves, the cut is recomputed, allowing the flat topic representation to adapt over time. Consequently, CobwebTM supports both a flat topic model for benchmarking and a full hierarchical representation that enables exploration across levels of abstraction.

4 Lifelong Topic Modeling

We first evaluate the performance of CobwebTM in a lifelong topic modeling setting. The primary objective is to assess the model's ability to maintain coherent topics, ensure stability across time steps, and adapt to new data without catastrophic forgetting or the need for extensive retraining.

4.1 Experimental Setup

Datasets. We utilize three datasets suited for temporal analysis: the **Spatiotemporal News Dataset** (Jomaa, 2025), the **Stack Overflow Dataset** (Movshovitz-Attias et al., 2013), and the **TweetNER7 Dataset** (Ushio et al., 2022). These datasets represent varying document lengths and stream velocities. The Spatiotemporal News and TweetNER7 datasets are temporally ordered to reflect real-world streams with intermittent topics, while the Stack Overflow dataset is randomly shuffled to simulate an approximately uniform incremental topic distribution. Detailed preprocessing steps and dataset statistics are provided in Appendix B.1.

Baselines. We compare CobwebTM against a range of incremental and static baselines. First, we evaluate **Online LDA** (Hoffman et al., 2010), an incremental variational Bayes variant of Latent Dirichlet Allocation. Second, we include **Lifelong NTM** (Gupta et al., 2020), a neural topic model extending DocNADE with Elastic Weight Consolidation and experience replay. Third, we employ **BERTopic (Incremental)** pipelines using incremental clustering algorithms, specifically **DBSTREAM** and **MiniBatchKMeans**. Finally, to benchmark against non-incremental upper bounds, we evaluate **BERTopic (Re-fit)** pipelines that retrain **HDBSCAN** and **KMeans** from scratch on the accumulated corpus at each time step.

Implementation Details. We use the embedding model RoBERTa Large (dimension size of 1,024) (Liu et al., 2019) for all pipelines which require it to ensure consistent feature spaces. We vary the initial batch size (default 2,000 documents, sensitivity analysis at 500). Unlike neural baselines, CobwebTM does not require large batch sizes for stability, so we fix the successive batch sizes at a middle ground of 125 documents. Topics between consecutive batches are matched using a greedy alignment strategy based on cosine similarity of topic embeddings.

Evaluation Metrics. We evaluate topic quality using the **Topic Coherence (C_v)** measure (Röder et al., 2015), an indirect confirmation metric that combines an NPMI-style word co-occurrence statistic with a context-based similarity score computed over sliding windows.

$$C_v = \frac{2}{N(N-1)} \sum_{i<j} \cos(\vec{v}(w_i), \vec{v}(w_j))$$

where $\vec{v}(w_i) = \{\text{NPMI}(w_i, w_k)\}_{k=1}^{N}$.

Unlike purely count-based coherence measures, C_v incorporates distributional information from the reference corpus, yielding scores that are comparable across batches. Coherence is computed over all data observed up to the current batch. We measure topic stability across consecutive batches using the **Adjusted Rand Index (ARI)** (Greene et al., 2014), computed between document–topic assignments from the previous and current batches. Higher ARI values indicate that topic assignments remain consistent over time, reflecting stable topic structure under incremental updates. To quantify semantic drift of topics across batches, we compute **Topic Centroid Drift (TCD)** for matched topics. For each topic, we represent its semantic centroid using topic embeddings and define drift as one minus the cosine similarity between the current and previous centroids. The reported score is the mean drift across all matched topics. Values close to zero indicate minimal semantic change and stable topics. Lastly, to ensure that we achieve topics that are interpretable, we calculate **Intruder Similarity Score (ISIM)**. For each topic, we insert a random word and calculate the average cosine similarity between a set of topic words and the intruder word, with lower averages corresponding to tighter and more unique topics.

We average results across three trials for each dataset on each configuration we report for CobwebTM. We set `leaf_ratio` $= 0.15$ to adapt natural outlier pruning, and `max_clusters` $= 1.3 \cdot K$, where K is the recommended number of clusters for each dataset as provided by the original paper.

4.2 Results

As shown in Figures 2, 3, 4, and 5, and in Tables 1, 2, and 3 CobwebTM outperforms or performs competitively with baselines in C_v, ARI, TCD, and ISIM across all three datasets. We analyze specific results below.

Comparison to BERTopic. Our comparisons to BERTopic pipelines boil down to the difference of the clustering algorithms. We find that DBSTREAM is extremely volatile on semantically dense datasets, resulting in poor coherence throughout batches. While BERTopic with MiniBatchKMeans initially outperforms CobwebTM in the TweetNER benchmark, CobwebTM shows more growth across all three datasets and eventually surpasses all methods in final performance, shown in Tables 1, 2, and 3. Additionally, Cobweb has no learning parameters, with the only two parameters being user-specified for granularity decisions in flat topic modeling.

Comparison to Neural Methods. The neural baseline exhibits degraded performance on the datasets, likely due to memory constraints and the datasets' large technical vocabulary. Even after pruning to the most frequent words, Lifelong NTM struggled to disambiguate topics beyond the initial batch as vocabulary diversity increased, as shown in Figure 2 and Figure 4. This behavior reflects a broader limitation of neural topic models, which must learn word semantics from the corpus itself

Method	C_{vf}	$\Delta C_v\%$	ARI	TCD	ISIM$_f$
CobwebTM (ours)	**0.741**	**110.82**	0.915	**0.000**	**0.191**
DBSTREAM	0.602	<u>71.55</u>	0.594	0.202	0.200
LifelongDocNADE	0.281	−11.61	0.479	0.369	<u>0.194</u>
MiniBatchKMeans	0.413	−0.89	**0.952**	0.236	0.215
OnlineLDA	0.317	7.06	<u>0.948</u>	<u>0.033</u>	0.216
Refit-HDBSCAN	<u>0.673</u>	7.35	0.944	0.053	0.196
Refit-KMeans	0.369	−6.49	0.549	0.120	0.224

Table 1: Lifelong topic modeling results on **TweetNER**. Best in bold, second-best underlined.

Method	C_{vf}	$\Delta C_v\%$	ARI	TCD	ISIM$_f$
CobwebTM (ours)	**0.613**	**26.09**	**0.997**	**0.000**	0.207
DBSTREAM	0.457	−3.67	0.012	0.155	0.188
LifelongDocNADE	0.184	−75.07	0.122	0.684	0.206
MiniBatchKMeans	0.423	−9.67	<u>0.955</u>	0.245	<u>0.176</u>
OnlineLDA	0.520	10.71	0.897	<u>0.029</u>	0.200
Refit-HDBSCAN	0.425	1.81	0.735	0.050	**0.157**
Refit-KMeans	<u>0.551</u>	<u>15.98</u>	0.416	0.205	0.224

Table 2: Lifelong topic modeling results on **Stack Overflow** Dataset. Best in bold, second-best underlined.

rather than leveraging pretrained embeddings. Our approach, which combines pretrained encoder representations with symbolic learning, enables more stable incremental performance and reduced susceptibility to catastrophic forgetting.

Metrics Over Time. CobwebTM shows strong increases in performance by coherence unlike other methods, as shown in Tables 1, 2, and 3 and in Figures 3 and 4, indicating that incremental aggregation of documents does not hinder topic construction and maintenance. Additionally, CobwebTM has strong TCD and ARI between batches across all datasets, highlighting a quality of clustering stability and the ability to create new topics when needed without harming the sanctity of topics from previous batches.

Strong Temporal Clustering Stability. The Adjusted Rand Index (ARI) (Greene et al., 2014) measures the consistency of topic assignments across batches, reflecting a model's resistance to topic drift and catastrophic forgetting. High ARI indicates that learned topics remain stable as new data arrive, while low ARI signals uncontrolled reorganization. As shown in Tables 1, 2, and 3, CobwebTM achieves consistently high ARI across datasets, with near-perfect scores on Stack Overflow (0.997) and Spatiotemporal News (0.984), demonstrating strong stability under incremental updates. Although MiniBatchKMeans scores slightly higher on TweetNER, its stability likely stems from fixed cluster constraints rather than adaptive topic evolution.

Robustness to Human-Centered Metrics. Our results, shown in Figure 5, are shown for 5 topic-words, averaged across 15 intruder words. CobwebTM achieves competitive scores to other models, though ranges of 0.02 to 0.06 across datasets suggests the metric is largely encoder-dependent at this level of comparison. We hypothesize that the tight margins are as a result of the semantically dense content in the datasets we leverage for testing incremental topic modeling, and these preliminary studies support that CobwebTM provides human-interpretable topics competitive with other topic models.

5 Hierarchical Topic Modeling

In the second set of experiments, we evaluate the quality of the hierarchical structures generated by CobwebTM. We benchmark against state-of-the-art hierarchical topic models to verify that our incremental construction yields meaningful taxonomies.

5.1 Experimental Setup

Datasets. We use three standard benchmarks: **20 Newsgroups** (Lang, 1995), **AG News** (Zhang et al., 2015), and **Stack Overflow** (Movshovitz-Attias et al., 2013).

Method	C_{vf}	$\Delta C_v\%$	ARI	TCD	ISIM$_f$
CobwebTM (ours)	**0.796**	**57.51**	**0.984**	**0.000**	0.208
DBSTREAM	0.418	−10.29	0.408	0.193	**0.191**
LifelongDocNADE	0.238	−49.49	0.127	0.462	0.196
MiniBatchKMeans	0.646	<u>40.37</u>	<u>0.962</u>	0.203	0.196
OnlineLDA	0.422	0.99	0.882	<u>0.053</u>	0.206
Refit-HDBSCAN	<u>0.657</u>	19.11	0.891	0.067	<u>0.193</u>
Refit-KMeans	0.614	29.11	0.326	0.229	0.221

Table 3: Lifelong topic modeling results on **Spatiotemporal News** Dataset. Best in bold, second-best underlined.

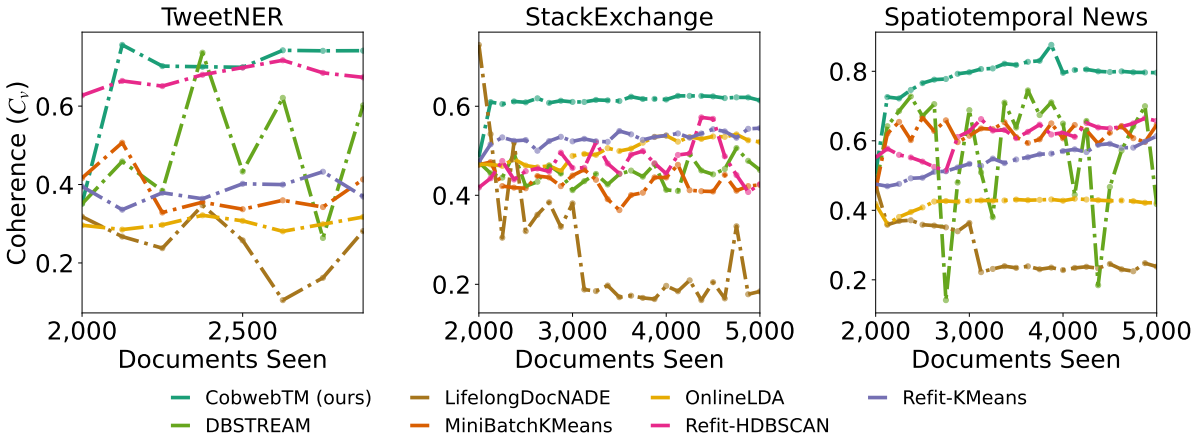


Figure 2: C_v comparison on the StackOverflow, Spatiotemporal News, and TweetNER dataset.

Baselines. We compare against three primary baselines: **TraCo** (Wu et al., 2024d), a neural model using Optimal Transport for topic regularization; **BoxTM** (Lu et al., 2024), a geometric approach modeling topics as hyper-rectangles, and **BERTopic (Hierarchical)** (Grootendorst, 2022), which uses agglomerative clustering on top of flat topics derived from HDBSCAN and KMeans.

Implementation Details. All embedding-based models use the same architectures and embedding dimensions as in Section 4.1. For the remaining models, we use the hyperparameters specified in their official implementations. Datasets for **BoxTM** and **TraCo** are preprocessed as described in Appendix B.2.

Evaluation Metrics. We assess the hierarchy using three metrics. **Topic Coherence (NPMI)** measures the semantic interpretability of individual topics using Normalized Pointwise Mutual Information (Isonuma et al., 2020). For word pairs within a topic, NPMI is defined as

$$\mathrm{NPMI}(w_i, w_j) = \frac{\log \frac{P(w_i,w_j)}{P(w_i)P(w_j)}}{-\log P(w_i, w_j)}.$$

We report the average NPMI score aggregated per level of the hierarchy.

Parent-Child Coherence (PCC) evaluates vertical semantic consistency by computing Cross-Level NPMI (Chen et al., 2021) between each child topic t and its parent $\pi(t)$. This metric averages NPMI scores over word pairs (w_c, w_p) with $w_c \in W_t$ and $w_p \in W_{\pi(t)}$, excluding overlapping words to ensure that coherence reflects meaningful specialization rather than redundancy.

Finally, **Sibling Topic Diversity (SD)** assesses horizontal distinctiveness among sibling topics using an adaptation of Topic Diversity (Dieng et al., 2020). For a set of sibling topics $\mathcal{S}(p)$ under a common parent p, SD is computed as the ratio of words appearing in exactly one sibling topic to the total number of unique words across all siblings.

5.2 Quantitative Results

Table 4 showcases quantitative comparisons across the three datasets using topic coherence (NPMI), parent–child coherence (PCC), and sibling diversity (SD).

Topic Coherence. CobwebTM attains the highest NPMI on all three datasets, achieving 0.206 on

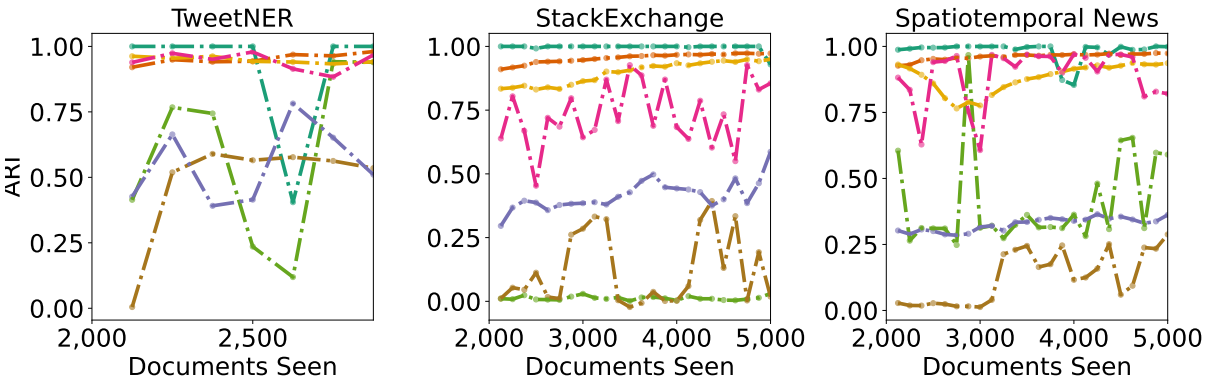


Figure 3: ARI comparison on the StackOverflow, Spatiotemporal News, and TweetNER dataset.

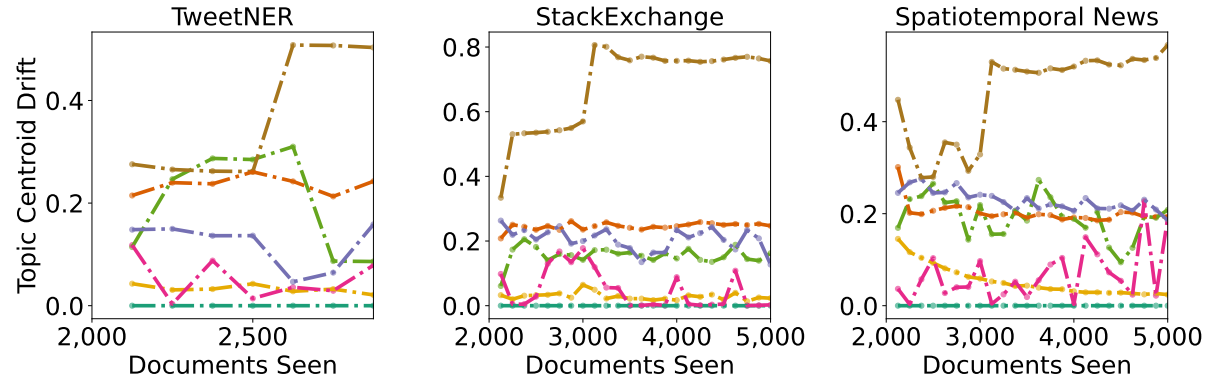


Figure 4: TCD comparison on the StackOverflow, Spatiotemporal News, and TweetNER dataset.

20 News Groups, 0.108 on AG News, and 0.131 on Stack Exchange as per Table 4. These results demonstrate that incremental, structure-aware hierarchy construction does not compromise topic interpretability. In contrast, BERTopic variants exhibit competitive but less consistent coherence, while TraCo and BoxTM suffer from substantially weaker topic quality, particularly on AG News and Stack Exchange.

Vertical Consistency. Parent–child coherence further highlights the advantages of CobwebTM. As shown in Table 4, CobwebTM achieves the highest PCC across all datasets, indicating that its child topics reliably specialize their parents. Post-hoc hierarchical approaches such as BERTopic display mixed behavior, while TraCo frequently produces negative PCC values, suggesting poor semantic alignment between hierarchical levels.

Sibling Diversity. While BoxTM achieves near-perfect sibling diversity, its low NPMI and modest PCC indicate that this separation often comes at the expense of semantic coherence, reflecting over-segmentation. In contrast, CobwebTM maintains high sibling diversity (SD ≥ 0.94 across datasets) while simultaneously preserving strong topic coherence and vertical consistency. This balance suggests that CobwebTM induces meaningful distinctions among sibling topics without fragmenting semantic structure.

Taken together, our results demonstrate that CobwebTM offers a more holistic solution to hierarchical topic modeling, effectively balancing interpretability, hierarchical alignment, and structural diversity. These findings underscore the promise of incremental, structure-aware learning for inducing high-quality topic hierarchies.

5.3 Qualitative Results

In this section, we visualize the hierarchies created by CobwebTM. Figure 1 shows sample hierarchies of topic summaries from the 20 Newsgroups and AG News datasets. The hierarchy captures meaningful semantic structure: parent nodes represent broad themes while children specialize into increasingly specific topics.

In the 20 Newsgroups hierarchy (Figure 1a), a technology-focused root topic is partitioned into co-

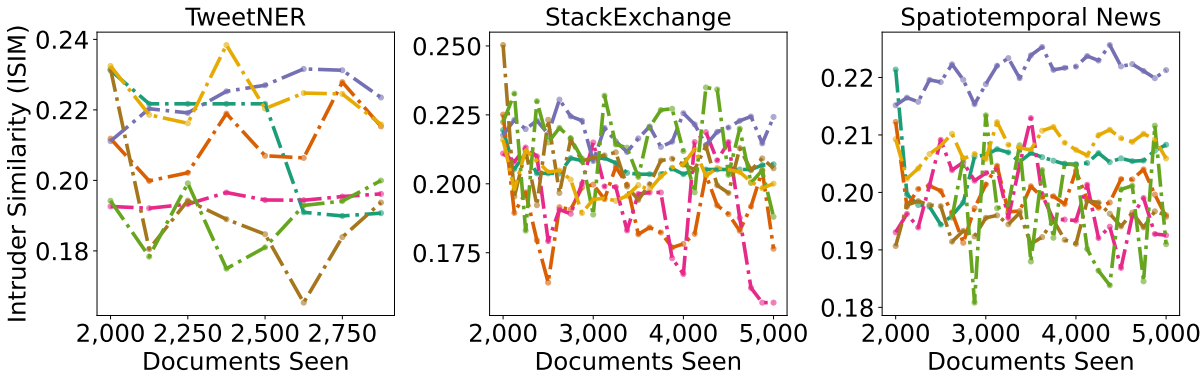


Figure 5: Intruder Similarity (ISIM) comparison on the StackOverflow, Spatiotemporal News, and TweetNER dataset.

Dataset	20 News Groups			AG News			Stack Exchange		
Method	NPMI	PCC	SD	NPMI	PCC	SD	NPMI	PCC	SD
BERTopic-HDBSCAN	0.161	0.045	0.600	0.101	**0.031**	0.913	0.130	0.036	0.891
BERTopic-KMeans	0.169	0.091	0.875	0.067	0.005	0.856	0.125	0.041	0.940
TraCo	0.155	-0.038	0.941	-0.107	-0.043	0.862	-0.040	-0.156	0.913
BoxTM	0.047	0.055	**0.993**	0.002	0.005	**0.998**	0.019	0.022	**0.996**
CobwebTM (ours)	**0.206**	**0.141**	0.958	**0.108**	0.027	0.942	**0.131**	**0.073**	0.959

Table 4: Comparison of hierarchical topic-modeling metrics across datasets.

herent subtopics related to memory specifications, operating systems, and graphical components. Importantly, CobwebTM is able to disentangle semantically similar terms that often cause confusion in embedding-based clustering. For example, within subtrees related to *windows–dos–drive*, lower-level topics remain focused on computing concepts rather than drifting toward unrelated senses such as automobile-related uses of "drive."

This behavior contrasts with hierarchies produced by common embedding-clustering pipelines such as BERTopic (using KMeans or HDBSCAN), where nearest-neighbor clustering can conflate semantically distinct senses of ambiguous terms. Because such methods rely primarily on local similarity, they implicitly treat all embedding dimensions as uniformly important, which can lead to mixed topic substructures.

Neural hierarchical topic models such as BoxTM and TraCo exhibit different behavior: while their hierarchies often contain lexically diverse terms, we observe weaker semantic coherence within some nodes. For instance, unrelated words like "butcher" and "boulder" may appear grouped together despite lacking a clear conceptual relationship.

In contrast, CobwebTM produces hierarchies that maintain stronger semantic consistency while separating ambiguous concepts. We attribute this behavior to Cobweb's probabilistic formulation: each node maintains variance estimates over embedding dimensions, effectively weighting feature importance when forming clusters. By maximizing category utility, CobwebTM favors partitions that reduce uncertainty while preserving coherent semantic structure across levels of the hierarchy.

6 Conclusion

In this paper, we introduced CobwebTM, a lifelong hierarchical topic model that incrementally constructs probabilistic concept hierarchies from streaming text. By leveraging pretrained language model embeddings, CobwebTM exploits the geometric structure of the embedding space to induce semantically coherent topic hierarchies without requiring task-specific hyperparameter tuning. The proposed framework naturally supports lifelong learning, allowing the hierarchy to evolve as new documents arrive. Extensive experiments across multiple hierarchical and lifelong topic modeling benchmarks demonstrate that CobwebTM consistently outperforms recent methods in both topic quality and hierarchical organization, highlighting its effectiveness as a scalable and adaptive topic modeling approach.

Limitations

While CobwebTM demonstrates strong empirical performance, several limitations remain. First, although the model incrementally induces hierarchical clusters in embedding space, topic word extraction relies on post hoc aggregation of documents at each node, rather than being directly generated during hierarchy construction. Second, CobwebTM depends on pretrained document embeddings; consequently, the quality of the learned hierarchy is constrained by the representational capacity of the underlying encoder, and semantic distinctions poorly captured in the embedding space may be lost. Third, the incremental clustering process is sensitive to document arrival order, particularly in non-stationary streams. Although local restructuring operations mitigate this effect, globally optimal hierarchies are not guaranteed. Finally, while well suited to lifelong learning, maintaining statistics over large hierarchies incurs growing memory and computational costs, potentially necessitating pruning, compression, or partitioning strategies in long-running deployments.

Future Work. A natural extension of CobwebTM is multimodal topic modeling. Because the framework operates on continuous representations, images and audio can be incorporated via contrastively trained vision–language models and text-based encoders, enabling heterogeneous data to be organized within a shared hierarchical topic space without modifying the underlying algorithm. Additionally, leveraging node-level statistics to generate topic summaries—without aggregating subtree documents—could improve interpretability and efficiency, while entropy and category utility measures may enable autonomous selection of appropriate topic granularity.

References

Nicki Barari, Xin Lian, and Christopher J MacLellan. 2024a. Avoiding catastrophic forgetting in visual classification using human concept formation. *CoRR*.

Nicki Barari, Xin Lian, and Christopher J. MacLellan. 2024b. Incremental concept formation over visual images without catastrophic forgetting. *Advances in Cognitive Systems*.

David M. Blei, Michael I. Jordan, Thomas L. Griffiths, and Joshua B. Tenenbaum. 2003a. Hierarchical topic models and the nested chinese restaurant process. In *Proceedings of the 17th International Conference on Neural Information Processing Systems*, NIPS'03, page 17–24, Cambridge, MA, USA. MIT Press.

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

Arian Bär, Pedro Casas, Lukasz Golab, and Alessandro Finamore. 2014. Dbstream: An online aggregation, filtering and processing system for network traffic monitoring. In *2014 International Wireless Communications and Mobile Computing Conference (IWCMC)*, pages 611–616.

Ziye Chen, Cheng Ding, Zusheng Zhang, Yanghui Rao, and Haoran Xie. 2021. Tree-structured topic modeling with nonparametric neural variational inference. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2343–2353, Online. Association for Computational Linguistics.

James E. Corter and Mark A. Gluck. 1992. Explaining basic categories: Feature predictability and information. *Psychological Bulletin*, 111(2):291–303.

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

Zhibin Duan, Dongsheng Wang, Bo Chen, Chaojie Wang, Wenchao Chen, Yewen Li, Jie Ren, and Mingyuan Zhou. 2021. Sawtooth factorial topic embeddings guided gamma belief network. *CoRR*, abs/2107.02757.

Douglas H. Fisher. 1987. Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, 2(2):139–172.

Mark A Gluck and James E Corter. 1985. Information, uncertainty, and the utility of categories. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 7.

Derek Greene, Derek O'Callaghan, and Pádraig Cunningham. 2014. How many topics? stability analysis for topic models. In *Machine Learning and Knowledge Discovery in Databases*, pages 498–513, Berlin, Heidelberg. Springer Berlin Heidelberg.

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

Anant Gupta, Karthik Singaravadivelan, and Zekun Wang. 2025. Hierarchical semantic retrieval with cobweb. *Preprint*, arXiv:2510.02539.

Pankaj Gupta, Yatin Chaudhary, Thomas Runkler, and Hinrich Schütze. 2020. Neural topic modeling with continual lifelong learning. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Matthew Hoffman, Francis Bach, and David Blei. 2010. Online learning for latent dirichlet allocation. In *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc.

Masaru Isonuma, Junichiro Mori, Danushka Bollegala, and Ichiro Sakata. 2020. Tree-Structured Neural Topic Model. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 800–806, Online. Association for Computational Linguistics.

Haidar Jomaa. 2025. Space-time minilm.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Sergei Koltcov, Vera Ignatenko, Maxim Terpilovskii, and Paolo Rosso. 2021. Analysis and tuning of hierarchical topic models based on renyi entropy approach. *Preprint*, arXiv:2101.07598.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. *Proceedings of the Twelfth International Conference on Machine Learning (ICML)*.

Xin Lian, Zekun Wang, and Christopher J. MacLellan. 2025. Efficient and scalable masked word prediction using concept formation. *Cognitive Systems Research*, 92:101371.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Yuyin Lu, Hegang Chen, Pengbo Mao, Yanghui Rao, Haoran Xie, Fu Lee Wang, and Qing Li. 2024. Self-supervised topic taxonomy discovery in the box embedding space. *Transactions of the Association for Computational Linguistics*, 12:1401–1416.

Yun Luo, Zhen Yang, Fandong Meng, Yafu Li, Jie Zhou, and Yue Zhang. 2025. An empirical study of catastrophic forgetting in large language models during continual fine-tuning. *Preprint*, arXiv:2308.08747.

Christopher J. MacLellan, Peter Matsakis, and Pat Langley. 2022. Efficient induction of language models via probabilistic concept formation. *Advances in Cognitive Systems*.

Christopher J. MacLellan and Harshil Thakur. 2021. Convolutional cobweb: A model of incremental learning from 2d images. *Advances in Cognitive Systems*.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. 2018. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861.

David Mimno, Wei Li, and Andrew McCallum. 2007. Mixtures of hierarchical topics with pachinko allocation. In *Proceedings of the 24th international conference on Machine learning*, pages 633–640.

Dana Movshovitz-Attias, Yair Movshovitz-Attias, Peter Steenkiste, and Christos Faloutsos. 2013. Analysis of the reputation system and user contributions on a question answering website: Stackoverflow. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '13, page 886–893, New York, NY, USA. Association for Computing Machinery.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake VanderPlas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Edouard Duchesnay. 2012. Scikit-learn: Machine learning in python. *CoRR*, abs/1201.0490.

Adler Perotte, Frank Wood, Noemie Elhadad, and Nicholas Bartlett. 2011. Hierarchically supervised latent dirichlet allocation. *Advances in neural information processing systems*, 24.

Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024. Topicgpt: A prompt-based topic modeling framework. *Preprint*, arXiv:2311.01449.

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*, WSDM '15, page 399–408, New York, NY, USA. Association for Computing Machinery.

Eleanor Rosch and Carolyn B. Mervis. 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7(4):573–605.

D. Sculley. 2010. Web-scale k-means clustering. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, page 1177–1178, New York, NY, USA. Association for Computing Machinery.

Simra Shahid, Tanay Anand, Nikitha Srikanth, Sumit Bhatia, Balaji Krishnamurthy, and Nikaash Puri. 2023. Hyhtm: Hyperbolic geometry based hierarchical topic models. *Preprint*, arXiv:2305.09258.

Asahi Ushio, Leonardo Neves, Vitor Silva, Francesco. Barbieri, and Jose Camacho-Collados. 2022. Named Entity Recognition in Twitter: A Dataset and Analysis on Short-Term Temporal Shifts. In *The 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, Online. Association for Computational Linguistics.

Felipe Viegas, Washington Cunha, Christian Gomes, Antônio Pereira, Leonardo Rocha, and Marcos Goncalves. 2020. CluHTM - semantic hierarchical topic modeling based on CluWords. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8138–8150, Online. Association for Computational Linguistics.

Zekun Wang, Ethan L Haarer, Nicki Barari, and Christopher J MacLellan. 2025. Taxonomic networks: A representation for neuro-symbolic pairing. *arXiv preprint arXiv:2505.24601*.

Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. 2024a. A survey on neural topic models: methods, applications, and challenges. *Artificial Intelligence Review*, 57(2).

Xiaobao Wu, Thong Nguyen, and Anh Tuan Luu. 2024b. A survey on neural topic models: methods, applications, and challenges. *Artificial Intelligence Review*, 57(2).

Xiaobao Wu, Thong Nguyen, Delvin Ce Zhang, William Yang Wang, and Anh Tuan Luu. 2024c. Fastopic: Pretrained transformer is a fast, adaptive, stable, and transferable topic model. *Preprint*, arXiv:2405.17978.

Xiaobao Wu, Fengjun Pan, Thong Nguyen, Yichao Feng, Chaoqun Liu, Cong-Duy Nguyen, and Anh Tuan Luu. 2024d. On the affinity, rationality, and diversity of hierarchical topic modeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in Neural Information Processing Systems (NeurIPS)*.

Yin Zheng, Yu-Jin Zhang, and Hugo Larochelle. 2013. A supervised neural autoregressive topic model for simultaneous image classification and annotation. *Preprint*, arXiv:1305.5306.

A Risks

Like all topic models, CobwebTM inherits and can amplify biases present in its training data, such as over-representing dominant viewpoints while marginalizing minority perspectives or sensitive topics. The hierarchical structure may further legitimize biased or stereotypical groupings by presenting them as coherent topics, which can mislead downstream analysis or decision-making. Additionally, because topic models abstract language into latent structures without grounding or normative judgment, they risk obscuring harmful associations in data and being misused to draw causal or normative conclusions from biased text corpora.

B Datasets and Preprocessing

B.1 Datasets

We use a diverse collection of datasets spanning news media, online forums, and social media to evaluate our models. This appendix provides a brief description of each dataset and the corresponding splits used in our experiments.

Spatiotemporal News Dataset We use the Spatiotemporal News Dataset (Jomaa, 2025), which consists of approximately 1.2 million English-language news articles collected from major news outlets across North America and annotated with temporal and geographic metadata. For our experiments, we randomly sample 5k documents from the test split.

Stack Overflow Dataset. We use the Stack Overflow Dataset (Movshovitz-Attias et al., 2013), which contains question-and-answer forum posts covering a broad range of technical topics in computer science and software engineering. We randomly sample 5k posts spanning diverse subject areas.

TweetNER7 Dataset. We use the TweetNER7 Dataset (Ushio et al., 2022), which consists of short-form posts from X.com (formerly Twitter) across a variety of topics. We use the provided test split, which contains approximately 2.8k tweets.

20 Newsgroups. The 20 Newsgroups dataset (Lang, 1995) contains 18,846 documents evenly distributed across 20 discussion groups and is a standard benchmark for topic modeling and text clustering. We use the version distributed through the `scikit-learn` library (Pedregosa et al., 2012).

AG News. We use the AG's News Topic Classification Dataset (AG News) (Zhang et al., 2015), which consists of news articles collected from Associated Press and Google News sources and categorized into four high-level topic classes. We randomly sample 50k documents from the training set and use the full test set consisting of approximately 7.6k documents.

All datasets used in our experiments are publicly available on the Hugging Face Hub and licensed for research use. While curated with privacy in mind, some datasets (e.g., Stack Overflow or TweetNER7) may contain identifiable or offensive content. We rely on the maintainers' preprocessing and licensing terms and do not perform additional filtering. No extra personally identifying information is collected, stored, or exposed.

B.2 Dataset Preprocessing

To preprocess the datasets, we follow the steps in (Wu et al., 2024d): (1) tokenize documents and convert them to lowercase; (2) remove numbers,

punctuations, and stopwords; (3) remove tokens with less than 3 characters.

C Ablation Studies

C.1 Encoder Ablations

We conduct ablation studies on the sentence encoder used to generate document embeddings. Our main experiments use `all-roberta-large-v1`, a RoBERTa Large model finetuned for information retrieval. To evaluate sensitivity to the embedding backbone, we also test two smaller encoders from the same family: `all-MiniLM-L12-v2` and `all-MiniLM-L6-v2`. Using models with similar training objectives but different parameter sizes allows us to isolate the effect of model capacity while controlling for architectural differences.

Table 8 and Figure 6 show that COBWEBTM is robust to encoder choice. `all-roberta-large-v1` achieves the strongest or near-strongest performance across most datasets, particularly on 20 Newsgroups and Stack Exchange. The MiniLM variants remain competitive despite their smaller size: `MiniLM-L6` performs comparably on 20 Newsgroups, while `MiniLM-L12` yields the best results on AG News. Overall, these results indicate that COBWEBTM maintains stable hierarchical structure and lifelong topic quality even with lightweight embedding models, with larger encoders providing modest but consistent gains. One important observation, consistent with prior work on Cobweb for retrieval (Gupta et al., 2025), is that we restrict our study to encoders trained with distance-based objectives. This design choice is motivated by the Cobweb likelihood formulation, which relies on mean squared distance to compute log probabilities. As a result, encoders optimized for embedding-space distance alignment are better suited to the model's underlying assumptions.

C.2 Batch Size Ablations

We also perform ablation studies with a reduced initial batch size for our lifelong topic modeling experiments. While real-world settings often contain a large corpus to pretrain a topic model with, instantiating it with stable topic definitions, there are many situations where a topic model has to be governed completely from scratch, necessitating smaller initial batch sizes.

We show the results of reducing the initial batch size to 500 on all three datasets in Figure 7. We see that COBWEBTM still comfortably outperforms state-of-the-art lifelong methods, owing to its fundamental design around piecemeal learning, and is able to balance constructing new topics with maintaining old topics.

C.3 UMAP Ablations

We perform ablation studies to investigate the effects of dimensionality reduction on our framework with respect to other topic modeling frameworks. Specifically, we investigate the effects of UMAP (McInnes et al., 2018) on the RoBERTA embeddings used in CobwebTM and all BERTopic pipelines. We compare no UMAP applied to UMAP with $d = 16, 128, 512$. As shown in figures 7, 5 and 6, Cobweb still performs best across all levels of UMAP, but there is no statistically significant difference in the results based on the level of UMAP applied, which is why we chose to omit it from the main paper.

C.4 Incremental Ablations

C.4.1 Holdout Ablation

To analyze the efficiency of incremental solutions in being able to create new topics to support, we ran an ablation with the TweetNER dataset, holding out the *person* category and streaming the rest of the documents as per the default training setup, with the last batch consisting exclusively of the *person* category. As shown in Figure 9, COBWEBTM, like the Refit pipelines, is able to correctly create a document for the new category, while the other incremental solutions fail.s

C.4.2 Cobweb Operation Analysis

We conducted an analysis of the frequency of the different node operations in Cobweb with each dataset, shown in figure 8. The most common operation is the "INSERT" operation, necessary for traversing the tree as a whole. Notably, the "INSERT" operation occurs at a log-likeAlso importantly, the MERGE and SPLIT operations each happen 10% of the time across batches, showcasing our method's ability to restructure clusters and create new clusters only when necessary. In this way, COBWEBTM is able to adapt its level of restructuring with respect to the size of its tree, a more robust solution than a fixed number of MERGEs and SPLITs. Additionally, NEW is only used to create a new leaf node for each document, which is why it happens exactly 125 times for each batch of 125 documents.

Dataset	StackExchange			
Model	None	16	128	512
COBWEBTM (ours)	**0.6131**	**0.6331**	**0.6227**	**0.6383**
BERTopic (MiniBatchKMeans)	0.4234	0.4377	0.4202	0.4393
BERTopic Refit (HDBSCAN)	0.4249	0.4829	0.4666	0.4713
BERTopic Refit (KMeans)	0.5507	0.5068	0.5110	0.5123
Lifelong NTM	0.1840	0.1621	0.1868	0.2057
BERTopic (DBSTREAM)	0.4569	0.4309	0.3872	0.4247
Online LDA	0.5199	0.5663	0.5332	0.5479

Table 5: Final C_v by UMAP dimensionality on StackExchange.

Dataset	TweetNER			
Model	None	16	128	512
COBWEBTM (ours)	**0.7414**	**0.7249**	**0.7581**	**0.7048**
BERTopic (MiniBatchKMeans)	0.4129	0.3531	0.3987	0.3807
BERTopic Refit (HDBSCAN)	0.6733	0.6501	0.6524	0.4558
BERTopic Refit (KMeans)	0.3694	0.4131	0.4094	0.4111
Lifelong NTM	0.2811	0.3042	0.3446	0.1684
BERTopic (DBSTREAM)	0.6019	0.3951	0.4655	0.5373
Online LDA	0.3171	0.3171	0.3171	0.3171

Table 6: Final C_v by UMAP dimensionality on TweetNER.

C.4.3 Low-Batch-Size Ablation

We ran an additional experiment with 5,000 documents from the StackExchange Dataset comparing our model's runtime to baselines, measuring the time taken to update topics for batches of size 10, with results shown in . The Cobweb algorithm does not have a batch-size parameter, training on one instance at a time, so at smaller batch-sizes, our model performs robustly while maintaining the best performance. Due to Cobweb's heuristic-guided best first search, it explores approximately $\log(n)$ nodes per update, making it robust as the size of the tree increases.

C.4.4 Corpus Size Ablations

We ran additional experiments measuring our model's performance with 20,000 documents and 50,000 documents from the StackExchange Dataset, which we report in Figures 11. As each new instance begins at the root node of the Cobweb tree and populates as a leaf through a single, greedily chosen path, COBWEBTM updates an average of O(log(N)) nodes with an insertion of each new document, minimizing our total time complexity and allowing us to retain robustness over long document streams. The results of the Refit Pipelines are not shown as they timed out after 15 hours of training.

D Use of Large Language Models

Large language models (LLMs) were used for grammar refinement, integration of new material, and as coding assistants for structuring and implementation support. All technical content and ideas were developed by the authors, and any LLM-generated output was subsequently modified and verified by the authors.

Dataset	Spatiotemporal			
Model	None	16	128	512
COBWEBTM (ours)	**0.7960**	**0.6948**	**0.6837**	0.6284
BERTopic (MiniBatchKMeans)	0.6460	0.6068	0.6403	0.6322
BERTopic Refit (HDBSCAN)	0.6572	0.5103	0.5190	0.5105
BERTopic Refit (KMeans)	0.6136	0.5568	0.5471	0.5516
Lifelong NTM	0.2381	0.1211	0.2331	0.1191
BERTopic (DBSTREAM)	0.4176	0.4109	0.4416	0.3964
Online LDA	0.4221	0.4163	0.4334	0.4185

Table 7: Final C_v by UMAP dimensionality on Spatiotemporal dataset.

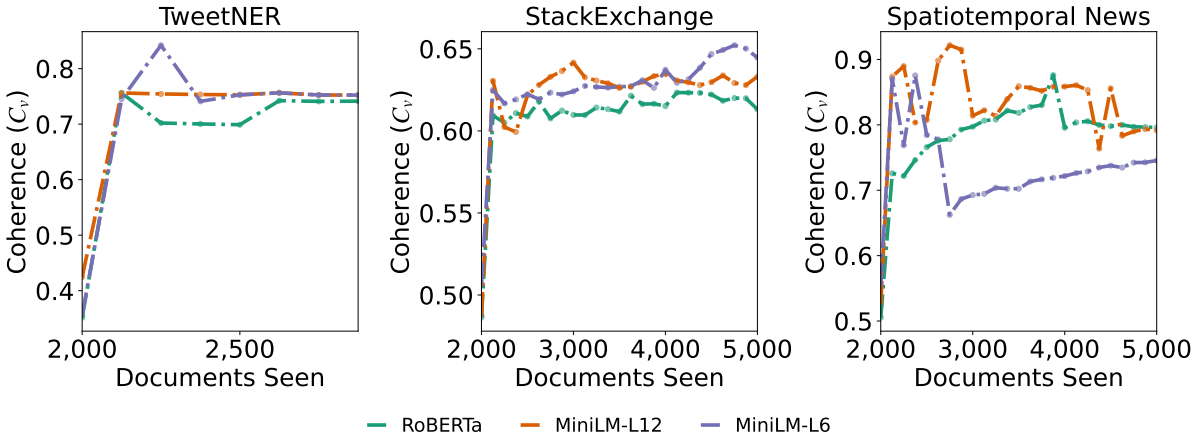
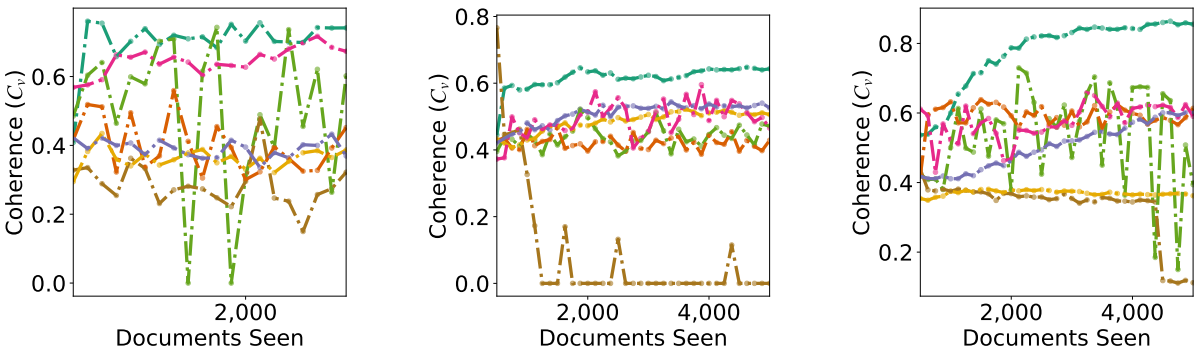


Figure 6: An ablation study to compare the lifelong results of using different embedding models for CobwebTM.

Dataset	20 News Groups			AG News			Stack Exchange		
Model	NPMI	PCC	SD	NPMI	PCC	SD	NPMI	PCC	SD
MiniLM-L6	0.205	0.130	0.935	0.104	0.023	0.924	0.127	0.070	0.952
MiniLM-L12	0.196	0.119	0.952	**0.109**	**0.031**	**0.959**	0.130	0.055	0.862
RoBERTa (ours)	**0.206**	**0.141**	**0.958**	0.108	0.027	0.942	**0.131**	**0.073**	**0.959**

Table 8: An ablation study to compare the hierarchical results of using different embedding models for CobwebTM.



(a) Tweetner Ablation (b) Stack Overflow Ablation (c) Spatiotemporal News Ablation

Figure 7: An ablation study to compare the lifelong results of CobwebTM to other methods across all datasets with a reduced initial batch size of 500 documents.

Dataset	TweetNER (Holdout: *person*)	
Model	Before Injection C_v	After Injection C_v
COBWEBTM (ours)	**0.6425**	**1.0000**
BERTopic (MiniBatchKMeans)	0.3812	0.0000
BERTopic Refit (HDBSCAN)	0.4845	**1.0000**
BERTopic Refit (KMeans)	0.3940	**1.0000**
Lifelong NTM	0.6120	0.0000
BERTopic (DBSTREAM)	0.3018	0.0000
Online LDA	0.3357	0.8132

Table 9: Holdout topic injection experiment on TweetNER dataset.

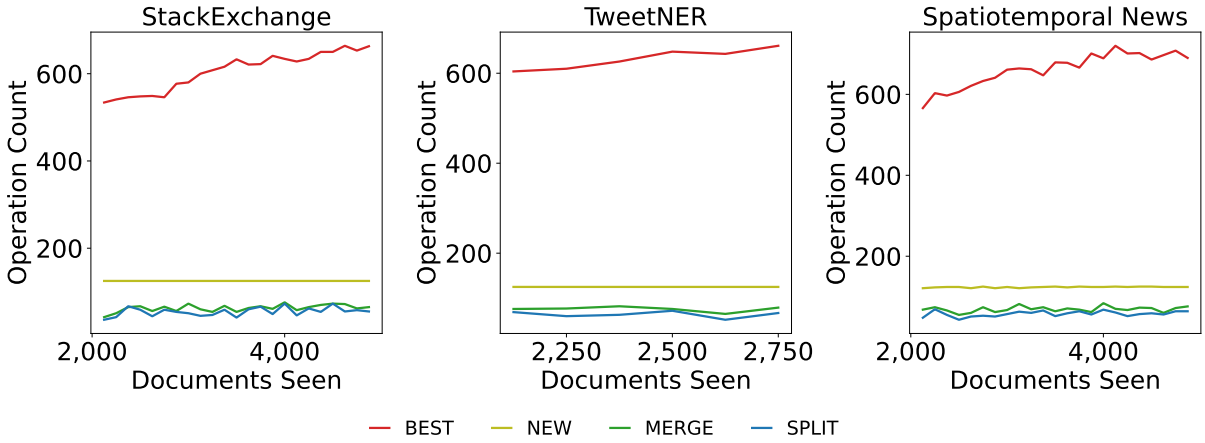


Figure 8: An ablation study to compare the amount of each Cobweb operation per-batch for all three datasets in CobwebTM.

Dataset	StackExchange (batch_size=10)	
Model	Total Time (s)	Final Batch (s)
COBWEBTM (ours)	382.94	1.37
BERTopic (MiniBatchKMeans)	362.81	1.52
BERTopic Refit (HDBSCAN)	16525.93	87.79
BERTopic Refit (KMeans)	10749.74	50.93
Lifelong NTM	657.44	2.62
BERTopic (DBSTREAM)	805.71	3.62
Online LDA	**25.17**	**0.07**

Table 10: Runtime comparison on StackExchange dataset.

Dataset	20k StackExchange		50k StackExchange	
Model	Start C_v	Final C_v	Start C_v	Final C_v
COBWEBTM (ours)	0.4932	<u>0.6573</u>	0.5010	<u>0.6441</u>
BERTopic (MiniBatchKMeans)	0.4810	0.4270	0.4958	0.4476
BERTopic Refit (HDBSCAN)				
BERTopic Refit (KMeans)				
Lifelong NTM	**0.7133**	0.5522	**0.7752**	0.0000
BERTopic (DBSTREAM)	0.4481	0.4722	0.4801	0.4144
Online LDA	0.4365	**0.6474**	0.4782	**0.5909**

Table 11: Coherence C_v comparison on 20k and 50k StackExchange datasets.