CobwebTM: Probabilistic Concept Formation for Lifelong and Hierarchical Topic Modeling

Anonymous ACL submission

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

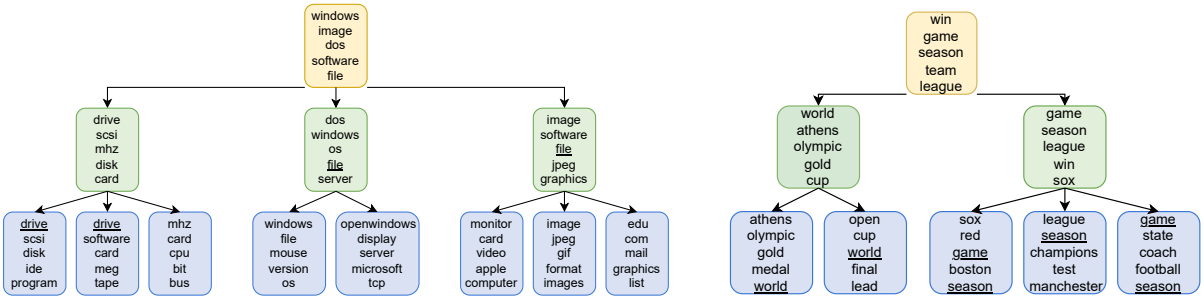
Topic modeling aims to uncover latent semantic structure in large collections of documents by grouping text into coherent topics. It is a fundamental tool for document organization, corpus exploration, and information retrieval, particularly in settings where labeled data is unavailable. As modern text corpora grow in scale, diversity, and temporal span, effective topic modeling increasingly requires unsupervised topic discovery, adaptability to streaming data, and the ability to 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 topic independence, and relies on bag-of-words representations that discard semantic similarity. These assumptions limit its ability to model imbalanced, correlated, or evolving topics, and make it poorly suited for lifelong or streaming settings. Recent advances in representation learning have led to neural topic models that leverage dense learned document embeddings (Zheng et al., 2013; Wu et al., 2024a). These approaches often achieve improved topic coherence and semantic expressiveness, 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. As a result, they struggle with lifelong learning, 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 this challenge 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 hierarchical organization.

Hierarchy is a natural and essential aspect of topic structure: real-world topics exist at multiple levels of granularity, from broad themes to fine-grained subtopics. While hierarchical topic models have been explored through Bayesian and neural approaches, many modern systems impose hierarchy post hoc or rely on fixed-depth latent structures, limiting flexibility and interpretability—particularly in dynamic settings.

In this work, we visit incremental concept for-



(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. Words that appear in multiple nodes at the same level are underlined to highlight shared semantic content across sibling topics.

mation as an alternative paradigm for topic modeling. We introduce CobwebTM, a lifelong and 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, and hierarchical topic modeling framework that performs unsupervised topic discovery over streaming text; (2) we show that probabilistic concept formation over embedding spaces enables robust lifelong topic modeling without catastrophic forgetting or fixed topic capacity; (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

The most widely used lifelong topic model is Online LDA (Hoffman et al., 2010), which updates global topics via mini-batch Variational Bayes. While it generalizes to out-of-domain data, it inherits LDA's bag-of-words assumption, requires a predefined number of topics, lacks topic restructuring, and does not retain long-term memory, forcing future data into existing topics.

Most neural topic models are trained in batch settings and are poorly suited for sequential updates without retraining (Wu et al., 2024b). They also suffer from catastrophic forgetting (Luo et al., 2025). Mitigation strategies such as elastic weight consolidation or replay (Gupta et al., 2020) reduce forgetting but still rely on fixed latent dimensions, limiting adaptation to out-of-domain data and new topics.

Some lifelong topic modeling approaches adopt incremental clustering over neural embeddings. BERTopic (Grootendorst, 2022) combines transformer embeddings, dimensionality reduction, clustering, and class-based topic representations. Lifelong variants replace static clustering with incremental methods such as DBStream (Bär et al., 2014) or Mini-Batch KMeans (Sculley, 2010), though these remain sensitive to parameters or fixed cluster assumptions.

2.2 Hierarchical Topic Modeling

Hierarchical topic models extend flat models like LDA (Blei et al., 2003b) by organizing topics across semantic levels. Early Bayesian nonparametric approaches such as hLDA (Blei et al., 2003a) and its variants (Mimno et al., 2007; Blei et al., 2010; Perotte et al., 2011) have been largely replaced by embedding-based methods. Examples include CluHTM (Viegas et al., 2020), HyHTM (Shahid et al., 2023), and BERTopic (Grootendorst, 2022), which construct hierarchies via clustering or linkage, often as post-hoc aggregation.

Neural Hierarchical Topic Models use VAEs (Kingma and Welling, 2013) to learn structured representations, including tree-based models (Isonuma et al., 2020), fixed-depth architectures (Duan et al., 2021), and geometric regularizations (Wu et al., 2024c; Lu et al., 2024). In contrast, our work revisits hierarchical induction through incremental conceptual clustering, showing that hierarchy can emerge from representation

geometry and retrieval-based similarity (Gupta et al., 2025) without explicit generative constraints.

2.3 Incremental Concept Formation

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

Cobweb (Fisher, 1987) incrementally constructs a probabilistic taxonomy via top-down sorting, functioning as a divisive concept formation method. Extensions (MacLellan et al., 2022; MacLellan and Thakur, 2021; Wang et al., 2025) adapt Cobweb to neural settings. Recent work demonstrates its robustness in vision and language tasks (Barari et al., 2024b,a; Lian et al., 2025), highlighting resistance to catastrophic forgetting and the ability to evolve topics over time without fixed batch sizes.

3 Methodology

We propose CobwebTM, a hierarchical topic modeling framework that leverages incremental concept formation to organize document embeddings into a semantic taxonomy. In contrast to batch-based clustering methods (e.g., k-Means, HDBSCAN), our approach constructs a tree-structured representation in an online manner, enabling topics to evolve as new documents arrive without retraining from scratch.

For each incoming batch, we first obtain document representations using a pretrained sentence transformer BERT-based (Devlin et al., 2019) model. These embeddings are subsequently passed to the Cobweb algorithm, detailed in the next section, which incrementally assigns each document to a leaf node in the hierarchy. Internal (non-terminal) nodes correspond to higher-level topics, with the documents contained in their respective subtrees defining the semantic scope of each topic.

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 over attributes with diagonal covariance,

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

where $\mu_c \in \mathbb{R}^D$ is the node mean and $\sigma_c^2 \in \mathbb{R}^D$ is a diagonal variance vector. The mean μ_c is updated incrementally as instances are assigned to the node, and the variance is taken to be a simple function of the node count N_c (i.e., a count-dependent constant), so that uncertainty decreases as more evidence accumulates. Concretely, each node stores the sufficient statistics needed to update μ_c online together with the cluster count N_c, which enables efficient maintenance of the Gaussian parameters throughout tree growth.

Cobweb constructs a hierarchy of concepts over the seen instances incrementally. Upon receiving a new document embedding x, the algorithm traverses the tree in a top-down best-first search, guided by maximizing *Category Utility* (CU) (Gluck and Corter, 1985; Corter and Gluck, 1992). Our utilized Cobweb adopts an *information-theoretic* formulation of CU as per (Barari et al., 2024b), which explicitly links feature predictability with informativeness. Let a parent node c_p have children $\mathcal{C}(c_p)$, and each node c maintain a diagonal-covariance Gaussian $p(\cdot \mid c) = \mathcal{N}(\mu_c, \mathrm{diag}(\sigma_c^2))$ with cluster count N_c. We define the empirical concept probability within the parent as

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

We define the node uncertainty as the differential entropy of its Gaussian:

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

where $\sigma_{c,d}^2$ is the (count-dependent) diagonal variance for dimension d in node c. Under this formulation, the category utility of a parent node c_p is the expected reduction in uncertainty from knowing the child concept:

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

During traversal, Cobweb selects the child that yields the highest CU increase under the corresponding local update, which is equivalent to maximizing the mutual information between the (child) concept assignment and the feature distribution conditioned on the parent.

Intuitively, maximizing CU favors concepts that substantially reduce feature uncertainty relative to their parent while maintaining sufficient support,

thereby balancing intra-class similarity and inter-class dissimilarity. For continuous attributes, maximizing CU is equivalent to maximizing the reduction in variance induced by the partition. As a result, Cobweb naturally determines both the depth and breadth of the resulting topic hierarchy without requiring a number of topics K.

At each node, Cobweb evaluates four possible operators to determine how x should be incorporated into the hierarchy. Specifically, x may be inserted into the best-matching existing child node by updating the child node's Gaussian parameters with x, used to create a new singleton child node, assigned to a merged node formed by combining the two best-matching children and re-averaging their means, or trigger a split operation in which the best-matching child is replaced by its own children, promoted one level up in the tree.

3.2 Topic Extraction

3.2.1 Hierarchical Topic Extraction

The resulting Cobweb tree represents a multi-level topic hierarchy. The root represents the average of the entire corpus, intermediate nodes represent broad topics, and leaves represent individual documents or fine-grained micro-topics.

To extract interpretable topics, we treat each node in the tree as a candidate topic. We employ a class-based TF-IDF (c-TF-IDF) procedure similar to Grootendorst (2022). For a given node C, we aggregate all documents in its subtree into a single large document. We then calculate the importance of word w in topic C as:

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

where $tf_{C,w}$ is the frequency of word w in topic C, f_w is the frequency of w across all topics, and A is the average number of words per topic. This formulation allows us to generate descriptive keywords for any node in the hierarchy.

3.2.2 Dynamic Flat Topic Extraction

While hierarchical structure provides rich granularity, most existing incremental topic models operate on a flat set of topics. To enable direct comparison with these baselines and support applications requiring static topic sets, CobwebTM introduces a dynamic method to extract a flat partition from the evolving hierarchy.

Since Cobweb is incremental, leaves (documents) are distributed across various depths of the tree based on their semantic specificity. To recover a coherent set of flat topics, we identify a "cut" through the tree that balances topic coherence with coverage. We traverse the hierarchy top-down and select nodes based on two criteria: the number of nodes in a layer must not exceed a `max_clusters` parameter, and the ratio of leaf nodes to total nodes in that layer must not exceed a `leaf_ratio` parameter.

This approach allows us to filter out high-level outliers (leaves near the root) while grouping deeper, semantically similar documents into robust clusters. By dynamically adjusting this cut as new data arrives, CobwebTM maintains a flat topic representation that evolves in real-time, bridging the gap between hierarchical concept formation and traditional flat incremental clustering.

Thus, CobwebTM provides both a flat topic model (by cutting the tree at a specific level or using Cobweb's "basic level" criteria) and a full hierarchical exploration tool, where users can zoom in from general concepts to specific discussions.

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. Lastly, 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.

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, and 4, CobwebTM outperforms all baselines, maintaining the best C_v and TCD across all three datasets, and is competitive in ARI in the TweetNER dataset while outperforming all methods in ARI in the other two datasets. This showcases its ability to retain topics over batches while accommodating new batches.

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 Figure 1. Additionally, Cobweb has no learning parameters, with the only two parameters being user-specified for granularity decisions in flat topic modeling. This highlights its value over traditional clustering topic modeling pipelines, most of which either require a minimum cluster size or number of clusters to be specified, and a dimensionality reduction algorithm to be tuned.

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 3. This behavior reflects a broader limitation of neural topic models, which must learn word semantics from the corpus itself rather than leveraging pretrained embeddings. Our approach, which combines pretrained encoder representations with symbolic learning, enables more stable incremental performance and reduced susceptibility to

Dataset	TweetNER			Stack Overflow			Spatiotemporal News		
Method	C_{vf}	$\Delta C_v\%$	ARI	C_{vf}	$\Delta C_v\%$	ARI	C_{vf}	$\Delta C_v\%$	ARI
CobwebTM (ours)	**0.741**	**110.82**	0.915	**0.613**	**26.09**	**0.997**	**0.796**	**57.51**	**0.984**
DBSTREAM	0.602	71.55	0.594	0.457	-3.67	0.012	0.418	-10.29	0.408
LifelongDocNADE	0.281	-11.61	0.479	0.184	-75.07	0.122	0.238	-49.49	0.127
MiniBatchKMeans	0.413	-0.89	**0.952**	0.423	-9.67	0.955	0.646	40.37	0.962
OnlineLDA	0.317	7.06	0.948	0.520	10.71	0.897	0.422	0.99	0.882
Refit-HDBSCAN	0.673	7.35	0.944	0.425	1.81	0.735	0.657	19.11	0.891
Refit-KMeans	0.369	-6.49	0.549	0.551	15.98	0.416	0.614	29.11	0.326

Table 1: Comparison of final topic coherence C_{vf}, coherence change $\Delta C_v\%$, and clustering quality (ARI, averaged across all batches) for lifelong topic modeling experiments on the TweetNER, Stack Exchange, and Spatiotemporal News datasets. Best results are shown in bold, and second-best results are underlined for each column.

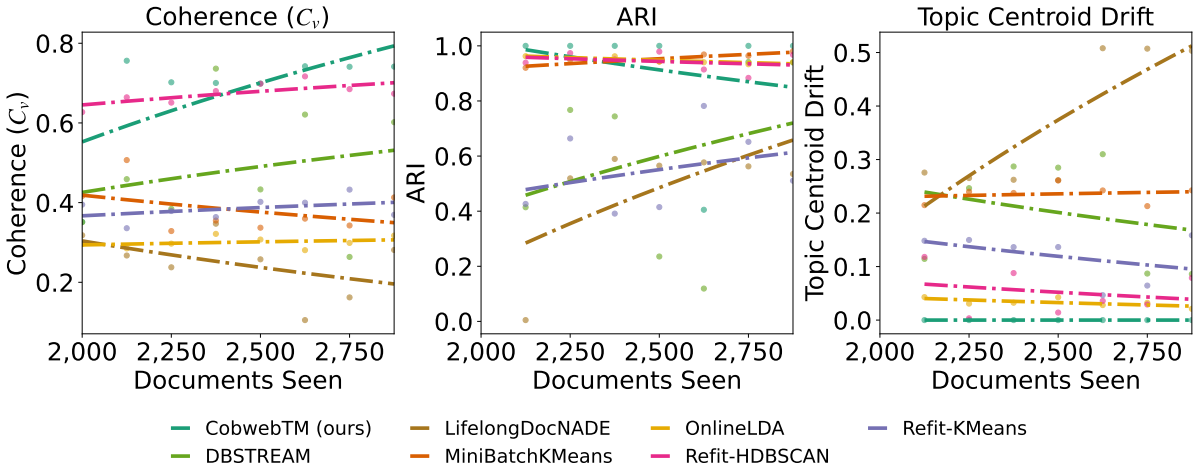


Figure 2: TweetNER results

catastrophic forgetting.

Metrics Over Time. CobwebTM shows strong increases in performance by coherence unlike other methods, as shown in Figure 1, 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 Table 1, 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.

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).

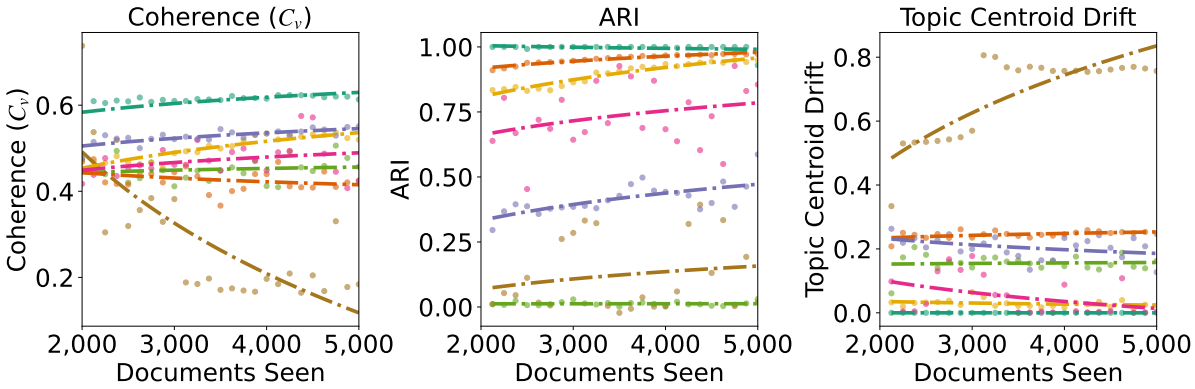


Figure 3: Stack Overflow results

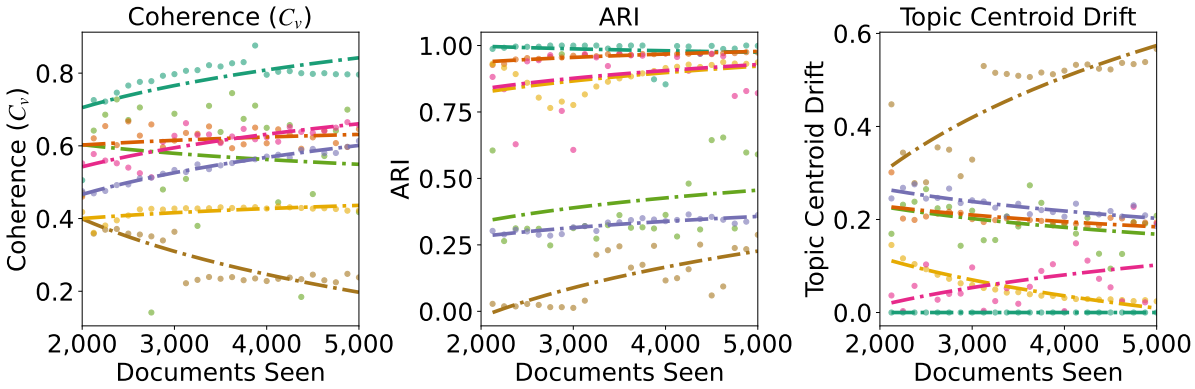


Figure 4: Spatiotemporal News results

Baselines. We compare against three primary baselines: **TraCo** (Wu et al., 2024c), 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, Sibling TD is computed as the ratio of words appearing in exactly one sibling topic to the total number of unique words across all siblings.

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 2: Comparison of hierarchical topic-modeling metrics across datasets.

5.2 Quantitative Results

Table 2 showcases quantitative comparisons across the three datasets using topic coherence (NPMI), parent–child coherence (PCC), and sibling diversity (SD). Overall, CobwebTM consistently achieves strong performance across all three metrics, indicating its ability to induce hierarchies that are simultaneously interpretable, semantically aligned across levels, and well-structured.

Topic Coherence. CobwebTM attains the highest NPMI on all three datasets, achieving 0.206 on 20 News Groups, 0.108 on AG News, and 0.131 on Stack Exchange as per Table 2. 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 2, 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 dataset and the AG News dataset. In addition to semantically relating subtopics strongly to their parent topics, Cobweb specializes in disentangling semantically similar topic summaries, shown especially by Figure 1a's partition of a technology-focused root topic into subtopics of specific realms of technology (memory specifications, operating systems, graphical elements). Cobweb's focus on Category Utility allows us to partition topics well asynchronously of the topic's breadth and depth.

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. Beyond fixed pretrained embeddings, future work may explore representation learning aligned with Cobweb's incremental hierarchy induction, for example by using hierarchical feedback signals to improve coherence and stability in lifelong settings. Finally, 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, Thomas L Griffiths, and Michael I Jordan. 2010. The nested chinese restaurant process and bayesian nonparametric inference of topic hierarchies. *Journal of the ACM (JACM)*, 57(2):1–30.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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*.

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*.

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.

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, Fengjun Pan, Thong Nguyen, Yichao Feng, Chaoqun Liu, Cong-Duy Nguyen, and Anh Tuan Luu. 2024c. 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.

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 125k 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 fil-

tering. No extra personally identifying information is collected, stored, or exposed.

B.2 Preprocessing

To preprocess the datasets, we follow the steps in (Wu et al., 2024c): (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 3 and Figure 5 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.

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 6. 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.

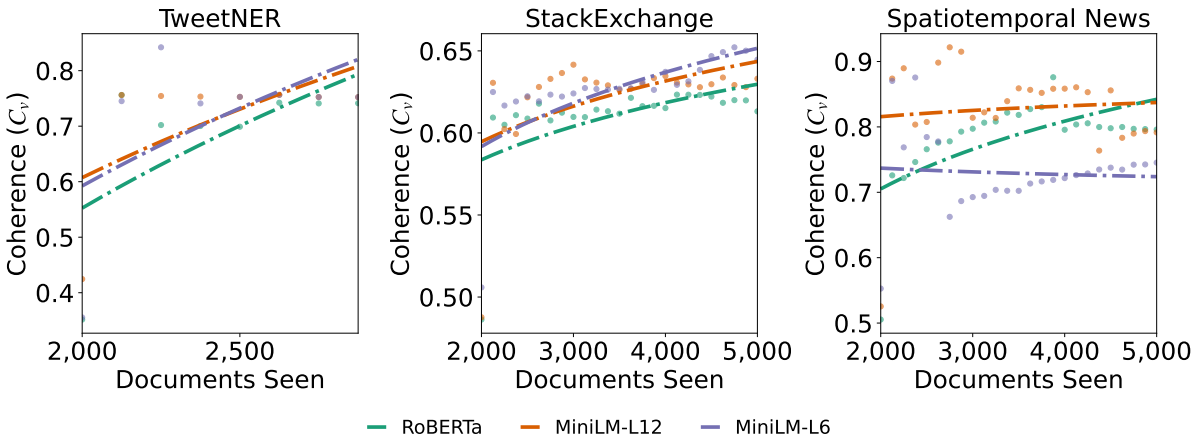
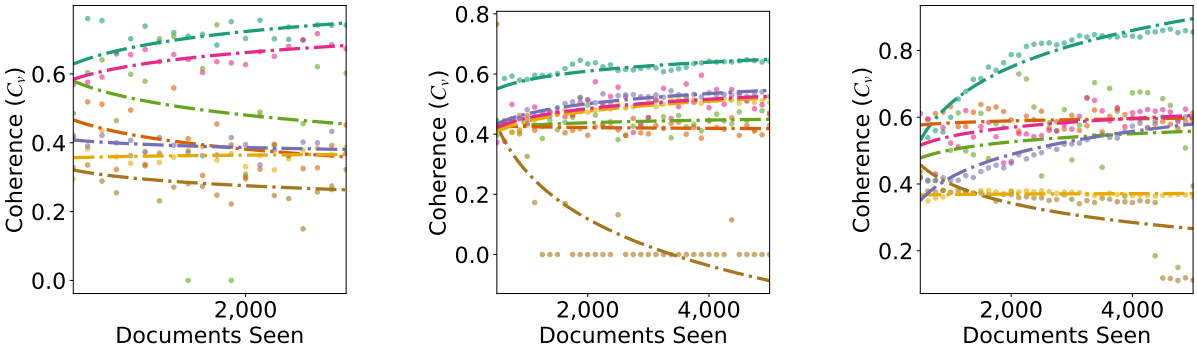


Figure 5: An ablation study to compare the lifelong results of using different embedding models for CobwebTM.

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 3: 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 6: 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.