Attentive Additively Regularized Topic Modeling for Local Context

Anonymous ACL submission

Abstract

Widely adopted probabilistic topic models ignore word order and represent each document as a mixture of topics. This limits their applicability to both short texts and long thematically segmented documents. We propose a local context topic model that combines ideas of attention and additive regularization (ARTM). We derive an EM-like algorithm with two built-in mechanisms for incorporating local context. By dropping the more computationally expensive of them, we make the algorithm simpler and faster without a substantial degradation in quality. The simplified version, called Attentive ARTM, can be considered as an extensible platform for topic modeling of sequential text beyond the bag-of-words assumption. Experiments on three datasets show that Attentive ARTM yields competitive document representations and substantially improves topic-boundary detection over bag-of-words topic models. The full model with coherence regularization consistently improves topic diversity, while the ablated variant is faster and often matches classification performance.

1 Introduction

Probabilistic topic models (PTMs) remain a useful tool for exploratory text analysis providing compact and interpretable representations of text collections. In models such as PLSA (Hofmann, 1999) and LDA (Blei et al., 2003), each topic is represented by a distribution over words, and each document is represented by a distribution over topics. The resulting topic-word distributions can be inspected directly, while document-topic vectors can be used in classification, retrieval, segmentation, corpus navigation, and other tasks.

Our research attacks two strong assumptions inherent in many probabilistic topic models. First, the bag-of-words assumption discards word order and represents a document only by term counts. Second, all word occurrences in a document are explained by the same document-level topic mixture. These assumptions make inference efficient, but they are often too coarse. Short documents may not contain enough evidence to estimate a reliable topic mixture. Long documents may contain several topical segments, so the topic of a particular word occurrence is better determined by its local context than by the entire document.

PTM research area is divided into three main approaches: optimizational, Bayesian, and neural.

Optimization-based formalism was introduced for probabilistic latent semantic analysis (PLSA) in (Hofmann, 1999). It is based on approximate stochastic matrix factorization, which is an ill-posed problem with infinitely many solutions. Such models require regularization to improve stability and generalization, and to incorporate additional data and constraints, if available.

Bayesian regularization long became the core mathematical technique in topic modeling, starting with latent Dirichlet allocation (LDA), proposed in (Blei et al., 2001, 2003). Later, hundreds of specialized models were developed on top of LDA: multimodal, multilingual, hierarchical, temporal, sparse, etc. for searching, classification, regression, segmentation, summarization, and other NLP tasks (Churchill and Singh, 2022).

Many Bayesian PTMs have been proposed to relax the bag-of-words assumption. The bigram topic model accounts for adjacent word pairs (Wallach, 2006). Biterm topic model (Yan et al., 2013) and word network topic model (Zuo et al., 2016) capture pairwise word co-occurrence in local contexts. SenLDA represents each sentence as a bag of words and each document as a bag of sentences (Balikas et al., 2016). Local LDA uses overlapping context windows, benefiting from word order only partially (Rahimi et al., 2022).

In the Bayesian approach, one estimates not the model parameters themselves, but their posterior distributions. Such a complication is both exces-

sive and restrictive from a practical point of view. For each new model, both the mathematical work of Bayesian inference and the programming work of learning algorithm implementation must be redone from scratch. Combining Bayesian PTMs to merge their functionality also encounters significant technical difficulties.

Additive regularization of topic models (ARTM) proposed in (Vorontsov, 2014) overcomes these difficulties by bringing Bayesian models back into the optimization framework (Vorontsov, 2023). ARTM framework allows one to transfer regularizers from one model to another, combine topic models by merging their regularizers, and share the learning algorithm code between multiple models. ARTM was implemented in open-source library `BigARTM`, which provides fast and modular extensible platform for learning PTMs from large text collections (Frei and Apishev, 2016).

Neural topic models (NTM) have been actively progressing in recent years. Models such as BERTopic (Grootendorst, 2022) and Contextual-Top2Vec (Angelov and Inkpen, 2024) represent a pipeline of ML/NLP procedures on top of an LLM, unlike holistic mathematical and algorithmic solution from the above approaches. The advantage of NTM follows from attention mechanism (Vaswani et al., 2017) that provides ready-made context-dependent embeddings of tokens, words, and sentences. Due to pre-training, LLM possesses knowledge of language and world not limited by the size of input collection. So, short texts and small corpora are no longer a problem for NTMs. The disadvantage of NTMs is that they are resource-intensive in terms of time, memory, and energy, especially for large collections.

In this paper, we propose a compromise solution within the ARTM optimization framework. The local context topic model inherits the speed and modular extensibility from ARTM, moving beyond the bag-of-words assumption to the approach inspired by attention mechanisms.

The contributions of this paper are:

- We propose local context topic model that depends only on the topic embeddings of terms and accounts for their order in the text.
- We derive EM-like algorithm for word topic embeddings and make it simpler and faster by eliminating unnecessary context statistics.
- We propose a set of evaluation metrics for comparing our model against bag-of-word models, including topic quality, document classification, robustness to short documents, synthetic boundary detection, and runtime.
- The code for both the algorithm and the evaluation metrics is made publicly available.

2 Bag-of-Words Topic Modeling

Let D be a collection of text documents. Each document $d \in D$ is a sequence $(w_1, \dots, w_{n_d})$ of n_d terms taken from the vocabulary W. Terms can be words or n-grams, depending on tokenization.

Under the bag-of-words hypothesis, a document is represented compactly as a multiset $d \subset W$, in which each element $w \in d$ occurs n_{dw} times.

Bag-of-words topic model explains the occurrence of term w in document d:

$$p(w \mid d) = \sum_{t \in T} p(w \mid t)\, p(t \mid d) = \sum_{t \in T} \varphi_{wt} \theta_{td}.$$

To estimate the parameters of the topic model $\varphi_{wt} = p(w \mid t)$ and $\theta_{td} = p(t \mid d)$ from the collection D, log-likelihood $L(\Phi, \Theta)$ is maximized, subject to probability distribution constraints:

$$\max\colon L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln p(w \mid d); \quad (1)$$

$$\sum_{w} \varphi_{wt} = 1, \quad \varphi_{wt} \geqslant 0, \quad \sum_{t} \theta_{td} = 1, \quad \theta_{td} \geqslant 0. \quad (2)$$

In ARTM framework, a weighted sum of regularization criteria $R_k(\Phi, \Theta)$ with non-negative *regularization coefficients* τ_k, $k = 1, \dots, K$ is added in order to regularize the ill-posed problem of matrix factorization, while keeping constraints (2):

$$\max\colon L(\Phi, \Theta) + \sum_{k=1}^{K} \tau_k R_k(\Phi, \Theta). \quad (3)$$

Define the norm operation, which maps a vector $(x_i)_{i \in I} \in \mathbb{R}^{|I|}$ to a discrete distribution $(p_i)_{i \in I}$:

$$p_i = \operatorname*{norm}_{i \in I}(x_i) = \frac{\max(0, x_i)}{\sum\limits_{j \in I} \max(0, x_j)}, \text{ for all } i \in I.$$

If $x_i \leqslant 0$ for all $i \in I$ then the result of norm is defined as zero vector.

Theorem 1 *(Vorontsov and Potapenko, 2015) Let the function $R(\Phi, \Theta)$ be continuously differentiable. Then a local extremum (Φ, Θ) of problem*

(3), (2) *satisfies the system of equations with auxiliary variables* $p_{tdw} = p(t \mid d, w)$, n_{wt}, *and* n_{td}:

$$p_{tdw} = \underset{t\in T}{\operatorname{norm}}\big(\varphi_{wt}\theta_{td}\big); \tag{4}$$

$$n_{wt} = \sum_{d\in D} n_{dw}p_{tdw}; \quad n_{td} = \sum_{w\in d} n_{dw}p_{tdw};$$

$$\varphi_{wt} = \underset{w\in W}{\operatorname{norm}}\left(n_{wt} + \varphi_{wt}\frac{\partial R}{\partial \varphi_{wt}}\right); \tag{5}$$

$$\theta_{td} = \underset{t\in T}{\operatorname{norm}}\left(n_{td} + \theta_{td}\frac{\partial R}{\partial \theta_{td}}\right). \tag{6}$$

Solving this system by the simple iteration method yields the Expectation–Maximization (EM) algorithm, which alternates the E-step (4) and M-step (5)–(6) until convergence.

The n-variables are interpreted as frequencies: n_{wt} counts how many times term w is assigned to topic t in the collection, n_{td} counts how many terms of document d are assigned to topic t.

The PLSA model (Hofmann, 1999) is a special case without regularization, $R = 0$.

The LDA model (Blei et al., 2003), when trained via maximum a posteriori probability, corresponds to a cross-entropy regularizer for smoothing or sparsing the columns of Φ and Θ, depending on the sign of the hyperparameters β_w and α_t, respectively: $R = \sum_{w,t} \beta_w \ln \varphi_{wt} + \sum_{t,d} \alpha_t \ln \theta_{td}$.

Convergence of the EM-algorithm to a stationary point of the optimization problem was proved in (Irkhin and Vorontsov, 2020) under easily verifiable assumptions about the regularizer.

It was also shown there that convergence can be improved without extra time or memory cost by substituting unbiased frequency estimates

$$\varphi^*_{wt} = \underset{w\in W}{\operatorname{norm}}(n_{wt}), \quad \theta^*_{td} = \underset{t\in T}{\operatorname{norm}}(n_{td}) \tag{7}$$

into the right-hand sides of the M-step (5)–(6) instead of φ_{wt} and θ_{td} correspondingly.

3 Local Context Topic Model

Let $\{w_1, \dots, w_n\}$ denote the concatenation of all documents in the text collection of size n with continuous numbering of terms.

Denote by $C_i \subset \{1, \dots, n\}$ the positions of terms that form the context (i. e. local neighborhood) of term w_i. The context can be defined in various ways: it may be short or long; left, right, or bidirectional; contiguous or excluding some terms; containing the central term w_i or excluding it. The context cannot cross document boundaries, i. e., it cannot include terms from other documents. The entire document may serve as the context; in this case, the model is expected to correspond to standard PLSA-based or LDA-based models.

Local context topic model predicts the occurrence of a term at position i given its context C_i:

$$p(w \mid C_i) = \sum_{t\in T} p(t \mid w)\frac{p(w)}{p(t)}p(t \mid C_i),$$

where contextual topic vector $p(t \mid C_i)$ is defined instead of $p(t \mid d)$ as a weighted average of topic vectors of its constituent terms:

$$p(t \mid C_i) \equiv \theta_{ti} = \sum_{c\in C_i} \alpha_{ci} p(t \mid w_c), \tag{8}$$

where *attention coefficient* α_{ci} indicates how important the context term w_c is for the central term w_i. These coefficients are nonnegative and normalized: $\alpha_{ci} \geqslant 0$, $\sum_{c\in C_i} \alpha_{ci} = 1$.

After substitution, the model takes the form of a function of the topic embeddings $\varphi_{tw} = p(t \mid w)$:

$$p(w \mid C_i) = \sum_{t\in T} \varphi_{tw}\frac{p(w)}{p(t)} \sum_{c\in C_i} \alpha_{ci}\varphi_{tw_c}, \tag{9}$$

where $p(t)$ is also a function of topic embeddings. By Bayes' theorem, the following identity must hold to ensure the model consistency:

$$p(t) = \sum_{w\in W} \varphi_{tw} p(w). \tag{10}$$

The regularized log-likelihood is maximized to estimate the model parameters $\varphi_{tw} = p(t \mid w)$, subject to probability distribution constraints:

$$\max\colon L(\Phi) = \sum_{i=1}^{n} \ln p(w_i \mid C_i) + R(\Phi); \tag{11}$$

$$\sum_{t\in T} \varphi_{tw} = 1, \quad \varphi_{tw} \geqslant 0. \tag{12}$$

As a solution, we propose an EM-like algorithm for a family of models called **Attentive ARTM**, or AARTM. Within the family, one make a choice of subsets C_i, attention coefficients α_{ci}, and combination of regularizers $R(\Phi) = \sum_k \tau_k R_k(\Phi)$. This framework inherits extensibility from ARTM for topic modeling beyond bag-of-words hypothesis.

$$\boxed{\begin{aligned} &p_{ti} = \underset{t\in T}{\operatorname{norm}}\big(\varphi_{tw_i}\theta_{ti}/p(t)\big); \qquad \text{(AARTM)} \\ &\theta_{ti} = \sum_{c\in C_i} \alpha_{ci}\varphi_{tw_c}; \quad p(t) = \sum_{w\in W} \varphi_{tw}p(w); \\ &\varphi_{tw} = \underset{t\in T}{\operatorname{norm}}\left(\sum_{i=1}^{n} p_{ti}[w_i = w] + \varphi_{tw}\frac{\partial R}{\partial \varphi_{tw}}\right). \end{aligned}}$$

Although the AARTM algorithm appears natural compared with its counterpart for bag-of-words models (4)–(6), it is in fact the result of a heuristic simplification of a strictly derived EM algorithm.

Theorem 2 *Let the function $R(\Phi)$ be continuously differentiable. Then a local extremum Φ of problem (11), (12) satisfies the system of equations with auxiliary variables $p_{ti} = p(t \,|\, C_i, w_i)$, $\theta_{ti} = p(t \,|\, C_i)$, N_{tw}, q_{wi}, and n_{tw}:*

$$p_{ti} = \operatorname*{norm}_{t\in T}\big(\varphi_{tw_i}\theta_{ti}/p(t)\big); \tag{13}$$

$$\theta_{ti} = \sum_{c\in C_i} \alpha_{ci}\varphi_{tw_c}; \;\; p(t) = \sum_{w\in W} \varphi_{tw}p(w);$$

$$q_{wi} = \sum_{c\in C_i} \alpha_{ci}[w_c{=}w]; \;\; N_{tw} = \sum_{i=1}^{n} q_{wi}\frac{p_{ti}}{\theta_{ti}}; \tag{14}$$

$$n_{tw} = \sum_{i=1}^{n} p_{ti}[w_i{=}w]; \;\; n_t = \sum_{i=1}^{n} p_{ti};$$

$$\varphi_{tw} = \operatorname*{norm}_{t\in T}\bigg(n_{tw} + \varphi_{tw}\frac{\partial R}{\partial \varphi_{tw}} + \varphi_{tw}N_{tw} - \varphi_{tw}n_t\frac{p(w)}{p(t)}\bigg). \tag{15}$$

The 3rd and 4th terms in the M-step (15) are novel compared to (5) and resemble additive regularization, though no regularizer really corresponds to them. Dropping them reduces the algorithm to (AARTM), but how to justify this?

Corollary 1 *Suppose that the context of a term is the entire document, the attention coefficients are uniformly distributed, and the unbiased estimate $\varphi^*_{tw} = \frac{n_{tw}}{n_w}$ from (7) is substituted into the M-step (15). Then the 3th term equals $\varphi^*_{tw}N_{tw} = n_{tw}$, where $N_{tw} = n_w$ does not depend on t.*

Corollary 2 *If the distribution $p(t)$ is introduced into the model as an external parameter and, instead of (10), is assumed not to depend on the model parameters, then the 4th term equals zero.*

Corollary 3 *If the unbiased estimates $\varphi^*_{tw} = \frac{n_{tw}}{n_w}$ and $p^*(t) = \frac{n_t}{n}$ are substituted into the M-step (15), then the 4th term equals n_{tw}.*

Thus, after substituting the unbiased estimates into the M-step, the 3rd term equals the 1st one if the context equals the document, while the 4th term either cancels with the 1st one, or equals zero if $p(t)$ is treated as an external parameter.

Hence, the 4th term can always be dropped.

The 3rd term is induced by the context. The variable q_{wi} equals the total weight of term w in context C_i. The variable N_{tw} accumulates term weights across all contexts C_i in which the central term w_i is significantly associated with topic t. The value N_{tw} is larger the more frequently term w occurs in the contexts of topic t. In fact, the 3rd term acts as an implicit co-occurrence regularizer. One can expect it to improve topic coherence and interpretability.

A similar distributional semantics mechanism is employed in word2vec language models and their followers (Mikolov et al., 2013; Pennington et al., 2014; Dai et al., 2015; Pagliardini et al., 2017), in word topic model (Chen, 2009), in biterm topic model (Yan et al., 2013), in word network topic model (Zuo et al., 2016) and its ARTM adaptation (Potapenko et al., 2017), and in coherence regularizers (Mimno et al., 2011; Newman et al., 2011).

Updating N_{tw} iterates over all terms in the context, which takes $O(|C_i| \cdot |T|)$ operations. Thus, 3rd term is the most time-consuming part of the M-step. Dropping it makes the algorithm faster.

Corollary 1 leads to $N_{tw}{=}n_w$ regardless of whether the collection is partitioned into documents, paragraphs, or sentences, and whether the documents are long or short. If the context C_i is defined by a wide sliding window around term w_i, and the documents are topically homogeneous, then the equality $N_{tw} \approx n_w$ holds approximately, which makes it possible to drop the 3rd term. In the experimental part of the work we test this hypothesis empirically.

Under the conditions of Corollary 1, the algorithm corresponds to the Θ-less bag-of-words ARTM model (Irkhin et al., 2020), in which the iterative M-step (6) for the matrix Θ is replaced with a simple and fast one-pass averaging of term embeddings over the document. In this way, the constraint $\Theta = f(\Phi)$ plays the role of a regularizer. As a result, overfitting is reduced, and the stability, interpretability, and distinctiveness of topics improves. Similar effects can be expected from AARTM family of models, which is more general.

In this work, we use positional coefficients $\alpha_{ci} = \operatorname*{norm}_{c\in C_i}\big((1-\gamma)^{|i-c|} \cdot [|i-c| \leqslant C]\big)$ with two parameters: the window size C and the effective context size $1/\gamma$ for exponential averaging.

We also consider two variants for handling the central term w_i: *self-aware* ($\alpha_{ii} > 0$) and *self-excluding* ($\alpha_{ii} = 0$).

Table 1: Dataset statistics after preprocessing.

Dataset	Train docs	Test docs	Classes	Vocabulary	Train tokens	Test tokens
20 Newsgroups	10,975	7,290	20	9,348	835,008	513,395
AG News	120,000	7,600	4	10,483	2,664,355	167,299
DBpedia14	559,975	69,997	14	11,329	12,887,091	1,609,248

4 Analogy with Transformer

The local context E-step transforms the context-independent vector $p(t \mid w_i)$ of term w_i into its contextual vector $p(t \mid w_i, C_i)$ in a way similar to neural language models, such as gated convolutional neural network (Dauphin et al., 2017) and query–key–value attention model (Vaswani et al., 2017). The *value* vectors of terms w_c from the context $c \in C_i$ are summed, weighted by attention coefficients α_{ci}, proportional to the proximity between the *key* vector of w_c and the *query* vector of w_i.

Despite the analogy, differences are important.

First, topic vectors are non-negative, normalized, interpretable, and sparse, whereas neural embeddings are unbounded, non-interpretable, and dense.

Second, in the attention model, the query, key, and value vectors are multiplied by learnable matrices of linear transformations. Such parameterization substantially enriches the model. However, our local context topic model is much simpler and has no learnable parameters yet.

Third, in our model, the parameters α_{ci} express only the positional proximity of terms w_c and w_i, while the attention model computes α_{ci} via their semantic similarity. This is partially compensated by the component-wise multiplication of the key vector $p(t \mid w_c)$ by the query vector $p(t \mid w_i)$, which acts as a filter, passing only those topics from w_c, that are significant for w_i. As a result, the sum is formed predominantly from the value vectors of terms, that are thematically close to w_i, which strengthens the analogy with attention model.

Moreover, this mechanism partially justifies the possibility of dropping the context-induced 3rd term from M-step (15), without degradation of the local-context properties of the model.

Continuing to exploit the analogy with attention models, we propose a *looping heuristic* inspired by the looped transformers (Giannou et al., 2023). Taking the vector $p(t \mid w_i, C_i)$ from E-step, we use it instead of $p(t \mid w_i)$, repeating the E-step L times. Below, we test this heuristic in experiments.

5 Experimental Setup

5.1 Datasets

We evaluate the models on three text classification datasets: 20 Newsgroups (Lang, 1995), AG News, and DBpedia14 (Zhang et al., 2015). The vocabulary is learned only from the training split. The preprocessing uses minimum document frequencies 10, 25, and 100 for the datasets accordingly; maximum document frequency 0.1; minimum token length 3; and maximum token length 20. Table 1 reports the corpus statistics after preprocessing.

5.2 Models

We evaluate two variants of our model: the full AARTM and AARTM-no-N, an ablation without the context-induced N_{tw} term in the M-step. We compare them with:

– optimization-based non-negative matrix factorization NMF (Lee and Seung, 1999);

– probabilistic topic models LDA (Blei et al., 2003) from Scikit-learn, biterm topic model BTM (Yan et al., 2013), PLSA as ARTM without regularization from `BigARTM` (Frei and Apishev, 2016);

– neural topic models: BERTopic (Grootendorst, 2022), combined topic model (Bianchi et al., 2021), Contextual-Top2Vec (Angelov and Inkpen, 2024).

By default, we use 100 topics for all models. If the model returns fewer topics after clustering or topic reduction, then we evaluate the actually returned topics. Our models use $C = 100$, $\gamma = 0.01$, one attention loop, self-excluding contexts, 50 iterations, and tolerance 10^{-4}.

5.3 Metrics

Topic coherence. We report NPMI@10 and C_v@10 (Röder et al., 2015) on the top 10 words of each topic. NPMI is computed from document-level word co-occurrence training data. C_v is computed on the tokenized training documents.

Topic diversity. We report TD@25, the fraction of unique words among all top-25 topic words:

$$\mathrm{TD@25} = \Big|\bigcup_{t=1}^{T} \mathcal{W}_t^{(25)}\Big| \;/\; \sum_{t=1}^{T} |\mathcal{W}_t^{(25)}|.$$

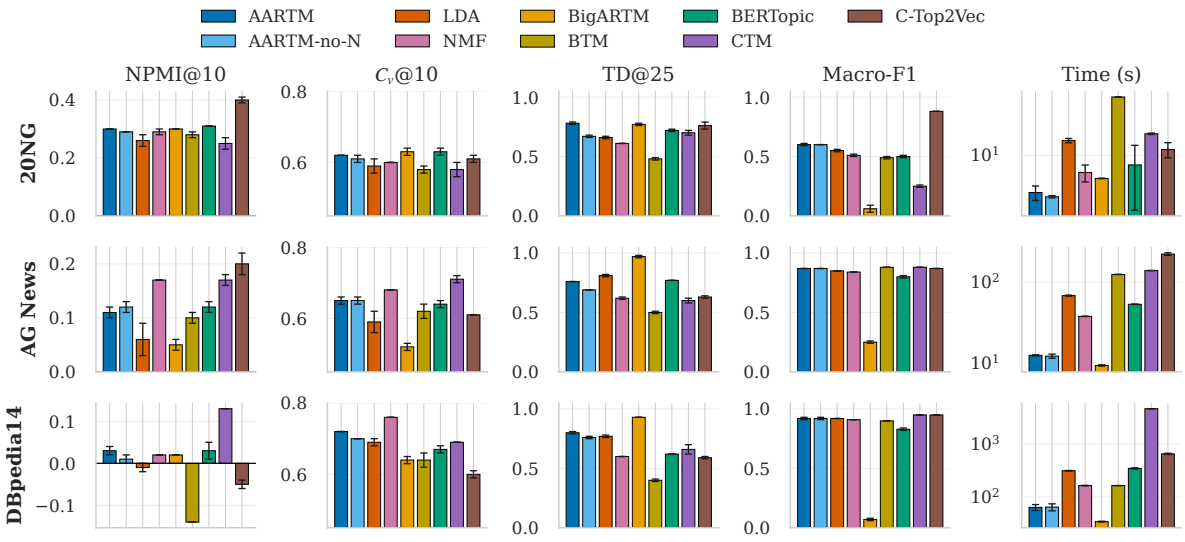

Figure 1: Main results

Document classification. We infer document-level topic vectors and train a logistic regression classifier. For AARTM variants, document vectors are obtained by averaging token-level topic posteriors within each document. We report macro-F1.

Perplexity. For ablations of the local-context models, we report token-level perplexity of the local context topic model (9). When normalized perplexity is shown, values are rescaled within each dataset-specific sweep for visualization.

Runtime. We report wall-clock training time in seconds. Preprocessing pipelines differ across baselines and are not included in runtime calculation.

6 Main Results

Figure 1 shows the main results. AARTM is not uniformly the best method in document classification: pretrained contextual baselines, such as Contextual-Top2Vec, obtain the strongest Macro-F1 on some datasets. This is expected, since AARTM is trained only from the collection.

The topic-quality metrics show a mixed picture. AARTM obtains competitive C_v coherence and high topic diversity, but it does not dominate NPMI; other models can be stronger depending on the dataset. The full AARTM generally produces more diverse topics than AARTM-no-N.

Thus, the advantage of AARTM is the combination of competitive document vectors, local topic assignments, high topic diversity, and low time.

7 Parameter Optimization and Ablation

Context window size. Figure 2 shows that too short contexts lead to worse coherence and perplexity since the model observes not enough co-occurrence data. Increasing the window size improves all metrics, after which the gains saturate. This suggests that the model can benefit from moderately broad window, without precise tuning of its size. Our default recommendation is to use a long but exponentially decayed context with $C = 100$, $\gamma = 0.01$, which gives robust performance across datasets.

Self-aware context. Figure 3 compares self-excluding and self-aware contexts. Including a term in its own context leads to predictive leakage and model overfitting. Perplexity, Macro-F1, and C_v deteriorate, while the topic diversity TD@25 approaches its maximum value, also indicating model degradation. Default recommendation is to use self-excluding contexts.

Number of E-step loops. Figure 4 shows that increasing the number of loops L usually decreases Macro-F1 and does not consistently improve coherence. We have not succeeded in finding the EM operating mode that benefits from loops, inspired by looped transformers (Giannou et al., 2023), not the EM theory. Default recommendation is $L = 1$.

Co-occurrence implicit regularizer. Figure 5 shows the role of the 3rd term of M-step (15). Adding this regularization term improves topic diversity, but has little effect on other metrics. Default recommendation is to drop it and benefit from faster (AARTM) framework.

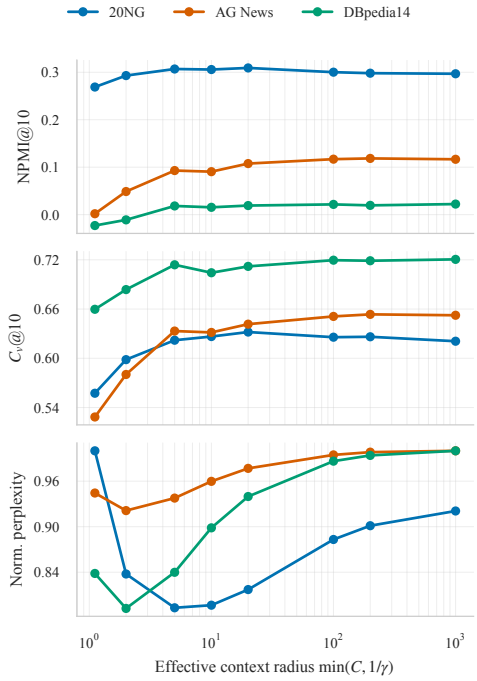

Figure 2: Optimization of context window size

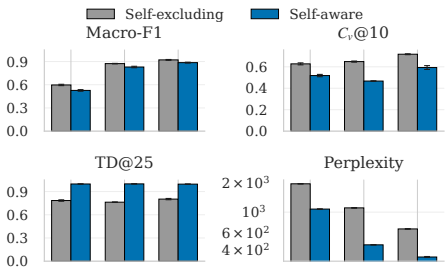

Figure 3: Comparing self-excluding and self-aware

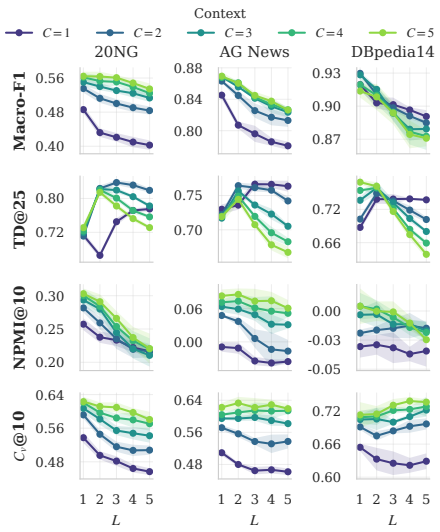

Figure 4: Optimization of the number of E-step loops

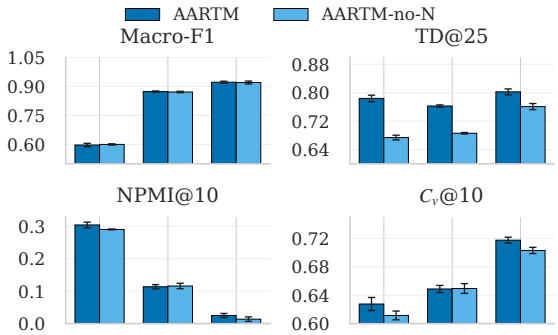

Figure 5: Ablation of co-occurrence implicit regularizer

8 Auxiliary tasks

Robustness to Short Texts. Figure 6 shows how the document classification Macro-F1 metric depends on the document size. Each model is trained on the full training corpus and evaluated on truncated test documents. AARTM degrades more slowly than LDA and NMF. This supports the intuition that the local-context inference captures more useful information from text than the bag-of-words one.

Boundary Detection. The local context topic model produces token-level topic posteriors and can therefore be used for text segmentation tasks that require local topic information. We evaluate this property on a synthetic topic-boundary detection task. Each example is constructed by concatenating two segments sampled from different classes. The true boundary is the concatenation point.

For every candidate boundary j, we compute average topic vectors in fixed windows to the left and to the right of j. The predicted boundary is the position with the largest Hellinger distance between the two window averages. We use 500 synthetic examples per dataset and report mean absolute error and Hit@10.

Figure 7 shows that the proposed model substantially outperforms windowed LDA and NMF representations. The improvement is large across all three datasets, indicating that token-level contextual posteriors capture sharp local topic changes.

9 Conclusion

We introduced a probabilistic topic model of local context with additive regularization and a simple attention mechanism. We derived the EM-like learning algorithm and proved that it has two built-in mechanisms for taking into account the local context. First is weighted averaging of topic vec-

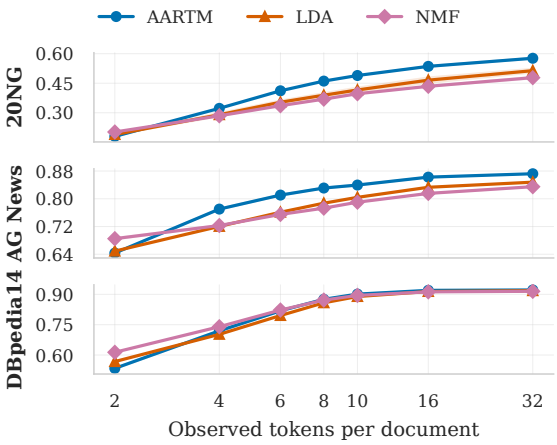

Figure 6: Macro-F1 depends on the document size

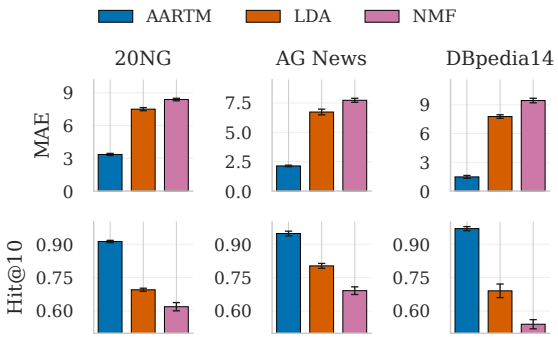

Figure 7: Synthetic boundary detection

tors over the context. Second is smoothing topic vectors for frequently co-occurring pairs of terms. The second mechanism is computationally more expensive, having the form of an optional regularizer. We substantiated, both theoretically and experimentally, that it can be dropped, making the algorithm simpler and faster without significantly reducing the quality of the model. We presented a simplified version, called Attentive ARTM or AARTM, as an extensible platform for topic modeling of sequential text, beyond the bag-of-words assumption.

Our experiments on 20 Newsgroups, AG News, and DBpedia14 have shown that AARTM is competitive with conventional topic models and substantially more local than document-level bag-of-words models. Pretrained contextual baselines can achieve stronger document classification accuracy, but AARTM offers a combination of strengths: low training time, interpretable topic-word distributions, token-level contextual topic assignments, and high extensibility due to additive regularization.

The code of our model and experiments that we report on in this work is available as open source at GitHub.

Future work includes: (a) learnable parametrization of attention weights; (b) combining AARTM with additive regularizers, based on existing AART experience, including many types of aware topic models such as multimodal, multilingual, hierarchical, temporal, semi-supervised, and others; (c) alignment of topic vectors using pre-trained embeddings from large language models while preserving interpretability.

10 Limitations

The proposed model uses fixed attention weights that depend only on token distance. This makes the model simple and efficient, but it also limits its expressiveness: the model cannot learn that some context tokens should be more important than others based on semantic similarity rather than positional proximity.

In this work, we did not study the interaction of local context with additive regularization, leaving this for future research. It seems natural and promising to use regularizers for decorrelation, topic segmentation, and alignment of topic vectors with pre-trained embeddings from LLM.

Finally, our experiments focus on English classification datasets. Additional evaluation on longer multi-topic documents, multimodal and multilingual collections, and human-rated topic interpretability would be needed to fully characterize the model.

References

Dimo Angelov and Diana Inkpen. 2024. Topic modeling: Contextual token embeddings are all you need. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13528–13539. Association for Computational Linguistics.

Georgios Balikas, Massih-Reza Amini, and Marianne Clausel. 2016. On a topic model for sentences. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '16, pages 921–924, New York, NY, USA. ACM.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International*

Joint Conference on Natural Language Processing, Volume 2: Short Papers, pages 759–766.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2001. Latent dirichlet allocation. In *Advances in Neural Information Processing Systems 14, NIPS 2001, December 3–8, 2001, Vancouver, British Columbia, Canada]*, pages 601–608. MIT Press.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Berlin Chen. 2009. Word topic models for spoken document retrieval and transcription. 8(1):2:1–2:27.

Rob Churchill and Lisa Singh. 2022. The evolution of topic modeling. *ACM Comput. Surv.*, 54(10s).

Andrew M. Dai, Christopher Olah, and Quoc V. Le. 2015. Document embedding with paragraph vectors. In *NIPS Deep Learning Workshop*.

Yann N. Dauphin, Angela Fan, Michael Auli, and David Grangier. 2017. Language modeling with gated convolutional networks. In *Proceedings of the 34th International Conference on Machine Learning – Volume 70*, ICML'17, pages 933–941. JMLR.org.

Oleksandr Frei and Murat Apishev. 2016. Parallel non-blocking deterministic algorithm for online topic modeling. In *AIST'2016, Analysis of Images, Social networks and Texts*, volume 661, pages 132–144. Springer International Publishing Switzerland, Communications in Computer and Information Science (CCIS).

Angeliki Giannou, Shashank Rajput, Jy-Yong Sohn, Kangwook Lee, Jason D. Lee, and Dimitris Papailiopoulos. 2023. Looped transformers as programmable computers. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 11398–11442. PMLR.

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

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, New York, NY, USA. ACM.

Ilya Irkhin, Viktor Bulatov, and Konstantin Vorontsov. 2020. Additive regularization of topic models with fast text vectorization. *Computer Research and Modeling*, 12(6):1515–1528.

Ilya Irkhin and Konstantin Vorontsov. 2020. Convergence of the algorithm of additive regularization of topic models. *Trudy Instituta Matematiki i Mekhaniki UrO RAN*, 26(3):56–68.

Ken Lang. 1995. Newsweeder: Learning to filter netnews. In *Proceedings of the 12th International Conference on Machine Learning*, pages 331–339.

Daniel D. Lee and H. Sebastian Seung. 1999. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

David Mimno, Hanna M. Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP'11, pages 262–272, Stroudsburg, PA, USA. Association for Computational Linguistics.

David Newman, Edwin V. Bonilla, and Wray L. Buntine. 2011. Improving topic coherence with regularized topic models. In *Advances in Neural Information Processing Systems 24*, pages 496–504.

Matteo Pagliardini, Prakhar Gupta, and Martin Jaggi. 2017. Unsupervised learning of sentence embeddings using compositional n-gram features. *CoRR*, abs/1703.02507.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Anna Potapenko, Artem Popov, and Konstantin Vorontsov. 2017. Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks. In *Communications in Computer and Information Science, vol 789. AINL-6: Artificial Intelligence and Natural Language Conference, St. Petersburg, Russia, September 20-23, 2017*, pages 167–180. Springer, Cham.

Marziea Rahimi, Morteza Zahedi, and Hoda Mashayekhi. 2022. A probabilistic topic model based on short distance co-occurrences. *Expert Systems with Applications*, 193:116518.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Konstantin Vorontsov. 2014. Additive regularization for topic models of text collections. *Doklady Mathematics*, 89(3):301–304.

Konstantin Vorontsov. 2023. Rethinking probabilistic topic modeling from the point of view of classical non-bayesian regularization. In *Data Analysis and Optimization*, pages 397–422. Springer.

Konstantin Vorontsov and Anna Potapenko. 2015. Additive regularization of topic models. *Machine Learning, Special Issue on Data Analysis and Intelligent Optimization with Applications*, 101(1):303–323.

Hanna M. Wallach. 2006. Topic modeling: Beyond bag-of-words. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 977–984, New York, NY, USA. ACM.

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A biterm topic model for short texts. In *Proceedings of the 22-nd International Conference on World Wide Web*, WWW'13, pages 1445–1456, Geneva, Switzerland. International World Wide Web Conferences Steering Committee.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in Neural Information Processing Systems 28*.

Yuan Zuo, Jichang Zhao, and Ke Xu. 2016. Word network topic model: A simple but general solution for short and imbalanced texts. *Knowledge and Information Systems*, 48(2):379–398.

A Proofs

Lemma 1 *(Vorontsov, 2023). Let the function $f(\Omega)$ be continuously differentiable with respect to a set of vectors $\Omega = (\omega_j)_{j\in J}$, $\omega_j = (\omega_{ij})_{i\in I_j}$. If ω_j is a local extremum vector of the mathematical programming task*

$$\max\colon f(\Omega),\ \sum_{i\in I_j}\omega_{ij} = 1,\ \omega_{ij} \geqslant 0,\ i \in I_j,\ j \in J$$

and if $\omega_{ij}\frac{\partial f}{\partial \omega_{ij}} > 0$ for some i, then ω_j satisfies the equations

$$\omega_{ij} = \operatorname*{norm}_{i\in I_j}\left(\omega_{ij}\frac{\partial f}{\partial \omega_{ij}}\right). \tag{16}$$

Proof of Theorem 2 is based on applying Lemma 1 to problem (11), (12).

Let us write the functional to be maximized, replacing the index t with s:

$$L(\Phi) = \sum_{i=1}^{n}\ln\left(\sum_{s\in T}\varphi_{sw_i}\frac{p(w_i)}{p(s)}\theta_{si}\right) + R(\Phi).$$

First, compute three useful partial derivatives:

$$\frac{\partial\varphi_{sw_i}}{\partial\varphi_{tw}} = [w_i{=}w][t{=}s];$$

$$\frac{\partial\theta_{si}}{\partial\varphi_{tw}} = \frac{\partial}{\partial\varphi_{tw}}\left(\sum_{c\in C_i}\alpha_{ci}\varphi_{sw_c}\right)$$

$$= \sum_{c\in C_i}\alpha_{ci}[w_c{=}w][t{=}s] = q_{wi}[t{=}s];$$

$$\frac{\partial p(s)}{\partial\varphi_{tw}} = \frac{\partial}{\partial\varphi_{tw}}\left(\sum_{v\in W}\varphi_{sv}p(v)\right) = p(w)[t{=}s].$$

To apply the lemma, let us find the partial derivatives of $L(\Phi)$, using the differentiation rule $\left(\frac{ab}{c}\right)' = \frac{a'b}{c} + \frac{ab'}{c} - \frac{abc'}{c^2}$:

$$\varphi_{tw}\frac{\partial L}{\partial\varphi_{tw}} - \varphi_{tw}\frac{\partial R}{\partial\varphi_{tw}} =$$

$$= \sum_{i=1}^{n}\frac{p(w_i)}{p(w_i\,|\,C_i)}\varphi_{tw}\sum_{s\in T}\frac{\partial}{\partial\varphi_{tw}}\left(\frac{\varphi_{sw_i}\theta_{si}}{p(s)}\right)$$

$$= \sum_{i=1}^{n}\frac{p(w_i)}{p(w_i\,|\,C_i)}\varphi_{tw}$$

$$\times\left(\frac{\theta_{ti}[w_i{=}w]}{p(t)} + \frac{\varphi_{tw_i}q_{wi}}{p(t)} - \frac{\varphi_{tw_i}\theta_{ti}p(w)}{p^2(t)}\right)$$

$$= \sum_{i=1}^{n}\frac{\varphi_{tw_i}\theta_{ti}p(w_i)}{p(w_i\,|\,C_i)\,p(t)}$$

$$\times\left([w_i{=}w] + \varphi_{tw}\frac{q_{wi}}{\theta_{ti}} - \varphi_{tw}\frac{p(w)}{p(t)}\right)$$

$$= \sum_{i=1}^{n}p_{ti}\left([w_i{=}w] + \varphi_{tw}\frac{q_{wi}}{\theta_{ti}} - \varphi_{tw}\frac{p(w)}{p(t)}\right)$$

$$= n_{tw} + \varphi_{tw}N_{tw} - \varphi_{tw}n_t\frac{p(w)}{p(t)}.$$

Collecting all the obtained equations into a system yields the statement of the theorem.

The theorem is proved.

Proof of Corollary 1. Denote the sequence of term indices of document d in the continuous numbering by $I_d = [i_d, \ldots, i_d + n_d - 1]$.

By assumption, $\alpha_{ci} = \frac{1}{n_d}[c \in I_d]$.

Then, the topic vector θ_{ti} of the local context equals the topic vector θ_{td} of the document d:

$$\theta_{ti} = \sum_{c\in C_i}\alpha_{ci}\varphi_{tw_c} = \frac{1}{n_d}\sum_{c\in C_i}[c \in I_d]\varphi_{tw_c}$$

$$= \frac{1}{n_d}\sum_{w\in d}n_{dw}\varphi_{tw} = \frac{n_{td}}{n_d} = \theta_{td}.$$

Table 2: Example topics from AARTM.

Dataset	Topic label	Top words
20NG	religion	church, authority, catholic, pope, tradition, orthodox, modern, churches
20NG	science	science, theory, scientific, test, universe, physical, results, scientists, physics
20NG	finances	money, cost, pay, tax, million, insurance, costs, billion, dollars, spend
AG News	internet	internet, service, online, users, access, mail, customers
AG News	countries	india, pakistan, indian, kashmir, pakistani, delhi, musharraf, sri, lanka
AG News	air transportation	air, airlines, airline, boeing, airways, airport, flight, aircraft, plane
DBpedia14	literature	novel, book, published, written, story, fiction, first, author, short
DBpedia14	films	film, directed, starring, drama, produced, comedy, stars, written, based
DBpedia14	medicine	hospital, medical, health, center, medicine, care, memorial, cancer, doctor

Then, q_{wi} equals the frequency estimate of the probability of term w in the document:

$$q_{wi} = \sum_{c \in C_i} \alpha_{ci} [w_c = w] = \frac{1}{n_d} \sum_{i \in I_d} [w_i = w] = \frac{n_{dw}}{n_d}.$$

Then, N_{tw}, after substituting $n_{td} = \sum_{i \in I_d} p_{ti}$, equals n_w, losing its dependence on topic t:

$$\begin{aligned} N_{tw} &= \sum_{i=1}^{n} q_{wi} \frac{p_{ti}}{\theta_{ti}} \\ &= \sum_{d \in D} \frac{n_{dw}}{n_d} \left(\sum_{i \in I_d} p_{ti} \right) \frac{n_d}{n_{td}} = n_w. \end{aligned}$$

Multiplying N_{tw} by the unbiased estimate $\varphi^*_{tw} = \frac{n_{tw}}{n_w}$, we get the equality $\varphi^*_{tw} N_{tw} = n_{tw}$, which completes the proof.

Proof of Corollary 2 repeats, step by step, the proof of Theorem 2.

Proof of Corollary 3 follows immediately from substitution of the unbiased estimates:

$$\varphi^*_{tw} n_t \frac{p(w)}{p^*(t)} = \frac{n_{tw}}{n_w} n_t \frac{n_w}{n_t} = n_{tw}.$$

B Self-Aware Context and Posterior Sharpening

Let p_i be the current topic distribution for token i, q_i be the neighbor-only context distribution, and λ be the normalized self-context weight. Then

$$\theta_i = q_i, \qquad \text{self-excluding} \qquad (17)$$

$$\theta_i = \lambda p_i + (1 - \lambda) q_i, \qquad \text{self-aware} \qquad (18)$$

The update is:

$$p^{\text{new}}_{it} \propto \frac{p_{it} \theta_{it}}{n_t} \qquad (19)$$

Compare the two topics a and b. Without self-awareness:

$$\frac{p^{\text{new}}_{ia}}{p^{\text{new}}_{ib}} = \frac{p_{ia}}{p_{ib}} \frac{q_{ia}}{q_{ib}} \frac{n_b}{n_a} \qquad (20)$$

With self-awareness:

$$\frac{p^{\text{new}}_{ia}}{p^{\text{new}}_{ib}} = \frac{p_{ia}}{p_{ib}} \frac{\lambda p_{ia} + (1 - \lambda) q_{ia}}{\lambda p_{ib} + (1 - \lambda) q_{ib}} \frac{n_b}{n_a} \qquad (21)$$

Now compare the self-aware ratio to the self-excluding ratio:

$$\frac{R_{\text{aw}}}{R_{\text{ex}}} = \frac{1 - \lambda + \lambda r_{ia}}{1 - \lambda + \lambda r_{ib}}, \quad r_{it} = \frac{p_{it}}{q_{it}} \qquad (22)$$

Therefore $R_{\text{aw}} > R_{\text{ex}}$ whenever $r_{ia} > r_{ib}$.

If topic a is already dominating in the token compared to its surrounding context, then self-aware context increases the odds of topic a against topic b even more. This increases topic odds for already-favored topics and decreases them for weak topics.

C Qualitative Analysis

Table 2 shows examples of topics produced by the attentive model. We checked whether topics are interpretable by human inspection. Here, we selected a few topics from one run on each dataset and assigned a topic label to each topic manually.

D Runtime

All runtime experiments were run on Nvidia GeForce RTX 4090, with 24Gb VRAM, JAX version 0.10.0, and CUDA version 13.2.1-1.

Table 3 reports runtime benchmarks for 50 training iterations with batch size 10,000 on 20 Newsgroups. We separately report the first compiled run and steady runs after warm-up. The no-N_{tw} variant

Table 3: Runtime benchmark

Model	T	C	Compile +run, sec	Steady run, sec	Tokens/sec
With N_{wt}	10	10	1.5	1.4 ± 0.02	≈650,000
	10	100	1.7	1.6 ± 0.01	≈574,000
	10	1000	5.4	5.3 ± 0.03	≈180,000
	100	10	1.9	1.8 ± 0.01	≈514,000
	100	100	4.3	4.2 ± 0.02	≈227,000
	100	1000	96.3	95.8 ± 0.63	≈10,100
No N_{wt}	10	10	1.4	1.3 ± 0.01	≈700,000
	10	100	1.4	1.3 ± 0.01	≈700,000
	10	1000	2.9	2.3 ± 0.02	≈337,000
	100	10	1.7	1.6 ± 0.01	≈574,000
	100	100	3.6	3.1 ± 0.02	≈271,000
	100	1000	76.2	75.7 ± 0.2	≈12,900

is consistently faster because it does not compute the transposed attention statistic required for N_{tw}.

Runtime increases with both the number of topics and the context radius, as expected from the finite-window implementation. This differs from the exponential moving-average implementation, which can compute attention in time linear in sequence length and independent of an explicit window radius. In this work we use finite windows because they allow direct control over the context size and straightforward batching.