Is Neural Topic Modelling Better than Clustering? An Empirical Study on Clustering with Contextual Embeddings for Topics

Anonymous ACL submission

Abstract

Recent work incorporates pre-trained word embeddings such as BERT embeddings into Neural Topic Models (NTMs), generating highly coherent topics. However, with high-quality contextualized document representations, do we really need sophisticated neural models to obtain coherent and interpretable topics? In this paper, we conduct thorough experiments showing that directly clustering high-quality sentence embeddings with an appropriate word selecting method can generate more coherent and diverse topics than NTMs, achieving also higher efficiency and simplicity.

1 Introduction

Topic modelling is an unsupervised method to uncover latent semantic themes among documents (Boyd-Graber et al., 2017). Neural topic models (NTMs) (Miao et al., 2016; Srivastava and Sutton, 2017) incorporating neural components have significantly advanced the modelling results than the traditional Latent Dirichlet Allocation (LDA; Blei et al. 2003). Later, contextualized word and sentence embeddings produced by pre-trained language models such as BERT (Devlin et al., 2019) have demonstrated the state-of-the-art results in multiple Natural Language Processing (NLP) tasks (Xia et al., 2020), which attracts attentions from the topic modelling community. Recent work has successfully incorporated these contextualized embeddings into NTMs, showing improved topic coherence than conventional NTMs that use Bag-of-Words (BoW) as document representations (Bianchi et al., 2021a,b; Jin et al., 2021). Despite the promising performance, existing NTMs are generally based on a variational autoencoder framework (VAE; Kingma and Welling 2013), which suffers from hyper-parameters tuning and computational overheads (Zhao et al., 2021). Moreover, the integration of the pre-trained embeddings to the standard VAE framework adds additional model complexity. With high-quality contextualized document representations, do we really need sophisticated NTMs to obtain coherent and interpretable topics?

Recent work (Aharoni and Goldberg, 2020; Sia et al., 2020; Thompson and Mimno, 2020) has shown that directly congregating contextualized embeddings can get semantically similar word or document clusters. Specifically, Sia et al. (2020) cluster *vocabulary-level* word embeddings and obtain top words from each cluster using weighing and re-ranking, while Thompson and Mimno (2020) consider polysemy and perform *token-level* clustering. However, the use of term frequency (TF) to select topic words fails to capture the semantics of clusters precisely because words with high frequency may be common across different clusters. In addition, they only compare the performance with the traditional LDA while ignoring the promising NTMs proposed recently.

Is neural topic modelling better than simple embedding clustering? This work compares the performance of NTMs and contextualized embedding-based clustering systematically. We employ a straightforward framework for clustering. In addition, we explore different strategies to select topic words for clusters. We evaluate our approach on three datasets with various text lengths.

Our contributions are as follows: First, we find that directly clustering high-quality sentence embeddings can generate as good topics as NTMs, providing a simple and efficient solution to uncover latent topics among documents. Second, we propose a new topic word selecting method, which is the key to producing highly coherent and diverse topics. Third, we show that the clustering-based model is robust to the length of documents and the number of topics. Reducing the embedding dimensionality negligibly affects the performance but saves runtime.

2 Models

This study compares embedding clustering-based models with LDA and a series of existing NTMs as follows. Implementation details are supplied in Appendix A.

LDA (Blei et al., 2003): the representative traditional topic model in history.

ProdLDA (Srivastava and Sutton, 2017): a prominent NTM that employs the VAE to reconstruct the BoW representation.

CombinedTM (Bianchi et al., 2021a): extends ProdLDA by concatenating the contextualized SBERT (Reimers and Gurevych, 2019) embeddings with the original BoW as the new input to feed into the VAE framework.

ZeroShotTM (Bianchi et al., 2021b): also builds upon ProdLDA, but it replaces the original BoW with SBERT embeddings entirely.

BERT+KM (Sia et al., 2020): a clustering-based method that uses TF to weight and re-rank words to obtain topic words.

Our Methods: we use a simple clustering framework with contextualized embeddings for topic modelling, as shown in Figure 1. We first encode pre-processed documents to obtain contextualized sentence embeddings through pre-trained language models. After that, we lower the dimension of the embeddings before applying clustering methods (e.g., K-Means; KM) to group similar documents. Each cluster will be regarded as a topic. Finally, we adopt a weighting method to select representative words as topics.

We believe that high-quality document embeddings are critical for clustering-based topic modelling. We thus experiment with different embeddings including BERT, RoBERTa (Liu et al., 2019), and SBERT. We also adopt SimCSE (Gao et al., 2021), a recently proposed sentence embeddings of contrastive learning, that has shown the state-of-the-art performance on multiple semantic textual similarity tasks. Both supervised and unsupervised SimCSE are investigated in our experiment (e.g., Table 2).

Pre-trained contextualized sentence embeddings often have high dimensionalities. To reduce the computational cost, we apply the Uniform Manifold Approximation Projection (UMAP) (McInnes et al., 2018) in our implementation to reduce the dimensionality while maintaining the essential information of the embeddings. We find that reducing dimensionality before clustering has a negligible impact on performance (Section 4.4).

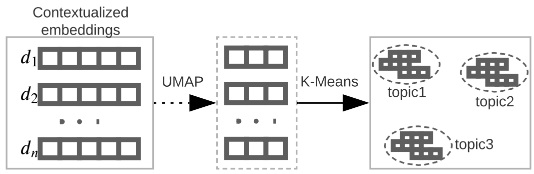

Figure 1: Architecture of our method. Reducing embedding dimension is optional but can save runtime (see Section 4.4).

We cluster the dimension-reduced sentence embeddings using K-Means because of its efficiency and simplicity. Semantically close documents are gathered together, and each cluster is supposed to represent a topic.

3 Topic Words for Clusters

Once we have a group of clustered documents, selecting representative topic words is vital to identify semantics of topics. Inspired by Term Frequency-Inverse Document Frequency (TFIDF) (Ramos et al., 2003), we explore several weighting metrics to obtain topic words in clusters. Let $n_{t,d}$ be the frequency of word t in document d, $\sum_{t'} n_{t',d}$ be the total words' frequency in the document, and D be the entire corpus. TFIDF is defined as $\mathbf{TFIDF} = \frac{n_{t,d}}{\sum_{t'} n_{t',d}} \cdot \log\left(\frac{|D|}{|\{d \in D : t \in d\}|}\right)$. While capturing the word importance across the entire corpus, TFIDF ignores that semantically similar documents have been grouped together. To address this issue, we consider two alternative strategies. First, we concatenate the documents within a cluster to be a single long document and calculate the term frequency of each word in each cluster:

$$\mathbf{TF_i} = \frac{n_{t,i}}{\sum_{t'} n_{t',i}} \tag{1}$$

where $n_{t,i}$ is the frequency of word t in cluster i, $\sum_{t'} n_{t',i}$ is the total word frequency in the cluster. Second, for each cluster i, we apply TFIDF:

$$\mathbf{TFIDF_i} = \frac{n_{t,d_i}}{\sum_{t'} n_{t',d_i}} \cdot \log\left(\frac{|D_i|}{|\{d \in D_i : t \in d\}|}\right) \tag{2}$$

where n_{t,d_i} denotes the frequency of word t in document d, which is in cluster i, and $|D_i|$ is the number of documents in cluster i.

Besides the two local cluster-based strategies, we further incorporate the global word importance with local term frequency within each cluster:

$$\mathbf{TFIDF} \times \mathbf{TF_i} = \mathbf{TFIDF} \cdot \mathbf{TF_i} \tag{3}$$

and we combine the global word importance with term frequency across clusters:

$$\mathbf{TFIDF} \times \mathbf{IDF_i} = \mathbf{TFIDF} \cdot \log \left(\frac{|K|}{|\{t \in K\}|} \right) \quad (4)$$

where $|K|$ is the number of clusters and $|\{t \in K\}|$ is the number of clusters that word t appears.

4 Experiments

4.1 Datasets

We adopt three datasets of various text lengths in our experiments, namely 20Newsgroups[1], M10 (Lim and Buntine, 2015), and BBC News (Greene and Cunningham, 2006). We follow OCTIS (Terragni et al., 2021) to pre-process these raw datasets. The statistics of the datasets are shown in Table 1.

Dataset	D	V	L	N_d
20Newsgroups	16,309	1,612	20	48
M10	8,355	1,696	10	5.9
BBC News	2,225	2,949	5	120

Table 1: Statistics of the pre-processed datasets, where D denotes the total number of documents, V denotes the vocabulary size, L denotes the number of corpus categories, and N_d denotes the average number of words per document.

4.2 Evaluation Metrics

We evaluate the topic quality in terms of both topic diversity and topic coherence: Topic Diversity (*TU*) (Nan et al., 2019) measures the uniqueness of the words across all topics; Normalized Pointwise Mutual Information (*NPMI*) (Newman et al., 2010) measures topic coherence internally using a sliding window to count word co-occurrence patterns; Topic Coherence (C_V) (Röder et al., 2015) is a variant of *NPMI* that uses the one-set segmentation to count word co-occurrences and the cosine similarity as the similarity measure.

4.3 Results & Analysis

We report the main results in Table 2. For the complete results using different embeddings, please refer to Appendix B.

Directly clustering high-quality sentence embeddings can generate good topics. From Table 2, it can be observed that SBERT and SimCSE-based clustering models achieve the best averaged topic coherence among the three datasets while maintaining remarkable topic diversities. Conversely, clustering RoBERTa achieves similar or worse results than contextualized NTMs. The results suggest that contextualized embeddings are essential to get high-quality topics.

Topic words weighting method is vital. We can see in Figure 2 that inappropriate word selecting methods ($\mathbf{TFIDF} \times \mathbf{TF_i}$ and $\mathbf{TF_i}$) lead to worse topic coherence than the contextualized NTMs (i.e., CombinedTM and ZeroShotTM), and even the BoW-based ProdLDA. Moreover, from Table 2, BERT+KM adopt TF to obtain top words for each cluster, which ignores that the words may also be prevalent in other clusters, thus having poor topic diversities. Instead, our proposed method, $\mathbf{TFIDF} \times \mathbf{IDF_i}$, considers the locally important words and globally infrequent words at the same time. We provide more comparison of the word selecting methods in Section 4.4.

Clustering-based topic models are robust to various lengths of documents. From Table 2 and Figure 2, we find that clustering-based models with high-quality embeddings (SBERT and SimCSE) consistently perform better than conventional LDA and NTMs, especially on the short text dataset M10, even with different word selecting methods.

4.4 Ablation Studies

We further investigate the impact of the topic word selecting methods, different embedding dimensionalities, as well as the topic numbers.

Topic word selecting methods. Table 3 shows the comparison between different word weighting methods. $\mathbf{TFIDF} \times \mathbf{IDF_i}$ achieves significantly better results among all methods. This indicates that $\mathbf{TFIDF}$ marks out the important words to each document in the entire corpus, while $\mathbf{IDF_i}$ penalizes the common words in multiple clusters. Conversely, the other three methods ignore that frequent words in a cluster may also be prevalent in other clusters, hence selecting such words leading to low topic diversities. A further analysis in Appendix C also supports the observation.

Embedding dimensionality reduction. We apply UMAP to reduce the dimensionality of the sentence embeddings before clustering. As shown in Figure 3, the embeddings dimensionality negligibly affects topic quality for all word selecting methods. However, reducing to a lower dimensionality decreases the computational runtime (Appendix D).

[1] http://qwone.com/~jason/20Newsgroups/

	20Newsgroups			M10			BBC News		
Model	*TU*	*NPMI*	C_V	*TU*	*NPMI*	C_V	*TU*	*NPMI*	C_V
LDA	0.717	0.040	0.511	0.681	-0.177	0.336	0.312	-0.014	0.357
ProdLDA	0.736	0.045	0.574	0.650	-0.260	0.432	0.702	-0.044	0.540
CombinedTM(SBERT$_{\texttt{base}}$)	0.700	0.065	0.601	0.581	0.001	0.443	0.606	0.042	0.639
ZeroShotTM(SBERT$_{\texttt{base}}$)	**0.729**	0.069	0.614	0.633	-0.056	0.433	0.699	-0.050	0.531
BERT$_{\texttt{base}}$+KM†	0.346	0.065	0.521	0.484	0.116	0.588	0.529	0.111	0.637
BERT$_{\texttt{base}}$*	0.562	0.118	0.649	0.763	0.146	0.725	0.689	0.129	0.700
RoBERTa$_{\texttt{large}}$*	0.404	0.014	0.440	0.669	0.001	0.506	0.673	0.046	0.555
BERT$_{\texttt{base}}$+UMAP*	0.589	0.128	0.671	0.794	0.159	0.706	0.716	0.135	0.716
RoBERTa$_{\texttt{large}}$+UMAP*	0.463	0.054	0.499	0.636	0.046	0.513	0.706	0.077	0.632
SBERT$_{\texttt{base}}$*	0.668	0.126	0.658	0.832	0.164	0.742	0.727	0.137	0.719
SRoBERTa$_{\texttt{base}}$*	0.670	0.128	0.654	0.815	0.149	0.713	0.719	0.131	0.699
SBERT$_{\texttt{base}}$+UMAP*	0.679	0.139	0.690	0.841	0.192	0.715	**0.749**	0.142	**0.730**
SRoBERTa$_{\texttt{base}}$+UMAP*	0.680	0.138	0.684	0.830	0.192	0.722	0.747	0.135	0.716
Unsup-SimCSE(BERT$_{\texttt{base}}$)*	0.677	0.147	0.694	0.831	0.180	**0.750**	0.730	0.142	0.722
Unsup-SimCSE(BERT$_{\texttt{base}}$)+UMAP*	0.692	0.139	0.685	**0.851**	**0.206**	0.744	0.733	**0.146**	0.729
Sup-SimCSE(BERT$_{\texttt{base}}$)*	0.721	**0.151**	**0.702**	0.829	0.180	0.746	0.736	0.143	0.720
Sup-SimCSE(BERT$_{\texttt{base}}$)+UMAP*	0.714	0.146	0.698	0.815	0.202	0.730	0.739	0.143	0.724

Table 2: Topic coherence (*NPMI* and C_V) and topic diversity (*TU*) of the top 10 words. All results are averaged across the 5 settings of topic number ($K = \{\text{ground truth}, 25, 50, 75, 100\}$). Best results are in bold. †: we use the method from (Sia et al., 2020). $*$: our methods adopt $\mathbf{TFIDF} \times \mathbf{IDF_i}$ (Eq. 4) to select topic words. Dimensionality: `base`: 768, `large`: 1024.

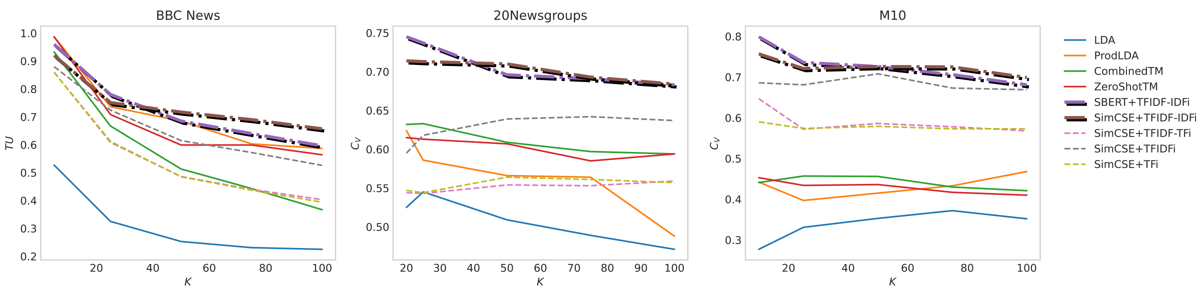

Figure 2: Topic coherence (C_V) and diversity (*TU*) of different models over different topic number K. Cluster models use SBERT$_{\texttt{base}}$+UMAP and Sup-SimCSE(BERT$_{\texttt{base}}$)+UMAP.

Method	Avg *TU*	Avg *NPMI*	Avg C_V
$\mathbf{TF_i}$	0.442	0.081	0.555
$\mathbf{TFIDF_i}$	0.508	0.110	0.626
$\mathbf{TFIDF} \times \mathbf{TF_i}$	0.438	0.078	0.551
$\mathbf{TFIDF} \times \mathbf{IDF_i}$	**0.689**	**0.145**	**0.702**

Table 3: Comparison between different topic word selecting methods on 20Newsgroups using Unsup-SimCSE(RoBERTa$_{\texttt{base}}$)+UMAP with $K = 30$.

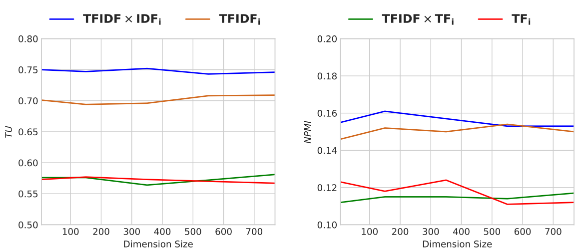

Figure 3: Topic coherence and diversity over different embedding dimensions on BBC News using Unsup-SimCSE(RoBERTa$_{\texttt{base}}$)+UMAP with $K = 30$.

Topic numbers K. We investigate the impact of the different number of topics K on the performance of the models. Figure 2 plots the trends of *TU* and C_V on three datasets. We observe that the *TU* of clustering-based topic models, especially the models using $\mathbf{TFIDF} \times \mathbf{IDF_i}$, decrease slowly compared to others when K increases. The similar trend can be observed for topic coherence, while the C_V of LDA and NTMs either fluctuates significantly or stays at a low level.

5 Conclusion

We conduct a thorough empirical study to show that a clustering-based method can generate commendable topics as long as high-quality contextualized sentence embeddings are used, together with an appropriate topic word selecting strategy. Compared to neural topic models, clustering-based models are more simple, efficient and robust to various document lengths and topic numbers.

References

Roee Aharoni and Yoav Goldberg. 2020. Unsupervised domain clusters in pretrained language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7747–7763, Online. Association for Computational Linguistics.

Federico Bianchi, Silvia Terragni, and Dirk Hovy. 2021a. 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, Online. Association for Computational Linguistics.

Federico Bianchi, Silvia Terragni, Dirk Hovy, Debora Nozza, and Elisabetta Fersini. 2021b. Cross-lingual contextualized topic models with zero-shot learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1676–1683, Online. Association for Computational Linguistics.

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

Jordan L Boyd-Graber, Yuening Hu, David Mimno, et al. 2017. *Applications of topic models*, volume 11. Now Publishers Incorporated.

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.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*.

Derek Greene and Pádraig Cunningham. 2006. Practical solutions to the problem of diagonal dominance in kernel document clustering. In *Proceedings of the 23rd international conference on Machine learning*, pages 377–384.

Yuan Jin, He Zhao, Ming Liu, Lan Du, and Wray Buntine. 2021. Neural attention-aware hierarchical topic model. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1042–1052, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Kar Wai Lim and Wray Buntine. 2015. Bibliographic analysis with the citation network topic model. In *Proceedings of the Sixth Asian Conference on Machine Learning*, volume 39 of *Proceedings of Machine Learning Research*, pages 142–158, Nha Trang City, Vietnam. PMLR.

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. *arXiv preprint arXiv:1907.11692*.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

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

Feng Nan, Ran Ding, Ramesh Nallapati, and Bing Xiang. 2019. Topic modeling with Wasserstein autoencoders. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6345–6381, Florence, Italy. Association for Computational Linguistics.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108, Los Angeles, California. Association for Computational Linguistics.

Juan Ramos et al. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer.

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

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

Suzanna Sia, Ayush Dalmia, and Sabrina J. Mielke. 2020. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1728–1736, Online. Association for Computational Linguistics.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. *arXiv preprint arXiv:1703.01488*.

Silvia Terragni, Elisabetta Fersini, Bruno Giovanni Galuzzi, Pietro Tropeano, and Antonio Candelieri. 2021. OCTIS: Comparing and optimizing topic models is simple! In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 263–270, Online. Association for Computational Linguistics.

Laure Thompson and David Mimno. 2020. Topic modeling with contextualized word representation clusters. *arXiv preprint arXiv:2010.12626*.

Patrick Xia, Shijie Wu, and Benjamin Van Durme. 2020. Which *BERT? A survey organizing contextualized encoders. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7516–7533, Online. Association for Computational Linguistics.

He Zhao, Dinh Phung, Viet Huynh, Yuan Jin, Lan Du, and Wray Buntine. 2021. Topic modelling meets deep neural networks: A survey. *arXiv preprint arXiv:2103.00498*.

A Configuration Details

We implement LDA and NTMs based on OCTIS (Terragni et al., 2021) [2] and use their default settings. Specifically, ProdLDA, CombinedTM, and ZeroShotTM share the same configurations, i.e. one hidden layer with 100 neurons, ADAM optimizer and Momentum as 0.99; we randomly dropout 20% hidden units; we run 100 epochs of each model, and the batch size is 64. For BERT+KM, we follow Sia et al. (2020) by reducing embedding dimension to 50 using Principal Component Analysis (PCA) and adopting TF to select words. For our methods, we reduce embedding dimension to 5 using UMAP. We use BERT, RoBERTa, and SBERT embeddings provided by HuggingFace [3], and SimCSE embeddings provided from its official Github [4].

B Complete Results

We present the complete comparison between different contextualized embeddings in Table 4.

C Comparison of Topic Words

We run Sup-SimCSE(RoBERTa$_{\texttt{base}}$)+UMAP on 20Newsgroup and show the differences of topic diversities produced by distinct word selecting methods in Table 5. It is clear that $\mathbf{TFIDF_i}$ and $\mathbf{TF_i}$ tend to choose common words across multiple topics.

D Runtime

We compare the model runtime between the contextualized NTM CombinedTM and clustering-based models. We reduce the dimensionality of the sentence embeddings to 50 using UMAP. All models run on NVIDIA T4 GPU. Results are in Table 6.

Model	Runtime
CombinedTM	149s
SBERT(BERT$_{\texttt{base}}$)	113s
SBERT(BERT$_{\texttt{base}}$)+UMAP to dim=50	101s

Table 6: Runtime comparison on 20Newsgroups with $K = 30$. Results are averaged across 5 runs.

[2] https://github.com/MIND-Lab/OCTIS
[3] https://huggingface.co/models
[4] https://github.com/princeton-nlp/SimCSE

	20Newsgroups			M10			BBC News		
Model	*TU*	*NPMI*	C_V	*TU*	*NPMI*	C_V	*TU*	*NPMI*	C_V
LDA	0.717	0.040	0.511	0.681	-0.177	0.336	0.312	-0.014	0.357
ProdLDA	0.736	0.045	0.574	0.650	-0.260	0.432	0.702	-0.044	0.540
CombinedTM	0.700	0.065	0.601	0.581	0.001	0.443	0.606	0.042	0.639
ZeroShotTM	**0.729**	0.069	0.614	0.633	-0.056	0.433	0.699	-0.050	0.531
$\text{BERT}_{\texttt{base}}$	0.562	0.118	0.649	0.763	0.146	0.725	0.689	0.129	0.700
$\text{BERT}_{\texttt{large}}$	0.550	0.116	0.646	0.743	0.138	0.715	0.684	0.132	0.705
$\text{RoBERTa}_{\texttt{base}}$	0.385	0.028	0.464	0.634	-0.008	0.480	0.671	0.098	0.646
$\text{RoBERTa}_{\texttt{large}}$	0.404	0.014	0.440	0.669	0.001	0.506	0.673	0.046	0.555
$\text{BERT}_{\texttt{base}}$+KM†	0.346	0.065	0.521	0.484	0.116	0.588	0.529	0.111	0.637
$\text{BERT}_{\texttt{base}}$+UMAP	0.589	0.128	0.671	0.794	0.159	0.706	0.716	0.135	0.716
$\text{BERT}_{\texttt{large}}$+UMAP	0.563	0.126	0.662	0.751	0.176	0.681	0.721	0.139	0.720
$\text{RoBERTa}_{\texttt{base}}$+UMAP	0.434	0.063	0.522	0.640	0.091	0.547	0.710	0.106	0.664
$\text{RoBERTa}_{\texttt{large}}$+UMAP	0.463	0.054	0.499	0.636	0.046	0.513	0.706	0.077	0.632
$\text{SBERT}_{\texttt{base}}$	0.668	0.126	0.658	0.832	0.164	0.742	0.727	0.137	0.719
$\text{SBERT}_{\texttt{large}}$	0.674	0.135	0.673	0.844	0.168	0.752	0.718	0.134	0.714
$\text{SRoBERTa}_{\texttt{base}}$	0.670	0.128	0.654	0.815	0.149	0.713	0.719	0.131	0.699
$\text{SRoBERTa}_{\texttt{large}}$	0.649	0.115	0.640	0.823	0.155	0.735	0.696	0.122	0.694
$\text{SBERT}_{\texttt{base}}$+UMAP	0.679	0.139	0.690	0.841	0.192	0.715	0.749	0.142	**0.730**
$\text{SBERT}_{\texttt{large}}$+UMAP	0.681	0.139	0.691	0.836	0.203	0.723	0.744	0.136	0.725
$\text{SRoBERTa}_{\texttt{base}}$+UMAP	0.680	0.138	0.684	0.830	0.192	0.722	0.747	0.135	0.716
$\text{SRoBERTa}_{\texttt{large}}$+UMAP	0.680	0.131	0.670	0.799	0.196	0.700	0.728	0.121	0.705
Unsup-SimCSE($\text{BERT}_{\texttt{base}}$)	0.677	0.147	0.694	0.831	0.180	0.750	0.730	0.142	0.722
Unsup-SimCSE($\text{BERT}_{\texttt{large}}$)	0.700	0.145	0.693	0.832	0.182	0.750	0.728	0.135	0.714
Unsup-SimCSE($\text{RoBERTa}_{\texttt{base}}$)	0.696	0.142	0.682	0.823	0.164	0.726	0.731	0.137	0.700
Unsup-SimCSE($\text{RoBERTa}_{\texttt{large}}$)	0.722	0.147	0.694	0.812	0.171	0.734	0.736	0.142	0.711
Unsup-SimCSE($\text{BERT}_{\texttt{base}}$)+UMAP	0.692	0.139	0.685	**0.851**	**0.206**	0.744	0.733	0.146	0.729
Unsup-SimCSE($\text{BERT}_{\texttt{large}}$)+UMAP	0.694	0.145	0.698	0.843	0.200	0.721	0.736	0.128	0.709
Unsup-SimCSE($\text{RoBERTa}_{\texttt{base}}$)+UMAP	0.689	0.145	0.703	0.843	0.192	0.726	0.747	0.130	0.701
Unsup-SimCSE($\text{RoBERTa}_{\texttt{large}}$)+UMAP	0.717	0.146	0.701	0.813	0.190	0.710	0.752	0.138	0.713
Sup-SimCSE($\text{BERT}_{\texttt{base}}$)	0.721	0.151	0.702	0.829	0.180	0.746	0.736	0.143	0.720
Sup-SimCSE($\text{BERT}_{\texttt{large}}$)	0.706	**0.155**	**0.709**	0.833	0.189	**0.762**	0.744	0.146	**0.730**
Sup-SimCSE($\text{RoBERTa}_{\texttt{base}}$)	0.718	0.145	0.693	0.829	0.170	0.734	0.738	0.140	0.715
Sup-SimCSE($\text{RoBERTa}_{\texttt{large}}$)	0.716	0.148	0.696	0.826	0.179	0.742	0.751	**0.147**	0.726
Sup-SimCSE($\text{BERT}_{\texttt{base}}$)+UMAP	0.714	0.146	0.698	0.815	0.202	0.730	0.739	0.143	0.724
Sup-SimCSE($\text{BERT}_{\texttt{large}}$)+UMAP	0.721	0.150	0.704	0.834	**0.206**	0.728	0.750	0.145	0.729
Sup-SimCSE($\text{RoBERTa}_{\texttt{base}}$)+UMAP	0.709	0.144	0.700	0.822	0.195	0.711	0.752	0.142	0.723
Sup-SimCSE($\text{RoBERTa}_{\texttt{large}}$)+UMAP	0.708	0.147	0.701	0.818	0.189	0.704	**0.754**	0.145	0.725

Table 4: Topic coherence (*NPMI* and C_V) and topic diversity (*TU*) of the top 10 words. All results are averaged across the 5 number of topics ($K = \{\text{ground truth}, 25, 50, 75, 100\}$). Each model is averaged over 5 runs. Best results are in bold. †: we use the method from (Sia et al., 2020), which uses PCA to reduce embedding dimensionality and TF to select words. For other clustering-based models, we use KM to cluster embeddings and $\mathbf{TFIDF} \times \mathbf{IDF_i}$ (Eq. 4) to select topic words. Dimensionality: `base`: 768, `large`: 1024.

Topic	Weighting Method	Topic Words
Topic 1	$\mathbf{TFIDF \times IDF_i}$	car bike ride engine brake tire drive mile road front
	$\mathbf{TFIDF_i}$	car bike **good** brake drive **make** ride **time** engine tire
	$\mathbf{TF_i}$	car bike **good** drive **make** **time** engine ride back **year**
Topic 2	$\mathbf{TFIDF \times IDF_i}$	armenian turkish people kill israeli genocide village jewish war government
	$\mathbf{TFIDF_i}$	armenian people turkish genocide government **make** israeli kill **time** village
	$\mathbf{TF_i}$	people armenian turkish **make** kill government **time** **year** state child

Table 5: Comparison of topic words generated using different weighting methods when $K = 30$. Repeated words across topics are marked with an underline. Incoherent words are in bold.