

# Human in the loop: How to effectively create coherent topics by manually labeling only a few documents per class

Anton Thielmann, Christoph Weisser, Benjamin Säfken

Clausthal University of Technology

LREC-Coling 2024  
20-25 May



TU Clausthal  
Clausthal University of Technology

- Topic modeling has undergone a great deal of development.
- Traditional methods rely on unsupervised approaches (e.g. LDA, ETM, CTM, BERTopic, etc.) (Blei et al., 2003, Dieng et al., 2020, Bianchi et al., 2021, Grootendorst, 2022)
- Few-shot methods achieve remarkable results in various supervised label-scarce settings.
  - The metrics of interest are in this case not the coherence of clusters, but model accuracy, F1 score, or precision

- Manually labeling a few documents might be an attractive option for unsupervised tasks such as document clustering.



**Richard Socher** ✓

@RichardSocher



Rather than spending a month figuring out an unsupervised machine learning problem, just label some data for a week and train a classifier.

11:47 PM · Mar 10, 2017

- By leveraging pre-trained sentence transformers (Reimers and Gurevych, 2019) and class-based term frequency inverse document frequency (tf-idf) for topic extraction, we can generate coherent topics with only a few labeled documents per class.

## Contributions:

- 1 Introduction of a Document Classification and Topic Extraction (DCTE) approach.
- 2 Benchmarking against traditional models and demonstrating superior topic coherence.

## Definitions

- Vocabulary of words:  $V = \{w_1, \dots, w_n\}$
- Corpus of documents:  $D = \{d_1, \dots, d_M\}$
- Document representation: Each document  $d_i$  as a sequence of words  $d_i = [w_{i1}, \dots, w_{in_i}]$ , where  $w_{ij} \in V$  and  $n_i$  is the length of  $d_i$ .

## Document and Topic Embeddings

- Document embeddings:  $\mathcal{D} = \{\delta_1, \dots, \delta_M\}$ , where  $\delta_i$  is the vector representation of  $d_i$ .
- Topics: Each topic  $t_k$  as a probability distribution over  $V$ , expressed as  $(\phi_{k,1}, \dots, \phi_{k,n})^T$  with  $\sum_{i=1}^n \phi_{k,i} = 1$ .

- Label a small subset of documents:  $\{d_1, \dots, d_k\}$ .
  - We find that as little as one document per class can be sufficient.
- Fine-tune a classification model on this labelled subset using SETFIT (Tunstall et al., 2022)
  - $\mathcal{D} = \{\delta_1, \dots, \delta_M\}$  are hence corpora specific finetuned document embeddings
- Apply the trained model to classify the remaining documents and extract the topics.

- Utilize class-based tf-idf to identify significant terms within each cluster:

$$\text{tf-idf}(w|c) = \frac{\text{frequency}(w_c)}{n_c} \cdot \log \left( \frac{N}{\sum_j w_j} \right)$$

- Here,  $w_c$  is the frequency of word  $w$  in class  $c$ ,  $n_c$  is the number of words in class  $c$ , and  $N$  is the total number of documents.

# Experimental Setup and Data

- data sources: 20 Newsgroups, BBC News, and M10.
- hyperparameter tuning for all benchmark models **except** for DCTE
  - All models are fit using OCTIS (Terragni et al., 2021)
  - We use *all-MiniLM-L6-v2* (Reimers and Gurevych, 2019) where applicable for all models

- To evaluate the topics, we use normalised pointwise mutual information (NPMI) coherence scores (Lau et al., 2014).

$$NPMI(t_k) = \sum_{j=2}^N \sum_{i=1}^{j-1} \frac{\log \frac{P(w_i, w_j)}{P(w_i)P(w_j)}}{-\log(P(w_i, w_j))}$$

- Complete documents as window sizes

# Benchmark Model Comparison

| Model             | 20 News      | BBC           | M10           |
|-------------------|--------------|---------------|---------------|
| LDA               | 0.096        | -0.214        | -0.218        |
| NeuralLDA         | 0.046        | -0.357        | -0.55         |
| ProdLDA           | 0.161        | -0.099        | <b>-0.09</b>  |
| BERTopic          | -0.10        | 0.044         | -0.303        |
| BERTopic*         | 0.128        | <b>0.2068</b> | -0.126        |
| K-means           | 0.115        | 0.0648        | -0.134        |
| ETM               | -0.089       | -0.077        | -0.188        |
| CTM               | <b>0.205</b> | -0.002        | -0.213        |
| DCTE <sup>1</sup> | <b>0.221</b> | <b>0.20</b>   | <b>-0.015</b> |
| DCTE <sup>2</sup> | <b>0.163</b> | <b>0.153</b>  | <b>-0.054</b> |
| DCTE <sup>3</sup> | 0.117        | 0.103         | -0.146        |
| DCTE <sup>4</sup> | <b>0.186</b> | <b>0.117</b>  | <b>-0.119</b> |

\* Evaluating only the top 50% coherent topics.

- 1 Most coherent model using 20, 5, and 10 randomly drawn labeled samples respectively.
- 2 Best model with only one labeled training sample per class.
- 3 Average coherence with one labeled training sample per class.
- 4 Average coherence with four labeled training samples per class.

**Table:** NPMI coherence scores for tested models on three benchmark datasets. Models achieving top coherence are highlighted.

# Results

| Topic | DCTE      |           | CTM        |         |
|-------|-----------|-----------|------------|---------|
|       | Religion  | Space     | Religion   | Space   |
|       | god       | space     | conclusion | year    |
|       | jesus     | launch    | science    | mission |
|       | church    | satellite | atheist    | launch  |
|       | christian | mission   | church     | orbit   |
|       | religion  | orbit     | atheism    | solar   |
|       | belief    | moon      | truth      | make    |
|       | word      | data      | religion   | space   |
|       | atheist   | science   | tradition  | moon    |
|       | faith     | earth     | christian  | planet  |
|       | people    | rocket    | argument   | surface |
| NPMI  | 0.385     | 0.41      | 0.324      | 0.111   |

**Table:** Topics created with DCTE and CTM and the respective NPMI coherence scores for the 20 NG dataset.

# Results: Random Draw Sampling

| Dataset | Samples | Average       | Max           | Std. Dev.   |
|---------|---------|---------------|---------------|-------------|
| 20 News | 20      | <b>0.185</b>  | <b>0.221</b>  | $\pm 0.080$ |
|         | 40      | 0.108         | 0.144         | $\pm 0.024$ |
|         | 60      | 0.122         | 0.190         | $\pm 0.032$ |
|         | 80      | 0.145         | 0.190         | $\pm 0.045$ |
|         | 100     | 0.162         | 0.200         | $\pm 0.047$ |
| BBC     | 5       | 0.097         | <b>0.200</b>  | $\pm 0.084$ |
|         | 10      | <b>0.192</b>  | 0.152         | $\pm 0.059$ |
|         | 15      | 0.115         | 0.124         | $\pm 0.054$ |
|         | 20      | 0.081         | 0.139         | $\pm 0.024$ |
|         | 25      | 0.121         | 0.115         | $\pm 0.020$ |
| M10     | 10      | -0.178        | <b>-0.015</b> | $\pm 0.119$ |
|         | 20      | -0.153        | -0.120        | $\pm 0.021$ |
|         | 30      | -0.142        | -0.078        | $\pm 0.033$ |
|         | 40      | -0.098        | -0.033        | $\pm 0.055$ |
|         | 50      | <b>-0.094</b> | -0.054        | $\pm 0.035$ |

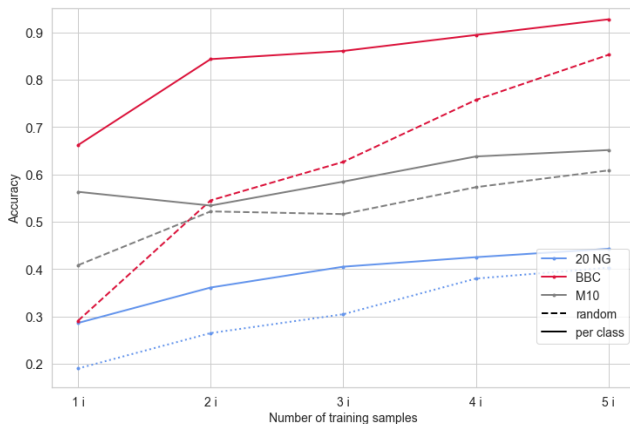
**Table:** Experimental results for random draw sampling. The average NPMI coherence, maximum scores, and standard deviations over 5 runs are shown.

# Results: Per Class Sampling

| Dataset | Samples per Class | Average       | Max           | Std. Dev.   |
|---------|-------------------|---------------|---------------|-------------|
| 20 News | 1                 | 0.117         | 0.163         | $\pm 0.036$ |
|         | 2                 | 0.115         | 0.196         | $\pm 0.046$ |
|         | 3                 | 0.137         | 0.189         | $\pm 0.061$ |
|         | 4                 | <b>0.186</b>  | <b>0.208</b>  | $\pm 0.021$ |
|         | 5                 | 0.164         | 0.192         | $\pm 0.027$ |
| BBC     | 1                 | 0.103         | 0.153         | $\pm 0.032$ |
|         | 2                 | 0.133         | 0.187         | $\pm 0.047$ |
|         | 3                 | 0.107         | 0.180         | $\pm 0.041$ |
|         | 4                 | 0.117         | <b>0.191</b>  | $\pm 0.040$ |
|         | 5                 | <b>0.142</b>  | 0.186         | $\pm 0.022$ |
| M10     | 1                 | -0.146        | -0.078        | $\pm 0.037$ |
|         | 2                 | <b>-0.115</b> | <b>-0.054</b> | $\pm 0.039$ |
|         | 3                 | -0.158        | -0.107        | $\pm 0.032$ |
|         | 4                 | -0.119        | -0.103        | $\pm 0.015$ |
|         | 5                 | -0.121        | -0.098        | $\pm 0.027$ |

**Table:** Experimental results for per class sampling. The average NPMI coherence, maximum scores, and standard deviations over 5 runs are shown.

- Random sampling vs per class sampling effect on accuracy



- With improved few-shot methods, labeling only a few datapoints is an additional option for topic modeling
  - Especially in highly specialized domains

## **Future Work:**

- Test other few-shot methods
- Include more sophisticated topic extraction methods suited for multi-labeled and multi-topic documents
- The code for DCTE is available at:  
<https://github.com/AnFreTh/STREAM>

Thank you for watching!

I look forward to discussing this with you in person at the conference.

*See you at LREC-Coling 2024!*



D. M. Blei, A. Y. Ng, & M. I. Jordan.

Latent dirichlet allocation.

*Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.



A. B. Dieng, F. J. R. Ruiz, & D. M. Blei.

Topic modeling in embedding spaces.

*Transactions of the Association for Computational Linguistics*,  
8:439–453, 2020.



F. Bianchi, S. Terragni, & D. Hovy.

Pre-training is a hot topic: Contextualized document embeddings  
improve topic coherence.

*Proceedings of ACL/IJCNLP 2021*.



M. Grootendorst.

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



N. Reimers & I. Gurevych.

Sentence-bert: Sentence embeddings using siamese bert-networks.  
*EMNLP 2019.*



L. Tunstall, et al.

Efficient few-shot learning without prompts.  
*arXiv preprint arXiv:2209.11055, 2022.*



S. Terragni, et al.

Octis: Comparing and optimizing topic models is simple!  
*EACL 2021 System Demonstrations, pages 263–270.*



J. Lau, D. Newman, & T. Baldwin.

Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality.