Topics in the Haystack: Enhancing Topic Quality through Corpus Expansion

Anton Thielmann
Institute of Mathematics
Clausthal University of Technology
Anton.thielmann@tu-clausthal.de

Arik Reuter
Institute of Mathematics
Clausthal University of Technology
Arik.reuter@gmx.de

Quentin Seifert
Spatial Data Science and
Statistical Learning
University of Göttingen
Quentinedward.seifert
@uni-goettingen.de

Elisabeth Bergherr
Spatial Data Science and
Statistical Learning
University of Göttingen
Elisabeth.bergherr
@uni-goettingen.de

Benjamin Säfken
Institute of Mathematics
Clausthal University of Technology
Benjamin.saefken@tu-clausthal.de

Extracting and identifying latent topics in large text corpora have gained increasing importance in Natural Language Processing (NLP). Most models, whether probabilistic models similar to Latent Dirichlet Allocation (LDA) or neural topic models, follow the same underlying approach of topic interpretability and topic extraction. We propose a method that incorporates a deeper

Action Editor: Wei Lu. Submission received: 22 June 2023; revised version received: 20 December 2023; accepted for publication: 27 December 2023.

https://doi.org/10.1162/coli_a_00506

understanding of both sentence and document themes, and goes beyond simply analyzing word frequencies in the data. Through simple corpus expansion, our model can detect latent topics that may include uncommon words or neologisms, as well as words not present in the documents themselves. Additionally, we propose several new evaluation metrics based on intruder words and similarity measures in the semantic space. We present correlation coefficients with human identification of intruder words and achieve near-human level results at the word-intrusion task. We demonstrate the competitive performance of our method with a large benchmark study, and achieve superior results compared with state-of-the-art topic modeling and document clustering models. The code is available at the following link: `https://github.com/AnFreTh/STREAM`.

1. Introduction

Identifying latent topics in large text corpora is a central task in Natural Language Processing (NLP). With the ever-growing availability of textual data in virtually all languages and about every possible topic, automated topic extraction is gaining increasing importance. Hence, the approaches are manifold. A comprehensive overview over current approaches is, for example, given in Vayansky and Kumar (2020) and Barde and Bainwad (2017). For almost all models, a topic is intuitively defined by a set of words with each word having a probability of occurrence for the given topic. Different topics can share words, and a document can be linked to more than one topic. Generative probabilistic models, such as Probabilistic Latent Semantic Analysis (Hofmann 2001) and Latent Dirichlet Allocation (LDA) (Blei, Ng, and Jordan 2003), are still widely used and inspired multiple adaptations as several studies (e.g., Agarwal and Chen 2010; Blei, Griffiths, and Jordan 2010; Chien, Lee, and Tan 2018; Ramage et al. 2009; Rosen-Zvi et al. 2012) all draw heavily from word co-occurrences. Due to its popularity and general good performance on benchmark datasets, the interpretation of a topic from LDA is seldom challenged. Neural topic models (e.g., Dieng, Ruiz, and Blei 2020; Wang, Zhou, and He 2019; Bianchi, Terragni, and Hovy 2020) further improve upon the existing methods by integrating *word-embeddings* or variational autoencoders (Srivastava and Sutton 2017) into the modeling approach, but still heavily rely on the ideas from Blei, Ng, and Jordan (2003).

New methods that challenge the typical idea of topic modeling also integrate *word-* and *document-embeddings* (Miles et al. 2022; Angelov 2020; Grootendorst 2022; Sia, Dalmia, and Mielke 2020; Thielmann, Weisser, and Säfken 2022). However, improvement over the current state of the art is usually measured in terms of performance as determined by evaluation metrics on standard benchmark datasets. While older models were still evaluated using likelihood-based perplexity metrics (Lafferty and Blei 2005; Larochelle and Lauly 2012; Rosen-Zvi et al. 2012), empirical results showed a negative correlation between perplexity-based metrics and human evaluation of a topic model (Chang et al. 2009). Additionally, Chang et al. (2009) first introduced the idea of **intruder words**. According to this idea, a topic is considered *coherent* or simply put, *good*, if a randomly chosen word, not belonging to that topic, can clearly be identified by humans. As human evaluation of models is cost and time intensive, researchers used new evaluation methods that correlated with human evaluation (Lau, Newman, and Baldwin 2014; Newman et al. 2010). Hoyle et al. (2021) even found no contemporary model at all that used human feedback as a form of model evaluation. Newer models were hence evaluated using coherence scores (Angelov 2020; Dieng, Ruiz, and Blei

2020; Grootendorst 2022; Sia, Dalmia, and Mielke 2020; Srivastava and Sutton 2017). However, Hoyle et al. (2021) found severe flaws in coherence scores. First, they find that coherence scores exaggerate differences between models and, second, they validate the findings from Bhatia, Lau, and Baldwin (2017) and find much lower Pearson correlations between automated coherence scores and human evaluation as compared with Lau, Newman, and Baldwin (2014). New evaluation methods as proposed by Weisser et al. (2023), for example, often lack any form of human verification.

We identify two shortcomings in the current state-of-the-art in topic modeling and document clustering. The first is the significant gap in validated automatic evaluation methods for topic models. The second stems from the continued reliance on evaluation methods based on word co-occurrences and outdated definitions of topics from older models. Current methods rely on limited corpora from which the topic representations are created. However, integrating larger corpora into the modeling process can enhance topic quality by including contextually relevant words that were missing from the original corpus.

1.1 Contributions

The contributions of this paper are hence twofold and can be summarized as follows:

- We propose Context Enhanced Document Clustering (CEDC) which, with only a few adaptations, integrates linguistic ideas into its modeling. Soft-clustering on the document level is integrated, such that P(document|topic) is modeled.
- We introduce new topic modeling performance metrics. The effectiveness of the proposed metrics is validated by demonstrating impressive correlations with human judgment.
- We conduct a benchmark study comparing the presented approach to state-of-the-art topic modeling and document clustering methods. The presented approach outperforms common benchmark models on both coherence scores and the presented new metrics for topic evaluation.
- Our findings illustrate that even without any hyperparameter tuning, CEDC can achieve superior performance compared with existing state-of-the-art topic models that have undergone extensive hyperparameter optimization.

The remainder of the paper is structured as follows: First, we give an overview over related methods. Second, a short introduction into the used linguistic ideas and the definition of topics is presented. Third, we investigate the design of questionnaires employed for the evaluation of topic models, providing insights into the essential factors and considerations inherent in constructing robust questionnaire designs. Fourth, the method of extracting latent topics from documents, incorporating the aforementioned definitions, is presented. Fifth, new evaluation metrics are introduced and validated by presenting correlations with human annotators. Sixth, the proposed model is applied to two common data sets and compared with state-of-the-art topic models. Finally, a discussion of the limitations as well as a conclusion is given in Sections 7 and 8.

2. Related Literature

2.1 Bayesian Generative Topic Models

Topic modeling has been dominated by Bayesian graphical models. Probabilistic Latent Semantic Analysis (PLSA) (Hofmann 2001) and especially LDA (Blei, Ng, and Jordan 2003) being some of the most notable examples. In the context of these generative models and related structured probabilistic models (Blei and Lafferty 2007; Mazarura, De Waal, and de Villiers 2020), a word w is interpreted as a discrete token in the vocabulary $\mathbf{V}$. The corpus comprises all available documents, where a document d is a sequence of words mostly represented as a bag-of-words vector. The K topics, denoted as $t_1, \cdots, t_K$, are modeled as categorical probability distributions over the vocabulary. Additionally, the document-specific topic distributions $Cat(\boldsymbol{\theta}^{(d)})$, are parameterized by $\boldsymbol{\theta}^{(d)} \in \mathbb{R}^K$ for each document d.

Most Bayesian generative models follow the algorithmic structure of LDA with slight adjustments as given in Algorithm 1, where $Cat(\cdot)$ refers to the categorical distribution and $Dir(\cdot)$ refers to the Dirichlet distribution. α and β are thus the distributional parameters for the Dirichlet and the categorical distribution, respectively, and must be specified in advance or optimized according to an optimization criterion. Adaptations of this generative model comprise several techniques: the inclusion of (1) variational autoencoders (Srivastava and Sutton 2017); (2) contextualized word embeddings (Das, Zaheer, and Dyer 2015); and (3) contextualized document embeddings (Bianchi, Terragni, and Hovy 2020). Further adaptations including word, document, or topic embeddings have also achieved remarkable results, slightly adjusting the overall generative process (e.g., Dieng, Ruiz, and Blei 2020).

2.2 Clustering-based Topic Models

While the early models like LDA and PLSA suffered from the restrictions imposed by bag-of-words representations, contextualized embeddings offered further possibilities beyond the inclusion into generative modeling (Dieng, Ruiz, and Blei 2020; Bianchi, Terragni, and Hovy 2020). The general idea behind these approaches is to leverage pre-trained word or document embeddings. These, often-times dimensionality-reduced representations are subsequently clustered. Each document or word cluster then represents an individual topic. For example, Sia, Dalmia, and Mielke (2020) demonstrate the efficacy of this conceptually simple approach by applying various centroid-based

Algorithm 1 Topic Modeling Algorithm

Input: Corpus of documents D, hyperparameters $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}$.
Output: Topic assignments and word selections for each document.
for each document d in the corpus D **do**
    Choose a topic distribution: $\boldsymbol{\theta}^{(d)} \sim Dir(\boldsymbol{\alpha})$
    for each word index $i = 1, \ldots, l_d$ in d **do**
        Choose a topic: $t_i^{(d)} \sim Cat(\boldsymbol{\theta}^{(d)})$
        Choose a word: $w_i^{(d)} \sim Cat(\boldsymbol{\beta}_{t_i^{(d)}})$
    end for
end for

clustering algorithms to word embeddings. Coherent topics are retrieved from the word clusters.

Angelov (2020) utilizes joint embeddings of words and documents and leverages HDBSCAN (McInnes, Healy, and Astels 2017) to create clusters of documents from which topics are extracted. Closely following Angelov (2020), Grootendorst (2022) uses sentence transformers (Reimers and Gurevych 2019) to obtain document embeddings. Term-frequency inverse-document frequency (tf-idf) scores (Salton 1989) are used to obtain scores for the probabilities of words under a given topic. Miles et al. (2022) also leverage sentence transformers (Reimers and Gurevych 2019) but uses Particle Swarm Optimization to identify latent clusters in the corpus.

While being algorithmically much simpler than classical topic models, these document clustering schemes have proven extremely effective, often outperforming the computationally more demanding generative models (Sia, Dalmia, and Mielke 2020; Grootendorst 2022; Angelov 2020; Thielmann, Weisser, and Säfken 2022).

3. On the Nature of Topics

While there have been numerous approaches to extracting latent topics from large text corpora, little effort has been made in adapting those models to more refined definitions of a topic. We propose a topic model that follows ideas from linguistic definitions of topics (Davison 1982, 1984). We present two simple ideas from linguistic theory in order to construct more humanly interpretable topics:

(i) A word that most accurately expresses the topic of a document may not necessarily occur in that document.

(ii) Only using nouns and noun phrases is more appropriate for representing understandable topics.

Idea (i) closely follows (Guijarro 2000): "a topic is, above all, a textual category that is determined by the context and not by purely formal or structural aspects." Therefore, the topic of a document or even a sentence may go beyond the mere occurrence of all the words in that document. That is, a word that most accurately expresses the topic of a document may not necessarily occur in that document. We leverage a simple example from a *New York Times* headline to demonstrate that:

> *"Lehman had to die so Global Finance could live"*

That sentence pertains to the financial crisis and the collapse of the Lehman Brothers bank, but neither phrase is explicitly mentioned. A bag-of-words model that only considers words present in the document corpus would not be able to accurately capture the document's topic. Contextually relevant words, even if not present in the document, can provide better representations. Figure 1 shows the described example. Comparing the cosine distance in a reduced embedding space between the complete embedded sentence (TEXT) and each embedded word or phrase demonstrates how words and phrases not occurring in that text can be a meaningful summary of that text. "Banking crisis" is a more meaningful representation of the sentence than, for example, "global" and lies closer to the text in the semantic space.

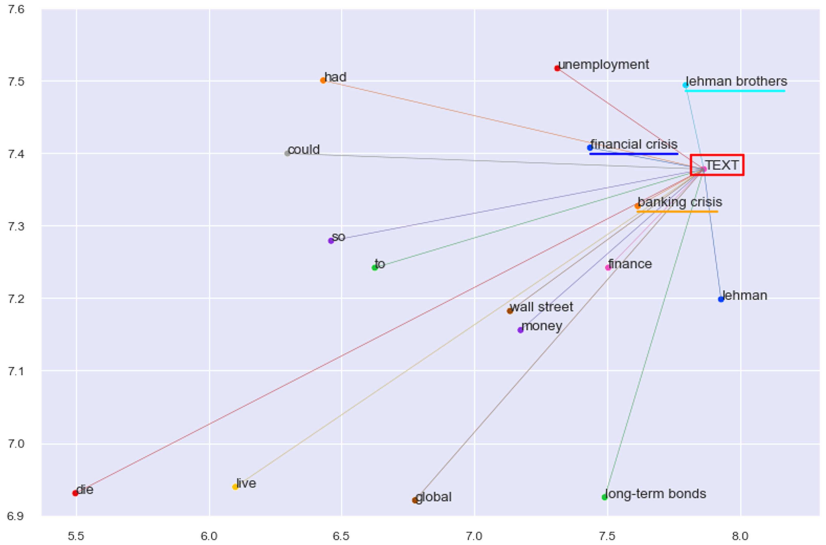

Figure 1
The word best representing a sentence (or document) does not necessarily need to be included in that text. The figure represents a *New York Times* headline from the financial crisis in 2009: "*Lehman had to die so Global Finance could live*." All words present in that text and additional words are mapped into a high dimensional feature space. The dimensions are reduced to visually demonstrate, that words not occurring in that sentence, e.g., *banking crisis*, are better suited to summarize that sentence than words present in the sentence, e.g., *global*.

Common topic models (e.g., Blei, Ng, and Jordan 2003; Dieng, Ruiz, and Blei 2020; Bianchi, Terragni, and Hovy 2020; Das, Zaheer, and Dyer 2015), as well as document clustering methods (e.g., Grootendorst 2022; Angelov 2020; Sia, Dalmia, and Mielke 2020), face a limitation in that they only consider words that appear in the reference corpus when generating topic representations. This limitation can lead to incorrect topic interpretations, as shown in the example above. Through expanding the reference corpus and leveraging pre-trained embedding models, we make sure that "the indispensability of frame knowledge for understanding texts" (Beghtol 1986) is accounted for.

Point (ii) closely follows Beghtol (1986), after whom one of the features of generalized titles is the absence of verbal forms. Following the idea that a title is the highest macroproposition of a textual unit (Beghtol 1986), we apply this idea to the construction of topics and hence propose to only consider nouns and noun phrases for the proposed method of topic extraction. Martin and Johnson (2015) already demonstrated improvements in topic quality when focusing on nouns, but reduced the complete corpus to nouns only, losing any semantic or syntactic information hidden in the documents.

3.1 Human Topic Evaluation

The findings from Hoyle et al. (2021) show that current state-of-the-art automated topic model evaluation does not represent human judgment of topic quality. However, human topic evaluation is inherently complicated as human evaluation is highly subjective. Moreover, as demonstrated in Section 3 above, topics can be highly complex and theoretical constructs. Simply relying on human judgment to determine whether

a topic is *good* or not does not encompass all of a topic's properties. The documents that should be represented by the topic or topic diversity are seldom accounted for (Newman et al. 2010; Lund et al. 2019; Clark et al. 2021). Clark et al. (2021) even question human judgment alltogether; however, the used questionnaire design not only does not provide a midpoint but additionally can strongly induce a bias in preference due to a highly biasing follow-up question (Clark et al. 2021) (see, e.g., Lehman et al. 1992). Newman et al. (2010), for example, use the straightforward approach of letting humans rate the created topics quality. Choosing a 3-point scale for model evaluation, however, can induce unreliability of responses (Krosnick 2018). Simple assessment of topic quality therefore does not suffice and creating great questionnaires and adequately operationalizing what researchers are interested in is notably important.

Adapting questions and tasks to the complicated nature of topics can result in promising questionnaire designs. Lund et al. (2019), for example, introduced a topic-word matching task, weighting and selecting answers from participants who have a high confidence and performed well on test questions. Choosing that approach reduces ambiguity in answers, but also induces a bias towards highly confident participants and neglects the subtle differences in perceived quality from humans. Promising results are also achieved with further refined questionnaire designs. Bhatia, Lau, and Baldwin (2017, 2018) introduce document-level topic model evaluation leveraging the intruder-topic task, also introduced in Chang et al. (2009). In this task, participants are presented with a list of words that are related in some way. The task typically involves presenting a series of word lists, with each list containing a set of related words except for one "intruder" word that does not belong to the category or topic (e.g., *Apple, Orange, Pineapple, Bicycle, Banana, Mango*). The participants' objective is to identify and quickly recognize the intruder word within each list. Thus, the intruder-task, if cleverly designed and using "intruder" words from different topics, can account for topic diversity and even could account for a topic adequately representing a set of documents, by sampling "intruder" words directly from the set of documents.

4. Methodology

Given the seminal works of Grootendorst (2022) and the results shown by Thielmann, Weisser, and Säfken (2022), we propose a simple yet highly effective document clustering and topic extraction method. The pseudo algorithm for the complete model can be seen in Algorithm 2. The proposed method expands a given base-corpus by enriching extracted document clusters with nouns from an external expansion-corpus and can be summarized in a simplified manner as follows: First, the given corpus of documents is embedded using contextualized transformer embeddings. For example Bianchi, Terragni, and Hovy (2020) showed that contextualized embeddings can improve topic quality significantly. Second, the dimensionality of the embedded documents is reduced to alleviate the curse of dimensionality caused by the typically large number of dimensions in text embeddings. Third, the embedded documents are clustered using a Gaussian mixture model (GMM). The central results from this clustering step are the document clusters in general, which can be interpreted as topics among the documents, and the centroids of the found clusters more specifically. Please note that we do not enhance the documents during clustering, thereby ensuring that no erroneous expansion can compromise the quality of the topics. See Section 4.2 for different approaches. Subsequently, all nouns existing in the base-corpus as well as all nouns in the extension-corpus are embedded into the same embedding space as previously the documents

using the same text-embedding model as for the texts in the base corpus. After that, we select the prototypical words representing each topic as the words with the most similar embedding to the centroid representing the topic. This allows us to not only obtain soft document-topic-assignment scores via the similarity of documents to centroids of topics, but to also obtain scores for the most likely words given a specific topic. Note that the words representing a topic among documents in the base corpus can also come from the extension corpus provided they enhance the word-level representation of a found topic. Last, a cleaning step can be performed to remove overly similar terms from the topics. Through the simple step of reference corpus expansion and leveraging soft clustering, we can significantly improve upon previous document clustering and topic extraction methods (Grootendorst 2022; Angelov 2020).

More formally and extensively, the proposed approach can be described as follows: Let $\mathrm{V} = \{w_1, \ \dots \ , w_n\}$ be the vocabulary of words and $\mathrm{D} = \{d_1, \ \dots \ , d_M\}$ be a corpus (i.e., a collection of documents). Each document is a sequence of words $d_i = \left[w_{i1}, \dots, w_{in_i}\right]$ where $w_{ij} \in V$ and n_i denotes the length of document d_i. Further, let $\mathcal{D} = \{\boldsymbol{\delta}_1, \ \dots \ , \boldsymbol{\delta}_M\}$ be the set of documents represented in the embedding space, such that $\boldsymbol{\delta}_i$ is the vector representation of d_i and let $\mathcal{W} = \{\boldsymbol{\omega}_1, \ \dots \ , \boldsymbol{\omega}_n\}$ be the vocabulary's representation in the same embedding space. Hence, each word w_i in the embedding space represented as $\boldsymbol{\omega}_i \in \mathbb{R}^L$ has the same dimensionality L as a document vector $\boldsymbol{\delta}_i \in \mathbb{R}^L$. There are different representations of topics, but mostly a topic t_k from a set of topics $T = \{t_1, \ \dots \ , t_K\}$ is represented as a discrete probability distribution over the vocabulary (Blei, Ng, and Jordan 2003), such that t_k is often expressed as $(\phi_{k,1}, \dots, \phi_{k,n})^T$ and $\sum_{i=1}^{n} \phi_{k,i} = 1$ for every k, where $\phi_{k,n} \in [0, 1]$. Thus, $\boldsymbol{\phi_k} = (\phi_{k,1}, \dots, \phi_{k,n})$ simply describes the probability vector over the vocabulary for topic k.[1]

Based upon the idea expressed in Section 3, we form clusters from the documents embeddings, $\mathcal{D}$, and subsequently extract topics, t_k, that represent these clusters best. Hence, after transforming the raw documents into document vectors, they are clustered. Due to the curse of dimensionality (Aggarwal, Hinneburg, and Keim 2001) we reduce the dimensionality of the document embeddings before clustering using UMAP (McInnes, Healy, and Melville 2018), closely following Angelov (2020) and Grootendorst (2022). However, we allow each document to belong to more than one cluster resulting in document topic matrices $\boldsymbol{\theta}$ and word topic matrices $\boldsymbol{\beta}$, similar to LDA (Blei, Ng, and Jordan 2003). The documents are clustered with a GMM (Reynolds 2009), as it not only allows for soft-clustering, but also has the advantage of optimizing hyperparameters via, for instance, the Akaike information criterion or the Bayesian information criterion. As a result, CEDC, in contrast to others (Angelov 2020; Grootendorst 2022; Sia, Dalmia, and Mielke 2020), offers not only word-topic distributions but also document-topic distributions.

4.1 Topic Extraction

To find the words that best represent the corpus' topics, we first extract the centroids of the k clusters, $\boldsymbol{\mu}_k \in \mathbb{R}^L$, in the original embedding space. Second, we filter the given vocabulary for nouns and enhance this vocabulary by a specified external vocabulary of nouns, resulting in a new enriched dictionary $\hat{V} = \{w_1, \ \dots \ , w_n, w_{n+1}, \ \dots \ , w_{n+z}\}$. The word vectors $\boldsymbol{\omega}_i$ closest to $\boldsymbol{\mu}_k$ in the embedding space are the words that represent cluster k's centroid best (Angelov 2020) and are thus selected as the prototypical words

1 See table A.1 in the Appendix for a complete variable and notation list.

for the corresponding topic. Note that it could happen that a word represents a topic ideally for $w \notin V$ but $w \in \hat{V}$. To select the words best representing a topic, we compute the cosine similarity between every word in $\hat{V}$ and all cluster centroids in the embedding space. For a single word w, its embedding $\boldsymbol{\omega}$ and a single cluster with centroid $\boldsymbol{\mu}$, we hence compute:

$$sim(\boldsymbol{\omega}, \boldsymbol{\mu}) = \frac{\boldsymbol{\omega} \cdot \boldsymbol{\mu}}{\|\boldsymbol{\omega}\|\|\boldsymbol{\mu}\|} \tag{1}$$

where $\boldsymbol{\omega} \cdot \boldsymbol{\mu} = \sum_{i=1}^{L} \omega_i \mu_i$ and

$$\|\boldsymbol{\omega}\|\|\boldsymbol{\mu}\| = \sqrt{\sum_{i=1}^{L} (\omega_i)^2} \sqrt{\sum_{i=1}^{L} (\mu_i)^2}$$

L denotes the vectors dimension in the feature space which is identical for $\boldsymbol{\omega}$ and $\boldsymbol{\mu}$.

To avoid having words in a topic that are semantically overly similar as, for example, *economics* and *economy*, each topic can be *cleaned*. The cosine similarity between the top Z words contained in a topic can be computed and all words that exceed a certain threshold, for example, 0.85,[2] are removed in descending order of the similarity with the clusters centroid. An additional advantage of the corpus expansion is the possibility to model documents in one language, but create topics in a different language, when using a multi-language embedding model.

Cleaning the topics based on a similarity threshold offers maximum flexibility. Depending on the task and the preferences, one could thus create topics with maximally divergent words. In our experiments we find a threshold of 0.85 to be reasonable and delete all words that have a cosine similarity score greater than or equal to 0.85 compared against other words present in the topic in descending order.

4.2 Corpus Expansion

In the realm of topic modeling, leveraging additional corpora has long been used to address issues associated with short documents, such as tweets. A straightforward yet effective solution involves merging similar short documents into longer ones, better suited for classical algorithms like LDA (Mehrotra et al. 2013). For example, Kant, Weisser, and Säfken (2020) offer an approach for aggregating tweets based on common hashtags, a strategy also utilized by Luber et al. (2021). On the other hand, Thielmann et al. (2021) and Thielmann, Weisser, and Krenz (2021) use expansion corpora before actual topic modeling to combat severe imbalances in their corpora. Bicalho et al. (2017) present a framework for extending short documents in topic modeling, although it only uses words already present in the main vocabulary. A similar conceptual approach is adopted by Zheng, Liu, and San Wong (2018). These methods, all involving the enrichment of documents, are susceptible to errors, as it is crucial to ensure the correctness of this enrichment. In cases where document enrichment is flawed, the topic model may generate topics that bear no relation to the original corpus, failing to uncover its latent themes and instead presenting erroneous extended topics.

2 The cosine similarity between the words "economy" and "economies," using the paraphrase-MiniLM-L6-v2 embedder (Reimers and Gurevych 2019) is, for instance, 0.9.

Algorithm 2 CEDC

Input:
- Corpus D $= \{d_1, \ldots, d_M\}$
- Vocabulary V $= \{w_1, \ldots, w_n\}$
- Embedding model $\mathcal{M}$ (e.g., All-MiniLM-L12-v2)
- Hyperparameters: K

Initialization:
- Embed documents D $\rightarrow \mathcal{D} = \{\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_M\}$
- Embed vocabulary V $\rightarrow \mathcal{V} = \{\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_n\}$

Dimensionality Reduction:
- Use UMAP to reduce the dimensionality of the documents, $\mathcal{D} \in \mathbb{R}^L \rightarrow \tilde{\mathcal{D}} \in \mathbb{R}^j$, $j < m$

Document Clustering:
- Gaussian Mixture (GMM) on $\tilde{\mathcal{D}}$
- Identify clusters $\mathbf{C}_1, \mathbf{C}_2, \ldots, \mathbf{C}_K$

Cluster centroids:
for $k = 1$ to K **do**
    - Calculate cluster centroids $\boldsymbol{\mu}_k \in \mathbb{R}^L$ in the original embedding space
end for

Enhance Vocabulary:
- Enhance candidate word vocabulary $\hat{V} = \{w_1, \ldots, w_n, w_{n+1}, \ldots, w_{n+z}\}$ and embed enhanced vocabulary $\hat{V} \rightarrow \hat{V} = \{\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_n, \boldsymbol{\omega}_{n+1}, \ldots, \omega_{n+z}\}$

Topic Extraction:
for $k = 1$ to K **do**
    - Calculate distance between $\boldsymbol{\mu}_k$ and $\hat{V}$ in the original embedding space with: $sim(\boldsymbol{\omega}, \boldsymbol{\mu}) = \frac{\boldsymbol{\omega} \cdot \boldsymbol{\mu}}{\|\boldsymbol{\omega}\| \|\boldsymbol{\mu}\|}$
    - Store the similarity scores in vector α_k
end for
for $k = 1$ to K **do**
    - $\hat{\alpha}_k = \left(\frac{\alpha_k}{max(\alpha_k)}\right)$
    $(\phi_{k,1}, \ldots, \phi_{k,n})^T$ = Sort($\hat{\alpha}_k$, *descending*)
end for

Output:
- Topics

In contrast, CEDC diverges significantly from the aforementioned methods in its conceptual approach. Notably, it refrains from expanding the documents themselves at any stage, guaranteeing the integrity of the topic modeling process. Document expansion occurs only after modeling, when candidate words that best represent a topic are selected. For a concise overview of the method, please refer to the pseudo-algorithm provided in Algorithm 2.

5. Evaluation

Given the described approach, we are effectively losing any idea of co-occurence based coherence for model evaluation. The words best describing a cluster of documents or

topic do not necessarily have to occur together often in documents. In fact, a word capturing the topic of a single document optimally does not necessarily have to be contained in that same document. Additionally, by enhancing the corpus, it might be possible that neologisms are the words best representing a topic. Imagine, for example, a set of documents being equally about software and hardware issues. The neologism *software-hardware* would be an understandable and reasonable word describing that topic, but would perform poorly in any word co-occurrence-based evaluation measure.

For evaluation, we hence propose new, non-word co-occurrence-based measures and use existing measures leveraging word embeddings (Terragni, Fersini, and Messina 2021). We validate the intruder-based metrics by computing correlations with human annotations. See the Appendix for a comprehensive overview over all introduced and used metrics.

5.1 Topic Expressivity (EXPRS)

First, we propose a novel measure inherently representing the meaningfulness of a topic. For that, we leverage stopwords, which, as widely recognized, fulfill a grammatical purpose, but transport nothing about the meaning of a document (Salton 1989; Wilbur and Sirotkin 1992). Hence, we compute the vector embeddings of all stopwords and calculate a centroid embedding. Subsequently, we compute the cosine similarity between a topic centroid and the stopword centroid (see Figure 2).

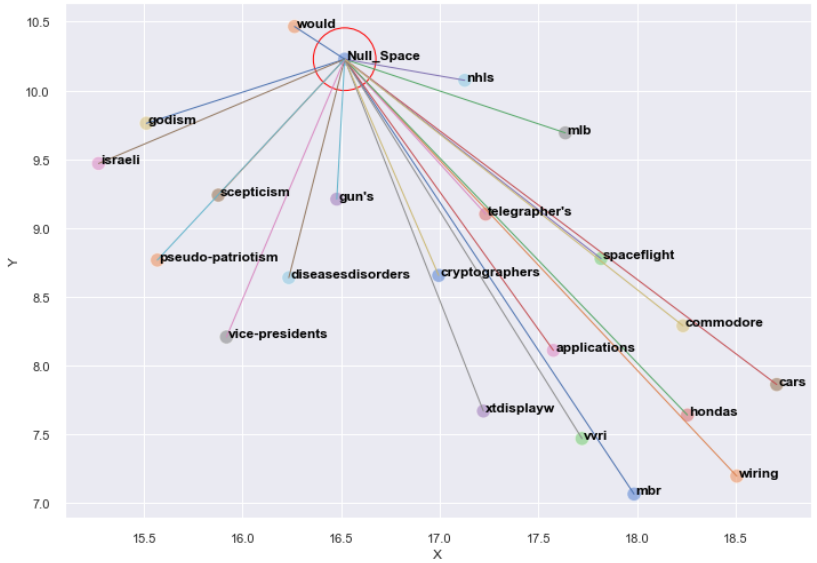

Figure 2
The expressivity of a model is captured by averaging over the topic centroids cosine similarity to the null space, defined as the centroid of all embedded stopwords. For visualization the vector dimensions are heavily reduced, but the overall expressivity is still visualized. Due to the dimensionality reduction, the axes are just labelled "X" and "Y," respectively. The visualized topics are created from the 20 Newsgroups data set with the CEDC method and a single topic, "would," created with a LDA model. The topic's top word is annotated at the topic's position in the reduced embedding space.

The weighted topic vector centroid, $\boldsymbol{\gamma}_k$, is computed by taking the top Z words and normalizing their weights, such that $\sum_{i=1}^{Z} \phi_{k,i} = 1$. The complete vector is hence computed as $\boldsymbol{\gamma}_k = \frac{1}{Z} \sum_{i=1}^{Z} \phi_{k,i} \boldsymbol{\omega_i}$ and the overall metric, which we call the model's **expressivity**, where we sum over all K topics, is defined as:

$$EXPRS(\boldsymbol{\gamma}, \boldsymbol{\psi}) = \frac{1}{K} \sum_{i=1}^{K} sim(\boldsymbol{\gamma}_i, \boldsymbol{\psi}) \tag{2}$$

with $\boldsymbol{\psi}$ being the centroid vector representation of all *stopwords*. Note that $\boldsymbol{\gamma}_i \neq \boldsymbol{\mu}_i$, as $\boldsymbol{\mu}_i$ is the centroid of the document cluster and $\boldsymbol{\gamma}_i$ is the centroid of topic t_i. Note also that the metrics results can differ depending on the choice of stopwords. However, this also allows for flexible domain specific adaptations where one would like to automatically evaluate a topics expressivity dependent on a custom set of stopwords.

5.2 Embedding Coherence (COH)

A measure, generally introduced by Aletras and Stevenson (2013) and reformulated by Fang et al. (2016) resembling classical coherence scores, is constructed by computing the similarity between the top Z words in a topic. While Aletras and Stevenson (2013) compute the word vectors using word co-occurrences, we follow Fang et al. (2016) and use the created word embeddings. In contrast to classical coherence, we compute the similarity between every top-Z word in the topic and do not implement a sliding-window approach. Hence, for Z words, we sum over $\frac{Z(Z-1)}{2}$ cosine similarities:

$$COH(t_k) = \sum_{i=1}^{Z-1} \sum_{j=i+1}^{Z} sim(\boldsymbol{\omega}_i, \boldsymbol{\omega}_j) \tag{3}$$

where the overall average coherence of a model is hence computed as:

$$\frac{2}{K(Z-1)Z} \sum_{k=1}^{K} COH(t_k)$$

Note that Terragni et al. (2021) additionally normalize the word embeddings, COH^{pw}, before computing the similarity scores for a more stable metric.

5.3 Word Embedding-based Weighted Sum Similarity (WESS)

A metric representing the diversity or the similarity between the topics of a topic model was introduced by Terragni, Fersini, and Messina (2021) as the word embedding-based weighted sum similarity and is slightly adjusted for comparing models with a different number of topics as:

$$WESS(T) = \frac{(K-1)K}{2} \sum_{i=1}^{K-1} \sum_{j=i+1}^{K} sim(\boldsymbol{\gamma}_i, \boldsymbol{\gamma}_j) \tag{4}$$

where $\boldsymbol{\gamma}_i$ represents the weighted topic centroid for topic i. While this metric certainly captures the similarity between topics, it does also reflect the diversity of the model. Hence, if $WESS(T)$ is close to 1, the model would have created topics that are extremely similar to one another.

Additionally, we propose three different new metrics, leveraging the idea of intruder words (Chang et al. 2009) and similarly integrating an idea of topic diversity. First, a metric that is based upon unweighted topic centroids.

5.4 Intruder Shift (ISH)

Given the top Z words from a topic, we calculate the topic's unweighted centroid, denoted as $\tilde{\boldsymbol{\gamma}}_i$. Subsequently, we randomly select a word from that topic and replace it with a randomly selected word, from a randomly selected different topic. The centroid of the resulting words is again computed, denoted as $\hat{\boldsymbol{\gamma}}_i$. Given a coherent topic and generally diverse topics, one would expect a larger shift in the topic centroids. Therefore we calculate the **intruder shift** of every topic and average over the number of topics:

$$ISH(T) = \frac{1}{K}\sum_{i=1}^{K} sim(\tilde{\boldsymbol{\gamma}}_i\hat{\boldsymbol{\gamma}}_i) \tag{5}$$

Hence, one would expect a coherent and diverse topic model to have a lower ISH score than an incoherent and non-diverse topic model.

5.5 Intruder Accuracy (INT)

The second intruder-word based metric follows the classical approach of identifying an intruder word more closely. Given Z top words of a topic, we again randomly select an intruder word from a randomly drawn topic. Subsequently, we calculate the cosine similarity for every possible pair of words within the set of the top Z words. Then we calculate the cosine similarity of each top word and the intruder $\hat{\boldsymbol{\omega}}$. Finally, our metric reports the fraction of top words to which the intruder has the least similar word embedding.

$$INT(t_k) = \frac{1}{Z}\sum_{i=1}^{Z} 1(\forall j : sim(\boldsymbol{\omega}_i, \hat{\boldsymbol{\omega}}) < sim(\boldsymbol{\omega}_i, \boldsymbol{\omega}_j)) \tag{6}$$

Thus we return the number of words from the set where the farthest word from them in the embedding space is the intruder word, divided by the number of words, Z, taken into account (see Figure 3 for a visualization).

5.6 Average Intruder Similarity (ISIM)

As a last metric, we propose the average cosine similarity between every word in a topic and an intruder word:

$$ISIM(t_k) = \frac{1}{Z}\sum_{i=1}^{Z} sim(\boldsymbol{\omega}_i, \hat{\boldsymbol{\omega}}) \tag{7}$$

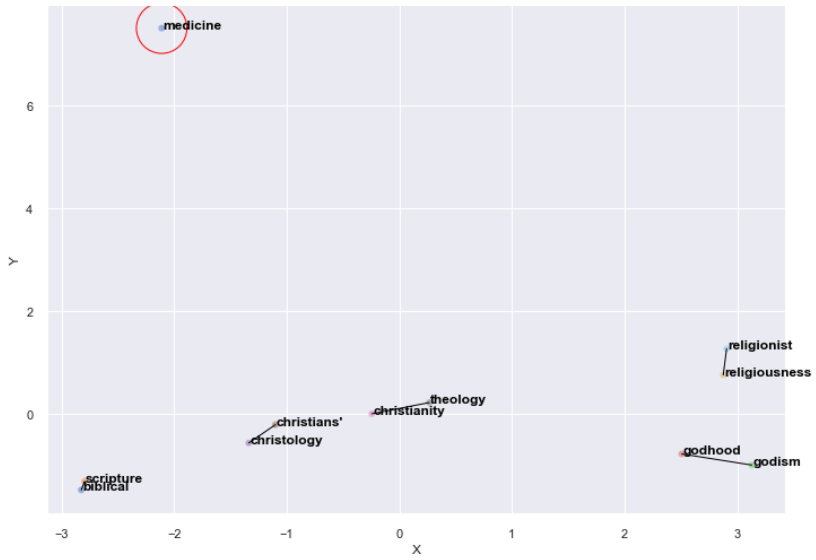

Figure 3
The intruder word detection in the embedding space. A topic, covering "religion," and an intruder word, "medicine," are plotted with heavily reduced dimensions, using principal component analysis. The intruder word clearly separates from the otherwise coherent topic, even in a two-dimensional space. Due to the dimension reduction, the axis are just labelled with "X" and "Y," respectively. The topic is again created with the CEDC method on the 20 Newsgroups data set.

To account for any induced randomness in the metrics *ISH*, *INT*, and *ISIM* due to the random choice of a particular intruder from a particular topic, we propose to calculate those metrics multiple times with differently chosen random intruder words and subsequently average the results. Hence, the robustness against the specific selection of intruder words is increased.

5.7 Validation of Metrics

To validate the intruder word based evaluation metrics we take the publicly available data from Chang et al. (2009). Similar to Lau, Newman, and Baldwin (2014) we compute the metrics over all topics and all models provided in Chang et al. (2009) for the 20 Newsgroups dataset. However, for clear interpretability, we reduce all words that include hyphens, due to the representations from Chang et al. (2009). Hence, we compute the metrics for 7,004 topics in total. We compute the accuracy of the metrics in terms of the *true* intruder and the humanly detected intruder for all metrics as well as the Pearson-*r*. While the important measures are here the correlation with the human annotations, reporting the correlations with the *true* intruder word ensures that the metrics are not inherently biased towards machine selection. For the accuracy, we consider a pre-selected or human-selected intruder to be correctly identified, if the score for this word is the lowest or highest, respectively, among all displayed top words, and for the correlation we consider the average coherence between a human-selected or an intruder-word and the other displayed words in a topic compared with those scores for the rest of the displayed words. The results are shown in Table 1. For all results it must

Table 1
Metric evaluation: Accuracy and Pearson correlation with the reported *true* (Intruder) and humanly selected (Human) intruder word from Chang et al. (2009) for all models and all topics on the 20 Newsgroups dataset. The three best results for the human correlation and accuracy are marked in bold. One can see that the metric evaluation for different embedding models produces impressive results, given the correlation between participants of 0.77. The paraphrase-MiniLM-L6-v2 performs best, considering *INT* and *ISIM*, closely followed by the GloVe model.[3]

	Accuracy		Correlation	
Score	Intruder	**Human**	Intruder	**Human**
Paraphrase-MiniLM-L6-v2				
ISH	0.613	0.512	0.526	0.492
INT	0.722	**0.622**	0.775	**0.728**
ISIM	0.810	**0.686**	0.574	**0.539**
Multi-qa-mpnet-base-dot-v1				
ISH	0.675	0.573	0.598	0.567
INT	0.700	**0.604**	0.751	0.708
ISIM	0.791	0.672	0.543	0.511
All-MiniLM-L12-v2				
ISH	0.766	**0.652**	0.519	**0.591**
INT	0.677	0.58	0.723	0.687
ISIM	0.766	0.652	0.519	0.490
All-mpnet-base-v2				
ISH	0.763	**0.652**	0.626	**0.592**
INT	0.661	0.577	0.727	0.689
ISIM	0.763	0.652	0.511	0.482
All-distilroberta-v1				
ISH	0.766	**0.652**	0.625	**0.592**
INT	0.677	0.587	0.729	0.687
ISIM	0.766	0.652	0.519	0.490
word2vec GoogleNews				
ISH	0.413	0.335	0.338	0.302
INT	0.719	0.603	0.774	**0.715**
ISIM	0.820	**0.684**	0.554	**0.506**
Glove Wikipedia				
ISH	0.622	0.506	0.496	0.439
INT	0.750	**0.634**	0.786	**0.727**
ISIM	0.808	**0.677**	0.595	**0.549**

be noted that the human answers have some ambiguity in them. As reported by Lau et al. (2014), the Pearson-*r* between the human answers was 0.77.

Hence, the results for *INT* with a maximum correlation of 0.728 is highly credible and outperforms the reported correlations (Lau, Newman, and Baldwin 2014) for coherence evaluation metrics. Interestingly, *ISIM* performs best when considering the

3 As embedding models we consider the Paraphrase-MiniLM-L6-v2 model (Reimers and Gurevych 2019), the All-MiniLM-L12-v2 model (Wang et al. 2020), the All-mpnet-base-v2 model (Song et al. 2020), the Multi-qa-mpnet-base-dot-v1 model (Song et al. 2020), and the All-distilroberta-v1 model (Liu et al. 2019) as well as a word2vec model pre-trained on the GoogleNews corpus and a GloVe model pre-trained on a Wikipedia corpus.

Table 2
Metric evaluation for previously proposed metrics: Accuracy and Pearson correlation with the reported *true* (Intruder) and humanly selected (Human) intruder word from Chang et al. (2009) for all models and all topics on the 20 Newsgroups dataset. Here, we report the results for NPMI coherence (*NPMI*), Embedding Coherence (COH^{pw}), Word Embedding-based Centroid Coherence (*WECC*), and Contextualized Pointwise-Mutual-Information (*CPMI*).

	Accuracy		Correlation	
Score	Intruder	**Human**	Intruder	**Human**
NPMI	0.787	0.655	0.617	0.468
NPMI *	0.381	0.312	0.457	0.277
COH^{pw}	0.305	0.253	0.206	0.109
WECC	0.565	0.454	0.445	0.353
CPMI	0.06	0.058	0.080	0.057
INT	0.722	*0.622*	**0.775**	**0.728**
ISIM	**0.810**	**0.686**	0.574	*0.539*

* Using only 20,000 documents from the reference corpus as for CPMI.

accuracy for the *true* intruder word, but significantly worse when considering the human selected word. We find that, independent of the chosen model, the newly introduced metrics strongly outperform the results reported by Lau, Newman, and Baldwin (2014) at the topic-level with reported Pearson correlations of around $r = 0.6$.

Additionally, we compare the accuracy and correlation with four other existing metrics. More specifically, we compute results for normalized pointwise mutual information (*NPMI*) Coherence (Lau, Newman, and Baldwin 2014), as well as embedding coherence (COH^{pw}) and word embedding-based centroid coherence (*WECC*) (Terragni et al. 2021). The results are shown in Table 2. See the Appendix for a comprehensive overview over all of the mentioned evaluation metrics. As a base corpus for metrics leveraging additional documents, we use a Wikipedia dump with a size of around 1.6 million documents. We use the provided implementation from the OCTIS package with default parameters (Terragni et al. 2021). Note that for COH^{pw} and *WECC* the used Word2Vec (Le and Mikolov 2014) embeddings are normalized by dividing by the sum of the vector entries before computing the cosine similarity, which we find significantly decreases the quality of those metrics as compared to not including this sum-based normalization. We also include the recently proposed contextualized point-wise mutual information (*CPMI*) metric proposed by Rahimi et al. (2023), where the classical word probabilities from NPMI are exchanged with an estimate for the probability of words in context based on a pre-trained masked language model. We set the segment length and segment step parameter to 40. Because this metric is highly computationally demanding, we are only able to use a fraction of Wikipedia comprising around 20,000 documents. One can also note that even with this reduced corpus size, computing the *CPMI* scores for the benchmark study takes around 23 hours using an A100 graphics card when evaluating the word-pair-likelihoods with a BERT model (Devlin et al. 2018) as proposed by Rahimi et al. (2023). Note that all other metrics are computed in less than 3 minutes on inferior CPUs.

Interestingly, we find very bad correlations between *CPMI* and the human evaluation. However, Rahimi et al. (2023) report very low correlations between classical *NPMI* and *CPMI* as well as surprisingly low scores for *CPMI* for otherwise well performing

topics.[4] Additionally, Rahimi et al. (2023) also find lower correlations between intruder words detected by a chatbot and CPMI compared with NPMI, which is validated in our results. This is also confirmed when we use the same reference corpus of 20,000 documents for NPMI that we used for CPMI. While the performance of NPMI with a smaller reference corpus is drastically reduced, it is still superior to CPMI in terms of intruder metrics which resonates with Rahimi et al. (2023).

Overall, we find that the introduced metrics *INT* and *ISIM* achieve the highest correlations with human evaluation. We find a comparably lower correlation for NPMI compared to Lau, Newman, and Baldwin (2014), since they use a different Wikipedia source corpus for computing the metric but are very similar to the results reported by Stammbach et al. (2023) and Hoyle et al. (2021). However, even the reported correlation scores of around 0.6 by Lau, Newman, and Baldwin (2014) are lower than the correlations for *INT* of 0.73.

5.7.1 Heuristic Analysis. While we compute correlation scores between the presented metrics and human evaluation, the metric *Topic Expressivity* is more heuristically motivated. The general idea is to punish models that create garbage topics more severely as this tends to be an issue when using word- or document clustering techniques (e.g., for BERTopic [Grootendorst 2022]). In addition to Figure 2, a simple example can motivate the validity. Consider two topics:

> Topic 1:
>
> Spacecraft, Neil Armstrong, Orbit, Spaceship, Satellite, Nasa, Solarwind, Apollo 11, Rocket, Moon
>
> Topic 2:
>
> is, in, and, with, have, we, are, about, from, has

Computing the NPMI coherence, without adjusting for stopwords or word length and using the 20 Newgroups corpus as the reference corpus, would lead to the following NPMI coherence scores: Topic $1 = -0.34908$ and Topic $2 = 0.2041$. However, Topic 2 does not capture any semantic meaning whereas Topic 1 clearly is about outer space. The bad NPMI coherence scores for Topic 1 stem from the fact that neither *Neil Armstrong* nor *Apollo 11* occur in the reference corpus. The topic expressivity, however, would severely punish Topic 2, with scores of 0.31 and 0.9, respectively (smaller scores are better). Note, that we adjusted the NPMI coherence measure in our benchmark study to correctly account for stopwords, by filtering the reference corpus, respectively. However, more or less meaningless words such as *without* are not accounted for in, for example, the nltk stopword list (Bird, Klein, and Loper 2009). Thus, the introduced measure offers a heuristic solution to penalize topics that uncover little semantic meaning.

6. Results

To evaluate the proposed model, we compare the model results with different benchmark models. We also demonstrate the validity of our two hypotheses on corpus expansion and noun phrases stated in Section 2.

4 Topic: god, christian, people, believe, jesus. Reported score: 0.017.

As a corpus expanding the reference corpus in CEDC for topic extraction, we use the Brown corpus taken from nltk (Bird, Klein, and Loper 2009), which we also use for filtering the vocabulary for noun-phrases. Note that for our applications we do not account for *n*-grams, which could further improve the results of CEDC. We compute the proposed metrics from Section 5 except for the *ISH* metric due to its inferior performance on the intruder word detection task (Table 1). Additionally, we compute NPMI scores (Lau, Newman, and Baldwin 2014) with the input corpus as the reference corpus and topic diversity (WESS) and word embedding pairwise coherence scores (COH^{pw}) using the OCTIS framework (Terragni et al. 2021). All word embedding-based metrics are computed with the paraphrase-MiniLM-L6-v2 model (Reimers and Gurevych 2019) due to the results from Table 1, except for *WESS* and COH^{pw} where we use OCTIS' default pre-trained word2vec (Le and Mikolov 2014) model. The word2vec model is trained on the GoogleNews corpus. The number of top words, Z, taken into account for the metrics *EXPRS*, *COH*, *WESS*, *INT*, and *ISIM* is 10. For *INT* and *ISIM*, we randomly select an intruder word from a randomly selected topic 50 times and report the averages.

To confirm our two hypotheses from Section 2 that expanding the reference corpus and only considering nouns for topic extraction can increase the topic quality, we perform several analyses. We compare the presented method with and without reference corpus expansion and with and without noun phrase filtering. The averaged results over 3 datasets can be seen in Table 3.

6.1 Hypothesis I: Corpus Expansion

Our results confirm our hypothesis that expanding the reference corpus leads to creating better topics depicted by nearly all metrics. Unsurprisingly, we find that NPMI coherence scores, only using the reference corpus for computing the coherence, are decreased when expanding the reference corpus during topic extraction. Additionally, we find that using a smaller pre-trained model for computing the metrics, as the leveraged word2vec (Le and Mikolov 2014) model for $COH^{pw}$and *WESS*, also shows a decrease in performance when expanding the reference corpus. That is presumably due to the smaller vocabulary size used in these models.

To demonstrate that CEDC actually takes words from the expansion corpus to create the topics, we analyze how many of the top words are actually taken from the expansion corpus; see Table 4. Over all 4 datasets, always using the brown corpus as the expansion corpus, CEDC creates topics with around 50% of words taken from the expansion corpus.

6.2 Hypothesis II: Noun Phrases

We find that the noun-based models perform worse than the models that consider all types of words and for the different embedding models used to construct the evaluation metrics. However, we find that when cleaning the topics the topic quality increases when using only nouns as compared to using all word types. Additionally we find that expanding the reference corpus and only considering nouns achieves better performance than no expansion and using all word types.

6.3 Benchmarks

As comparison models, we use BERTopic (Grootendorst 2022) and Top2Vec (Angelov 2020) as closely related models and representatives of clustering based topic models;

Table 3
Comparison of noun-based topic extraction vs. non-noun-based model extraction for the CEDC model. The reported metrics are averaged over the results for three datasets, the 20 Newsgroups dataset, the BBC News dataset, and the M10 dataset. All datasets are taken from OCTIS. All models are fitted using the all-MiniLM-L6-v2 model (Reimers and Gurevych 2019). Given the results from Table 1, paraphrase-MiniLM-L6-v2 is used for the embedding based evaluation metrics. We report the baseline metrics for a model not using an expanded corpus and using all word types and report the differences to that baseline. We find that especially expanding the reference corpus leads to better topics, represented by nearly all metrics. As expected, the NPMI coherence scores are considerably worse, when expanding the reference corpus. That is due to the fact that we used the original corpus the models where fit on as the NPMI coherence reference corpus. Additionally, we find that only considering nouns for topic words can increase the evaluation metrics, especially when we clean the topics.

	Coherence Measures			Diversity Measures			Intruder Measures	
Model	NPMI ↑	COH^{pw} ↑	COH ↑	TOP DIV ↑	WESS ↓	EXPRS ↓	ISIM ↓	INT ↑
CEDC*	0.016	0.430	0.427	0.783	0.377	0.459	0.184	0.719
CEDC+	−0.757	−0.079	+0.07	+0.08	−0.069	−0.061	−0.019	+0.085
CEDC**	−0.018	+0.011	−0.014	−0.013	+0.016	+0.007	+0.004	−0.031
CEDC**+	−0.70	−0.073	+0.011	+0.052	−0.046	−0.043	−0.027	+0.050
Cleaned with Similarity Threshold of 0.85								
CEDC*	0.014	0.433	0.421	0.775	0.386	0.467	0.189	0.708
CEDC+	−0.752	−0.095	+0.042	+0.077	−0.060	−0.055	−0.026	+0.066
CEDC**	−0.016	+0.010	+0.013	+0.003	−0.012	+0.004	±0	+0.021
CEDC**+	−0.689	−0.081	+0.032	+0.055	−0.048	−0.045	−0.029	+0.045

* Baseline.
** Only nouns.
\+ Expanded.

Table 4
Percentage of words from topic top-words taken from the expansion corpus for the CEDC method.

	20NG	Reuters	M10	BBC	Average
All words	73.3%	26.3%	57.5%	72.0%	57.3%
Only Nouns	64.8%	17.7%	50.5%	63.0%	49.0%

LDA (Blei, Ng, and Jordan 2003) as a model not leveraging pre-trained embeddings; CTM (Bianchi, Terragni, and Hovy 2020) as a generative probabilistic model leveraging pre-trained embeddings; CTMNeg (Adhya et al. 2023) using CTM as the base model; a simple K-means model, closely following the architecture from Grootendorst (2022), but replacing HDBSCAN with a K-means clustering approach; ETM (Dieng, Ruiz, and Blei 2020) leveraging word2vec (Le and Mikolov 2014); NeuralLDA; and ProdLDA (Srivastava and Sutton 2017). All models are fit using the OCTIS framework (Terragni et al. 2021). Where applicable the same pre-trained embedding model as for CEDC,

Table 5
Average rank table over all datasets and all metrics. We find that simple corpus expansion outperforms all hyperparameter tuned benchmark models. On average, we find document clustering methods to perform remarkably well with the best average *traditional* model being CTMNeg based on CTM (Bianchi, Terragni, and Hovy 2020; Adhya et al. 2023).

Model	Avg. Rank ↓
K-means	6.2
BERTopic†	8.3
Top2Vec†	7.2
TOP2Vec	5.7
LDA	9.0
ProdLDA	8.7
NeuralLDA	10.4
ETM	9.2
CTM	7.4
CTMNeg	4.8
CEDC^{+}	4.1
CEDC*	5.5
CEDC^{*+}	4.5

† HDBSCAN results with > 10 topics.
* Only nouns.
$^{+}$ Expanded topic corpus.

all-MiniLM-L6-v2 (Reimers and Gurevych 2019), is used. Note that we perform extensive hyperparameter tuning for all models except for CEDC. For comparing CEDC with other models we use 4 standard benchmark datasets, 20 Newsgroups, Reuters (Lewis 1997), BBC News, and M10, as shown in Tables 6 and B.3–B.4. We fix the number of topics to the *true* number of topics of 20, 90, 10, and 5, respectively. We compute average rank scores for all models over all datasets. Table 5 shows the average rank over all models, all datasets, and all metrics. CEDC performs best and topic expansions clearly improve topic quality. A complete average rank table over all models and datasets can be found in the Appendix, Table B.1. For all tested models, we use the same pre-trained embedding model all-MiniLM-L6-v2 (Reimers and Gurevych 2019), where applicable. NPMI coherence scores are calculated as presented by Lau, Newman, and Baldwin (2014). For the best possible comparison, we use the same dimensionality reduction for CEDC as is used in Toc2Vec (Angelov 2020) and BERTopic (Grootendorst 2022). Hence, we use UMAP (McInnes, Healy, and Melville 2018) and reduce the dimensions to 5, explicitly using the same hyperparameters as done in the mentioned models. The same is done for the simple K-means model.

For LDA, ProdLDA, NeuralLDA, ETM, CTMNeg, and CTM, we optimize over various hyperparameters with Bayesian optimization as provided by the OCTIS package (Terragni et al. 2021). We use model perplexity, measured based on the evidence lower bound of a validation sample of documents, as the objective function in order to not rely on metrics, such as NPMI coherence or WESS, that measure either cohesion or separation of topics. LDA is optimized over the parameters of the two symmetric Dirichlet priors on the topic-specific word distribution and the document-specific

Table 6
Benchmark results on the 20 Newsgroups and Reuters datasets. All models are fit using the all-MiniLM-L6-v2 pre-trained embedding model (Reimers and Gurevych 2019) where applicable. paraphrase-MiniLM-L6-v2 is used for the evaluation metrics ISIM, INT, TOP DIV, and EXPRS. For the metrics available in OCTIS we use the default embeddings which are pre-trained word2vec embeddings on the Google News corpus. Extensive hyperparameter tuning is performed for the comparison models (see Appendix). All models, except BERTopic and Top2Vec, are fit with a pre-specified number of 20 or 90 topics respectively. BERTopic and Top2Vec detect the *optimal* number of topics automatically, hence we fit the model as intended by the authors. However, we additionally fit a K-means model using the class based tf-idf topic extraction method from BERTopic with 20 and 90 topics, respectively, and hierarchically reduce the number of topics in Top2Vec.

	20 Newsgroups							
	Coherence Measures			Diversity Measures			Intruder Measures	
Model	NPMI ↑	COH^{pw} ↑	COH ↑	TOP DIV ↑	WESS ↓	EXPRS ↓	ISIM ↓	INT ↑
K-means	0.080	0.081	0.289	0.312	0.920	0.466	0.138	0.414
BERTopic†	0.033	0.039	0.244	0.362	0.607	0.499	0.151	0.280
Top2Vec†	0.164	0.080	0.341	0.370	0.288	0.472	0.156	0.513
Top2Vec	0.158	0.100	0.384	0.346	0.825	0.442	0.152	0.654
LDA	−0.141	0.031	0.260	0.281	0.875	0.447	0.181	0.275
ProdLDA	−0.003	0.064	0.247	0.344	0.835	0.518	0.157	0.243
NeuralLDA	−0.187	0.011	0.210	0.590	0.820	0.685	0.193	0.131
ETM	−0.514	0.038	0.274	0.634	0.265	0.695	0.259	0.197
CTM	−0.069	0.027	0.251	0.360	0.725	0.533	0.167	0.301
CTMNeg	0.051	0.473	0.264	0.895	0.328	0.497	0.151	0.313
$CEDC^{+}$	−0.893	0.364	0.523	0.925	0.234	0.368	0.130	0.886
$CEDC^{*}$	0.156	0.443	0.414	0.775	0.352	0.460	0.171	0.742
$CEDC^{*+}$	−0.807	0.342	0.460	0.885	0.256	0.380	0.126	0.832

	Reuters							
	Coherence Measures			Diversity Measures			Intruder Measures	
Model	NPMI ↑	COH^{pw} ↑	COH ↑	TOP DIV ↑	WESS ↓	EXPRS ↓	ISIM ↓	INT ↑
K-means	−0.139	0.042	0.209	0.441	0.578	0.531	0.151	0.179
BERTopic†	−0.158	0.039	0.202	0.475	0.584	0.556	0.152	0.167
Top2Vec†	−0.240	0.067	0.340	0.504	0.159	0.407	0.206	0.304
Top2Vec	−0.168	0.075	0.367	0.505	0.271	0.388	0.216	0.376
LDA	−0.822	0.025	0.387	0.533	0.394	0.660	0.364	0.172
ProdLDA	−0.650	0.005	0.256	0.441	0.299	0.573	0.203	0.197
NeuralLDA	−0.446	0.013	0.209	0.645	0.920	0.733	0.196	0.129
ETM	−0.920	0.008	0.486	0.676	0.096	0.671	0.467	0.190
CTM	−0.602	0.012	0.285	0.441	0.362	0.617	0.237	0.209
CTMNeg	−0.521	0.402	0.234	0.580	0.412	0.571	0.189	0.172
$CEDC^{+}$	−0.581	0.167	0.489	0.539	0.316	0.320	0.172	0.695
$CEDC^{*}$	−0.252	0.198	0.421	0.458	0.365	0.344	0.176	0.610
$CEDC^{*+}$	−0.252	0.179	0.431	0.483	0.356	0.339	0.172	0.643

† HDBSCAN results with $>$ 20 or 90 topics, respectively.
* Only nouns.
+ Expanded topic corpus.

topic distribution. For ProdLDA, NeuralLDA, CTMNeg, and CTM, the learning rate parameter, as well as the number of layers and the number of neurons per layer in the inference network, are considered. Finally, for ETM, we tune the learning rate, the number of hidden units in the encoder, and the embedding size.

Since BERTopic and Top2Vec are highly insensitive to different hyperparameter settings of the underlying HDBSCAN algorithm and also do not provide a way to measure the (marginal) likelihood of data, we choose the default hyperparameters for those models. While finding the optimal hyperparameters for these models might improve their performances compared to the models where we implemented hyperparameter tuning, the same is true for CEDC.

For CEDC we do not implement any form of hyperparameter tuning. Hence, the GMM is fit using scikit-learns default parameters. The convergence threshold for the Expectation Maximization (EM) algorithm is 0.0001; each component has its own general covariance matrix and 1e-6 is added to the covariance diagonals for regularization purposes. The maximum number of iterations in the EM algorithm is set to 100 and K-means is used to initialize the weights. Hence, the results achieved by CEDC could be further optimized, for example, by optimizing GMM with respect to the Bayesian- or Akaike Information Criterion. Additionally, the pre-trained embedding could be fine-tuned, which is true for all models leveraging pre-trained embeddings and could additionally improve the models' performance (Thielmann, Weisser, and Säfken 2022; Bianchi, Terragni, and Hovy 2020).

As expected, the models closely related to CEDC also perform well. However, while Top2Vec, BERTopic, and the used K-means model are closely related to the proposed CEDC, CEDC achieves much better results concerning all metrics. Notably, CTM demonstrates exceptional performance on smaller datasets (please refer to the Appendix material). Since CTM leverages pre-trained document embeddings, the performance improvement compared to, for example, ETM when regarding small corpora is to be expected. CTMNeg exhibits remarkable proficiency in terms of topic diversity due to its utilization of negative sampling, inherently yielding dissimilar topics. Given that we exclusively utilize the source corpus for the computation of NPMI coherence, it is unsurprising that CEDC exhibits considerably diminished performance when considering the expansion corpus. This occurs because the topics generated include words that are absent in the evaluation corpus. Thus, CEDC* without corpus expansion outperforms CEDC$^+$ in terms of NPMI coherence. Notably, CEDC* performs quite similarly across almost all metrics in comparison to Top2Vec and K-means, as they share conceptual similarities. Additionally, CEDC$^+$ consistently demonstrates strong performance for all metrics for all datasets, consistently achieving an average rank within the top four for all metrics except NPMI. CEDC$^+$ performs best in 5 of 8 metrics and achieves an average rank of 3.0 when not considering NPMI coherence. Furthermore, our findings do not support the notion that models employing a hard clustering approach significantly underperform on a multi-label dataset like Reuters, when compared to models incorporating soft clustering techniques (as observed in CTM/ETM vs. Top2Vec/BERTopic results).

This is also confirmed when taking a closer look at the extracted topic words (see Tables C.1–C.2). While CTMNeg also creates coherent topics, it also inflates topics with words that are ambiguous or rather unrelated to the overal theme, such as demonstrated in topic 15: [make, church, people, time, work, president, day, thing, give, job]. The same is true for ProdLDA or K-means. This becomes especially evident when looking at the 20 Newsgroups dataset and the *encryption* topic.

CEDC: [secure, encryption, security, encrypt, privacy]
CTMNeg: [key, chip, government, encryption, clipper]
ProdLDA: [algorithm, escrow, government, encryption, agency]
K-means: [key, encryption, chip, clipper, escrow]

Table 7
Two selected topics (*Sports* and *Space*) for the best performing topic models accross all metrics as well as for a *bad* performing model.

Topic	Words
CEDC	game, league, player, play, baseball, sport, pitch, hockey, team, bat
CEDC	orbit, satellite, solar, planet, shuttle, mission, earth, rocket, moon, plane
CTMNeg	lose, playoff, hockey, fan, baseball, watch, play, shot, devil, ranger
CTMNeg	image, space, format, mission, satellite, datum, send, orbit, include, shuttle
K-means	game, team, player, play, season, win, score, year, hit, goal
K-means	space, launch, orbit, satellite, mission, earth, solar, moon, shuttle, planet
Top2Vec	baseball, league, hockey, playoff, pitch, sport, game, ball, team, bat
Top2Vec	space, moon, solar, orbit, opportunity, government, proposal, mission, planet, technology
ProdLDA	game, team, win, player, muslim, play, playoff, genocide, turkish, pen
ProdLDA	mission, orbit, flight, station, launch, fuel, moon, solar, surface, year

From the top-5 words it is relatively hard to conclude that the CTMNeg topic is about encryption. Whereas the other words, such as *key* and *chip*, certainly are related to *encryption* they are so only in the context knowledge of what this topic is about. Table 7 shows two selected topics from several models fit on the 20 Newsgroups dataset. The coloring scheme is performed based on cosine similarity in the embedding space. We find that ProdLDA, the worst performing model in terms of automated evaluation metrics, also quite obviously mixes up topics—for example, *sports* and *mideastern politics*. On the other hand, especially CEDC and K-means create topics where each word clearly belongs to the overall detected themes.

7. Conclusion

We develop a novel model for topic extraction beyond the mere occurrence of words in the reference corpus. We are able to show that expanding the reference corpus improves model performance. Additionally, we can confirm that restricting the word types for topic extraction by only considering nouns can also lead to improved topic quality, under certain conditions. CEDC outperforms commonly used state-of-the-art topic models on multiple benchmark datasets, even in cases where the comparison models underwent extensive hyperparameter tuning while no hyperparameter tuning was performed for CEDC.

Given that almost all newly introduced topic models are evaluated automatically (Hoyle et al. 2021), automatic evaluation metrics are of utmost importance. Hoyle et al. (2021) even postulated that automatic topic model evaluation is broken, as the current used metrics have overall low correlations with human judgment of topic quality. Hoyle et al. (2022) even go a step further and argue that neural topic modeling altogether is broken. However, in their comprehensive analysis they fail to address human evaluation or intruder word-based topic evaluation. Looking at Tables C.1–C.3 we cannot conclude that neural topic models perform worse than non-neural models and

find that neural topic models can achieve great results also for automated evaluation metrics that strongly align with human judgment. Instead, we find that, for instance, LDA while also creating concise topics creates topics that lack any specific meaning, such as: [time, thing, lot, good, bad, make, feel, pretty, real, experience]. We present multiple novel evaluation metrics closely following state-of-the-art human evaluation of topic model quality and achieve great correlations with human evaluation. We greatly improve upon the correlation with human evaluation compared to the currently most often used metric, NPMI, achieving correlations of around $r = 0.73$ compared to NPMI correlations of $r = 0.63$. The proposed approach of using word embeddings and cosine similarity achieves impressive results given the overall lower agreement between human responses (Pearson-$r = 0.77$).

Additionally, we introduce a novel evaluation metric, based upon the centroid cluster of *stopwords* in the embedding space. Given the approach of enhancing the reference corpus, the described model might be especially useful when evaluating short texts or identifying sparsely represented topics in a corpus (Thielmann, Weisser, and Krenz 2021; Thielmann et al. 2021). Through the inherent sparsity of the data, the words best describing a topic might not be included in the reference corpus and an enhancement could thus greatly improve the creation of topics.

8. Limitations

Automated evaluation of topic model quality is inherently difficult. That difficulty is considerably increased by the fact there is no gold standard or even a ground truth for the quality of a topic. Chang et al. (2009) introduced the reasonable approach of evaluating the coherence of a set of words with intruder-words. However, one cannot expect 100% agreement between people when it comes to judging whether a word is an intruder word in a topic. The proposed evaluation metrics achieve impressive results with human annotations; they cannot, however, reflect human ambiguity or extreme subtlety in perceived topic quality. Additionally, as all evaluation metrics based upon human evaluation and hence experimental results achieved with human participants, the metrics might reflect a selection bias (WEIRD) (Henrich, Heine, and Norenzayan 2010). Further embedding models could be evaluated and tested and larger human evaluation studies could be conducted.

Recent findings about the dominance of certain dimensions in transformer embeddings (Timkey and van Schijndel 2021) suggest an inherent bias in transformer embeddings that could negatively affect similarity measures in the semantic space. Our results do not suggest that such a bias negatively influences the modeling results; however, this study does not look into the dimensionality effects, which could be the topic of further research.

Moreover, the creation of transformer models solely for the purpose of topic extraction that emphasize, for example, the beginnings of phrases due to their increased importance to the underlying topics of a subsubsection (Kieras 1980, 1981) could greatly improve upon the existing methods.

Appendix A. Supplemental Methodology

To make reading easier, we provide a full notation list. All used variables and their notation can be found here.

Table A.1
Variable list.

V	Vocabulary
D	Corpus
M	Number of documents in the corpus
d_i	Document i
w_i	Word i in V
$\boldsymbol{\omega}_i$	Word i represented in the embedding space
$\boldsymbol{\delta}_i$	Document i represented in the embedding space
$\hat{\boldsymbol{\delta}}_i$	d_i represented in the reduced embedding space
t_k	Topic k
T	Set of topics
$\phi_{k,i}$	Probability of word i in topic k
$\boldsymbol{\gamma}_k$	Topic centroid vector of topic k
$\boldsymbol{\mu}_k$	Mean of document cluster k
$\boldsymbol{\theta}$	Document cluster/topic matrix
$\boldsymbol{\beta}$	Word cluster/topic matrix
ψ	Null Space/centroid of all stopwords

All modeling steps from the proposed method are presented here in extensive form. First, the target corpus should be embedded. This can be done, either using contextualized transformer embeddings, as, for example, Bianchi, Terragni, and Hovy (2020) showed that contextualized embeddings can improve topic quality. However, approaches as used by Sia et al. (2020) where every word is embedded singularly and the documents are represented as centroid vectors of all occurring words are also possible. Second, the dimensions of the embedded documents, $\boldsymbol{\delta_i}$, are reduced due to the curse of dimensionality. Afterwards, the reduced embeddings, $\hat{\boldsymbol{\delta}}_i$, are clustered, for example, using GMM such that soft clustering is possible. The centroids for each document cluster, $\boldsymbol{\mu}_k$, are computed. Next, the corpus is filtered for nouns and all nouns present in the corpus supplemented by all nouns present in an expansion corpus are embedded. Note that here the same embedding procedure must be chosen as for the documents (see, e.g., Angelov 2020; Grootendorst 2022). Then, the similarity between all candidate words and all document cluster centroids is computed. Based on the candidate embeddings and the similarity to the document clusters $\boldsymbol{\mu}_k$, the topic centroids $\boldsymbol{\gamma}_k$ are computed and similar to LDA, we get a document topic matrix, $\boldsymbol{\theta}$, and a word topic matrix, $\boldsymbol{\beta}$. Last, a cleaning step can be performed to remove overly similar words from the topics.

Appendix B. Additional Benchmark Results

In addition to the 20 Newsgroups and Reuters dataset, we fit all models on the M10 and BBC News datasets. Both datasets are taken from OCTIS (Terragni et al. 2021). CEDC again outperforms most other models on nearly all metrics. Interestingly, CTM achieves good results for the BBC News dataset, which is comparably small with <2,000 documents. For the M10 dataset, which is composed of scientific papers and hence a more *difficult* dataset, we find that topic expansion strongly improves the model performance.

Table B.1
Benchmark results on the BBCNews dataset. All models are fit using the all-MiniLM-L6-v2 pre-trained embedding model (Reimers and Gurevych 2019) where applicable. paraphrase-MiniLM-L6-v2 is used for the evaluation metrics ISIM, INT, TOP DIV, and EXPRS. For the metrics available in OCTIS we use the default embeddings which are pre-trained word2vec embeddings on the Google News corpus. Extensive hyperparameter tuning is performed for the comparison models (See Appendix). All models, except BERTopic and Top2Vec, are fit with a pre-specified number of 5 topics. BERTopic and Top2Vec detect the optimal number of topics automatically, hence we fit the model as intended by the authors. However, we additionally fit a K-means model using the class based tf-idf topic extraction method from BERTopic with 5 topics and hierarchically reduce the number of topics in Top2Vec.

	Coherence Measures			Diversity Measures			Intruder Measures	
Model	NPMI	COH^{pw}	COH	TOP DIV	WESS	EXPRS	ISIM	INT
K-means	4.8	6.8	7.6	7.6	8.2	7.0	1.5	5.8
BERTopic†	5.5	9.0	11.4	9.5	8.5	9.5	3.4	10.0
Top2Vec†	5.8	6.5	5.8	10.0	7.2	6.5	10.0	5.5
Top2Vec	4.8	5.0	4.5	7.5	7.2	4.0	8.2	4.2
LDA	6.5	11.2	8.8	7.8	8.5	6.8	11.1	11.1
ProdLDA	7.0	9.6	10.8	7.5	8.5	10.5	6.8	9.2
NeuralLDA	9.2	11.0	12.6	5.4	11.8	12.8	8.0	12.8
ETM	7.8	11.2	6.0	7.5	7.0	10.5	13.0	10.8
CTM	8.0	10.6	9.4	6.6	4.2	5.8	7.0	7.5
CTMNeg	4.8	1.0	7.8	2.1	4.8	6.8	3.9	7.6
$CEDC^{+}$	11.8	3.8	1.0	4.5	3.2	2.0	5.4	1.0
$CEDC^{*}$	5.1	2.0	3.2	8.2	7.2	6.0	8.6	3.2
$CEDC^{*+}$	10.1	3.2	2.2	6.8	4.5	3.0	4.1	2.2

† HDBSCAN results with > 10 topics.
* Only nouns.
+ Expanded topic corpus.

Table B.2
Average rank table over all datasets when not considering NPMI.

Model	Avg. Rank ↓
K-means	6.4
BERTopic†	8.8
Top2Vec†	7.4
TOP2Vec	5.8
LDA	9.3
ProdLDA	9.0
NeuralLDA	10.6
ETM	9.4
CTM	7.3
CTMNeg	4.8
$CEDC^{+}$	3.0
$CEDC^{*}$	5.5
$CEDC^{*+}$	3.7

† HDBSCAN results with > 10 topics.
* Only nouns.
+ Expanded topic corpus.

Table B.3
Benchmark results on the M10 dataset. All models are fit using the all-MiniLM-L6-v2 pre-trained embedding model (Reimers and Gurevych 2019) where applicable. paraphrase-MiniLM-L6-v2 is used for the evaluation metrics ISIM, INT, TOP DIV, and EXPRS. For the metrics available in OCTIS we use the default embeddings that are pre-trained word2vec embeddings on the Google News corpus. Extensive hyperparameter tuning is performed for the comparison models. All models, except BERTopic and Top2Vec, are fit with a pre-specified number of 10 topics. BERTopic and Top2Vec detect the *optimal* number of topics automatically, hence we fit the model as intended by the authors. However, we additionally fit a K-means model using the class-based tf-idf topic extraction method from BERTopic with 10 topics and hierarchically reduce the number of topics in Top2Vec.

	Coherence Measures			Diversity Measures			Intruder Measures	
Model	NPMI ↑	COH^{pw} ↑	COH ↑	TOP DIV ↑	WESS ↓	EXPRS ↓	ISIM ↓	INT ↑
K-means	−0.108	0.063	0.254	0.940	0.354	0.458	0.149	0.320
BERTopic†	−0.318	0.056	0.231	0.628	0.424	0.514	0.165	0.219
Top2Vec†	−0.345	0.083	0.315	0.060	0.547	0.478	0.220	0.326
Top2Vec	−0.270	0.100	0.335	0.780	0.496	0.454	0.198	0.484
LDA	−0.176	0.035	0.244	0.830	0.330	0.440	0.208	0.177
ProdLDA	−0.251	0.074	0.222	0.970	0.425	0.508	0.170	0.220
NeuralLDA	−0.571	0.030	0.186	0.373	0.582	0.581	0.185	0.118
ETM	−0.204	0.044	0.255	0.330	0.591	0.500	0.268	0.151
CTM	−0.322	0.060	0.239	0.950	0.247	0.353	0.172	0.271
CTMNeg	−0.152	0.447	0.256	1.0	0.261	0.378	0.157	0.311
$CEDC^{+}$	−0.8411	0.338	0.512	0.855	0.322	0.383	0.179	0.827
$CEDC^{*}$	−0.5762	0.441	0.419	0.770	0.394	0.420	0.193	0.719
$CEDC^{*+}$	−0.8033	0.358	0.451	0.825	0.339	0.395	0.166	0.818

† HDBSCAN results with > 10 topics.
* Only nouns.
+ Expanded topic corpus.

Since we only consider the training corpora for the NPMI metric, CEDC has an inherent disadvantage in this metric. When we compare overall average rankings over all metrics except for NPMI, the advantage of CEDC over the benchmark models becomes even more apparent, which is demonstrated by the average performance rankings omitting NPMI scores shown in Table B.2.

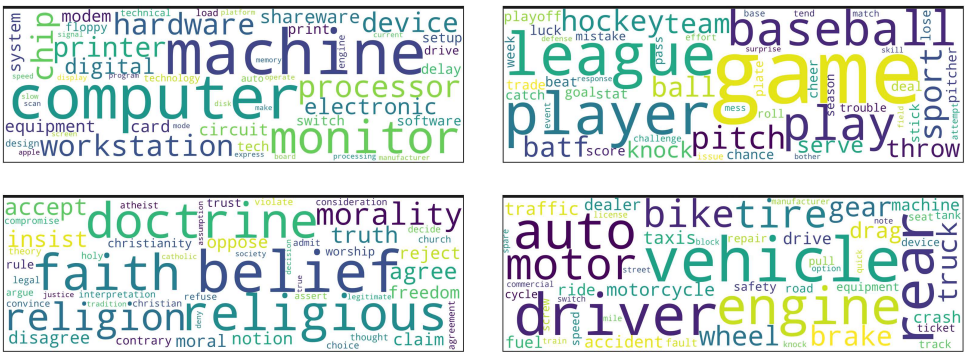

Figure B.1
Four created topics with CEDC from the 20 Newsgroups dataset.

Table B.4
Benchmark results on the BBC News dataset. All models are fit using the all-MiniLM-L6-v2 pre-trained embedding model (Reimers and Gurevych 2019) where applicable. paraphrase-MiniLM-L6-v2 is used for the evaluation metrics ISIM, INT, TOP DIV, and EXPRS. For the metrics available in OCTIS we use the default embeddings which are pre-trained word2vec embeddings on the Google News corpus. Extensive hyperparameter tuning is performed for the comparison models. All models, except BERTopic and Top2Vec, are fit with a pre-specified number of 10 topics. BERTopic and Top2Vec detect the *optimal* number of topics automatically, hence we fit the model as intended by the authors. However, we additionally fit a K-means model using the class based tf-idf topic extraction method from BERTopic with 5 topics and hierarchically reduce the number of topics in Top2Vec.

	Coherence Measures			Diversity Measures			Intruder Measures	
Model	NPMI ↑	COH^{pw} ↑	COH ↑	TOP DIV ↑	WESS ↓	EXPRS ↓	ISIM ↓	INT ↑
K-means	−0.868	0.088	0.333	1.000	0.297	0.490	0.139	0.667
BERTopic†	−0.307	0.053	0.232	0.623	0.423	0.513	0.166	0.218
Top2Vec†	−0.339	0.082	0.314	0.059	0.542	0.477	0.218	0.329
Top2Vec	−0.324	0.097	0.334	0.920	0.419	0.435	0.173	0.528
LDA	−0.150	0.029	0.208	0.840	0.447	0.480	0.202	0.098
ProdLDA	−0.290	0.050	0.212	0.960	0.484	0.541	0.171	0.199
NeuralLDA	−0.460	0.077	0.190	1.000	0.574	0.558	0.170	0.136
ETM	−0.184	0.043	0.249	0.600	0.510	0.489	0.252	0.182
CTM	−0.299	0.050	0.232	1.000	0.236	0.369	0.148	0.241
CTMNeg	−0.194	0.454	0.263	1.000	0.278	0.499	0.150	0.344
CEDC^{+}	−0.851	0.351	0.456	0.810	0.368	0.444	0.186	0.701
CEDC*	0.055	0.440	0.402	0.765	0.433	0.518	0.202	0.602
CEDC^{*+}	−0.772	0.373	0.403	0.795	0.397	0.474	0.181	0.656

† HDBSCAN results with > 5 topics.
* Only nouns.
+ Expanded topic corpus.

Appendix C. Experimental Setup

For all tested models, we use the same pre-trained embedding model all-MiniLM-L6-v2 (Reimers and Gurevych 2019), where applicable. NPMI coherence scores are calculated as presented by Lau, Newman, and Baldwin (2014). For the best possible comparison, we use the same dimensionality reduction for CEDC as is used in Doc2Vec (Angelov 2020) and BERTopic (Grootendorst 2022). Hence, we use UMAP (McInnes, Healy, and Melville 2018) and reduce the dimensions to 5, explicitly using the same hyperparameters as done in the mentioned models. The same is done for the simple K-means model.

Table C.1
The CEDC model and CTMNeg fit on the 20 Newsgroups dataset. The topic extraction corpus for CEDC is expanded with the brown corpus taken from the nltk package (Bird, Klein, and Loper 2009).

Topic	Words
	CEDC
1	game, league, player, play, baseball, sport, pitch, hockey, team, bat
2	application, program, software, workstation, code, window, file, programming, print, tool
3	bullet, firearm, weapon, attack, shoot, kill, action, armed, protect, protection
4	homosexual, homosexuality, sexual, insist, reject, accept, morality, contrary, disagree, oppose
5	machine, chip, circuit, electronic, hardware, equipment, device, computer, workstation, processor
6	vehicle, auto, engine, rear, tire, driver, truck, motor, wheel, bike
7	israeli, conflict, oppose, attack, peace, struggle, arab, turkish, armenian, kill
8	action, consideration, complain, oppose, bother, rule, issue, policy, insist, accept
9	complain, respond, response, consideration, suggestion, idea, bother, challenge, influence, accept
10	orbit, satellite, solar, planet, shuttle, mission, earth, rocket, moon, plane
11	mailing, mail, send, email, contact, message, telephone, address, customer, request
12	printer, print, font, format, digital, make, manufacture, manufacturer, machine, workstation
13	sell, sale, purchase, offer, brand, customer, supply, vendor, deal, price
14	send, inform, publish, message, newsgroup, reader, mailing, post, topic, mail
15	lose, result, score, loss, beat, challenge, division, note, gain, fall
16	belief, faith, doctrine, accept, truth, religion, notion, religious, trust, interpretation
17	hardware, computer, device, drive, machine, monitor, electronic, chip, shareware, modem
18	patient, complain, care, affect, effect, issue, treat, suffer, response, treatment
19	interpretation, truth, assert, argue, claim, consideration, logic, insist, complain, belief
20	secure, encryption, security, encrypt, privacy, protect, protection, scheme, enforcement, access
	CTMNeg
1	key, chip, government, encryption, clipper, security, algorithm, secure, encrypt, law
2	state, gun, law, weapon, crime, fire, society, batf, government, kill
3	car, buy, ride, engine, bike, speed, problem, turn, back, brake
4	image, space, format, mission, satellite, datum, send, orbit, include, shuttle
5	human, belief, life, religion, faith, exist, evidence, word, science, claim
6	study, year, health, patient, medical, drug, disease, effect, doctor, treatment
7	game, play, year, point, team, score, season, player, good, hit
8	sell, price, sale, condition, shipping, buy, offer, pay, interested, manual
9	window, run, version, server, support, problem, display, software, set, client
10	people, kill, armenian, woman, time, child, live, day, man, fire
11	entry, error, output, program, line, problem, set, window, write, remark
12	drive, scsi, card, speed, controller, problem, disk, ide, board, fast
13	screen, advance, mouse, draw, print, convert, character, driver, monitor, video
14	armenian, turkish, genocide, muslim, population, greek, jewish, history, state, political
15	make, church, people, time, work, president, day, thing, give, job
16	interested, address, reply, newsgroup, fax, student, mailing, mail, contact, news
17	light, water, energy, temperature, orbit, turn, air, battery, large, side
18	thought, understand, portion, practice, speak, express, opinion, spread, aren, frequently
19	lose, playoff, hockey, fan, baseball, watch, play, shot, devil, ranger
20	portion, longer, frequently, due, introduction, primarily, consist, poor, improve, variety

Table C.2
ProdLDA and K-means fit on the 20 Newsgroups dataset.

Topic	Words
	ProdLDA
1	image, system, format, file, software, processing, graphic, quality, package, analysis
2	unique, permission, importance, complaint, portion, weekend, previously, extreme, storage, gather
3	paper, topic, helpful, article, advance, reader, permission, author, progress, reply
4	connect, card, port, pin, connector, controller, monitor, scsi, bus, driver
5	algorithm, escrow, government, encryption, agency, chip, clipper, key, scheme, secret
6	good, year, time, game, ride, bike, hit, run, bag, blue
7	window, screen, problem, run, font, menu, default, driver, display, error
8	unique, importance, remark, unknown, combine, portion, closely, extreme, behavior, precisely
9	game, team, win, player, muslim, play, playoff, genocide, turkish, pen
10	people, work, time, fire, make, kill, armenian, soldier, building, tear
11	mission, orbit, flight, station, launch, fuel, moon, solar, surface, year
12	motif, server, widget, system, mail, faq, mailing, client, programming, distribution
13	people, sin, man, church, love, make, pray, verse, give, life
14	unique, chain, shipping, storage, importance, condition, imagine, enable, portion, sale
15	advance, monitor, connect, mouse, board, multi, parallel, video, modem, download
16	unique, permission, thought, importance, possibly, combine, duty, violation, complaint, mess
17	cop, batf, knock, dog, joke, justify, funny, compound, armed, bat
18	make, atheism, point, belief, good, evidence, question, atheist, science, existence
19	car, problem, buy, drive, dealer, engine, bike, brake, tire, gear
20	people, disease, drug, health, medical, firearm, gun, patient, treatment, state
	K-means
1	modem, printer, mouse, print, port, serial, driver, fax, laser, problem
2	window, font, file, motif, application, run, display, program, server, color
3	religion, belief, atheist, faith, church, christian, atheism, exist, religious, sin
4	homosexual, homosexuality, man, moral, gay, sex, love, sin, church, word
5	car, bike, ride, engine, mile, oil, dog, tire, brake, road
6	patient, disease, doctor, medical, health, drug, treatment, food, study, pain
7	gun, government, firearm, weapon, law, people, crime, president, state, police
8	battery, sound, power, circuit, voltage, radio, audio, input, heat, output
9	game, team, player, play, season, win, score, year, hit, goal
10	key, encryption, chip, clipper, escrow, phone, government, algorithm, agency, security
11	image, file, program, format, version, user, software, entry, graphic, server
12	price, sale, sell, shipping, offer, condition, copy, cover, manual, include
13	drive, scsi, disk, controller, ide, hard, floppy, bus, boot, bio
14	science, evidence, theory, post, scientific, fact, point, claim, argument, context
15	batf, gas, compound, warrant, claim, court, evidence, start, law, tax
16	space, launch, orbit, satellite, mission, earth, solar, moon, shuttle, planet
17	post, money, delete, net, article, newsgroup, year, school, information, news
18	mail, address, send, list, request, email, post, software, message, phone
19	card, monitor, video, driver, memory, board, mhz, bit, vga, mode
20	armenian, turkish, israeli, jewish, kill, village, arab, genocide, russian, soldier

Table C.3
LDA fit on the 20 Newsgroups dataset.

Topic	Words
	LDA
1	problem, line, sound, power, work, current, ground, control, correct, radio
2	information, list, space, mail, system, send, address, launch, computer, datum
3	make, talk, work, job, money, question, president, spend, general, press
4	year, world, jewish, history, event, source, live, greek, ago, city
5	key, chip, bit, number, message, encryption, clipper, block, algorithm, system
6	game, win, play, team, year, player, good, season, hit, lose
7	force, war, israeli, plan, area, attack, military, policy, accord, peace
8	word, true, religion, man, life, church, love, make, belief, sin
9	post, read, question, good, find, book, write, answer, article, reply
10	drive, card, system, driver, disk, work, run, memory, scsi, video
11	buy, price, sell, good, offer, cost, pay, sale, interested, include
12	small, water, large, effect, high, make, gas, theory, side, air
13	people, kill, armenian, child, woman, man, die, turkish, dead, burn
14	file, window, image, program, version, application, color, display, run, server
15	car, phone, company, engine, technology, product, front, market, mile, big
16	time, thing, lot, good, bad, make, feel, pretty, real, experience
17	care, drug, year, study, increase, doctor, number, patient, disease, medical
18	gun, law, weapon, police, crime, criminal, person, fire, firearm, bill
19	start, back, leave, turn, happen, stop, bike, time, guy, call
20	people, government, state, group, case, individual, society, idea, free, personal

Appendix D. Summary of Existing and Proposed Metrics

- **NPMI Coherence (NPMI):**
  - NPMI Coherence measures the average normalized pointwise mutual information between words in a topic. The word (co)occurrence probabilities are estimated based on document chunks of a reference corpus.
  - Established metric for traditional topic models.
  - No embeddings are required.
  - Needs a potentially large reference corpus.
  - Results depend on hyperparameters for estimating the word (co)occurrence probabilities.
  - Can yield inconsistent or biased results, especially for Neural Topic Models.
- **Embedding Coherence (COH):**
  - Embedding Coherence measures the average cosine similarity of word pairs within a topic.
  - Simple, easily explainable metric.

  - Requires word embeddings.
    - Choice of word embeddings is a hyperparameter.

- **Word Embedding-based Centroid Coherence (WECC):**
  - WECC computes the average cosine similarity of words to the centroid of the words of their topic.
  - Can take advantage of word embeddings.
  - Lower computational complexity than COH.
  - Requires word embeddings.
    - Choice of word embeddings is a hyperparameter.

- **Contextualized Pointwise Mutual Information (CPMI):**
  - The CPMI metric works similarly to the NPMI metric. Here the similarity of two words is assessed based on likelihood ratios by a BERT model.
  - Utilizes the powerful BERT model.
  - Potentially very high computational cost.
  - High computational cost can lead to poor performance of this metric since the reference corpus has to be restricted.

- **Topic Expressivity (EXPRS):**
  - Topic Expressivity is computed as the cosine similarity between the centroid of the embeddings of the stopwords and the centroid of the topic.
  - Novel metric for measuring the expressivity of a topic.
  - Requires word embeddings.
    - Choice of word embeddings is a hyperparameter.
  - Results depend on the choice of stopwords.

- **Word Embedding-based Weighted Sum Similarity (WESS):**
  - The diversity metric WESS measures the average cosine similarity between the centroids of a given collection of topics.

  - Simple metric for assessing the diversity of a clustering.
  - Requires word embeddings.
    - Choice of word embeddings is a hyperparameter.

- **Intruder Shift (ISH):**
  - Intruder shift assesses how much a topic's centroid changes when the embedding of an unrelated intruder word is added.
  - Intruder-based metric.
  - Intuitive explanation.
  - Requires word embeddings.
    - Choice of word embeddings is a hyperparameter.
  - Depends on the choice of intruder (averaging over many choices alleviates this issue).

- **Intruder Accuracy (INT):**
  - For a given topic and a given intruder word, Intruder Accuracy is the fraction of topwords to which the intruder has the least similar embedding among all topwords.
  - Intruder-based metric.
  - High correlation with human assessments.
  - Requires word embeddings.
    - Choice of word embeddings is a hyperparameter.
  - Depends on the choice of intruder (averaging over many choices alleviates this issue).

- **Average Intruder Similarity (ISIM):**
  - ISIM measures the average cosine similarity of topwords of a topic to an intruder word.
  - Intruder-based metric.
  - Intuitive explanation.
  - High correlation with human assessments.

– Requires word embeddings.

– Choice of word embeddings is a hyperparameter.

– Depends on the choice of intruder (averaging over many choices alleviates this issue).

Acknowledgments

Funding by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) within project 450330162 is gratefully acknowledged.

References

Adhya, Suman, Avishek Lahiri, Debarshi Kumar Sanyal, and Partha Pratim Das. 2023. Improving contextualized topic models with negative sampling. *arXiv preprint arXiv:2303.14951*.

Agarwal, Deepak and Bee-Chung Chen. 2010. FLDA: Matrix factorization through latent Dirichlet allocation. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, pages 91–100. https://doi.org/10.1145/1718487.1718499

Aggarwal, Charu C., Alexander Hinneburg, and Daniel A. Keim. 2001. On the surprising behavior of distance metrics in high dimensional space. In *International Conference on Database Theory*, pages 420–434. https://doi.org/10.1007/3-540-44503-X_27

Aletras, Nikolaos and Mark Stevenson. 2013. Evaluating topic coherence using distributional semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers*, pages 13–22.

Angelov, Dimo. 2020. Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

Barde, Bhagyashree Vyankatrao and Anant Madhavrao Bainwad. 2017. An overview of topic modeling methods and tools. In *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 745–750. https://doi.org/10.1109/ICCONS.2017.8250563

Beghtol, Clare. 1986. Bibliographic classification theory and text linguistics: Aboutness analysis, intertextuality and the cognitive act of classifying documents. *Journal of Documentation*, 42(2):84–113. https://doi.org/10.1108/eb026788

Bhatia, Shraey, Jey Han Lau, and Timothy Baldwin. 2017. An automatic approach for document-level topic model evaluation. *arXiv preprint arXiv:1706.05140*. https://doi.org/10.18653/v1/K17-1022

Bhatia, Shraey, Jey Han Lau, and Timothy Baldwin. 2018. Topic intrusion for automatic topic model evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 844–849. https://doi.org/10.18653/v1/D18-1098

Bianchi, Federico, Silvia Terragni, and Dirk Hovy. 2020. Pre-training is a hot topic: Contextualized document embeddings improve topic coherence. *arXiv preprint arXiv:2004.03974*. https://doi.org/10.18653/v1/2021.acl-short.96

Bicalho, Paulo, Marcelo Pita, Gabriel Pedrosa, Anisio Lacerda, and Gisele L. Pappa. 2017. A general framework to expand short text for topic modeling. *Information Sciences*, 393:66–81. https://doi.org/10.1016/j.ins.2017.02.007

Bird, Steven, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc.

Blei, David M., 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. https://doi.org/10.1145/1667053.1667056

Blei, David M. and John D. Lafferty. 2007. A correlated topic model of science. *The Annals of Applied Statistics*, 1(1):17–35. https://doi.org/10.1214/07-AOAS114

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

Chang, Jonathan, Sean Gerrish, Chong Wang, Jordan Boyd-Graber, and David Blei. 2009. Reading tea leaves: How

humans interpret topic models. *Advances in Neural Information Processing Systems*, 22.

Chien, Jen Tzung, Chao-Hsi Lee, and Zheng-Hua Tan. 2018. Latent Dirichlet mixture model. *Neurocomputing*, 278:12–22. https://doi.org/10.1016/j.neucom.2017.08.029

Clark, Elizabeth, Tal August, Sofia Serrano, Nikita Haduong, Suchin Gururangan, and Noah A. Smith. 2021. All that's 'human' is not gold: Evaluating human evaluation of generated text. *arXiv preprint arXiv:2107.00061*. https://doi.org/10.18653/v1/2021.acl-long.565

Das, Rajarshi, Manzil Zaheer, and Chris Dyer. 2015. Gaussian LDA for topic models with word embeddings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 795–804. https://doi.org/10.3115/v1/P15-1077

Davison, Alice. 1982. A systematic definition of sentence topic. *Center for the Study of Reading Technical Report; no. 264*.

Davison, Alice. 1984. Syntactic markedness and the definition of sentence topic. *Language*, 60(4):797–846. https://doi.org/10.1353/lan.1984.0012

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dieng, Adji B., Francisco J. R. Ruiz, and David M. Blei. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453. https://doi.org/10.1162/tacl_a_00325

Fang, Anjie, Craig Macdonald, Iadh Ounis, and Philip Habel. 2016. Using word embedding to evaluate the coherence of topics from Twitter data. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1057–1060. https://doi.org/10.1145/2911451.2914729

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

Guijarro, Arsenio Jesús Moya. 2000. Towards a definition and hierarchization of topic. *Talk and Text: Studies on Spoken and Written Discourse*, edited by A. Rothwell, A. Guijarro & J. Albentosa, pages 97–116.

Henrich, Joseph, Steven J. Heine, and Ara Norenzayan. 2010. Most people are not weird. *Nature*, 466(7302):29–29. https://doi.org/10.1038/466029a, PubMed: 20595995

Hofmann, Thomas. 2001. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1):177–196. https://doi.org/10.1023/A:1007617005950

Hoyle, Alexander, Pranav Goel, Andrew Hian-Cheong, Denis Peskov, Jordan Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken? The incoherence of coherence. *Advances in Neural Information Processing Systems*, 34.

Hoyle, Alexander, Pranav Goel, Rupak Sarkar, and Philip Resnik. 2022. Are neural topic models broken? *arXiv preprint arXiv:2210.16162*. https://doi.org/10.18653/v1/2022.findings-emnlp.390

Kant, G., C. Weisser, and B. Säfken. 2020. TTLocVis: A Twitter topic location visualization package. *Journal of Open Source Software*, 5(54). https://doi.org/10.21105/joss.02507

Kieras, David E. 1980. Initial mention as a signal to thematic content in technical passages. *Memory & Cognition*, 8(4):345–353. https://doi.org/10.3758/BF03198274, PubMed: 7421575

Kieras, David E. 1981. Topicalization effects in cued recall of technical prose. *Memory & Cognition*, 9(6):541–549. https://doi.org/10.3758/BF03202348, PubMed: 7329234

Krosnick, Jon A. 2018. Questionnaire design. In *The Palgrave Handbook of Survey Research*. Springer, pages 439–455. https://doi.org/10.1007/978-3-319-54395-6_53

Lafferty, John and David Blei. 2005. Correlated topic models. *Advances in Neural Information Processing Systems*, 18.

Larochelle, Hugo and Stanislas Lauly. 2012. A neural autoregressive topic model. *Advances in Neural Information Processing Systems*, 25.

Lau, Jey Han, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539. https://doi.org/10.3115/v1/E14-1056

Le, Quoc and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International Conference on Machine Learning*, pages 1188–1196.

Lehman, Darrin R., Jon A. Krosnick, Robert L. West, and Fan Li. 1992. The focus of judgment effect: A question wording effect due to hypothesis confirnation bias. *Personality and Social Psychology Bulletin*, 18(6):690–699. https://doi.org/10.1177/0146167292186005

Lewis, David D. 1997. Reuters-21578 text categorization collection data set. UCI Machine Learning Repository. https://doi.org/10.24432/C52G6M

Liu, Yinhan, 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*.

Luber, Mattias, Anton Thielmann, Christoph Weisser, and Benjamin Säfken. 2021. Community-detection via hashtag-graphs for semi-supervised NMF topic models. *arXiv preprint arXiv:2111.10401*.

Lund, Jeffrey, Piper Armstrong, Wilson Fearn, Stephen Cowley, Courtni Byun, Jordan Boyd-Graber, and Kevin Seppi. 2019. Automatic evaluation of local topic quality. *arXiv preprint arXiv:1905.13126*. https://doi.org/10.18653/v1/P19-1076

Martin, Fiona and Mark Johnson. 2015. More efficient topic modelling through a noun only approach. In *Proceedings of the Australasian Language Technology Association Workshop 2015*, pages 111–115.

Mazarura, J., A. De Waal, and P. de Villiers. 2020. A gamma-poisson mixture topic model for short text. *Mathematical Problems in Engineering*, pages 1–17. https://doi.org/10.1155/2020/4728095

McInnes, Leland, John Healy, and Steve Astels. 2017. HDBSCAN: Hierarchical density based clustering. *The Journal of Open Source Software*, 2(11). https://doi.org/10.21105/joss.00205

McInnes, Leland, John Healy, and James Melville. 2018. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*. https://doi.org/10.21105/joss.00861

Mehrotra, R., S. Sanner, W. Buntine, and L. Xie. 2013. Improving LDA topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 889–892. https://doi.org/10.1145/2484028.2484166

Miles, Samuel, Lixia Yao, Weilin Meng, Christopher M. Black, and Zina Ben Miled. 2022. Comparing PSO-based clustering over contextual vector embeddings to modern topic modeling. *Information Processing & Management*, 59(3):102921. https://doi.org/10.1016/j.ipm.2022.102921

Newman, David, 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.

Rahimi, Hamed, Jacob Louis Hoover, David Mimno, Hubert Naacke, Camelia Constantin, and Bernd Amann. 2023. Contextualized topic coherence metrics. *arXiv preprint arXiv:2305.14587*.

Ramage, Daniel, David Hall, Ramesh Nallapati, and Christopher D. Manning. 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 248–256. https://doi.org/10.3115/1699510.1699543

Reimers, Nils and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using siamese BERT-networks. *arXiv preprint arXiv:1908.10084*. https://doi.org/10.18653/v1/D19-1410

Reynolds, Douglas A. 2009. Gaussian mixture models. *Encyclopedia of Biometrics*, 741:659-663. https://doi.org/10.1007/978-0-387-73003-5_196

Rosen-Zvi, Michal, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. 2012. The author-topic model for authors and documents. *arXiv preprint arXiv:1207.4169*.

Salton, Gerard. 1989. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley.

Sia, Suzanna, Ayush Dalmia, and Sabrina J. Mielke. 2020. Tired of topic models? Clusters of pretrained word embeddings make for fast and good topics too! *arXiv preprint arXiv:2004.14914*. https://doi.org/10.18653/v1/2020.emnlp-main.135

Song, Kaitao, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33:16857–16867.

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

Stammbach, Dominik, Vilém Zouhar, Alexander Hoyle, Mrinmaya Sachan, and Elliott Ash. 2023. Re-visiting automated topic model evaluation with large language models. *arXiv preprint arXiv:2305.12152*. https://doi.org/10.18653/v1/2023.emnlp-main.581

Terragni, Silvia, 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. https://doi.org/10.18653/v1/2021.eacl-demos.31

Terragni, Silvia, Elisabetta Fersini, and Enza Messina. 2021. Word embedding-based topic similarity measures. In *International Conference on Applications of Natural Language to Information Systems*, pages 33–45. https://doi.org/10.1007/978-3-030-80599-9_4

Thielmann, Anton, Christoph Weisser, and Astrid Krenz. 2021. One-class support vector machine and LDA topic model integration—evidence for AI patents. In *Soft Computing: Biomedical and Related Applications*. Springer, pages 263–272. https://doi.org/10.1007/978-3-030-76620-7_23

Thielmann, Anton, Christoph Weisser, Astrid Krenz, and Benjamin Säfken. 2021. Unsupervised document classification integrating web scraping, one-class SVM and LDA topic modelling. *Journal of Applied Statistics*, pages 574–591. https://doi.org/10.1080/02664763.2021.1919063, PubMed: 36819086

Thielmann, Anton, Christoph Weisser, and Benjamin Säfken. 2022. Human in the loop: How to effectively create coherent topics by manually labeling only a few documents per class. *arXiv preprint arXiv:2212.09422*.

Timkey, William and Marten van Schijndel. 2021. All bark and no bite: Rogue dimensions in transformer language models obscure representational quality. *arXiv preprint arXiv:2109.04404*. https://doi.org/10.18653/v1/2021.emnlp-main.372

Vayansky, Ike and Sathish A. P. Kumar. 2020. A review of topic modeling methods. *Information Systems*, 94:101582. https://doi.org/10.1016/j.is.2020.101582

Wang, Rui, Deyu Zhou, and Yulan He. 2019. ATM: Adversarial-neural topic model. *Information Processing & Management*, 56(6):102098. https://doi.org/10.1016/j.ipm.2019.102098

Wang, Wenhui, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.

Weisser, Christoph, Christoph Gerloff, Anton Thielmann, Andre Python, Arik Reuter, Thomas Kneib, and Benjamin Säfken. 2023. Pseudo-document simulation for comparing LDA, GSDMM and GPM topic models on short and sparse text using Twitter data. *Computational Statistics*, 38(2):647–674. https://doi.org/10.1007/s00180-022-01246-z, PubMed: 37223721

Wilbur, W. John and Karl Sirotkin. 1992. The automatic identification of stop words. *Journal of Information Science*, 18(1):45–55. https://doi.org/10.1177/016555159201800106

Zheng, Chu Tao, Cheng Liu, and Hau San Wong. 2018. Corpus-based topic diffusion for short text clustering. *Neurocomputing*, 275:2444–2458. https://doi.org/10.1016/j.neucom.2017.11.019