Quantifying Topic Model Influence on Text Layouts based on Dimensionality Reductions

Daniel Atzberger[1] [a], Tim Cech[2] [b], Willy Scheibel[1] [c],
Jürgen Döllner[1] [d], and Tobias Schreck[3] [e]
[1] *Hasso Plattner Institute, Digital Engineering Faculty, University of Potsdam, Germany*
[2] *Digital Engineering Faculty, University of Potsdam, Germany*
[3] *Graz University of Technology, Austria*
{*daniel.atzberger, willy.scheibel*}*@hpi.uni-potsdam.de*, {*tim.cech, doellner*}*@uni-potsdam.de, tobias.schreck@cgv.tugraz.at*

Keywords: Topic Modeling, Dimensionality Reductions, Text Spatializations.

Abstract: Text spatializations for text corpora often rely on two-dimensional scatter plots generated from topic models and dimensionality reductions. Topic models are unsupervised learning algorithms that identify clusters, so-called topics, within a corpus, representing the underlying concepts. Furthermore, topic models transform documents into vectors, capturing their association with topics. A subsequent dimensionality reduction creates a two-dimensional scatter plot, illustrating semantic similarity between the documents. A recent study by Atzberger et al. has shown that topic models are beneficial for generating two-dimensional layouts. However, in their study, the hyperparameters of the topic models are fixed, and thus the study does not analyze the impact of the topic models' quality on the resulting layout. Following the methodology of Atzberger et al., we present a comprehensive benchmark comprising (1) text corpora, (2) layout algorithms based on topic models and dimensionality reductions, (3) quality metrics for assessing topic models, and (4) metrics for evaluating two-dimensional layouts' accuracy and cluster separation. Our study involves an exhaustive evaluation of numerous parameter configurations, yielding a dataset that quantifies the quality of each dataset-layout algorithm combination. Through a rigorous analysis of this dataset, we derive practical guidelines for effectively employing topic models in text spatializations. As a main result, we conclude that the quality of a topic model measured by coherence is positively correlated to the layout quality in the case of Latent Semantic Indexing and Non-Negative Matrix Factorization.

1 INTRODUCTION

Topic Models (TMs) are a class of unsupervised learning algorithms for analyzing the semantic structure of collections of documents, so-called text corpora (Crain et al., 2012). By analyzing patterns of co-occurring words within the documents, TMs extract concepts – so-called topics – as clusters in the vocabulary. Thereby, topics are given as vectors, whose components express the relevance of the respective term for the topic; in many cases, a human-interpretable concept can be derived from the most relevant words within a topic. Furthermore, TMs represent each document as a vector that describes its semantic composition. Besides their wide use in the NLP domain, e.g., for text classification (Aggarwal and Zhai, 2012a), text summarization (Nenkova and McKeown, 2012), or text clustering (Aggarwal and Zhai, 2012b), TMs are also used for the visualization of text corpora using a map-metaphor (Kucher and Kerren, 2019). The underlying two-dimensional scatter plot, which determines the position for each document, originates from applying a dimensionality reduction (DR).

The quality of the two-dimensional layout depends on the choice of the TM, the DR, and the respective hyperparameters. Existing works dealing with the representation of high-dimensional data by two-dimensional scatter plots usually do not consider TMs (Espadoto et al., 2021; Vernier et al., 2020), even though TMs are essential in many visualizations. Contrasting this, Atzberger et al. (2023) showed in their benchmark study, that applying a TM for text corpora as a particular case of high dimensional data can lead

[a] https://orcid.org/0000-0002-5409-7843
[b] https://orcid.org/0000-0001-8688-2419
[c] https://orcid.org/0000-0002-7885-9857
[d] https://orcid.org/0000-0002-8981-8583
[e] https://orcid.org/0000-0003-0778-8665

to higher quality layouts. However, in their experiments, the hyperparameters of the TM are fixed for each dataset. Therefore, the full potential of TMs and the impact of their quality for text spatializations remains unclear.

This paper presents an extension of the benchmark proposed by Atzberger et al. (2023). The new benchmark is given by a quadruple $\mathcal{B} = (\mathcal{D}, \mathcal{L}, Q_{TM}, Q_{DR})$ containing (1) a set of text corpora $\mathcal{D}$, (2) a set of layout algorithms $\mathcal{L}$ that are combinations of TMs and DRs, (3) a set of quality metrics Q_{TM} that capture quality aspects of TMs, and (4) a set of quality metrics Q_{DR} that capture aspects related to the accuracy and perception. By evaluating the benchmark on a computational cluster, we generate two datasets: one containing the quality scores of 35 topic models and one containing the quality scores of more than 6000 different layouts. By analyzing the result datasets, we investigate the impact of the quality of TMs on the quality of the layout. Our analysis shows that a higher quality score for TMs result in better layouts concerning accuracy and perception in case of Latent Semantic Indexing and Non-Negative Matrix Factorization.

The remaining part of this work is structured as follows: we give an overview of the related work in Section 2. Our benchmark is detailed in Section 3, and implementation details are presented in Section 4. The results are analyzed in Section Section 5 and discussed in Section 6. We conclude this work and present directions for future work in Section 7.

2 RELATED WORK

We cover three aspects that are related to our considerations: (1) benchmark studies that evaluate the accuracy of DRs, (2) benchmark studies that evaluate the perception capabilities of DRs, and (3) approaches for quantifying and exploring the quality of TMs.

Benchmarking Dimensionality Reductions for Accuracy. Different benchmarks have been proposed to derive guidelines for the effective use of DRs for visualization tasks. Those benchmarks usually comprise a set of datasets, DRs, and quality metrics for quantifying the accuracy of DRs. Thereby, accuracy approximates how well high-dimensional structures, e.g., neighborhoods, are preserved in the lower-dimensional representation (Behrisch et al., 2018). The execution of a benchmark results in a dataset, which is then the subject of further analysis, e.g., which DR shows the best results for a given dataset.

van der Maaten et al. (2009) were the first to apply this methodology to compare non-linear DRs' performance with PCA. In a similar study, Gisbrecht and Hammer (2015) focused on the performance of non-linear DRs. Espadoto et al. (2021) introduced the first *large-scale* study, comprising 18 datasets, 44 DRs, and seven quality metrics for capturing the accuracy. From the results, the authors deduced that t-SNE overall shows the best performance. In a similar study, Vernier et al. (2020) analyzed the temporal stability using customized quality metrics. Furthermore, Vernier et al. (2021) developed two modifications of t-SNE that show great results with respect to temporal stability.

Even though these benchmarks contain text corpora as datasets, TMs are not considered as part of the layout algorithms. Atzberger et al. (2023) presented a benchmark containing five different text corpora and 52 layout algorithms originating from combining a TM and a subsequent DR. As a main result, the authors show that applying a TM improves the overall accuracy of the resulting layout. However, in their benchmark, the hyperparameters of each TM are fixed, and only the hyperparameters of the DRs are varying. Although best practices were applied and the results of the TMs were manually inspected, the potential of TMs remains unclear.

Benchmarking Dimensionality Reductions for Perception. Besides the accuracy, Atzberger et al. also analyze the perception capabilities of the resulting scatter plots. Thereby perception refers to the capability of a user to perceive clusters as introduced by Sedlmair et al. (2013).

The first benchmark that measures cluster separation metrics was proposed by Xia et al. (2023). Besides a purely quantitative assessment, the authors also performed a user study to compare DRs concerning different cluster analysis tasks, e.g., cluster identification, as done in a previous work (Xia et al., 2022).

Morariu et al. (2023) investigated in a user study whether quality metrics can describe the visual appearance of two-dimensional scatter plots. In a similar approach, Xia et al. (2021) collected a human-labeled dataset to train a neural network for modeling the human perception of visual clusters. A further work that relies on human judgments was presented by Wang et al. (2018), who developed a DR that aims at maximizing the perceived class separation. Our experiments follow the methodology proposed by Atzberger et al., who solely relied on class separation metrics and no human judgment.

Evaluating Topic Models. In most cases, TMs are evaluated using quantitative measures, e.g., perplexity or coherence measures (Röder et al., 2015). Alternatively, topics can be judged according to their interpretability by inspecting their most relevant words. An example of a visualization work that analyzes the quality of TMs that way is presented by Riehmann et al. (2019). Furthermore, visualization tools have been developed to support users in interpreting topics and discovering relations between them. For example, Sievert and Shirley (2014) presented *LDAvis*, which consists of a bubble chart together with a bar chart for exploring the topics of an LDA model. However, in most visualization papers that rely on TMs, TMs are treated as a "black box", without considering the results concerning quality measures or interpretability.

3 BENCHMARK

Our benchmark $\mathcal{B} = (\mathcal{D}, \mathcal{L}, Q_{TM}, Q_{DR})$ extends the benchmark proposed by Atzberger et al. by quality metrics for TMs. In the following we will present details on each of the four components.

Datasets. The set $\mathcal{D}$ contains four text corpora. The *20 Newsgroup*, *Reuters*, *Seven Categories*, and *Emails* datasets are standard datasets from *Kaggle*[1] and widely used to evaluate NLP algorithms. The *GitHub Projects* dataset that contains the source code of 653 software projects on GitHub, where all source code files have been merged into one document, presented by Atzberger et al., could not be included in our benchmark, as the computation of the coherence value for a TM would exceed the memory consumption. Various preprocessing operations are performed to remove words that do not have semantic meaning. Besides generic steps, e.g., removing stop words, dataset-specific actions are performed. After preprocessing, the text corpora are available as a *Document-Term Matrix* (DTM), i.e., the entry in cell (i, j) indicates the absolute frequency of the j-th word in the i-th document. Furthermore, each document is assigned to a unique category describing a higher-level concept. Details on the processing and implementation of the datasets can be found in our repository[2]. The characteristics of the four text corpora, containing the number of documents m, the size of the vocabulary n, and the number of categories k, are summarized in Table 1.

Layout Algorithms. In the DTM, each row describes a document, i.e., each document is represented as an n-dimensional vector containing the absolute frequencies of the words in the document. Since the semantic similarity between documents should be independent of their length, the similarity is measured by the cosine similarity. This basic document comparison model is called the *Vector Space Model* (VSM). The VSM only considers the absolute frequency of a term within a document and thus neglects the distribution of the word across all documents. By weighting the DTM according to the *term frequency-inverse document frequency* (tf-idf) scheme, terms that occur in only a few documents, and are thus of particular relevance to a document, are given a higher weight (Aggarwal and Zhai, 2012b). Specifically, the tf-idf of a term w in document d is given by

$$\text{tf-idf}(w,d) = \frac{n(w,d)}{\sum\limits_{d' \in C} n(w,d')} \cdot \log\left(\frac{|C|}{|\{d' \in C | w \in d'\}|}\right), \quad (1)$$

where $n(w,d)$ denotes the frequency of term w in document d. Typically, only a few terms from the vocabulary occur in a single document., the DTM is thus sparse. The basic idea of TMs is to detect clusters in the vocabulary that occur together in documents (Crain et al., 2012). *Latent Semantic Indexing* (LSI) is a TM that decomposes the (mxn)-dimensional DTM as the product of an (mxK)-dimensional document-topic matrix and a (Kxn)-dimensional topic-term matrix by applying a *Singular Value Decomposition* (SVD) (Deerwester et al., 1990). The number of topics K is a hyperparameter of the model. Similarly, *Non-Negative Matrix Factorization* (NMF) decomposes the DTM as a product of two matrices (Lee and Seung, 1999). Both methods can be applied to the tf-idf weighted DTM. The cosine similarity captures the similarity between the documents. *Latent Dirichlet Allocation* (LDA) is a probabilistic TM that assumes a generative process underlying a corpus (Blei et al., 2003). Each document is described as a distribution over the topics, which are, in turn given as distributions over the vocabulary. In addition to K, LDA requires the specification of two Dirichlet priors, α and β, which encode assumptions about the document-topic distribution and topic-word distribution, respectively. Since the documents are described as distributions, they are compared using the Jensen-Shannon distance.

Table 1: Characteristics for the four datasets in our benchmark containing the number of documents m, the size of the vocabulary n and the number of categories k.

Dataset	m	n	k
20 Newsgroup	11 314	6 672	20
Emails	9 111	6 992	4
Reuters	9 122	2 953	65
Seven Categories	3 127	11 373	7

[1] www.kaggle.com/

[2] DOI: 10.5281/zenodo.10040858

By applying a TM, each document is represented as a K-dimensional vector describing the expression in the topics. Thus, it requires a subsequent DR to represent the corpus as a two-dimensional scatter plot. *Multidimensional Scaling* (MDS) iteratively computes the positioning of the documents, which should represent the pairwise distances between the documents (Cox and Cox, 2008). *t-distributed Stochastic Neighbor Embedding* (t-SNE) is considered the best-known manifold learning algorithm and is known for preserving local structures well (van der Maaten and Hinton, 2008). Besides the specification of the learning rate and the number of iterations in the training algorithm, it requires the specification of the perplexity, which controls the trade-off between local and global structures. *Uniform Manifold Approximation and Projection* (UMAP) extends t-SNE to preserve global structures (McInnes et al., 2020). UMAP requires the specification of two hyperparameters, the number of neighbors as a trade-off between preserving local and global structures, and the minimal distance that controls how close data points can be grouped together in the two-dimensional layout. In any case, the DR can also be applied to the topics, and the document position aggregated according to its document representation (Atzberger et al., 2021).

Quality Metrics for Topic Models. According to Röder et al. (2015), "a set of statements or facts is said to be coherent, if they support each other", i.e., they seem to belong to each other concerning human interpretation. In the case of TMs, a statement is given by the most relevant terms within a topic. As human evaluations are expensive to produce, several coherence measures have been proposed to quantify a TM's quality concerning human interpretability. Röder et al. (2015) developed a four-stage pipeline that categorizes existing coherence measures and allows their combination, i.e., a quadruple specifies each coherence measure. In the first stage, the set of words is segregated into smaller pieces, e.g., word pairs. In the second stage, word probabilities are computed using a reference corpus, e.g., by dividing the number of documents in which the word occurs by the total number of documents. In the third stage, a confirmation measure derives how strongly a pair of words or subsets of words belong to each other based on their probabilities, which results in a vector description. Finally, the vector entries are aggregated to a final coherence score in the fourth stage. Our experiments evaluate the TMs using the pipeline C_V. The metric C_V has shown the best results in the study of Röder et al. (2015), which compares the coherence scores for topics that have been rated by humans in previous experiments (Aletras and Stevenson, 2013; Chang et al., 2009; Lau et al., 2014).

Table 2: Libraries used in our benchmark. Besides libraries providing TMs and DRs, we also rely on libraries for text preprocessing, e.g., the removal of stop words or lemmatization.

Algorithm	Library	Version
LSI, NMF, LDA	`Gensim`	4.2.0
Coherence Measures	`Gensim`	4.2.0
t-SNE, MDS	`Scikit-Learn`	1.2.1
UMAP	`UMAP-Learn`	0.5.3
General Text Preprocessing	`NLTK`	3.7
Lemmatization	`Spacy`	3.4.3

Quality Metrics for Dimensionality Reductions. To evaluate the effectiveness of DRs to preserve local and global structures of a given input data set in a two-dimensional scatter plot, different quality metrics have been proposed and utilized in several benchmark studies (Behrisch et al., 2018). In our study, we refer to the metrics that have also been used by Atzberger et al. (2023). The *Trustworthiness* α_T measures for each point in the 2D layout the percentage of points among the seven nearest neighbors (NN) that also belong to the seven NN in the input space, averaged over all points (Venna and Kaski, 2006). Vice versa, the *Continuity* α_C measures for each point in the input space the percentage among the seven NN that are also among the seven NN in the projected space, averaged over all points (Venna and Kaski, 2006). The *7-Neighborhood hit* α_{NH} measures the percentage of points with the same label among the seven NN, averaged over all points (Paulovich and Minghim, 2006). Our fourth metric is derived from the *Shephard Diagram*, a two-dimensional scatter plot that relates the pairwise distances in the high-dimensional input space to the Euclidean distances in the layout (Joia et al., 2011). The *Shephard Digram Correlation* α_{SDC} is given by the Spearman Rank Correlation of the Shephard Diagram and thus captures the global structure.

Unlike Atzberger et al. (2023), we only rely on the *Distance Consistency* β_{DC} to measure perception, as it reflects the idea of cluster separation better than combined with other metrics and has furthermore shown as most relevant in previous studies Sedlmair and Aupetit (2015). It measures the percentage of points whose category center, i.e., the average of all points in that category, is also its nearest category center in the input space (Sips et al., 2009). In an ideal scenario, the clusters are well separated to support cluster perception.

Table 3: Number of topics for each dataset evaluated in our experiments.

Dataset	k	$K \in \{a,b,c,d,e\}$
20 Newsgroup	20	20, 25, 30, 35, 40
Emails	4	8, 10, 12, 14, 16
Reuters	65	65, 82, 98, 114, 130
Seven Categories	7	14, 17, 21, 24, 28

Table 4: Range for the hyperparameters considered in our experiments.

DR	Parameter Name	Values
t-SNE	`n_iter`	250, 1000, 4000
t-SNE	`learning_rate`	28, 129, 599
t-SNE	`perplexity`	10–40 step size 10
UMAP	`min_dist`	0.25–0.75 step size 0.25
UMAP	`n_neighbors`	5, 10, 15, 20
MDS	`max_iter`	300

4 IMPLEMENTATION DETAILS

Our implementation is based on Python 3 and actively maintained libraries for topic modeling and DR, as listed in Table 2. Our computations are carried out on a computational cluster comprising ten AMD x64 HPE XL225n Gen10 (2 AMD EPYC 7742 processors, 512 GiB RAM, and 64 cores) and eleven AMD x64 Fujitsu RX2530 M5 (2 Intel Xeon Gold 5220S processors, 96 GiB RAM, and 32 cores). The cluster is managed using *Simple Linux Utility for Resource Management* (SLURM). For more details on our implementation, we refer to our repository.

The most relevant hyperparameter of a TM is the number of topics K, which depends on the dataset, as summarized in Table 3. In the case of the Emails and Seven Categories dataset, the number of categories k is relatively low. We therefore set the lower bound of K to $2k$ and its upper bound to $4k$. In the case of the 20 Newsgroups and Reuters dataset, in an ideal scenario of $K = k$, each topic would represent one category. We therefore set the lower bound to k and the upper bound to $2k$. LDA furthermore requires the specification of its two Dirichlet priors α and β. We let α vary over the values {symmetric, asymmetric, auto} as specified by Gensim and set β constant as suggested in the guidelines of Wallach et al. (2009). We apply LSI and NMF on the DTM and its tf-idf weighted variant. The value ranges for the hyperparameters for the DRs are specified in Table 4. By iterating over each combination of datasets, TM, and DR in a grid search, we generate a dataset comprising 6346 layouts.

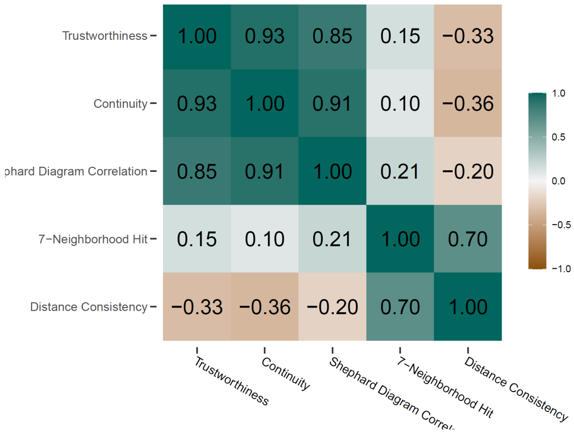

Figure 1: Heatmap showing the pairwise Pearson correlations between the five layout quality metrics using a diverging color scheme.

5 RESULTS

Evaluating our benchmark on a computational cluster results in a dataset containing more than 6000 layouts. For each entry, the accuracy and perception metrics are stored together with the coherence value of the underlying TM. We analyze the dataset in three steps: (1) we investigate the correlation of different quality metrics, (2) we investigate the impact factors on the topic model coherence, and (3) we show whether the coherence influences the quality of the resulting layout.

5.1 Correlation Analysis

Figure 1 shows the pairwise Pearson correlation between the metrics in Q_{DR}. The Trustworthiness α_T, Continuity α_C, and Shephard Diagram Correlation α_{SDC} show a strong correlation but do not correlate with the 7-Neighborhood hit α_{NH}. To weight both aspects of the accuracy equally, we define the aggregated accuracy metric α as:

$$\alpha = \frac{1}{2}\alpha_{NH} + \frac{1}{2}\left(\frac{\alpha_T + \alpha_C + 0.5\cdot(\alpha_{SDC}+1)}{3}\right), \quad (2)$$

where $0.5\cdot(\alpha_{SDC}+1)$ has replaced α_{SDC} to modify the value range from $[-1,1]$ to $[0,1]$. The Distance Consistency β_{DC} strongly correlates to α_{NH}, whereby both metrics are label-based. Since β_{DC} is the only perception metric we consider, we define

$$\beta = \beta_{DC}. \quad (3)$$

To illustrate α and β, Figure 2 shows four plots each for the 20 Newsgroup dataset. The four coherence measures strongly correlate. However, instead of aggregating them to a single metric, we solely rely on the coherence measure C_V as it has shown the best results in the study of Röder et al. (2015).

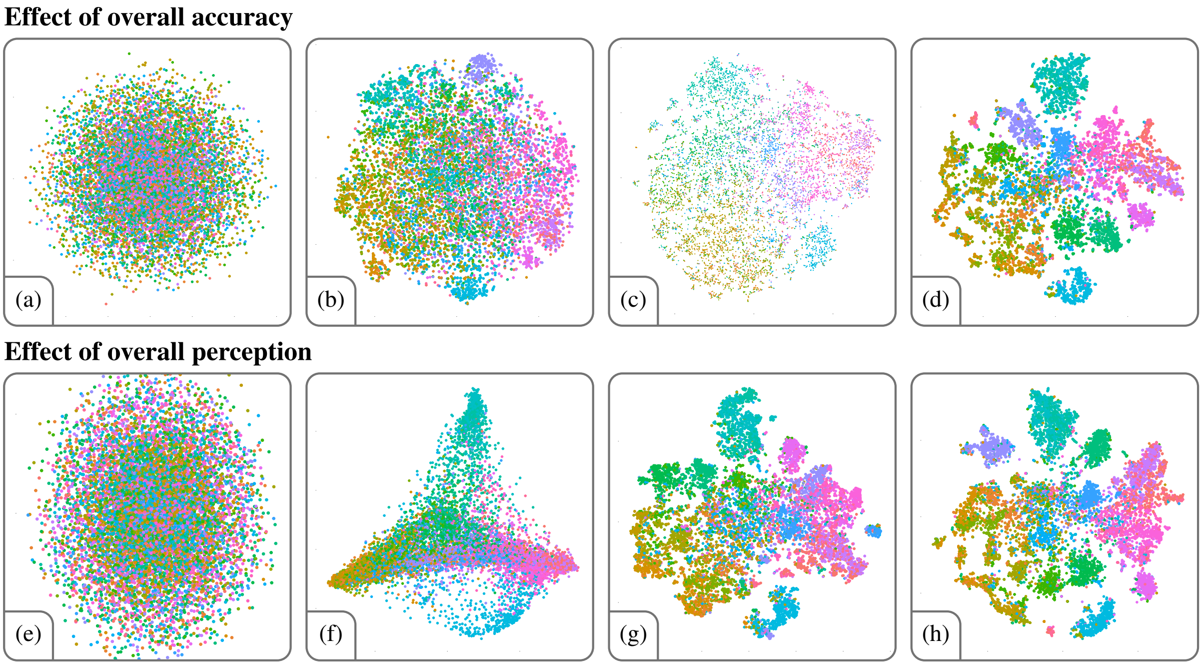

Figure 2: Scatter plots for the 20 Newsgroup corpus. The color represents the category of each document. The first row shows the effect of the overall accuracy ((a) $\alpha = 0.34$, (b) $\alpha = 0.47$, (c) $\alpha = 0.59$, (d) $\alpha = 0.71$); the second row illustrates the distance consistency ((e) $\beta = 0.05$, (f) $\beta = 0.22$, (g) $\beta = 0.40$, (h) $\beta = 0.56$).

5.2 Impact on Coherence

When applying TMs for NLP tasks, the concrete hyperparameter setting of the TM is often chosen based on a coherence value, i.e., different hyperparameter settings are evaluated, and the highest-scoring one is chosen. Figure 3 shows the quality of the TMs considered in our study in a bar chart. The height represents the value C_V of a TM for a specified number of topics $K \in \{a,b,c,d,e\}$ averaged over all four datasets. The number of topics $K \in \{a,b,c,d,e\}$, where $a < \cdots < e$ depends on the dataset, e.g, the value a is given by 20 in the case of 20 Newsgroup, 8 for the Emails dataset, 65 for the Reuters dataset, and 14 for the Seven Categories dataset.

For $K \geq b$, both variants of NMF perform best, followed by LDA and its versions and LSI performing worst. Only the case $K = a$ shows a different order. For LSI and its tf-idf weighted variant, the coherence strictly decreases with an increase of K. In the case of NMF and its tf-idf weighted variant, the coherence has a significant "jump" from $K = a$ to $K = b$ but then seems stable over K. LDA shows no clear pattern but stays within a small value range under variations of K. For every K, the LDA model with a symmetric prior α outperforms the version with the automatically learned one, and except for $K = e$ also the asymmetric one.

To summarize, when applying NMF, the tf-idf weighting does not improve the results; the best results are achieved for $K = b$. In the case of LSI and LDA, it is recommended to set $K = a$. Even though this does not significantly improve the coherence for LDA, it speeds up the training. The observation that K should be set to a also confirms the basic idea of a coherence measure to reflect the interpretability of a model, as we set $K = a$ to be the number of categories k (or in the particular case of very few categories $2k$) so that every topic could be linked to exactly one category.

5.3 Impact of Coherence on Layout Quality

We analyze the relationship between the coherence and the accuracy metric α and perception metric β, respectively; i.e., we explore whether the layout quality depends on the quality of the underlying TM. Our analysis relies on Kendall's tau, a correlation measure for ordinal data (Noether, 1981). Different from the Pearson correlation, Kendall's tau is a non-parametric measure that does not make assumptions on the underlying distribution of the data. Given two ordered sets $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$, the pairs (x_i, y_i) and (x_j, y_j) are said to be concordant if $x_i > x_j$ and $y_i > y_j$, and discordant otherwise. Let P denote the number of concordant pairs, Q the number of discordant pairs, T the number of ties in $\mathbf{x}$, and U the number of ties in $\mathbf{y}$;

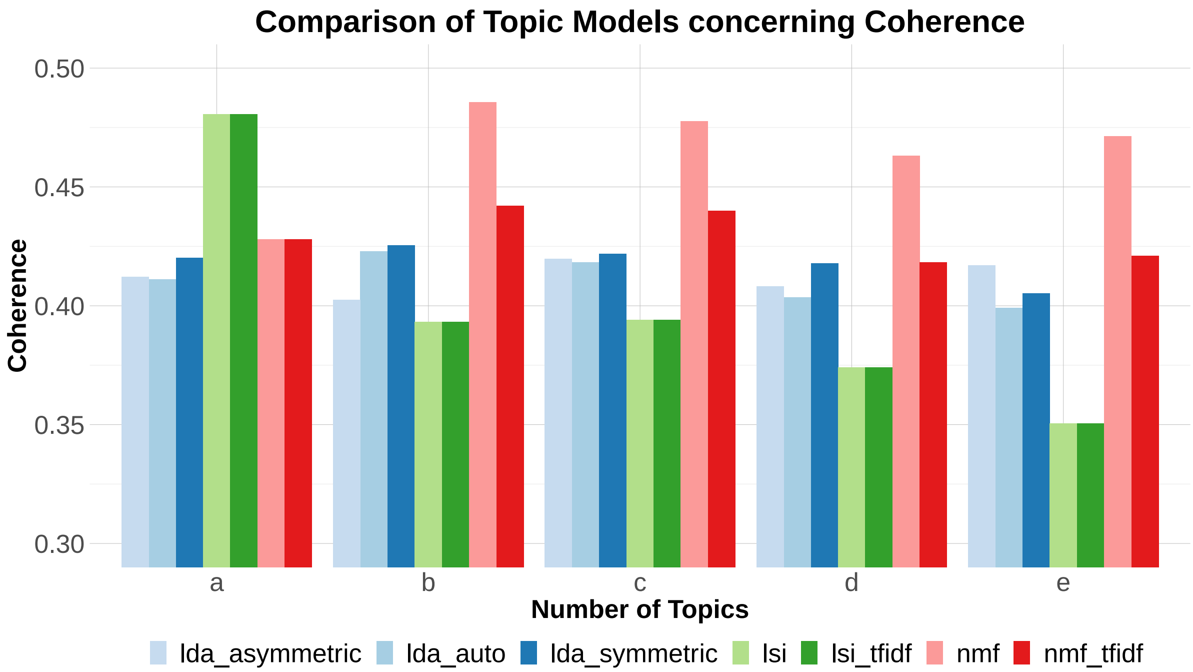

Figure 3: Grouped bar chart for comparing the coherence values of the different TMs and their variants: The color indicates the underlying TM and the saturation its variant. Note that the y-axis starts at 0.3 to support comparison.

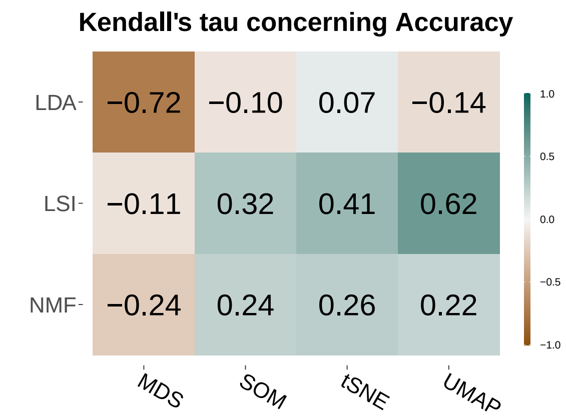

Figure 4: Kendall's tau for the sequences $\mathbf{x}_{T,P}$ and $\mathbf{y}^{\alpha}_{T,P}$, for three different TMs and four different DRs shown as a heatmap with a diverging color scheme.

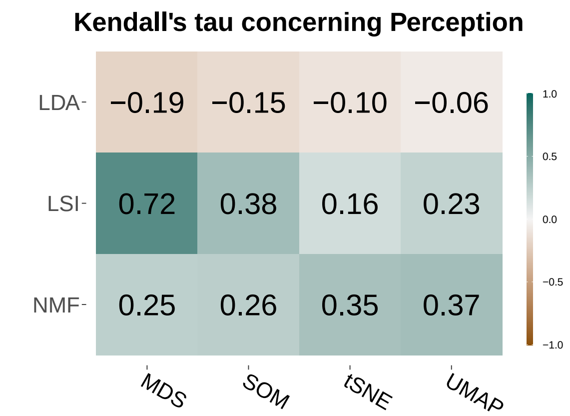

Figure 5: Kendall's tau for the sequences $\mathbf{x}_{T,P}$ and $\mathbf{y}^{\beta}_{T,P}$, for three different TMs and four different DRs shown as a heatmap with a diverging color scheme.

Kendall's tau is defined as

$$\tau = \frac{P-Q}{\sqrt{(P+Q+T)\cdot(P+Q+U)}}. \quad (4)$$

Kendall's tau ranges between $[-1,1]$, with 1 meaning that the two sequences have the same rankings.

For a given TM $T \in \{\text{LDA}, \text{LSI}, \text{NMF}\}$, and DR $P \in \{\text{MDS}, \text{SOM}, \text{ t-SNE}, \text{UMAP}\}$, let $\mathbf{x}_{T,P} = (x_1,\dots,x_n)$ denote the sequence of coherence values of fully parametrized TMs, which underly a layout of a dataset that originates from applying T as TM and P as DR. The length n is the number of evaluated hyperparameter combinations.Analogously, we define $\mathbf{y}^{\alpha}_{T,P} = (y^{\alpha}_1,\dots,y^{\alpha}_n)$ as the sequence of accuracy values α and $\mathbf{y}^{\beta}_{T,P} = (y^{\beta}_1,\dots,y^{\beta}_n)$ as the sequence of perception values β, where we assume the exact ordering as in the sequence $\mathbf{x}$. The results of Kendall's tau capturing the relationship between coherence and accuracy are shown in Figure 4, and the values of Kendall's tau capturing the relationship between coherence and perception are shown in Figure 5.

Concerning the accuracy α, LDA shows a very high ($|\tau| < 0.7$) negative to no ($|\tau| < 0.1$) correlation with any DR. However, as shown in Figure 3, the coherence for LDA models ranges within a very small range, which is neglected by the measure tau. Therefore, other changes to the model, e.g., the implementation or training method, could result in different observations. Concerning the perception β, LDA shows a low ($0.1 < |\tau| < 0.3$) negative to no correlation with any DR.

The coherence of an LSI model in combination with SOM, t-SNE, or UMAP shows a medium ($0.3 \leq |\tau| \leq 0.5$) to a high ($0.5 \leq |\tau| \leq 0.7$) positive correlation with the accuracy. Concerning perception, we observe a low ($0.1 \leq |\tau| \leq 0,3$) to a medium positive correlation. Surprisingly, LSI combined with MDS shows a very high correlation concerning β. However, as the results of Atzberger et al. (2023) showed, MDS performs worse than t-SNE and UMAP, and we therefore consider this observation as less relevant.

For NMF, the same basic pattern as for LSI is observed, but attenuated in the expressions of τ. We assume that this is because the coherence for topics $K \in \{b,c,d,e\}$ lies in a narrow range of values.

6 DISCUSSION

Based on our results, we formulate our main findings and discuss the threats to validity that underlie our argumentation.

Main Findings. Using Kendall's tau, we showed that the coherence of a TM is positively correlated with the accuracy and perception of the resulting layouts in the case of LSI and NMF in combination with t-SNE, UMAP, and SOMs. MDS shows a different pattern but is neglected, as the study of Atzberger et al. (2023) has shown that MDS performs worse than the others. For LDA, our experiments indicate a negative correlation. However, we suspect that this is because the coherence of the LDA models ranges within a small value range. This conjecture is emphasized by the observation that NMF shows a weaker correlation than LSI and also ranges in a smaller value range. Combining our findings with Atzberger et al.'s guidelines, we recommend using LSI in combination with t-SNE such that the coherence of the LSI model is maximized. In our experiments, LSI achieves its maximal coherence for $K = a$, i.e., when the number of topics matches the number of categories. This also aligns with the idea of a coherence metric to measure interpretability, as in the case $K = k$ each topic can be assigned to precisely one pre-defined category, which allows the user to cross-check the model by inspecting the topics.

Threats to Validity. Our major internal threat to validity lies in the design of our benchmark. For example, by only evaluating four datasets, it is unclear how transferable the results are for larger sets of text corpora. Vice versa, it is still being determined to what extent the results apply to specific datasets. Furthermore, the results depend on the chosen quality metrics, e.g., by selecting the distance consistency as the only perception metric, we have not considered other measures like the silhouette coefficient. Even though we evaluated different parametrizations of the layouts, other hyperparameters, e.g., the specific implementation and training method, might lead to other results. Also, the quality metrics have hyperparameters that need to be set by the user, e.g., the number of neighbors for the accuracy metrics. We consider implementation errors as the main external threat to validity. Even though we rely on actively maintained and widely used libraries, reviewed code, and did pair programming, we can not guarantee the absence of errors. To make our work more accessible and transparent and enable others to reproduce our results more quickly, we make our implementation open source.

7 CONCLUSIONS

Many text spatializations rely on a two-dimensional scatter plot, representing each document as a single point. Usually, these layouts are derived from applying a TM and a subsequent DR. Previous benchmark studies have shown that the generated layout of a text corpus layout differs strongly between the different DRs. Even though it is known that a TM can improve the layout algorithm, it is still being determined to what extent the quality of the TM affects the resulting layout. To address this issue, we proposed a benchmark $(\mathcal{D}, \mathcal{L}, Q_{TM}, Q_{DR})$ given as a quadruple of a set of text corpora $\mathcal{D}$, a set of layout algorithms $\mathcal{L}$ that originate from combining a TM and a DR, a set of coherence measures Q_{TM} to evaluate the quality of a TM, and a set of metrics Q_{DR} that quantify the accuracy and perception capabilities of a layout algorithm. By evaluating more than 6000 hyperparameter configurations, we derived a multivariate dataset for further analysis. Our results indicate that coherence is positively correlated to the accuracy metric α and the perception metric β in the case of LSI and NMF, in combination with t-SNE, UMAP, and SOMs. We see different directions for future work. Primarily, we plan to extend our benchmark to address our major internal threat to validity by including more datasets and layout algorithms. Besides accuracy and perception, other aspects of quality, e.g., temporal stability, could be quantified and taken into account, too. Furthermore, besides a quantitative study of text layout algorithms, a qualitative approach, which categorizes layouts according to their topological and geometrical properties, would be interesting. We expect that such a categorization would be beneficial for choosing DRs for specific analytics tasks.

ACKNOWLEDGEMENTS

We thank the anonymous reviewers for their valuable feedback. This work was partially funded by the German Ministry for Education and Research (BMBF) through grant 01IS22062 ("AI research group FFS-AI"). The work of Tobias Schreck was partially funded by the Austrian Research Promotion Agency (FFG) within the framework of the flagship project ICT of the Future PRESENT, grant FO999899544.

REFERENCES

Aggarwal, C. C. and Zhai, C. (2012a). A survey of text classification algorithms. In *Mining Text Data*, pages 163–222. Springer.

Aggarwal, C. C. and Zhai, C. (2012b). A survey of text clustering algorithms. In *Mining Text Data*, pages 77–128. Springer.

Aletras, N. and Stevenson, M. (2013). Evaluating topic coherence using distributional semantics. In *Proc. 10th International Conference on Computational Semantics*, IWCS '13, pages 13–22. ACL.

Atzberger, D., Cech, T., de la Haye, M., Söchting, M., Scheibel, W., Limberger, D., and Döllner, J. (2021). Software Forest: A visualization of semantic similarities in source code using a tree metaphor. In *Proc. 16th International Conference on Information Visualization Theory and Applications – Volume 3*, IVAPP '21, pages 112–122. INSTICC, SciTePress.

Atzberger, D., Cech, T., Scheibel, W., Trapp, M., Richter, R., Döllner, J., and Schreck, T. (2023). Large-scale evaluation of topic models and dimensionality reduction methods for 2d text spatialization. *IEEE Transactions on Visualization and Computer Graphics*.

Behrisch, M., Blumenschein, M., Kim, N. W., Shao, L., El-Assady, M., Fuchs, J., Seebacher, D., Diehl, A., Brandes, U., Pfister, H., Schreck, T., Weiskopf, D., and Keim, D. A. (2018). Quality metrics for information visualization. *Wiley/EG Computer Graphics Forum*, 37(3):625–662.

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

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., and Blei, D. (2009). Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems*, 22.

Cox, M. A. A. and Cox, T. F. (2008). Multidimensional scaling. In *Handbook of Data Visualization*, pages 315–347. Springer.

Crain, S. P., Zhou, K., Yang, S.-H., and Zha, H. (2012). Dimensionality reduction and topic modeling: From latent semantic indexing to latent Dirichlet allocation and beyond. In *Mining Text Data*, pages 129–161. Springer.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., and Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407.

Espadoto, M., Martins, R. M., Kerren, A., Hirata, N. S. T., and Telea, A. C. (2021). Toward a quantitative survey of dimension reduction techniques. *IEEE Transactions on Visualization and Computer Graphics*, 27(3):2153–2173.

Gisbrecht, A. and Hammer, B. (2015). Data visualization by nonlinear dimensionality reduction. *Wiley Data Mining and Knowledge Discovery*, 5(2):51–73.

Joia, P., Coimbra, D., Cuminato, J. A., Paulovich, F. V., and Nonato, L. G. (2011). Local affine multidimensional projection. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2563–2571.

Kucher, K. and Kerren, A. (2019). Text visualization revisited: The state of the field in 2019. In *Proc. European Conference on Visualization*, EuroVis '19, pages 29–31. EG.

Lau, J. H., Newman, D., and Baldwin, T. (2014). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *Proc. 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 530–539. ACC.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. *Springer Nature*, 401(6755):788–791.

McInnes, L., Healy, J., and Melville, J. (2020). UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv CoRR*, stat.ML(1802.03426):1–63. pre-print.

Morariu, C., Bibal, A., Cutura, R., Frénay, B., and Sedlmair, M. (2023). Predicting user preferences of dimensionality reduction embedding quality. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):745–755.

Nenkova, A. and McKeown, K. (2012). A survey of text summarization techniques. In *Mining Text Data*, pages 43–76. Springer.

Noether, G. E. (1981). Why Kendall tau? *Wiley Teaching Statistics*, 3(2):41–43.

Paulovich, F. and Minghim, R. (2006). Text map explorer: a tool to create and explore document maps. In *Proc. 10th International Conference on Information Visualisation*, IV '06, pages 245–251. IEEE.

Riehmann, P., Kiesel, D., Kohlhaas, M., and Froehlich, B. (2019). Visualizing a thinker's life. *IEEE Transactions on Visualization and Computer Graphics*, 25(4):1803–1816.

Röder, M., Both, A., and Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proc. 8th International Conference on Web Search and Data Mining*, WSDM '15, pages 399–408. ACM.

Sedlmair, M. and Aupetit, M. (2015). Data-driven evaluation of visual quality measures. *Wiley/EG Computer Graphics Forum*, 34(3):201–210.

Sedlmair, M., Munzner, T., and Tory, M. (2013). Empirical guidance on scatterplot and dimension reduction technique choices. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2634–2643.

Sievert, C. and Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proc. Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 63–70. ACL.

Sips, M., Neubert, B., Lewis, J. P., and Hanrahan, P. (2009). Selecting good views of high-dimensional data using class consistency. *Wiley/EG Computer Graphics Forum*, 28(3):831–838.

van der Maaten, L. and Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11):2579–2605.

van der Maaten, L., Postma, E., and van den Herik, J. (2009). Dimensionality reduction: a comparative review. Technical Report 009–005, Tilburg University, Tilburg Centre for Creative Computing, The Netherlands.

Venna, J. and Kaski, S. (2006). Visualizing gene interaction graphs with local multidimensional scaling. In *Proc. European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, ESANN '06, pages 557–562. ESANN.

Vernier, E. F., Comba, J. L. D., and Telea, A. C. (2021). Guided stable dynamic projections. In *Wiley/EG Computer Graphics Forum*, volume 40(3), pages 87–98.

Vernier, E. F., Garcia, R., Silva, I. d., Comba, J. L. D., and Telea, A. C. (2020). Quantitative evaluation of time-dependent multidimensional projection techniques. In *Wiley/EG Computer Graphics Forum*, volume 39(3), pages 241–252.

Wallach, H. M., Mimno, D., and McCallum, A. K. (2009). Rethinking LDA: Why priors matter. In *Proc. 22nd International Conference on Neural Information Processing Systems*, NIPS '09, pages 1973–1981.

Wang, Y., Feng, K., Chu, X., Zhang, J., Fu, C.-W., Sedlmair, M., Yu, X., and Chen, B. (2018). A perception-driven approach to supervised dimensionality reduction for visualization. *IEEE Transactions on Visualization and Computer Graphics*, 24(5):1828–1840.

Xia, J., Huang, L., Lin, W., Zhao, X., Wu, J., Chen, Y., Zhao, Y., and Chen, W. (2023). Interactive visual cluster analysis by contrastive dimensionality reduction. *IEEE Transactions on Visualization and Computer Graphics*, 29(1):734–744.

Xia, J., Lin, W., Jiang, G., Wang, Y., Chen, W., and Schreck, T. (2021). Visual clustering factors in scatterplots. *IEEE Computer Graphics and Applications*, 41(5):79–89.

Xia, J., Zhang, Y., Song, J., Chen, Y., Wang, Y., and Liu, S. (2022). Revisiting dimensionality reduction techniques for visual cluster analysis: An empirical study. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):529–539.