A Large-Scale Sensitivity Analysis on Latent Embeddings and Dimensionality Reductions for Text Spatializations

Daniel Atzberger, Tim Cech, Willy Scheibel, Jürgen Döllner, Michael Behrisch, and Tobias Schreck

Abstract—The semantic similarity between documents of a text corpus can be visualized using map-like metaphors based on two-dimensional scatterplot layouts. These layouts result from a dimensionality reduction on the document-term matrix or a representation within a latent embedding, including topic models. Thereby, the resulting layout depends on the input data and hyperparameters of the dimensionality reduction and is therefore affected by changes in them. Furthermore, the resulting layout is affected by changes in the input data and hyperparameters of the dimensionality reduction. However, such changes to the layout require additional cognitive efforts from the user. In this work, we present a sensitivity study that analyzes the stability of these layouts concerning (1) changes in the text corpora, (2) changes in the hyperparameter, and (3) randomness in the initialization. Our approach has two stages: data measurement and data analysis. First, we derived layouts for the combination of three text corpora and six text embeddings and a grid-search-inspired hyperparameter selection of the dimensionality reductions. Afterward, we quantified the similarity of the layouts through ten metrics, concerning local and global structures and class separation. Second, we analyzed the resulting 42 817 tabular data points in a descriptive statistical analysis. From this, we derived guidelines for informed decisions on the layout algorithm and highlight specific hyperparameter settings. We provide our implementation as a Git repository at hpicgs/Topic-Models-and-Dimensionality-Reduction-Sensitivity-Study and results as Zenodo archive at DOI:10.5281/zenodo.12772898.

Index Terms—Text spatializations, text embeddings, topic modeling, dimensionality reductions, stability, benchmarking

1 Introduction

Text data is generated in large amounts from various sources, such as social media platforms, product reviews, news articles, literature, and research articles. Thereby, text data can be distinguished between single documents, i.e., sequences of words from a vocabulary also known as terms, streams, or sets of documents [42]. The latter are called text corpora and are usually considered *Document-Term Matrices* (DTMs), which store the absolute term frequencies in the respective documents. One central question in analyzing a text corpus is providing an overview and displaying the semantic relatedness between the documents [75]. Several visualization approaches rely on a two-dimensional scatterplot, where each point represents a document, and the pairwise Euclidean distance between documents reflects their semantic similarity. Adjacent work in the field augments the two-dimensional scatterplot by utilizing cartographic metaphors, e.g., height fields, icons, or glyphs [31]. Such scatterplots are derived from a layout algorithm that applies a dimensionality reduction (DR) to the DTM directly or an intermediate latent embedding of the text corpus, which is, for example, derived from a topic model (TM) [57]. This visual representation of a text corpus provides the foundation for a user's *mental map*, i.e., an internal cognitive representation [7]. If the visualization differs, the mental map is updated, which requires cognitive efforts from the user.

Changes to the underlying scatterplot are impeding the information exploration process since "geometric variations, e.g., rotation, translation, ..., in the projection make the analysis for the user difficult. Internally, the human brain needs to revert these transformations in order to ease the comparison of projections" [24]. Consequently, the stability of the visualization mainly depends on the stability of the layout algorithm. Such stability of a layout algorithm comprises various aspects:

1. *Stability concerning input data*, i.e., small changes to the input data result in small changes to the layout. This is further known as *Visual-Data Correspondence* [39].
2. *Stability concerning hyperparameters*, i.e., small changes to the hyperparameters of the layout algorithm do not change the layout.
3. *Stability concerning randomness*, i.e., the layout is not affected by randomness in the initialization.

Furthermore, in the case of time-dependent text corpora, two more aspects are:

4. *Stability concerning corpus size*, i.e., in case of an incrementally growing corpus the positions of previous points are not affected.
5. *Stability concerning temporal coherence*, i.e., the changes in the scatterplot capture the evolution of the data points. This is a further aspect of Visual-Data Correspondence.

In presentations of visualization approaches for text corpora, aspects related to the stability of the text layout are rarely considered. For example, most layouts are derived from *t-distributed Stochastic Neighbor Embedding* (t-SNE), even though it is said to be highly sensitive to its hyperparameters [77]. Only a few studies have considered the stability of DRs, e.g., qualitative studies that inspect scatterplots [11, 24] or theoretical discussions [53]. Existing quantitative studies focus on two non-text data sets and, therefore, stability of text layout algorithms remains an open gap in the literature [30, 38].

The stability of algorithms and models can be analyzed in a sensitivity analysis, i.e., a quantitative study that analyzes the relative importance of input factors to the output [59]. In this work, we present a sensitivity analysis of layout algorithms for text corpora concerning changes to the input data, hyperparameters, and randomness. Our approach has two steps: (1) the data generation step and (2) the data analysis step. In the first step, we derive tabular datasets by measuring aspects related to local, global, and perceptual similarity between selected pairs of scatterplots. In total, we consider 38 941 scatterplots that are generated from three text corpora, six text embeddings, and four DRs by a grid-search-inspired hyperparameter selection of the

- *Daniel Atzberger, and Willy Scheibel are with University of Potsdam, Digital Engineering Faculty, Hasso Plattner Institute. E-mail: {firstname.lastname}@hpi.uni-potsdam.de*
- *Tim Cech, and Jürgen Döllner are with University of Potsdam, Digital Engineering Faculty. E-mail: {tcech,rico.richter.1,doellner}@uni-potsdam.de*
- *Michael Behrisch is with Utrecht University. E-mail: m.behrisch@uu.nl*
- *Tobias Schreck is with Graz University of Technology. E-mail: tobias.schreck@cgv.tugraz.at*

Received 31 March 2024; revised 1 July 2024; accepted 15 July 2024. Date of publication 17 September 2024; date of current version 29 November 2024. This article has supplementary downloadable material available at https://doi.org/10.1109/TVCG.2024.3456308, provided by the authors. Digital Object Identifier no. 10.1109/TVCG.2024.3456308

DRs. The scatterplots were created on a computation cluster using 50 nodes and an overall computation time of 50 000 hours. In the second step, we examine the similarities in a descriptive analysis, including a correlation analysis, statistical tests, and by visualizing the data distributions. Our study makes the following contributions to the field of text visualization and empirical analysis of layout algorithms:

1. A set of metrics that quantify the preservation of local and global structures, as well as cluster separation between two scatterplots with the same number of points.
2. Tabular datasets that capture the similarity between 42 817 pairs of scatterplots, that are derived from three text corpora by applying six embeddings and four DRs.
3. An analysis of the results concerning the stability of the layout algorithms and guidelines for their effective use.
4. The implementation and results of the entire pipeline provided as a git repository[1].

2 Related Work

We consider the following three aspects as related work: (1) previous studies that focused on the stability of DRs, (2) comparisons of DRs that propose guidelines for their effective use, and (3) studies that focus on the visual perception of scatterplots and allow for comparison of scatterplots using similarity measures.

2.1 Stability of Dimensionality Reductions

We further distinguish discussions on the stability of DRs into mathematical discussions, qualitative studies, and quantitative analyses. Nonato and Aupetit presented a survey on DRs, including a discussion on stability concerning input data and their capabilities to map new data points based on their mathematical internals [53]. García-Fernández et al. compared six DRs concerning the stability in the input data and hyperparameters by visually comparing their results on six datasets [24]. Similarly, Bredius et al. studied the stability of neural network projections - trained to approximate a DR - concerning changes in the input data [11]. Reinbold et al. applied k-order Voronoi diagrams to visualize the neighborhood preservation of several two-dimensional scatterplot representations of a high-dimensional dataset to analyze the stability of MDS and t-SNE concerning input data and hyperparameters [58]. Complimentary to qualitative studies, Khoder et al. presented a quantitative study on the stability of DRs regarding input data for hyperspectral images [38]. However, their metrics are specific to their domain and can not be adapted in our case. The most similar method to our work was applied by Hamad et al., who studied the stability of t-SNE by comparing sequences of scatterplots derived from smart home data [30]. Their quantification of stability relies on the Procrustes distance and a metric that captures neighborhood preservation. However, we use ten metrics for a more fine-granular quantification of similarity, four DRs, and six latent embeddings to address the specific domain of text corpora visualization.

In several application contexts, data is often only incrementally available, e.g., social media or sensor data. The capability of a DR to handle such data streams is called *out-of-sample capability* [20]. In order to preserve the mental map, the positions of previous points should not be affected by incoming data. Our terminology refers to this aspect as stability concerning corpus size. Existing evaluations of out-of-sample techniques focus on the overall distance and neighborhood preservation [16, 79] or runtime performance [16, 23, 79]. Xia et al. furthermore evaluated the stability of the layout using four metrics [79]. An overview of specialized out-of-sample techniques is presented by Neves et al. [16]. An evolution of data points over time is another form of temporal dependency (stability concerning temporal coherence). To visualize such evolution, static methods can be adapted, e.g., by using control points known as landmarks [53]. Furthermore, specialized DRs have been developed. For example, Rauber et al. presented a variant of t-SNE, where the loss function is adopted to consider temporal coherence [56]. Vernier et al. presented a quantitative framework for evaluating the temporal coherence of DRs by measuring the preservation of local and global structures [69], and used it to evaluate the quality of two novel variants of t-SNE [68]. Our work differs from previous work, as we specifically analyze the impact of text embeddings on the static stability of two-dimensional layouts.

[1] hpicgs/Topic-Models-and-Dimensionality-Reduction-Sensitivity-Study

2.2 Selecting Dimensionality Reductions

Besides stability, the capability of a DR to preserve local and global structures in a lower-dimensional representation, the so-called accuracy, is another quality aspect [53]. Several studies analyzed DRs concerning their accuracy to derive guidelines for their effective use, e.g., Fodor who presented a qualitative comparison of linear and non-linear DRs [22]. Further studies that focus on the mathematical principles of DRs were presented by Engel et al., who reviewed DRs from a visualization point of view [19], Gisbrecht and Hammer, who reviewed non-linear DRs [26], and Cunningham and Ghahramani, who presented a survey on linear DRs [15]. Complimentary to these theoretical discussions, the accuracy of DRs was evaluated in quantitative studies by using quality metrics [8]. van der Maaten et al. presented an early quantitative discussion on DRs [66]. In their study, the authors compared the accuracy of eleven non-linear DRs and PCA on five synthetic and five real-world datasets. Gove et al. presented a study on the influence of selected t-SNE hyperparameters and further presented a neural network to recommend better default settings for a given dataset [28]. Espadoto et al. presented a benchmark that is set up of 18 data sets, 44 DRs, and seven accuracy metrics [20]. With this broad sampling, the study has particular value for practitioners. Atzberger and Cech et al. followed their approach, focusing on text corpora and TMs [3, 4]. Their studies are based on benchmarks comprising a set of text corpora, layout algorithms that are combinations of text embeddings and DRs, as well as metrics for quantifying the accuracy and cluster separation. In an analysis that comprises more than 40k layouts, the authors showed that TMs improve the accuracy of text layouts. Our work follows the methodology of such quantitative studies based on a benchmark comprising a set of text corpora, layout algorithms, and metrics. However, even though we leverage parts of the implementation provided [4], our study is concerned with stability which is a different objective than accuracy and thus requires the use of different metrics.

In addition to accuracy, the suitability of a DR for specific tasks needs to be considered too. Etemadpour et al. compared the performance of four DRs in a user study to support abstract analytics tasks, such as cluster identification [21]. Their results showed that the optimal DR depends on the question and the "nature" of the data, e.g., whether the data is a text corpus or a set of images. Xia et al. formulated similar findings after conducting a user study to investigate which DRs are suitable for visual cluster analysis tasks, e.g., cluster identification, membership identification, distance comparison, and density comparison [81]. For visual class separation, Wang et al. developed a novel supervised linear DR, which aims to minimize a cost function that relies on class separation metrics [72]. Furthermore, Morariu et al. presented a model to predict human preferences based on scagnostics, cluster separability metrics, and accuracy metrics [51].

2.3 Visual Perception and Similarity Measures

Our analysis of the stability of text layouts is based on a quantification of similarity between scatterplots. Various metrics have been proposed to describe geometric structures in scatterplots, e.g., the class separation [61], and thus allow for a comparison. Among the most popular metrics are the scagnostic measures, which result from graph-theoretical characteristics [78]. However, user studies have shown that scagnostics are not aligned with human expectations for describing perceptual similarity [54, 73]. In addition, more complex models have been developed, which learn abstract representations of scatterplots based on human-labeled data. For example, Quadri et al. proposed a ranking model for optimizing designs of scatterplots for cluster identification [55]. Jeon et al. developed a regression model to estimate cluster ambiguity [34]. Ma et al. proposed a convolutional neural network to learn a representation of scatterplots for the quantification of similarity [47]. Similarly, Xia et al. presented another convolutional neural

Table 1: Characteristics for the three datasets containing the number of documents N, the size of the vocabulary n after preprocessing, the median size of the documents l, the number of categories k, and the number of topics K specified for the TMs, as well as the sparsity ratio $\gamma = 1 - u/Nn$, where u denotes the number of non-zero entries in the DTM.

Dataset	Source	N	n	l	k	K	γ
20 Newsgroup	scikit-learn.org/0.19/datasets/twenty_newsgroups.html	18 846	72 370	35	20	20	0.9993
Lyrics	kaggle.com/datasets/karnikakapoor/lyrics	10 995	32 758	135	4	12	0.9974
Seven Categories	kaggle.com/datasets/deepak711/4-subject-data-text-classification	3 142	34 947	198	7	14	0.9962

network for modeling human perception of visual clusters [80]. Those neural network approaches learn internal representations of scatterplots and allow for a comparison. However, the individual components have no semantic meaning and thus allow for no further reasoning. Lehmann and Theisel introduced an equivalence relation on two-dimensional scatterplots based on affine transformations [46]. By selecting a represent of each equivalence class, the number of scatterplots within a scatterplot matrix is reduced. In particular, this approach does not rely on any further metrics but allows for comparison.

3 Data Measurement

To analyze the stability of layout algorithms for text corpora, we selected three raw text corpora and defined and applied a suitable data processing pipeline. It is set up of four stages: (1) preprocessing and perturbations, (2) mapping into a latent space using text embeddings, (3) dimensionality reduction to the two-dimensional plane, and (4) pairwise comparison using similarity metrics.

3.1 Preprocessing & Perturbation

We selected the following three corpora for our study: *20 Newsgroup*, *Lyrics*, and *Seven Categories*. All documents of the corpora are written in the English language. The characteristics, as well as sources, of the text corpora are summarized in Table 1. We preprocessed each corpus as follows: In the first step, the documents of a corpus are tokenized, i.e., documents are split at whitespaces and stored as lists of terms. To remove terms that have no semantic meaning and reduce the vocabulary size, we removed all stopwords from the English language and lemmatized the vocabulary. In addition, we carried out corpus-specific preprocessing steps, such as the removal of email headers in the 20 Newsgroup corpus. For all details regarding the preprocessing, we refer to our git repository. After preprocessing, each corpus is represented by a DTM, i.e., each corpus is described by an Nxn-matrix, where N denotes the number of documents and n the vocabulary size. The entry in cell (i,j) gives the frequency of the j^{th} term in the i^{th} document. Furthermore, each document within a corpus is assigned one unique category that reflects its semantics, e.g., each document within the Seven Categories corpus is associated to either *Computer Science*, *History*, *Maths*, *Accounts*, *Physics*, *Biology*, or *Geography*. We selected these corpora as their categories are easily interpretable, and the extracted topics show a connection to these categories. Furthermore, the three corpora vary in the number of documents, vocabulary size, and categories.

To analyze the stability concerning changes in the input data, we apply synthetic perturbations to the DTMs. We define a jittering perturbation, which adds noise to the DTMs. Specifically, the entry in cell (i,j) is replaced by:

$$\mathrm{DTM}[i,j] = \max\{0, \mathrm{round}(\mathrm{DTM}[i,j] \cdot (1 + \varepsilon[i,j]))\}, \qquad (1)$$

where the matrix ε is a $n \times m$-matrix whose entries are randomly drawn from a uniform distribution on the interval $[-\lambda, \lambda]$ and $\lambda \in [0,1]$ controls the amount of jittering. We chose a relatively simple jittering function to avoid making any further assumptions.

3.2 Document Embeddings

Starting from the DTM representation, we further consider five additional document embeddings that are commonly used for corpus visualization and text analytics tasks, resulting in six latent representations for further analysis. From the perspective of the DTM representation, each document is represented as an n-dimensional vector containing the absolute frequencies of individual terms. This representation, along with the *cosine similarity*, constitutes the *Vector Space Model* (VSM) [14]. However, the DTM solely accounts for the absolute frequencies of terms within documents, disregarding whether these terms are prevalent across other documents in the corpus. Often, terms found in only a few documents indicate underlying concepts and hold significant relevance. By incorporating the *term frequency-inverse document frequency* (tf-idf) scheme, the VSM can be adapted to address this issue [1]. Specifically, the tf-idf of a term w in document d is given by the product of the term-frequency of the term w in d and the inverse document frequency of w in d, i.e.,

$$\text{tf-idf}(w,d) = \frac{n(w,d)}{\sum_{d' \in C} n(w,d')} \cdot \log\left(\frac{|C|}{|\{d' \in C | w \in d'\}|}\right), \qquad (2)$$

where $n(w,d)$ denotes the frequency of term w in document d, and C denotes the corpus. The DTM is usually a sparse matrix, i.e., most entries are zero due to documents containing only a fraction of the entire vocabulary. This observation is exemplified in Table 1, where the median lengths are significantly smaller than the vocabulary sizes.

TMs aim to provide a compressed representation of the DTM by grouping co-occurring words into topics. Topics are represented as vectors of size n, where the i^{th} entry reflects the influence of term w_i within the topic. Such representations often allow for the inference of human-interpretable concepts. For instance, *Latent Semantic Indexing* (LSI) employs *Singular Value Decomposition* (SVD) to decompose the DTM or its tf-idf weighted variant into a document-topic matrix and a topic-term matrix [17]. *Non-Negative Matrix Factorization* (NMF) approximates the DTM or its tf-idf weighted variant as a product of two matrices [44]. In the case of LSI and NMF, the documents are compared using the cosine similarity. *Latent Dirichlet Allocation* (LDA) is a probabilistic TM widely used in the visualization domain. LDA assumes a generative process underlying a corpus, resulting in topics represented as multinomial distributions over the vocabulary and documents represented as multinomial distributions over topics [9]. In measuring document similarity within LDA, the *Jensen-Shannon distance* is usually applied.

As a result of advances in GPU processing, deep learning models have been developed for learning high-dimensional continuous embeddings of corpora. For example, *Word2Vec* learns a representation for terms within a corpus by training a neural network that predicts the center word given its surroundings (*Continuos BOW Model*) or vice versa by predicting the surrounding terms given the central word (*Continuous Skip-gram Model*) [50]. *Doc2Vec* extends the concept of Word2Vec for entire documents by training a neural network to predict the next word given a document together with words (*Distributed Memory Model*) or a set of words given a document as input (*Distributed BOW Model*) [43]. The embeddings derived from Doc2Vec are learned in an iterative manner using back-propagation and allow for comparison using the cosine similarity. We further consider *Bidirectional Encoder Representations from Transformers* (BERT) - a deep learning model that is trained for two unsupervised NLP tasks and, as a result, generates high-dimensional representations for documents within a corpus [18]. By grouping similar documents according to their latent representations, BERT is known for generating well-interpretable topics using a class-based tf-idf weighting [29].

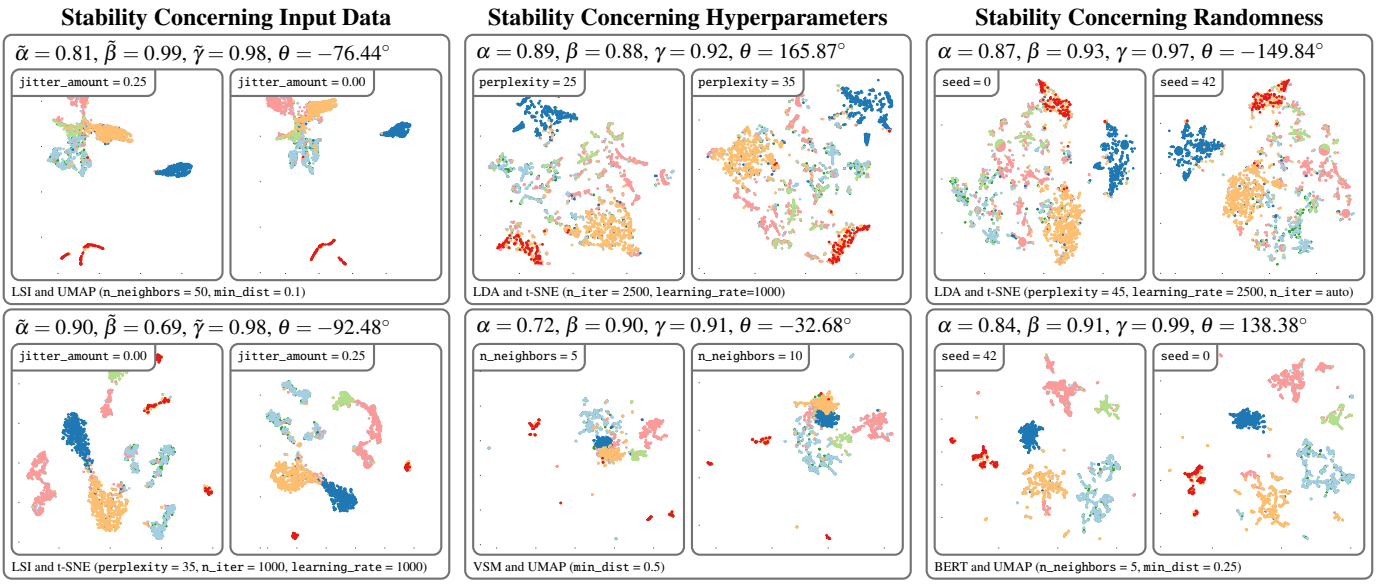


Fig. 1: Exemplary comparison of pairs of scatterplots. To analyze the stability concerning input data, we compare pairs of scatterplots that only differ in the amount of jitter applied to the DTM. To analyze the stability concerning hyperparameters, we compare pairs of scatterplots that differ in one hyperparameter setting with consecutive values. To analyze stability concerning randomness, we compare two layouts that only differ in their seeds.

3.3 Dimensionality Reductions

As a result of the second stage, each document is represented in a latent space. To further project the documents to the two-dimensional plane, we apply a DR. Thereby, we focus on DRs that are commonly used for text spatializations [4]: *Metric Multidimensional Scaling* (MDS), *Self-Organizing Maps* (SOMs), *t-distributed Stochastic Neighbor Embedding* (t-SNE), and *Uniform Manifold Approximation* (UMAP). Furthermore, t-SNE and UMAP are probably the most popular DRs among practitioners and also show the best results in terms of accuracy in previous studies [4, 20]. Although many more DRs exist, we limit our considerations to these four for capacity reasons.

MDS operates on a dissimilarity matrix of a dataset, aiming to generate a lower-dimensional representation where pairwise Euclidean distances reflect the entries in the dissimilarity matrix [13]. The positions of data points are iteratively computed by optimizing a stress function. The number of iterations constitutes a hyperparameter .

SOMs constitute a class of fully-connected two-layered neural networks where second-layer neurons are organized on a two-dimensional grid, whose width and height are determined by hyperparameters [40]. During training, input vectors activate neurons whose weight vectors are most similar to the input. The neuron of the highest activation determines the position within the grid. Weight adjustments during training minimize quantization errors, i.e., differences between input vectors and their best matching unit. For computational efficiency, we utilized *Principal Component Analysis* (PCA) to derive a lower-dimensional representation, that still captures 95% of the dataset's variance [36].

t-SNE is a DR aimed at preserving local structures within a dataset [65]. It operates by modeling a Gaussian distribution centered around each data point in the high-dimensional space, where the perplexity hyperparameter regulates the effective number of neighbors considered. The objective of t-SNE is to maintain neighborhood relationships in the low-dimensional representation using a t-distribution. Its iterative optimization process minimizes a stress function, which evaluates the dissimilarity between overall similarity scores derived from the respective distributions and the *Kullback-Leibler Divergence*.

UMAP was developed as an alternative to t-SNE to address its limitations, such as the difficulty in interpreting distances between clusters [48]. While conceptually similar to t-SNE, UMAP optimizes a stress function based on *Cross-Entropy* instead of Kullback-Leibler divergence. UMAP offers two hyperparameters: the number of neighbors, balancing local and global structures, and the minimal distance, regulating the proximity of data points in the two-dimensional layout.

Table 2: Metrics used in our study. We did not specify an optimum for rotation, e.g., 0, as the rotation can be carried out as a postprocessing step when comparing two layouts. All metrics are invariant under rotation.

Metric	Abbr.	Range	Optimum
Trustworthiness	α_T	[0,1]	1
Continuity	α_C	[0,1]	1
Mean Relative Rank Errors	α_{MM}, α_{MF}	[0,1]	1
Local Continuity Meta-Criterion	α_{LC}	[0,1]	1
Label Preservation	α_{LP}	[0,1]	1
Pearson's Correlation	β_{PC}	[-1,1]	1
Spearman's Rank Correlation	β_{SC}	[-1,1]	1
Cluster Ordering	β_{CO}	[-1,1]	1
Abs. Diff Distance Consistency	γ_{DC}	[0,1]	0
Rotation from Procrustes Analysis	θ_{PA}	[-180°,180°]	$\perp$

3.4 Comparison

Our analysis of stability requires a notion of similarity between scatterplots. From the study of the related work, we derived three different types of metrics for quantifying scatterplot similarity: (1) latent representations learned from neural networks, (2) perceptual similarity features, and (3) features that capture selected aspects, e.g., neighborhood preservation. We do not consider latent representations due to their lack of interpretability. Additionally, we omit scagnostics, as they do not align well with human judgment [54, 73]. Instead, we opted for metrics that quantify selected aspects of similarity, grouping them into local and global similarity, as well as cluster separation. To this end, we have adapted existing accuracy metrics, i.e., metrics that quantify the preservation of local and global structures of high-dimensional data in a low-dimensional representation [5]. We provide an overview in Table 2.

For quantifying local similarity, i.e., the preservation of neighborhoods, we consider six metrics. The *Trustworthiness* measure α_T is often used to quantify the accuracy of DRs [67]. It measures the pointwise-percentage of the k-Nearest-Neighbors (kNN) in the low-dimensional representation that also belong to the kNN in the high-dimensional representation, weighted by ranks and averaged over all points. Similarly, the *Continuity* α_C measures the proportion of points in the high-dimensional representation belonging to the kNN in the low-dimensional representation [67]. Both trustworthiness and continuity depend only on the pairwise dissimilarities of the points but not

on the positions of the points themselves. Therefore, both measures can also be used to compare two two-dimensional representations concerning the preservation of neighborhoods. The same consideration holds for the *Mean Relative Rank Errors* α_{MM} and α_{MF}, which are related to trustworthiness and continuity but with slightly different weightings [45], and the *Local Continuity Meta-Criterion* α_{LC}, which measures the pointwise intersection between the kNN of a point in the two scatterplots averaged over all points [12]. We further integrate the *Label Preservation* metric α_{LP}, which is similar to the α_{LC} but only considers the categories and not the positions. In any case, the local metrics require the specification of k, i.e., the number of neighbors considered. We choose $k = 7$ to be aligned with previous studies [3, 4, 20, 68, 69]. The preservation of neighborhoods between two scatterplots is highly relevant for the visualization, since close data points are assumed to be similar according to the *Gestalt principles* [76].

To quantify global similarity, we use three metrics. The *Spearman Rank Correlation* β_{SC} [62] and the *Pearson Correlation* β_{PC} [25] are derived from the *Shephard Diagram* from the two scatterplots. The Shephard diagram is a two-dimensional scatterplot whose points represent pairwise distances between two points in the first and second scatterplot [35]. In the case of a perfect match, the Shephard diagram would thus be a subset of a straight line through the origin. The similarity to the straight line is quantified by the two metrics. We further developed the *Cluster Ordering* metric β_{CO}, which relates the arrangement of categories between two scatterplots. The metric is given by the Pearson correlation of the pairwise distances between the categories centers; i.e., it relies on two graphs derived from the scatterplots similar to measures from the *Graph-based Family* [52]. All three metrics for the global similarity have a bounded value range. We do not consider metrics with unbounded range, e.g., *Normalized Stress* [41] or *Procrustes Distance* [27], as it is not evident how to normalize these measures across several text corpora and scales.

For our third set of metrics, we assess the efficacy of discerning given categories. Building upon the findings of Sedlmair and Aupetit, we use the *Distance Consistency* [60]. This metric evaluates the proportion of points within the projected two-dimensional space, where the associated category center, defined as the mean of all points in that category, coincides with its nearest category center with respect to the Euclidean distance [63]. To quantify the similarity between two scatterplots concerning class separation, we use the absolute difference γ_{DC} between their distance consistencies.

These ten similarity metrics are used in our sensitivity analysis as follows: To analyze the stability concerning input data, we compare scatterplots that differ only in the jitter amount, but with the other configurations fixed (*ceteris paribus*). To analyze the stability concerning hyperparameters, we pair scatterplots that differ in one hyperparameter setting with consecutive values. To analyze stability concerning randomness, the scatterplots differ only in the defined random seed. The selection process is illustrated in Figure 1. Our data processing pipeline results in three tabular datasets, where each entry is given by a pair of scatterplots and similarity measures from the ten metrics. As a post-processing step, one scatterplot can be rotated according to the rotation derived from Procrustes analysis. The angle is determined to minimize the Procrustes distance, i.e., the squared pairwise distances, between two scatterplots [37]. Before applying Procrustes analysis, we center both scatterplots such that the angle refers to rotation around the view center. We include the rotation to the tabular datasets.

3.5 Implementation

Our implementation is designed for various embeddings, DRs, and metrics and can be extended in the future. For each embedding and DR, we chose from specific implementations and hyperparameters. The large number of layout configurations required the use of a computing cluster. The project is freely available on GitHub for reproducibility and reuse[1]; the generated data is linked as Zenodo archive[2].

Table 3: Range for the hyperparameters considered in our experiments. Each configuration for one DR is combined with a dataset and TM.

DR	Parameter Name	Values
MDS	`max_iter`	100–300 step size 50
SOM	`m`	5–30 step size 5
SOM	`n`	5–30 step size 5
UMAP	`min_dist`	0.0, 0.1, 0.25, 0.5, 0.8, 0.99
UMAP	`n_neighbors`	2, 5, 10, 20, 50, 100, 200
t-SNE	`learning_rate`	10, 28, 129, 359, 1000, auto
t-SNE	`n_iter`	1000, 2500, 5000, 10000
t-SNE	`perplexity`	5–55 step size 10

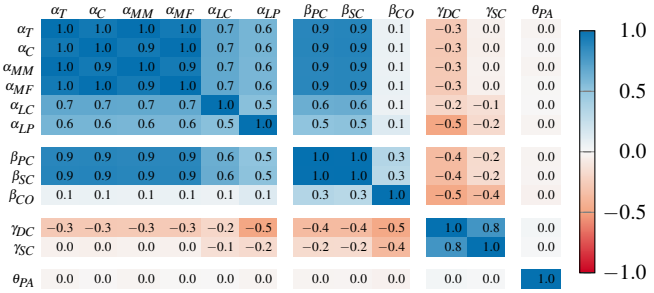


Fig. 2: Heatmap showing the pairwise correlations between the similarity metrics using a diverging color scheme. We additionally show the correlation with the *Silhouette Coefficient*, which is another cluster separation metric. Metrics that correlate nearly perfect, i.e., $\alpha_T, \alpha_C, \alpha_{MM}, \alpha_{MF}$ as well as β_{PC}, β_{SC} are considered as one metric by taking their averages. Note: the local and global similarity measures show a negative correlation to the class separation measures, as they have opposite optimums.

3.5.1 Software Dependencies

The implementation is based on Python 3.10 and depends on actively maintained libraries that are popular among practitioners for the embeddings and DRs. Our text preprocessing pipeline relies on *NLTK* (3.7) and *spaCy* (3.4.3) for lemmatization. For topic modeling and Doc2Vec, we use *Gensim* (4.2.0). The pretrained BERT models are provided by the *Sentence Transformer* library (2.2.2). For t-SNE and MDS, we use the implementation provided by *scikit-learn* (1.2.1); for UMAP, we use *umap-learn* (0.5.3); and for SOMs, we utilize the *sparse-som* library (0.6.1) [49]. For the similarity metrics, we adopted the approaches and implementations by Atzberger and Cech et al. [4] and *ZADU* [33].

3.5.2 Hyperparameter Settings

We selected values for the hyperparameters of the DRs following the documentation of the respective library and the original papers. The value ranges for the hyperparameters for the DRs are specified in Table 3. For each corpus-embedding combination, we used a fixed configuration for the embedding, i.e., we did not iterate over the embedding's hyperparameters. When applying TMs, we set the number of topics K to the number of categories k in the case of the 20 Newsgroup corpus, $K = 2k$ for the Seven Categories corpus, and $K = 3k$ for the Lyrics corpus, since the latter two have relatively few categories. We followed best practices in choosing the TM's hyperparameters and inspected the topics of each trained TM to ensure interpretable topics [70, 71]. The extracted topics are provided in the supplemental material. For BERT, we chose the two pre-trained models, `all-mpnet-base-v2` and `all-distilroberta-v1`, as they have shown the highest scores for the sentence embedding task[3]. Doc2Vec requires the specification of the embedding dimension and the number of iterations. Following best practices, the trained models have an accuracy of above 95% to predict the nearest neighbors of the inferred documents to be themselves among the three nearest neighbors[4].

[2]Zenodo archive DOI:10.5281/zenodo.12772898

[3]sbert.net/docs/pretrained_models.html

[4]radimrehurek.com/gensim/auto_examples/tutorials/run_doc2vec

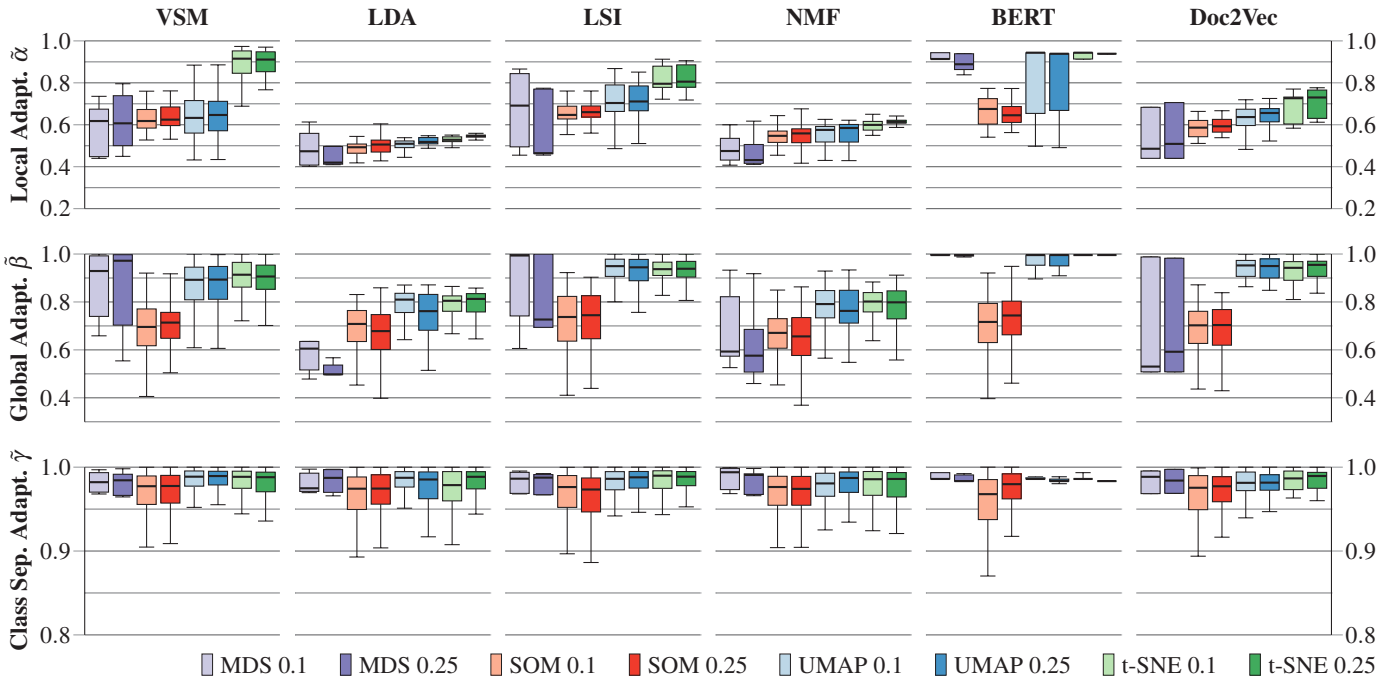


Fig. 3: Results of the first experiment to quantify the stability concerning changes to the input data. The hue of each bar indicates the DR, and the intensity indicates the amount of jitter applied to the DTM. The metrics $\tilde{\alpha}$, $\tilde{\beta}$, and $\tilde{\gamma}$ quantify how well the layout algorithm adapts to changes to the DTM, with 1 being optimal. The visualization indicates that BERT, in combination with t-SNE, best reflects changes to the DTM concerning $\tilde{\alpha}$ and $\tilde{\beta}$, resulting in improvements compared to the VSM. Note: The vertical axis ranges differ between the three metrics $\tilde{\alpha}$, $\tilde{\beta}$, and $\tilde{\gamma}$.

3.5.3 Computational Cluster

The setup of the computational cluster is similar to the previous study of Atzberger and Cech et al. [4]. Specific to this study, we used jobs with a RAM limit of 40 GB and used up to 29 nodes to optimize for a higher throughput of jobs. In total, our experiments had a run time of over 49 000 CPU hours. From the targeted 40 860 layouts, the cluster could compute 38 941 layouts (95.3 %). The unsuccessful computations can be accounted to timeouts after 30 hours (120 layouts) and general abortions due to exceeding RAM, planned downtime, unplanned downtime, file system errors, etc. (1799 layouts).

4 Data Analysis & Results

In total, our three datasets that resulted from comparing pairs of scatterplots resulted in 42 817 data points. Due to interruptions in the computational cluster, not all layouts were created successfully. On average, each dataset contains 14 272 pairs of scatterplots. We first analyze the correlation between the metrics to derive an aggregated metric for local similarity, global similarity, and similarity concerning cluster separation. The distributions of data points are inspected using a series of boxplots, which allows us to analyze the three stability aspects covered in our study. In two binary tests, we analyze the effect of the tf-idf weighting scheme and the application of the DR on the topic representations.

4.1 Correlation of Similarity Metrics

To derive a higher-level overview of the similarity aspects, we aggregate several metrics to represent the three similarity aspects, e.g., by taking an average [20, 51, 68, 69]. However, relying on the average carries the risk of overweighting within one aspect, e.g., in case most metrics are strongly correlated. Alternatively, taking a weighted average based on pairwise correlations counteracts this, but conversely, it might outweigh a single metric that is not correlated to the others at all [3]. Figure 2 shows the pairwise correlations of the ten similarity metrics, as well as the *Silhouette Coefficient* as a further class separation metric and the rotation that is derived from Procrustes analysis. To derive the correlations, we randomly selected 3000 pairs of scatterplots, equally distributed over the corpora.

Regarding local similarity, there is a strong positive correlation between all metrics; the first four metrics correlate nearly perfectly. Therefore, we averaged the first four metrics. Thus, we define the aggregated local similarity metric α as

$$\alpha = \frac{1}{3}\left(\frac{\alpha_T + \alpha_C + \alpha_{MM} + \alpha_{MF}}{4} + \alpha_{LC} + \alpha_{LP}\right). \tag{3}$$

Regarding global similarity, the Spearman and Pearson correlation measures correlate perfectly. Therefore, we consider their average and pair it with the cluster ordering metric. Furthermore, we apply an affine transformation to all three metrics, to translate their value ranges to [0,1]. Thus, we define the global similarity measure β as

$$\beta = \frac{1}{2}\left(\frac{0.5\cdot(\beta_{PC}+1)+0.5\cdot(\beta_{SC}+1)}{2} + \frac{1}{2}(\beta_{CO}+1)\right). \tag{4}$$

To quantify changes to the class separability, we use the absolute difference between the distance consistencies between two scatterplots. As γ_{DC} has its optimum at 0, we define the metric γ as:

$$\gamma = 1 - \gamma_{DC} \tag{5}$$

4.2 Stability Concerning Input Data

Starting from the three similarity metrics α, β, and γ, we first analyze the stability of text layout algorithms concerning changes to the input data. For this, we applied jittering on the DTM and compared scatterplots that result from the same layout algorithm with the same hyperparameter configurations. A stable layout algorithm should reflect the changes to the DTM in the layout. Given a layout algorithm Φ, we quantify the adaptability of the layout algorithm to changes in the local structures of the DTM using the following metric:

$$\tilde{\alpha} = 1 - |\alpha(\mathrm{DTM}, \mathrm{DTM}^{\mathrm{jitter}}) - \alpha(\Phi(\mathrm{DTM}), \Phi(\mathrm{DTM}^{\mathrm{jitter}}))|, \tag{6}$$

where $\alpha(\mathrm{DTM}, \mathrm{DTM}^{\mathrm{jitter}})$ denotes the local similarity between the DTM and its jittered variant, and $\alpha(\Phi(\mathrm{DTM}), \Phi(\mathrm{DTM}^{\mathrm{jitter}}))$ denotes the local similarity between the two scatterplots $\Phi(\mathrm{DTM})$ and

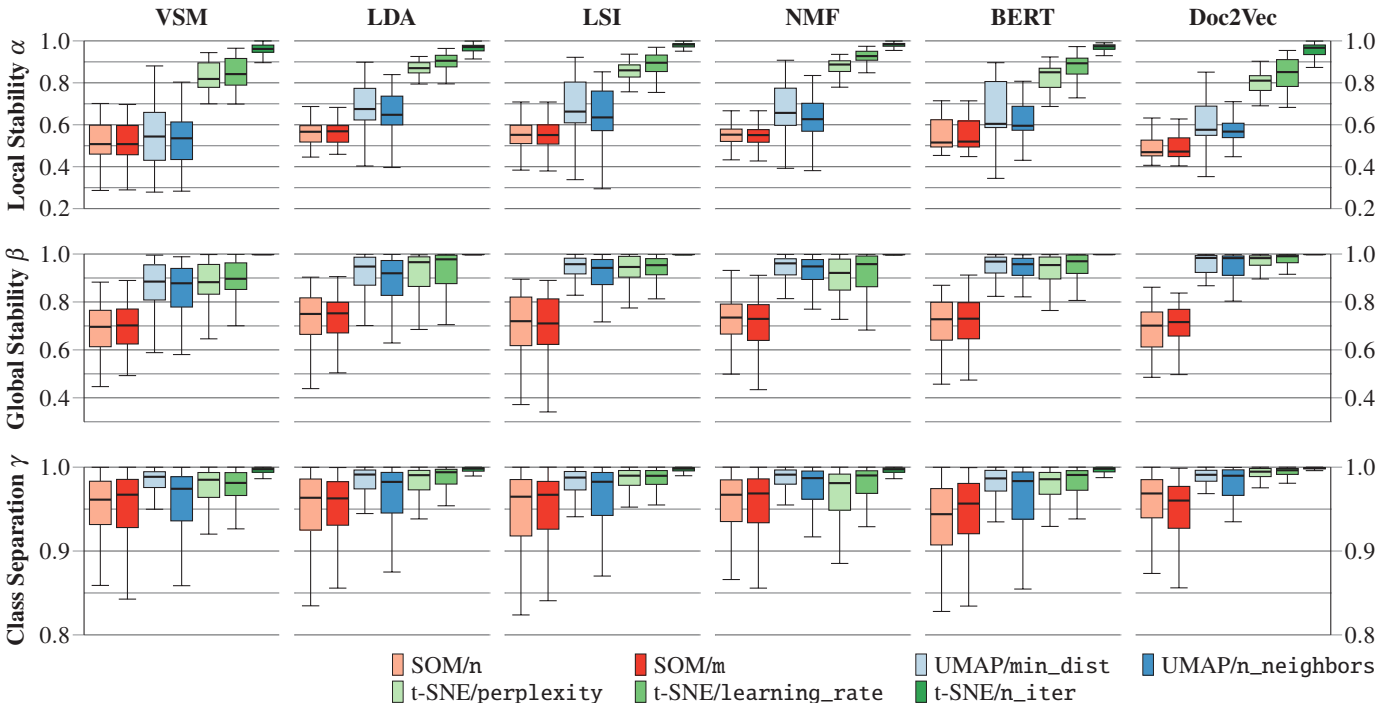


Fig. 4: Results of the second experiment to quantify the stability concerning hyperparameters. The hue of each bar indicates the DR and the intensity indicates a specific hyperparameter that is varied. LDA, LSI, and NMF, in combination with t-SNE, show the highest stability concerning changes to the hyperparameters. Note: The vertical axis ranges differ between the three metrics α, β, and γ.

$\Phi(\text{DTM}^{\text{jitter}})$. Analogously, we define $\tilde{\beta}$ to quantify how well the layout algorithm adapts to changes in the DTM concerning global structures and $\tilde{\gamma}$ how well changes concerning cluster separation are captured. In the case of a value of 1, the scatterplots reflect changes to α, β, and γ perfectly. The results are shown in Figure 3.

t-SNE best reflects changes to local structures of the DTM, as it shows the highest values for $\tilde{\alpha}$. Furthermore, in most cases, t-SNE and UMAP adapt best to changes in the DTM's global structures measured by $\tilde{\beta}$. All DRs show appropriate changes regarding cluster separation measured by $\tilde{\gamma}$, but SOMs show the largest range across all embeddings.

Changes to the DTM are reflected poorly in the case of layouts based on LDA. We assume that LDA extracts similar topics for both the DTM and its jittered variant and, therefore, represents both very similarly in the latent space. Thus, the dissimilarity of the DTM and its jittered variant is not reflected in the latent space, which means that the resulting scatterplots do not depict the desired change either after applying a DR. NMF shows the same effect as LDA. Doc2Vec also leads to decreases in $\tilde{\alpha}$ but reflects global changes well in the case of t-SNE and UMAP. BERT and LSI reflect changes in global structures well and improve the results shown by the VSM. However, only BERT shows improvements concerning $\tilde{\alpha}$ compared to the VSM for all four DRs. In summary, BERT in combination with t-SNE best reflects changes to the input data.

4.3 Stability Concerning Hyperparameters

In the second experiment, we analyze the stability of the layout algorithms concerning small changes to the hyperparameters. Figure 4 shows the pairwise similarities between pairs of scatterplots that differ in consecutive values in exactly one hyperparameter. We omit MDS since, in 100 percent of all cases, MDS has converged after 200 iterations, i.e., the pairwise similarities are one. Among the remaining three DRs, SOMs are the most sensitive to changes in their hyperparameters (referring to their median) concerning α, β, and γ. The hyperparameters height and width have nearly the same impact due to the symmetry of the grid structure of the SOM layout. UMAP shows a significant sensitivity concerning local similarity α but is much more stable regarding global similarity β. Changes to the minimum distance do not affect the cluster separation γ strongly. t-SNE shows the highest scores for all three metrics. From the small interquartile distance of the boxes displaying the number of iterations, we deduce that the algorithm converges quite early. The perplexity parameter has the most significant impact on t-SNE layouts' stability. But still, it is more stable than SOMs and UMAP. Our results contradict the widespread statement that t-SNE creates unstable layouts. However, the metrics used ignore perceptual differences due to rotation, as they are invariant under rotation.

Each embedding improves the stability concerning changes in the hyperparameters. The improvements through LDA, LSI, and NMF are comparable in each case, i.e., we see no clear benefit in choosing one over the other. However, the TMs outperform BERT and Doc2Vec in terms of α. The dimensions of the latent spaces are significantly lower for the three TMs than in the case of Doc2Vec (50) and the two BERT models (768 for both embeddings). We assume that higher dimensionality can result in stronger distortions due to the DRs. To summarize, t-SNE in combination with a topic model shows the most stable behavior concerning changes to its input parameters.

4.4 Stability Concerning Randomness

By specifying a random seed of the DR, the layouts are reproducible across multiple runs. Since different initializations might result in different local optima, two scatterplots derived from the same layout algorithm can thus differ. Figure 5 shows the values for the similarity metrics between scatterplots that only differ in the random seed.

The results show that differences in the random seed can be very considerable, e.g., in the case of MDS concerning α. t-SNE shows the most stable behavior concerning α by far, i.e., the neighborhoods are represented similarly regardless of the selected seed. UMAP, followed by t-SNE, achieves the highest stability concerning global structures. MDS, UMAP, and t-SNE show stable behavior concerning γ, whereas SOMs can lead to larger changes.

For UMAP and t-SNE, embeddings can further increase the stability concerning α. Furthermore, embeddings improve the global stability in the case of UMAP. The stability concerning class separation shows high values across all embeddings. Overall, we favor t-SNE due to the strong dominance concerning α. In combination with LDA, the local stability can further be improved.

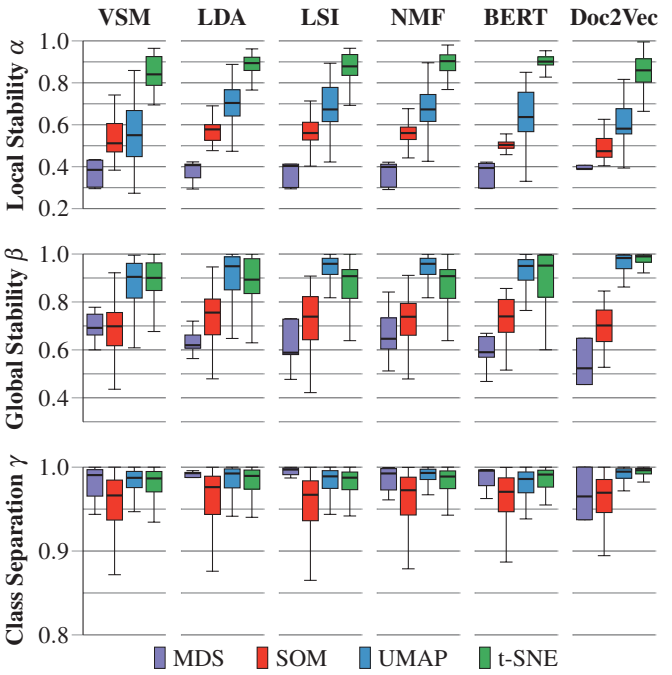


Fig. 5: Results of the third experiment to quantify the stability concerning randomness. The color of each bar indicates the DR underlying the layout algorithm. Overall, LDA in combination with t-SNE shows the best results. Note: The vertical axis ranges differ between the three metrics α, β, and γ.

Table 4: Results of the binary test for the null hypothesis "The tf-idf weighting scheme does not improve the stability concerning input data (S1), hyperparameters (S2), or randomness (S3)."

		S1			S2			S3		
TM	DR	$\tilde{\alpha}$	$\tilde{\beta}$	$\tilde{\gamma}$	α	β	γ	α	β	γ
VSM	MDS	0.50	0.03	0.94	0.95	0.95	0.95	0.94	0.30	0.94
	SOM	0.45	0.99	0.66	0.24	0.99	0.47	0.42	1.00	0.84
	t-SNE	0.00	0.00	0.01	0.00	0.99	0.03	0.00	1.00	0.23
	UMAP	0.00	0.00	0.05	0.00	0.16	0.24	0.00	0.09	0.39
LSI	MDS	0.65	0.65	0.01	1.00	1.00	1.00	0.05	0.98	0.99
	SOM	0.00	1.00	0.99	0.00	0.99	0.99	0.00	0.99	0.99
	t-SNE	0.00	1.00	0.05	0.00	0.99	0.99	1.00	0.02	1.00
	UMAP	0.00	0.99	0.99	0.00	0.99	0.99	0.00	1.00	1.00
NMF	MDS	0.05	0.01	0.21	1.00	1.00	1.00	0.05	0.00	0.57
	SOM	0.07	0.05	0.92	0.00	0.99	0.99	0.00	1.00	0.77
	t-SNE	1.00	0.99	0.00	0.00	1.00	0.99	1.0	1.0	1.0
	UMAP	0.00	0.00	0.99	0.00	0.99	0.99	0.00	0.97	1.00

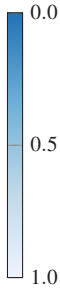
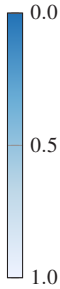


Table 5: Results of the binary test for the null hypothesis "Equation (7) does not improve the stability concerning input data (S1), hyperparameters (S2), or randomness (S3)."

		S1			S2			S3		
TM	DR	$\tilde{\alpha}$	$\tilde{\beta}$	$\tilde{\gamma}$	α	β	γ	α	β	γ
LDA	MDS	0.92	0.42	0.15	0.69	0.69	0.69	1.00	0.07	0.00
	SOM	0.00	0.00	0.12	0.13	0.00	0.09	0.16	0.00	0.01
	t-SNE	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	UMAP	0.00	0.00	0.03	0.00	0.00	0.30	0.03	0.00	0.10
LSI	MDS	1.00	0.50	1.00	1.00	1.00	1.00	1.00	0.70	1.00
	SOM	0.75	0.11	0.87	0.05	0.41	0.76	0.01	0.04	0.78
	t-SNE	0.09	0.86	0.09	0.26	0.48	0.70	0.89	0.77	0.06
	UMAP	0.85	0.55	0.45	0.55	0.83	0.50	0.01	0.02	0.27
NMF	MDS	0.29	0.01	0.29	1.00	1.00	1.00	0.15	0.00	0.15
	SOM	0.00	0.07	0.99	0.00	0.93	0.99	0.00	0.99	0.81
	t-SNE	0.99	0.01	0.63	0.00	1.00	0.99	0.48	1.00	1.00
	UMAP	0.00	0.00	0.99	0.00	0.99	0.99	0.00	0.94	0.94



4.5 Binary Tests

In our previous experiments, we aggregated different layout algorithms that share the same embedding and DR. However, the VSM, LSI, and NMF can be applied to either the DTM or its tf-idf weighted variant. Atzberger and Cech et al. have empirically shown that the tf-idf weighting improves layout accuracy and perception [4]. We analyze whether this observation also transfers to stability using a binary test. For this, we compare the aggregated stability metrics $\tilde{\alpha}$, $\tilde{\beta}$, and $\tilde{\gamma}$ in the case of stability concerning input data, and α, β, and γ in the case of the other two stability aspects of two pairs of scatterplots $(\Phi_1(\text{DTM}), \Phi_2(\text{DTM}))$ and $(\Phi_1(\text{DTM}^{\text{tf-idf}}), \Phi_2(\text{DTM}^{\text{tf-idf}}))$, where Φ_1 and Φ_2 are layout algorithms with specified hyperparameters that were compared in the respective experiment. For all possible combinations of such pairs of tuples, we determine the occurrences n_α, n_β, and n_γ in which the tf-idf weighted variant shows larger values. Assuming that the tf-idf weighting has no impact on the stability, the values n_α, n_β, and n_γ are distributed according to a binomial distribution with probability 0.5. Given this distribution, we determine the probability of our observations for n_α, n_β, and n_γ. In the case, of a small probability, we reject the hypothesis, since a large value would be unlikely in the case of probability 0.5. Our results are summarized in Table 4. A hypothesis is usually rejected if the probability is smaller than 0.05. In this case, the improvement made using the tf-idf scheme is statistically validated.

In the case of the VSM, we see that the tf-idf weighting in combination with t-SNE and UMAP improves all three stability aspects in terms of preserving local structures. Furthermore, for both t-SNE and UMAP, it supports the layout algorithm to adapt to changes in the input data. Also, for LSI, the tf-idf weighting scheme improves stability regarding changes in the input data and hyperparameters in combination with t-SNE and UMAP concerning local structures. The combination of NMF and UMAP profits from the tf-idf weighting in all three stability aspects concerning the preservation of local structures. We want to point out that a high probability does not indicate the validity of the null hypothesis. For example, the high probability for LSI in combination with t-SNE concerning stability concerning randomness might be due to the small value range of the similarity measures, as shown in Figure 5.

Our second binary test concerns the input for the DR. Most visualization approaches apply the DR to the document representations of the corpus in the latent space [4]. In the case of a TM, the components of the document representations describe the importance of a topic within the document. However, the topics might have different similarities, which are not taken into account when applying the DR directly. Atzberger et al. proposed an alternative by applying the DR on the topics themselves and deriving the document position in the two-dimensional plane as a linear combination, i.e., the position $\bar{d}$ of a document d is given by

$$\bar{d} = \sum_{j=1}^{K} \theta_j \bar{\phi}_j, \tag{7}$$

where $\theta = (\theta_1, \ldots, \theta_K)$ denotes the topic representation of d, and $\bar{\phi}_1, \ldots, \bar{\phi}_K$ denotes the positions of the topics after application of a DR [2]. Analogously, we compute the probabilities for the null hypothesis that Equation (7) does not improve the three stability aspects. The results are shown in Table 5.

LDA, in combination with t-SNE, benefits from Equation (7) across all three stability aspects in terms of local and global structures, as well as class separation. In the case of LSI and NMF, the results are not that obvious. We suspect that in the case of LSI and NMF, the extracted topics are "more orthogonal" to each other since these two TMs rely on eigenvectors. In the case of LDA, this is not the case, and therefore, the dissimilarities between the topics differ more, which is emphasized in Equation (7).

5 Discussion

From the results of our evaluation, we see certain combinations of text embeddings and DRs that are particularly suitable for generating stable two-dimensional layouts for text corpora. We use these observations together with the findings of Atzberger and Cech et al. [3,4] to derive guidelines for the effective combination of text embeddings and DRs. However, our analysis, as well as the guidelines derived from it, are subject to threats to validity.

5.1 Main Findings

A visualization designer has to make numerous design decisions when creating text spatializations. A fundamental one is whether the DR should be applied to the DTM or the embedding of the corpus in a latent space. From the three sensitivity analyses, we have shown that text embeddings can improve all three aspects of stability. As such, we conclude that:

G1 We recommend using a text embedding to increase the stability concerning input data, hyperparameters, and randomness.

Furthermore the three sensitivity analyses show, that depending on the specific stability aspect, the text embeddings differ in their performance. Therefore, depending on the stability aspect that is to be optimized, we recommend the following embeddings:

G2-S1 We recommend BERT when optimizing for stability concerning input data.

G2-S2 We recommend LDA, LSI, and NMF when optimizing for stability concerning hyperparameters.

G2-S3 We recommend LDA when optimizing for stability concerning randomness.

When applying LDA, we derive from the corresponding binary tests:

G3 We recommend applying an aggregation according to Equation (7) when applying LDA.

When not optimizing for one specific stability aspect, but for all three, we further consider the weaknesses of each embedding. Since BERT shows poor results concerning accuracy [4], and LDA is very sensitive to changes in the input parameters, we conclude:

G4 We recommend using LSI as text embedding when optimizing for all three stability aspects.

In all experiments, t-SNE showed the best results, especially in terms of the preservation of local structures. We therefore conclude:

G5 We recommend using t-SNE as the dimensionality reduction.

In particular, these guidelines align with the recommendations concerning accuracy from previous studies [3, 4].

5.2 Threats to Validity

Our results depend on specific choices and have thus threats to validity. We see two major areas: (1) the sampling used in the data measurement step and (2) errors in the implementation and execution.

For the data measurement, we selected text corpora, text embeddings, DRs, and similarity metrics. In any of the four categories, we had to select a subset among many possibilities. Our results rely on three text corpora. A priori, it is unclear to what extent the patterns in the boxplot visualizations depend on the specific corpora. We added the boxplot visualizations for each corpus to the supplemental material. In any case, we see that the patterns in the individual boxplot visualizations are similar to the aggregated view. Therefore, our argumentation from section 4 is still valid individually. We assume additional corpora will not affect our results, particularly the derived guidelines. For each corpus-embedding combination, we fixed the hyperparameters of the embedding algorithm following best practices. We furthermore inspected the resulting topics of the TMs and compared them to the given categories to ensure that the model is of high quality. It is unclear if guideline G2 transfers to the case where more variants of each embedding are evaluated. Nevertheless, one of our main findings, that embeddings can improve the stability concerning changes to the input data, the hyperparameters, and the random seed, was validated in our experiments by using embeddings following best practices. Lastly, even our similarity metrics required the specification of hyperparameters, such as the number of points k to be considered as nearest neighbors. In choosing $k = 7$, we followed previous studies and emphasized the metrics to capture local similarity.

For the entire data processing pipeline, we used actively maintained libraries that are widely used among practitioners. However, we can not guarantee that these libraries have no bugs and do not change their behavior across releases. Furthermore, our implementations, e.g., the similarity metrics, could carry defects. We addressed this by using only code reviewed by at least one co-author and pair-programming sessions. Finally, we provide our entire code as a GitHub repository to allow for transparency. Due to errors in the cluster, some layouts were not computed. Therefore, some scatterplot pairings could not be evaluated. It is unclear to what extent the missing values would affect the results.

6 Conclusions & Future Work

Many visualizations for text corpora rely on a two-dimensional scatterplot that is derived from applying a text embedding and a subsequent DR. Since changes to the layout require cognitive effort by the user, the stability of a layout algorithm needs to be considered by the visualization designer. In this study, we analyzed the stability of text layout algorithms concerning changes to the input data, hyperparameters, and randomness. For this, we measured the preservation of local and global structures and cluster separation between a large set of scatterplots that were derived from systematically iterating over the layout algorithms and the hyperparameters of the underlying DRs. Based on a correlation analysis of the similarity metrics, we aggregated them into three metrics to quantify the similarity concerning local structures, global structures, and class separation. Based on a detailed statistical analysis of the results, we analyzed the impact of the embedding algorithms and the DRs concerning the different stability aspects. We discussed our findings and derived guidelines for the effective use of text embeddings and DRs to generate text spatializations. Our work aims to address uncertainties when applying text embeddings and DRs for the visualization of text corpora. Furthermore, we hope practitioners and researchers consider our guidelines when applying latent embeddings and DRs. To draw a "big picture", we further see possible applications of our evaluation setup – particularly the metrics – for selecting DRs for exploring different embeddings, e.g., internal representations of neural network approaches. The findings from such experiments could be integrated into visualization approaches that aim to help users explain high-dimensional embeddings [10, 64, 74].

We see different directions for future work, e.g., by extending our experiments to address the major threats to validity. We plan to evaluate different configurations for each embedding to derive fine-granular insights into their impact on layout stability. Our approach for evaluating stability can further be adapted to quantify the stability of layouts of time-dependent text corpora. Such experiments would require additional, time-dependent embeddings and DRs. Furthermore, it would be interesting to measure the similarity of scatterplots by using additional metrics, e.g., Aupetit and Sedlmair presented a large set of class separation measures [6]. Quantifying stability using measures that best align with human similarity perception is particularly relevant concerning the preservation of the mental map. We plan to conduct a user study to verify to what extent our guidelines improve text spatializations for concrete analysis tasks. The results from such a user study might also lead to a deeper understanding of how humans perceive similarity in scatterplots. Furthermore, it would be interesting to see whether our findings for text embeddings might generalize to other types of data, including multidimensional tabular datasets. While our measurements do not directly address the stability of DRs in multidimensional tabular data, the latent embeddings used in our study could analogously represent tabular data, given their lower dimensionality and sparsity ratio. Finally, our pursuit of guidelines for text specializations motivates their generalization to multimodal corpora, i.e., sets of documents that include other modalities such as images. In particular, this would require new kinds of embeddings and a more complex description of latent embeddings across several data domains [32].

Acknowledgments

We thank the reviewers for their valuable feedback. This work was partially funded by the Federal Ministry of Education and Research, Germany through grant 01IS22062 and project 16KN086467 funded by the Federal Ministry for Economic Affairs and Climate Action of Germany. 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

[1] C. C. Aggarwal and C. Zhai. A survey of text classification algorithms. In *Mining Text Data*, pp. 163–222. Springer, 2012. doi: 10.1007/978-1-4614-3223-4_6 3

[2] D. Atzberger, T. Cech, M. de la Haye, M. Söchting, W. Scheibel, D. Limberger, and J. Döllner. Software Forest: A visualization of semantic similarities in source code using a tree metaphor. In *Proc. 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications – Volume 3*, IVAPP '21, pp. 112–122. SciTePress, Setúbal, Portugal, 2021. doi: 10.5220/0010267601120122 8

[3] D. Atzberger, T. Cech, W. Scheibel, J. Döllner, and T. Schreck. Quantifying topic model influence on text layouts based on dimensionality reductions. In *Proc. 19th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications – Volume 3*, IVAPP '24, pp. 593–602. SciTePress, Setúbal, Portugal, 2024. doi: 10.5220/0012391100003660 2, 5, 6, 8, 9

[4] D. Atzberger, T. Cech, W. Scheibel, M. Trapp, R. Richter, J. Döllner, and T. Schreck. Large-scale evaluation of topic models and dimensionality reduction methods for 2d text spatialization. *IEEE TVCG*, 30(1):902–912, 2024. doi: 10.1109/TVCG.2023.3326569 2, 4, 5, 6, 8, 9

[5] M. Aupetit. Visualizing distortions and recovering topology in continuous projection techniques. *Elsevier Neurocomputing*, 70(7–9):1304–1330, 2007. doi: 10.1016/j.neucom.2006.11.018 4

[6] M. Aupetit and M. Sedlmair. SepMe: 2002 new visual separation measures. In *Proc. Pacific Visualization Symposium*, PacificVis '16, pp. 1–8. IEEE, New York, 2016. doi: 10.1109/PACIFICVIS.2016.7465244 9

[7] F. Beck, M. Burch, S. Diehl, and D. Weiskopf. A taxonomy and survey of dynamic graph visualization. *Wiley CGF*, 36(1):133–159, 2017. doi: 10.1111/cgf.12791 1

[8] M. Behrisch, M. Blumenschein, N. W. Kim, L. Shao, M. El-Assady, J. Fuchs, D. Seebacher, A. Diehl, U. Brandes, H. Pfister, T. Schreck, D. Weiskopf, and D. A. Keim. Quality metrics for information visualization. *Wiley CGF*, 37(3):625–662, 2018. doi: 10.1111/cgf.13446 2

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003. doi: 10.5555/944919.944937 3

[10] A. Boggust, B. Carter, and A. Satyanarayan. Embedding Comparator: Visualizing differences in global structure and local neighborhoods via small multiples. In *Proc. 27th International Conference on Intelligent User Interfaces*, IUI '22, pp. 746–766. ACM, New York, 2022. doi: 10.1145/3490099.3511122 9

[11] C. Bredius, Z. Tian, A. Telea, R. N. Mulawade, C. Garth, A. Wiebel, U. Schlegel, S. Schiegg, and D. A. Keim. Visual exploration of neural network projection stability. In *Proc. Workshop on Machine Learning Methods in Visualisation for Big Data*, MLVis '22, pp. 1–5. EG, Eindhoven, 2022. doi: 10.2312/mlvis.20221068 1, 2

[12] L. Chen and A. Buja. Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis. *Taylor & Francis Journal of the American Statistical Association*, 104(485):209–219, 2009. doi: 10.1198/jasa.2009.0111 5

[13] M. A. A. Cox and T. F. Cox. Multidimensional scaling. In *Handbook of Data Visualization*, pp. 315–347. Springer, 2008. doi: 10.1007/978-3-540-33037-0_14 4

[14] S. P. Crain, K. Zhou, S.-H. Yang, and H. Zha. Dimensionality reduction and topic modeling: From latent semantic indexing to latent Dirichlet allocation and beyond. In *Mining Text Data*, pp. 129–161. Springer, 2012. doi: 10.1007/978-1-4614-3223-4_5 3

[15] J. P. Cunningham and Z. Ghahramani. Linear dimensionality reduction: Survey, insights, and generalizations. *Journal of Machine Learning Research*, 16(1):2859–2900, 2015. 2

[16] T. T. de Araújo Tiburtino Neves, R. M. Martins, D. B. Coimbra, K. Kucher, A. Kerren, and F. V. Paulovich. Fast and reliable incremental dimensionality reduction for streaming data. *Elsevier Computers & Graphics*, 102:233–244, 2022. doi: 10.1016/j.cag.2021.08.009 2

[17] S. Deerwester, S. T. Dumais, G. W. Furnas, T. K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41(6):391–407, 1990. doi: 10.1002/(SICI)1097-4571(199009)41:6<391::AID-ASI1>3.0.CO;2-9 3

[18] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proc. North-American Chapter of the Association for Computational Linguistics*, NAACL-HLT '19, pp. 4171–4186. ACL, Kerrville, TX, 2019. 3

[19] D. Engel, L. Hüttenberger, and B. Hamann. A survey of dimension reduction methods for high-dimensional data analysis and visualization. In *Proc. Workshop on Visualization of Large and Unstructured Data Sets: Applications in Geospatial Planning, Modeling and Engineering*, vol. 27, pp. 135–149. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany, 2012. doi: 10.4230/OASIcs.VLUDS.2011.135 2

[20] M. Espadoto, R. M. Martins, A. Kerren, N. S. T. Hirata, and A. C. Telea. Toward a quantitative survey of dimension reduction techniques. *IEEE TVCG*, 27(3):2153–2173, 2021. doi: 10.1109/TVCG.2019.2944182 2, 4, 5, 6

[21] R. Etemadpour, R. Motta, J. G. d. S. Paiva, R. Minghim, M. C. F. de Oliveira, and L. Linsen. Perception-based evaluation of projection methods for multidimensional data visualization. *IEEE TVCG*, 21(1):81–94, 2015. doi: 10.1109/TVCG.2014.2330617 2

[22] I. K. Fodor. A survey of dimension reduction techniques. Technical Report UCRL-ID-148494, Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, 2002. 2

[23] T. Fujiwara, J.-K. Chou, S. Shilpika, P. Xu, L. Ren, and K.-L. Ma. An incremental dimensionality reduction method for visualizing streaming multidimensional data. *IEEE TVCG*, 26(1):418–428, 2020. doi: 10.1109/TVCG.2019.2934433 2

[24] F. J. García-Fernández, M. Verleysen, J. A. Lee, I. Díaz Blanco, et al. Stability comparison of dimensionality reduction techniques attending to data and parameter variations. In *Proc. Workshop on Visual Analytics using Multidimensional Projections*, VAMP '13, pp. 5–9. EG, Eindhoven, 2013. doi: 10.2312/PE.VAMP.VAMP2013.005-009 1, 2

[25] X. Geng, D.-C. Zhan, and Z.-H. Zhou. Supervised nonlinear dimensionality reduction for visualization and classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 35(6):1098–1107, 2005. doi: 10.1109/TSMCB.2005.850151 5

[26] A. Gisbrecht and B. Hammer. Data visualization by nonlinear dimensionality reduction. *Wiley Data Mining and Knowledge Discovery*, 5(2):51–73, 2015. doi: 10.1002/widm.1147 2

[27] Y. Goldberg and Y. Ritov. Local procrustes for manifold embedding: a measure of embedding quality and embedding algorithms. *Springer Machine Learning*, 77:1–25, 2009. doi: 10.1007/s10994-009-5107-9 5

[28] R. Gove, L. Cadalzo, N. Leiby, J. M. Singer, and A. Zaitzeff. New guidance for using t-SNE: Alternative defaults, hyperparameter selection automation, and comparative evaluation. *Elsevier Visual Informatics*, 6(2):87–97, 2022. doi: 10.1016/j.visinf.2022.04.003 2

[29] M. Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv CoRR*, cs.CL(2203.05794), 2022. pre-print. doi: 10.48550/arXiv.2203.05794 3

[30] R. A. Hamad, E. Järpe, and J. Lundström. Stability analysis of the t-SNE algorithm for human activity pattern data. In *Proc. International Conference on Systems, Man, and Cybernetics*, SMC '18, pp. 1839–1845. IEEE, New York, 2018. doi: 10.1109/SMC.2018.00318 1, 2

[31] M. Hogräfer, M. Heitzler, and H.-J. Schulz. The state of the art in map-like visualization. *Wiley CGF*, 39(3):647–674, 2020. doi: 10.1111/cgf.14031 1

[32] Z. Huang, D. Witschard, K. Kucher, and A. Kerren. VA + Embeddings STAR: A state-of-the-art report on the use of embeddings in visual analytics. *Wiley CGF*, 42(3):539–571, 2023. doi: 10.1111/cgf.14859 9

[33] H. Jeon, A. Cho, J. Jang, S. Lee, J. Hyun, H.-K. Ko, J. Jo, and J. Seo. ZADU: A Python library for evaluating the reliability of dimensionality reduction embeddings. In *Proc. Conference on Visualization and Visual Analytics*, VIS '23, pp. 196–200. IEEE, New York, 2023. doi: 10.1109/VIS54172.2023.00048 5

[34] H. Jeon, G. J. Quadri, H. Lee, P. Rosen, D. A. Szafir, and J. Seo. CLAMS: A cluster ambiguity measure for estimating perceptual variability in visual clustering. *IEEE TVCG*, 30(1):770–780, 2024. doi: 10.1109/TVCG.2023.3327201 2

[35] P. Joia, D. Coimbra, J. A. Cuminato, F. V. Paulovich, and L. G. Nonato. Local affine multidimensional projection. *IEEE TVCG*, 17(12):2563–2571, 2011. doi: 10.1109/TVCG.2011.220 5

[36] I. Jolliffe. Principal component analysis. In *Encyclopedia of Statistics in Behavioral Science*. John Wiley & Sons, Ltd, 2005. doi: 10.1002/0470013192.bsa501 4

[37] D. G. Kendall. A survey of the statistical theory of shape. *Statistical Science*, 4(2):87–99, 1989. doi: 10.1214/ss/1177012582 5

[38] J. Khoder, R. Younes, and F. B. Ouezdou. Stability of dimensionality reduction methods applied on artificial hyperspectral images. In *ICCVG 2012: Computer Vision and Graphics*, vol. 7594 of *Lecture Notes in Computer Science*, pp. 465–474. Springer, Berlin, Heidelberg, 2012. doi: 10.1007/978-3-642-33564-8_56 1, 2

[39] G. Kindlmann and C. Scheidegger. An algebraic process for visualization design. *IEEE TVCG*, 20(12):2181–2190, 2014. doi: 10.1109/TVCG.2014.2346325 1

[40] T. Kohonen. Exploration of very large databases by self-organizing maps. In *Proc. International Conference on Neural Networks*, ICNN '97, pp. 1–6. IEEE, New York, 1997. doi: 10.1109/ICNN.1997.611622 4

[41] J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Springer Psychometrika*, 29(1):1–27, 1964. doi: 10.1007/BF02289565 5

[42] K. Kucher and A. Kerren. Text visualization techniques: Taxonomy, visual survey, and community insights. In *Proc. Pacific Visualization Symposium*, PacificVis '15, pp. 117–121. IEEE, New York, 2015. doi: 10.1109/PACIFICVIS.2015.7156366 1

[43] Q. Le and T. Mikolov. Distributed representations of sentences and documents. In *Proc. International Conference on Machine Learning*, ICML '14, pp. 1188–1196. Proceedings of Machine Learning Research, 2014. 3

[44] D. D. Lee and H. S. Seung. Learning the parts of objects by non-negative matrix factorization. *Springer Nature*, 401(6755):788–791, 1999. doi: 10.1038/44565 3

[45] J. A. Lee and M. Verleysen. Quality assessment of dimensionality reduction: Rank-based criteria. *Elsevier Neurocomputing*, 72(7–9):1431–1443, 2009. doi: 10.1016/j.neucom.2008.12.017 5

[46] D. J. Lehmann and H. Theisel. Optimal sets of projections of high-dimensional data. *IEEE TVCG*, 22(1):609–618, 2016. doi: 10.1109/TVCG.2015.2467132 3

[47] Y. Ma, A. K. H. Tung, W. Wang, X. Gao, Z. Pan, and W. Chen. ScatterNet: A deep subjective similarity model for visual analysis of scatterplots. *IEEE TVCG*, 26(3):1562–1576, 2020. doi: 10.1109/TVCG.2018.2875702 2

[48] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv CoRR*, stat.ML(1802.03426), 63 pages, 2020. pre-print. doi: 10.48550/arXiv.1802.03426 4

[49] J. Melka and J.-J. Mariage. Adapting self-organizing map algorithm to sparse data. In *Computational Intelligence*, pp. 139–161. Springer, 2019. doi: 10.1007/978-3-030-16469-0_8 5

[50] T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. *arXiv CoRR*, cs.CL(1301.3781), 12 pages, 2013. pre-print. doi: 10.48550/arXiv.1301.3781 3

[51] C. Morariu, A. Bibal, R. Cutura, B. Frénay, and M. Sedlmair. Predicting user preferences of dimensionality reduction embedding quality. *IEEE TVCG*, 29(1):745–755, 2023. doi: 10.1109/TVCG.2022.3209449 2, 6

[52] R. Motta, R. Minghim, A. de Andrade Lopes, and M. C. F. Oliveira. Graph-based measures to assist user assessment of multidimensional projections. *Elsevier Neurocomputing*, 150:583–598, 2015. doi: 10.1016/j.neucom.2014.09.063 5

[53] L. G. Nonato and M. Aupetit. Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment. *IEEE TVCG*, 25(8):2650–2673, 2019. doi: 10.1109/TVCG.2018.2846735 1, 2

[54] A. V. Pandey, J. Krause, C. Felix, J. Boy, and E. Bertini. Towards understanding human similarity perception in the analysis of large sets of scatter plots. In *Proc. CHI Conference on Human Factors in Computing Systems*, CHI '16, pp. 3659–3669. ACM, New York, 2016. doi: 10.1145/2858036.2858155 2, 4

[55] G. J. Quadri, J. A. Nieves, B. M. Wiernik, and P. Rosen. Automatic scatterplot design optimization for clustering identification. *IEEE TVCG*, 29(10):4312–4327, 2023. doi: 10.1109/TVCG.2022.3189883 2

[56] P. E. Rauber, A. X. Falcao, and A. C. Telea. Visualizing time-dependent data using dynamic t-SNE. In *Proc. Eurographics Conference on Visualization – Short Papers*, EuroVis '16, pp. 73–77. EG, Eindhoven, 2016. doi: 10.2312/eurovisshort.20161164 2

[57] S. Raval, C. Wang, F. Viégas, and M. Wattenberg. Explain-and-Test: An interactive machine learning framework for exploring text embeddings. In *Proc. Conference on Visualization and Visual Analytics*, VIS '23, pp. 216–220. IEEE, New York, 2023. doi: 10.1109/VIS54172.2023.00052 1

[58] C. Reinbold, A. Kumpf, and R. Westermann. Visualizing the stability of 2D point sets from dimensionality reduction techniques. *Wiley CGF*, 39(1):333–346, 2020. doi: 10.1111/cgf.13806 2

[59] A. Saltelli, M. Ratto, S. Tarantola, and F. Campolongo. Sensitivity analysis in practice: a guide to assessing scientific models. *Elsevier Reliability Engineering & System Safety*, 91(10–11):1109–1125, 2004. doi: 10.1016/j.ress.2005.11.014 1

[60] M. Sedlmair and M. Aupetit. Data-driven evaluation of visual quality measures. *Wiley CGF*, 34(3):201–210, 2015. doi: 10.1111/cgf.12632 5

[61] M. Sedlmair, A. Tatu, T. Munzner, and M. Tory. A taxonomy of visual cluster separation factors. *Wiley CGF*, 31(3pt4):1335–1344, 2012. doi: 10.1111/j.1467-8659.2012.03125.x 2

[62] S. Sidney. *Nonparametric Statistics for the Behavioral Sciences*. McGraw-Hill Professional, 2nd ed., 1988. 5

[63] M. Sips, B. Neubert, J. P. Lewis, and P. Hanrahan. Selecting good views of high-dimensional data using class consistency. *Wiley CGF*, 28(3):831–838, 2009. doi: 10.1111/j.1467-8659.2009.01467.x 5

[64] V. Sivaraman, Y. Wu, and A. Perer. Emblaze: Illuminating machine learning representations through interactive comparison of embedding spaces. In *Proc. 27th International Conference on Intelligent User Interfaces*, IUI '22, pp. 418–432. ACM, New York, 2022. doi: 10.1145/3490099.3511137 9

[65] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(11):2579–2605, 2008. 4

[66] L. van der Maaten, E. Postma, and J. van den Herik. Dimensionality reduction: a comparative review. Technical Report 009-005, Tilburg University, Tilburg Centre for Creative Computing, The Netherlands, 2009. 2

[67] J. Venna and S. Kaski. Visualizing gene interaction graphs with local multidimensional scaling. In *Proc. European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, ESANN '06, pp. 557–562. ESANN, 2006. 4

[68] E. F. Vernier, J. L. D. Comba, and A. C. Telea. Guided stable dynamic projections. *Wiley CGF*, 40(3):87–98, 2021. doi: 10.1111/cgf.14291 2, 5, 6

[69] E. F. Vernier, R. Garcia, I. d. Silva, J. L. D. Comba, and A. C. Telea. Quantitative evaluation of time-dependent multidimensional projection techniques. *Wiley CGF*, 39(3):241–252, 2020. doi: 10.1111/cgf.13977 2, 5, 6

[70] H. M. Wallach, D. Mimno, and A. K. McCallum. Rethinking LDA: Why priors matter. In *Proc. 22nd International Conference on Neural Information Processing Systems*, NIPS '09, pp. 1973–1981. Curran Associates, Inc., 2009. 5

[71] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno. Evaluation methods for topic models. In *Proc. 26th Annual International Conference on Machine Learning*, ICML '09, pp. 1105–1112. ACM, New York, 2009. doi: 10.1145/1553374.1553515 5

[72] Y. Wang, K. Feng, X. Chu, J. Zhang, C.-W. Fu, M. Sedlmair, X. Yu, and B. Chen. A perception-driven approach to supervised dimensionality reduction for visualization. *IEEE TVCG*, 24(5):1828–1840, 2018. doi: 10.1109/TVCG.2017.2701829 2

[73] Y. Wang, Z. Wang, T. Liu, M. Correll, Z. Cheng, O. Deussen, and M. Sedlmair. Improving the robustness of scagnostics. *IEEE TVCG*, 26(1):759–769, 2020. doi: 10.1109/TVCG.2019.2934796 2, 4

[74] Z. J. Wang, F. Hohman, and D. H. Chau. WizMap: Scalable interactive visualization for exploring large machine learning embeddings. In *Proc. 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pp. 516–523. ACL, Kerrville, TX, 2023. doi: 10.18653/v1/2023.acl-demo.50 9

[75] M. O. Ward, G. Grinstein, and D. Keim. *Interactive data visualization: foundations, techniques, and applications*. AK Peters/CRC Press, 2010. doi: 10.1201/9780429108433 1

[76] C. Ware. *Information Visualization: Perception for Design*. Morgan Kaufmann, 4th ed., 2019. 5

[77] M. Wattenberg, F. Viégas, and I. Johnson. How to use t-SNE effectively. Technical report, Distill, 2016. doi: 10.23915/distill.00002 1

[78] L. Wilkinson, A. Anand, and R. Grossman. Graph-theoretic scagnostics. In *Proc. Symposium on Information Visualization*, InfoVis '05, pp. 157–164. IEEE, New York, 2005. doi: 10.1109/INFVIS.2005.1532142 2

[79] J. Xia, L. Huang, Y. Sun, Z. Deng, X. L. Zhang, and M. Zhu. A parallel framework for streaming dimensionality reduction. *IEEE TVCG*, 30(1):142–152, 2024. doi: 10.1109/TVCG.2023.3326515 2

[80] J. Xia, W. Lin, G. Jiang, Y. Wang, W. Chen, and T. Schreck. Visual clustering factors in scatterplots. *IEEE CG&A*, 41(5):79–89, 2021. doi: 10.1109/MCG.2021.3098804 3

[81] J. Xia, Y. Zhang, J. Song, Y. Chen, Y. Wang, and S. Liu. Revisiting dimensionality reduction techniques for visual cluster analysis: An empirical study. *IEEE TVCG*, 28(1):529–539, 2022. doi: 10.1109/TVCG.2021.3114694 2