Dimensions of Similarity: Towards Interpretable Dimension-Based Text Similarity

Hans Ole Hatzel[*], Fynn Petersen-Frey, Tim Fischer and Chris Biemann

Universität Hamburg
ORCiD ID: Hans Ole Hatzel https://orcid.org/0000-0002-4586-7260,
Fynn Petersen-Frey https://orcid.org/0000-0001-6450-8100, Tim Fischer https://orcid.org/0000-0003-2560-5623,
Chris Biemann https://orcid.org/0000-0002-8449-9624

Abstract. This paper paves the way for interpretable and configurable semantic similarity search, by training state-of-the-art models for identifying textual similarity guided by a set of aspects or dimensions. The similarity models are analyzed as to which interpretable dimensions of similarity they place the most emphasis on. We conceptually introduce configurable similarity search for finding documents similar in specific aspects but dissimilar in others. To evaluate the interpretability of these dimensions, we experiment with downstream retrieval tasks using weighted combinations of these dimensions. Configurable similarity search is an invaluable tool for exploring datasets and will certainly be helpful in many applied natural language processing research applications.

1 Introduction

Traditional information retrieval methods heavily relied on interpretable token-based measures. More recently, embedding approaches have become increasingly popular. Recent advancements are made using transformer-based Siamese training setups that create embeddings for similarity search (e.g., [17]), often referred to as semantic similarity models. Unlike traditional methods, understanding why two documents (or a query and a document) are considered similar is non-trivial. While it intuitively follows from the training objective that paraphrases of the same sentence are deemed similar, it is not always clear how this property applies to a given pair of texts. In downstream applications, especially in interdisciplinary contexts, we repeatedly encountered questions of interpretability when suggesting semantic similarity search. We propose approaching this desideratum by modeling various dimensions of similarity, giving the user insight into which aspects of the documents are similar. Previous work [4] targets a similar issue for the specific domain of German poetry, demonstrating the need for more interpretable similarity measures for NLP applications. We hypothesize that existing similarity models place a significant focus on named entities which are often unaffected by paraphrases but are not helpful when, for example, searching for documents describing similar actions with entirely different named entities. Instead, we develop various models that focus on different aspects of the texts trained with the dataset provided as part of the SemEval 2022 Task 8 [2], which contains news articles that were annotated with regard to their similarity along seven dimensions. Additionally, we further analyze the dimensions and test the broader application to retrieval-focused downstream tasks.

We envision an application that allows for a configurable similarity search (see Figure 1) to focus on different aspects – for example, to retrieve documents that cover the same topic but from a different perspective as manifested in the style or tone – and brings interpretability to the results of similarity search (see Figure 2).

Our three key contributions are **(i)** training sentence similarity models on new data representing different aspects of similarity, presenting the first results for various dimensions on the SemEval dataset, and **(ii)** evaluating the use of our models on downstream tasks, thereby paving the way for **(iii)** our proposed user-configurable similarity search.

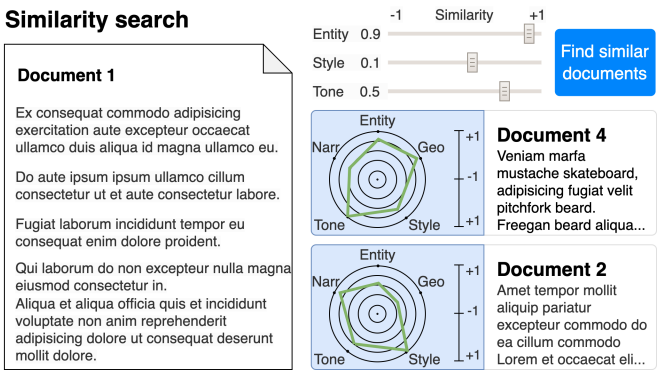

Figure 1. Search mockup. Left: current document. Right: configurable search and results that visualize its similarity across the various dimensions.

2 Background

Two currently popular architectures for text similarity comparisons are cross-encoders and bi-encoders. In both approaches, a language model is used. In a cross-encoder setup, one model receives both documents as input and produces a similarity score, whereas, in a bi-encoder, only one document is encoded at a time with the similarity being computed by comparing the output representations of various documents using an applicable similarity scoring function such as Euclidean distance or cosine similarity. As is already evident by this description, a cross-encoder can only compare two documents at a time, whereas the output representation comparison of the bi-encoder

[*] Corresponding Author. Email: hans.ole.hatzel@uni-hamburg.de

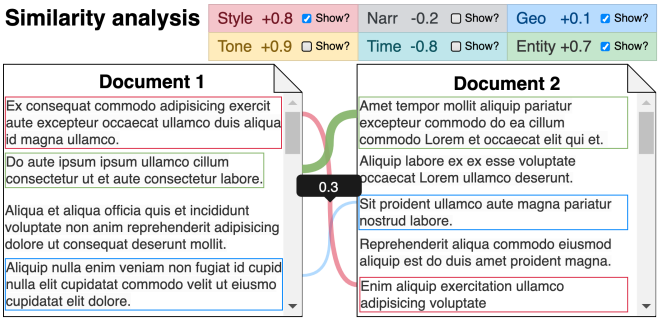

Figure 2. Analysis mockup. The similarity of two documents is broken down to sentence-wise dimension-based similarity.

easily scales to more documents making it the suitable architecture for corpus-wide similarity search. While bi-encoders have been shown to perform slightly worse than cross-encoders [17], the quadratic time requirement of cross-encoders for corpus-level comparisons makes bi-encoders the only suitable choice for efficient similarity searches on large corpora.

The SemEval 2022 Task 8 shared task [2] introduced a dataset of annotated news links, with the corresponding articles being scrapable from the Web. It consists of about 10,000 news article pairs of 18 language combinations. Further, the dataset introduces seven dimensions of similarity annotated for each pair of articles: *geography* (`GEO`), *entities* (`ENT`), *time* (`TIME`), *narrative* (`NAR`), *style* (`STYLE`), *tone* (`TONE`), and *overall* (`OVERALL`). In each pair, the scalar similarity along each dimension was rated as a whole number on a four-point scale (1–4) by multiple human annotators, with the average of multiple annotators being used as the gold target. Accordingly, the data contains non-integer real numbers in many cases. As randomly sampled article pairs are very likely to be entirely unrelated, various sampling techniques, such as the overlap in named entities, were employed to provide annotators with article pairs that have some degree of similarity. The dataset has two splits for evaluation: "English only" and "All data". Task participants were only evaluated on their performance in predicting the `OVERALL` dimension. The limited evaluation may, in part, have been motivated by the strong correlation between some classes; the scalar value describing the `NAR` class, for example (which focuses on narrative schema similarity), shares a Pearson correlation of ρ=.88 with the `OVERALL` class. For other class combinations, however, we see correlations as low as ρ=.22 (`TIME` and `STYLE`). Most participants did not utilize the extra dimensions, while few others [21, 8, 9] experimented with multi-label loss based on the additional dimensions. Only [8] reported scores for other dimensions (in the form of MSE loss), but their final correlation score was not competitive. Both bi-encoders and cross-encoder architectures were submitted as part of the task, with no clear winning architecture emerging in the evaluation of the system performances [2].

SentenceBERT is a Siamese network-based [17] approach for semantic sentence embeddings. In the recent past, various adaptations of the approach, such as SentenceBERT's language-agnostic counterpart LaBSE [7] have become quite popular in semantic search applications. The produced sentence embeddings allow a generic, blackbox-like semantic search across different languages using cosine similarity.

Ostendorff et al. [14] take a different approach to aspect-based similarity in scientific papers by comparing sections with the same titles using sentence similarity models, resulting in aspects of similarity such as *Results* or *Related Work*. Prior work exists in configurable search approaches (e.g., [11]), but we do not know of further work in the context of neural semantic similarity search.

3 Methodology

Toward our goal of a configurable similarity search system, we exploit the annotations of seven dimensions of similarity of the SemEval 2022 Task 8 dataset to train models that should be able to represent documents according to these different aspects. As we envision a retrieval system that scales to large document collections, we consider only bi-encoders in our experiments. Approximate nearest neighbor indices of document embeddings are typically utilized to develop fast and efficient document retrieval systems. While models with bi-encoder architecture can compute document embeddings, the cross-encoder architecture is not suitable for building such indices.

In the first step, we seek a suitable pretrained bi-encoder model to use as a base model for further fine-tuning on the individual dimensions (see Section 4.1). Here, we compare several state-of-the-art models and select the one that performs best in the multi-lingual setting.

Before continuing with fine-tuning this model on the seven dimensions of similarity, we want to understand which textual aspects this model focuses on. Consequently, in Section 4.2, we check our hypothesis that 'models pretrained on paraphrases focus mainly on named entities' by comparing these not finetuned models to an entity-embedding baseline.

Next, we train an individual model for each dimension of similarity as annotated in the SemEval dataset. This experiment (see Section 4.3) validates whether sentence similarity models are able to capture these dimensions and the resulting models could possibly be used in our envisioned application.

However, we understand that having seven different models requiring seven inference steps per document to obtain the different document representations is rather inefficient. This is the reason why, in the next step, we evaluate the use of multi-task learning (see Section 4.3) to develop a single model that produces embeddings for each dimension in a single forward pass. We achieve this by using a dimension-specific head that transforms the final embedding.

Finally, we move on to downstream experiments (see Section 5) to evaluate the use of our trained models on out-of-domain data. These experiments further serve as an exploration of the potential for configurable similarity search.

All our experimentation source-code is available online.[1]

4 Experiments using SemEval 2022 Task 8

We evaluate the performance of pretrained models on the SemEval 2022 Task 8 data and fine-tune existing models along all seven dimensions. The finetuned models will be evaluated in detail in our downstream experiments.

4.1 Preliminary experiments

We perform preliminary experiments to select a suitable pretrained model for our further experiments on learning different dimensions of similarity. We do not want to perform all experiments on multiple models due to resource and time constraints. We require the selected model to be competitive with the best participants of the SemEval Task 8 but still simple enough to be easily finetuned on the dataset.

Following those criteria, we start with an initial set of models, some of which were used by the original task participants, others

[1] https://github.com/uhh-lt/dimensions-of-similarity/

were selected as a point of comparison, and evaluate them without additional finetuning. As shown in Table 1, LaBSE performs best on the multi-lingual data but is outperformed by other models on the English split. While there are full systems in the shared task [2] that perform slightly better than a plain sentence embedding model by incorporating additional features on top of that model, the difference in performance is small compared to the difference in complexity and run-time efficiency. Based on these results and to optimize for multi-lingual capabilities, we selected a plain LaBSE model as the foundation for all further experiments.

Table 1. The pretrained LaBSE is outperformed on the English split of the dataset, but performs best in the multi-lingual scenario.

Pretrained Models	Overall ρ	
	en only	all languages
bert-base-multilingual-cased	0.54	0.35
stsb-xlm-r-multilingual	0.72	0.44
LaBSE	0.71	**0.65**
all-MiniLM-L6-v2	0.80	0.42
all-mpnet-base-v2	**0.82**	0.48

4.2 Baseline and Entity Focus

We suspect that existing similarity models exhibit a strong focus on named entities rather than other textual features. This would be a logical consequence of models being trained on paraphrases where entities can, in many cases, not be easily changed. To test this hypothesis – and to build a further baseline specifically for the *entity* dimension –, we explore how static word embeddings of just the entities compare with pretrained models, specifically LaBSE. Comparable performance of the sentence encoder with the static embedding model would indicate that a focus on entity mentions is plausible. We implemented the following procedure to develop a baseline that computes document similarities by focusing on entities. First, we represent a document only by such tokens that are considered entities by the SpaCy named entity tagger (all named entities with the tags `ORG`, `PERSON`, `GPE`, `LOC`, `LANGUAGE`, `PRODUCT`, `WORK_OF_ART` or `FAC`). Next, we use the English static fastText [1] embeddings to obtain a vector representation for each token. However, as the dataset contains not only English documents, we translate all non-English articles into English using EasyNMT[2], following [3]. This way, we can evaluate this fastText-based entity-focused baseline on the English-only and the multi-lingual setup. Given two documents a and b, next, we compute their tokens' pairwise cosine similarity, which results in a 2D matrix of shape $A \times B$ where A and B are the number of entity tokens that represent document a and b, respectively. To determine the similarity of a and b, we perform a greedy best match on this matrix by averaging each column's and each row's maximum. This is inspired by BERTScore's [22] precision and recall metric. Finally, we correlate the resulting score with the `ENT` dimension of the dataset.

The results in Table 2 show that by default LaBSE exhibits some focus on entities, as correlation with the `ENT` dimension is roughly on par with the `OVERALL` correlation. At the same time, the fastText baseline actually outperforms an un-finetuned LaBSE in both dimensions using just entities, as long as only English texts are considered. We attribute the much worse performance of fastText on all data to the translation quality provided by EasNMT. This baseline performs about on par with the Jaccard similarity based baselines in the shared task paper, which takes all tokens into account [2], but no numerical results are available for this baseline. We can say that a static-embedding entity approach actually outperforms LaBSE for entities in the English data. Concerning our hypothesis, while more inspection is needed, there appears to be some focus on entities with other aspects also taking an important role.

[2] https://github.com/UKPLab/EasyNMT

Table 2. Using just the fastText embeddings of entity mentions, we build a baseline that outperforms an un-finetuned LaBSE on English data.

Dataset	Model	Entities ρ	Overall ρ
English Split	LaBSE	0.67	0.66
	mBERT	0.34	0.36
	FastText	**0.75**	**0.68**
All Data	LaBSE	**0.61**	**0.64**
	mBERT	0.31	0.31
All Data (translated)	LaBSE	**0.57**	**0.58**
	mBERT	0.31	0.31
	FastText	0.53	0.48

4.3 Finetuning

Finetuning was performed in a Siamese-network setup, training on a cosine similarity loss using mean squared error to compare the cosine similarity of the Siamese network outputs with the training label, incentivizing (cosine-)similar embeddings for documents that were annotated as similar. We use the SentenceTransformer [17] library to perform the finetuning and set the training parameters to 3 epochs, 100 warmup steps and batch size of 32. The learning rate and optimizer were kept as default values of 2×10^{-5} and AdamW. In the labeled data, 1 represents high similarity along a specific dimension, whereas 4 represents extreme dissimilarity. We found that transforming the label scale had a large impact on performance and thus followed Di Giovanni et al. in this regard [3]. In fact introducing this label-scaling scheme was the single change, outside of pretrained model choice, that we found to have by far the largest positive impact on performance. Using Equation 1, we transform the labels from the [1, 4] range to [0.0, 1.0] instead of [-1.0, 1.0], as one would assume for cosine similarity.

$$f(x) = \frac{4 - x}{3} \tag{1}$$

We approach our goal – to obtain models that assess the similarity of documents with respect to seven dimensions of similarity – from two directions. Our first multi-model (MM) approach is to finetune one model per dimension of similarity. There was no need to change the model architecture of LaBSE for this approach. It results in 7 different models, each specialized for a certain dimension.

However, for our envisioned application of large-scale, configurable similarity search, this setup could be more efficient. Using individual models, each having its own set of weights, would require multiple inference runs to obtain all dimension-specific document embeddings.

Accordingly, we attempt in our second multi-task-learning (MTL) approach to unify all dimensions into a single multi-task model. Here, we changed the model architecture of LaBSE by adding multiple heads (one for each dimension of size 512). Further, we altered the training objective to optimize for all dimensions at once, while including an example of each dimension in each batch and keeping all other training parameters the same. This setup has the advantages of sharing the vast majority of parameters as well as requiring only a single inference run to obtain all dimension specific document embeddings.

We evaluate these two approaches on the multi-lingual eval split and compare it to the unfinetuned LaBSE (pretrained). The reported

scores denote the Pearson correlation with human judgments. Results are shown in Table 3.

Table 3. Pearson correlation with human judgments of our two approaches on the multi-lingual eval split. The multi-model (MM) finetuned LaBSE models outperform the multi-task-learning (MTL) model and the pretrained LaBSE on every dimension.

LaBSE Variant	GEO	ENT	TIME	NAR	STYLE	TONE	OVERALL
pretrained	.44	.64	.44	.65	.37	.44	.65
MM	**.62**	**.77**	**.52**	**.78**	**.56**	**.53**	**.78**
MTL	.53	.74	.51	**.78**	.48	.51	.76

We observe a clear improvement in LaBSE's performance when finetuning on each of the individual dimensions. For the GEO, ENT, NAR, OVERALL and STYLE dimensions there is a vast improvement over the pretrained model. TIME and TONE, however, are still worthy of improvement.

For the multi-task model, we see that for the dimensions OVERALL and NAR the performance is essentially equivalent but it lacks far behind on other dimensions. We attribute this to the fact that the multi-task model has trouble separating the different dimensions as the OVERALL and NAR dimensions are strongly correlated (and to a lesser extent also ENT) – a hypothesis that is supported by our results in Section 4.4.

The best participant in the shared task [21] achieved a Pearson correlation of 0.818 with OVERALL on the multi-lingual eval split using a cross-encoder architecture. However, for our use case of similarity search and retrieval, only the bi-encoder architecture is relevant. In this category, the best system [19] also used LaBSE in combination with data augmentation techniques as well as metadata-based features and achieved a Pearson correlation of 0.801 with OVERALL, which is only slightly above ours of 0.78 while using a considerably more complex system than our finetuned sentence embedding model. As a result, we use our own model for all further experiments in which we typically do not have access to all the same metadata (such as a release date) as in the SemEval setup.

4.4 Comparing human judgments and machine judgments

In this section, we analyze whether machines or humans can better distinguish between the different dimensions. Further, we try to identify which dimensions machines focus on compared to human judgments. To do so, we look at the correlations between all pairs of dimensions. For human judgments, we use the labels from the dataset to produce the pairwise Pearson correlations as shown in Figure 4 of [2]. For machine judgments, we calculate the pairwise Pearson correlations between two dimensions of two models finetuned on these dimensions. We performed this both for the individually finetuned models shown in Figure 3 and the MTL model shown in Figure 4. In case of the MTL model, the dimension-specific embeddings are used in the pairwise comparisons. Finally, we subtract the human judgement correlations from the machine correlations resulting in Figure 5 to identify how machines results differ from human judgments.

It is clearly visible that the machine judgments have a higher inter-dimensions correlation than human judgments, i.e., all values are positive. Further, it is most obvious that the MTL model has an even higher inter-dimensions correlation than the individual models. When

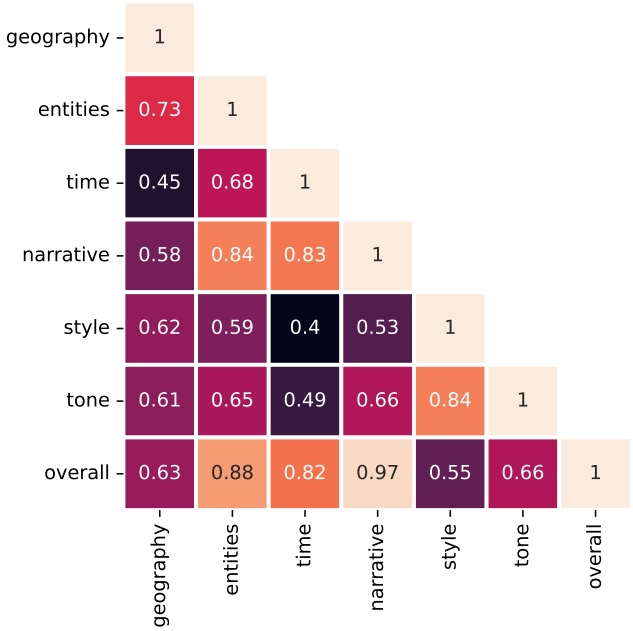

Figure 3. Correlation between the dimensions of the individual finetuned models. Inter-dimension correlations are considerably higher than for the human data judgements.

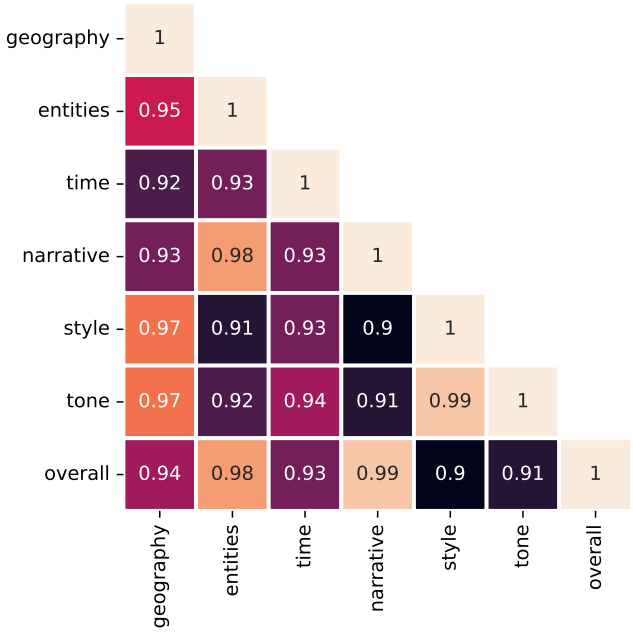

Figure 4. Correlation between the dimensions of the multi-task learning model. Inter-dimension correlations far surpass those of both human judgements and the individual models, indicating that this model, in which most parameters are shared across dimensions, is unable to distinguish between the dimensions.

looking at the individual models, it can be seen that they are especially worse at distinguishing STYLE and TONE than humans are. The visual outliers with the smallest correlation differences (OVERALL−NAR, ENT−NAR, ENT−OVERALL), are cases where the human judgments

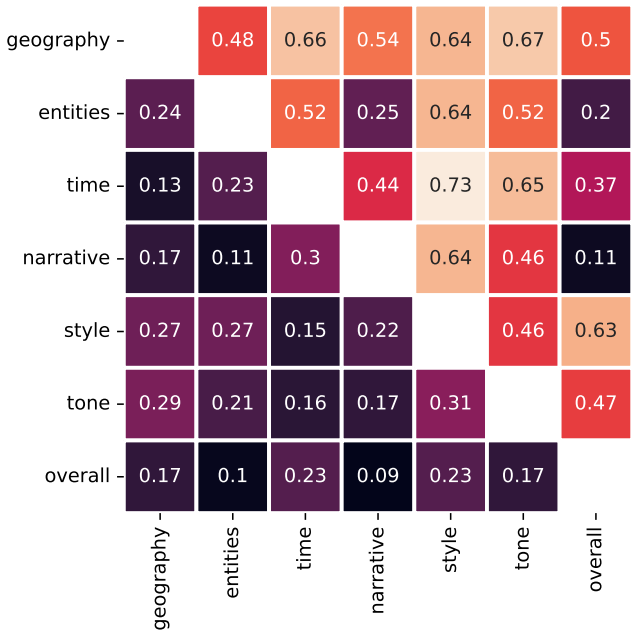

Figure 5. Correlation differences to human correlations: MTL model is on the top right, individual models are on the bottom left.

already strongly correlate. The same (but slightly larger) effect also can be observed for the models. From these experiments, we conclude that the finetuned models are worse at distinguishing the different dimensions than humans are. While the individual models are only slightly worse than the human judgments, the MTL model is almost unusable in this regard. More specifically, the inter-dimension correlation for the MTL model is extremely high (> 0.9) for all pairs. We verified this extensively to ensure it is not a bug in our code. We attribute this to the fact that the model cannot really change the majority of parameters to deal with specific dimensions. It is thus forced into a very similar behavior for all dimensions. Potential future remedies may be adjusted training curricula, currently we have one example of each dimension in each batch. Another potential avenue of improvement is the use of adapters [10] which would allow for sharing most parameters and only using a small set of dimension specific ones. Further improvements still may be achievable using adapter fusion [15] which in principle allows for sharing model capabilities across dimensions but would still require an inference pass for each dimension.

5 Down Stream Applications

In the typical downstream application of similarity search, paraphrase models are implicitly aligned with the users' intentions. In research applications, however, we observed that practitioners often desire a precise understanding of which aspects of texts are actually similar. Our approach is to model these aspects using the dimensions as annotated in the SemEval 2022 Task 8 dataset.

We envision an application used by domain experts to retrieve documents fulfilling certain similarity criteria. Such criteria could be configured by assigning weights to each dimension. The user could be presented with sliders controlling the similarity in each dimension ranging from -1, to +1 as depicted in Figure 1. As part of an exploratory research process, such an application would give instant feedback to changed weightings by displaying only those documents that score highest with respect to a current anchor document. For example, the user would able to retrieve news documents covering the same incident from different perspectives by looking for high similarity in the entity and time dimensions but low similarity in the style and tone dimensions.

In this specific example, we model the same incident by means of positive weights for the overall and entity dimensions and the different perspectives by a negative weight for the tone dimensions.

This task can be formulated as a weighted sum of similarities along dimensions of interest. More formally: given a pair of documents p and a weight vector w ($w \in \mathbb{R}^D$) that assigns a weight to each dimension $d \in D$, the weighted similarity $t(p, w)$ over all dimensions D is computed as follows:

$$t(p, w) = \sum_d w_d \cdot s_d(p) \tag{2}$$

where the similarity of a pair of documents in a certain dimension $s_d(p)$ can be estimated by our finetuned models described in the previous section.

In the following, we conduct experiments on various datasets, applying the suggested scheme from above. For all following experiments, we use the best versions of our LaBSE models that were finetuned on a single dimension. For these experiments, we assume correlations between specific dimensions in our finetuned model and labels created in other datasets. These assumed correlations are based on intuitive judgments and were not programmatically optimized in an effort to show the potential application of configurable searches. The goal of all following experiments is to understand how the dimensions relate to annotations in existing datasets, e.g., in our first experiment 5.1, we analyze if the `TONE` dimension is sufficiently correlated with ratings of product reviews. We are essentially modeling downstream applications for unseen datasets, simulating various use cases. Showing positive results in these experiments indicates that a configurable similarity search is feasible.

5.1 Product Reviews Experiments

For this first downstream experiment, we use the review dataset [6]. This dataset is a set of over 5 million product reviews collected from Amazon in 2014. Each product review has a rating that is based on a star-scaled system, where the highest rating has five stars, and the lowest rating has 1 star, generally meaning 'I love it.' and 'I hate it', respectively. While originally conceived for sentiment analysis task, we repurpose the dataset to simulate the following retrieval tasks:

1. Given a review with a certain star rating, we want to find reviews with the same rating for other products.
2. Given a review of a certain product, we want to find reviews of the same product, but with different ratings.

As the search space, we select the first 100 products that have 100 or more reviews and their first 100 reviews totaling 10,000 documents. We model the same product using the `ENT` and `OVERALL` dimensions, and we model the same star rating using the `TONE` dimension. In other words, when applying our finetuned models in this experiment, we assume increased similarity in the `ENT` and `OVERALL` dimension for reviews of the same product. In contrast, we expect increased similarity in the `TONE` dimension for reviews with the same star rating.

Table 4 shows the results of Task 1. Indeed, our model finetuned on the `TONE` dimension (finetuned-LaBSE-tone) outperforms the pretrained LaBSE and random baseline slightly. Additionally, we

Table 4. Mean average precision scores for retrieving a review of different products with the same star rating.

Finetuned	Model Name	MAP
✗	random	0.421
✗	LaBSE	0.451
✓	LaBSE-tone	0.458
✓	tone - overall	0.447
✓	2 · tone - overall	**0.464**

Table 5. Mean average precision scores for retrieving a review for the same product with a different rating.

Finetuned	Model Name	MAP
✗	random	0.006
✗	LaBSE	0.036
✓	LaBSE-overall	0.089
✓	overall + entities	0.088
✓	overall + entities - tone	**0.090**

configured the similarity search to lessen the impact of the OVERALL dimension by subtracting it, denoted as 'tone - overall', further improving the performance only when the TONE dimension is given more weight, denoted '2 · tone - overall'.

Table 5 shows the results of Task 2 and has generally way lower scores than Table 4 as the number of products is much larger than the number of ratings. This is also the main reason the random baseline performs so poorly. Again, our model finetuned on the OVERALL dimension (finetuned-LaBSE-overall) outperforms the pretrained LaBSE and random baseline. We also experimented with configuring the similarity search: We included the ENT dimension (overall + entities) and subtracted the TONE dimension (overall + entities - tone), which slightly improved the performance.

These experiments show a promising result: configuring the similarity search by adding or subtracting certain dimensions can improve the results. This is an indication that our embeddings can a) be useful outside the domain of news texts and can b) be configured in an interpretable way.

5.2 Text Register Experiment

We conduct a document retrieval experiment using the Corpus of Online Registers of English [12] dataset. This dataset consists of almost 50,000 documents crawled from the unrestricted open Web, which were manually annotated with a taxonomy of eight main categories and tens of subcategories. For this experiment, we sample 1000 documents labeled with "Opinion Narrative" and 1000 documents labeled with "Informational Narrative".

With this experiment we test whether our models can distinguish between these two categories by simulating a typical document similarity search scenario: Given a document of a certain category, retrieve more documents of the same category. We hypothesize that these categories should be best separated by the STYLE or TONE dimension, as informational articles mostly use formal language, whereas opinion articles may use colloquial language. We do not expect the OVERALL, ENT or GEO dimension to be good discriminators, as both categories contain narratives about various topics.

Table 6 confirms our hypothesis. A document retrieval system can find better results when utilizing a model finetuned on the STYLE dimension. While the difference in performance is not exceptional, we clearly observe that a document retrieval system returns better results

Table 6. Mean average precision scores for retrieving a narrative of the same category (opinion vs. informational).

Finetuned	Model Name	MAP
✗	LaBSE	0.574
✓	LaBSE-style	**0.630**
✓	LaBSE-tone	0.613
✓	LaBSE-narrative	0.539
✓	LaBSE-entities	0.529
✓	LaBSE-geography	0.529
✓	LaBSE-overall	0.534

when utilizing a finetuned model optimized for STYLE or TONE as opposed to using a model trained on paraphrases (pretrained-LaBSE).

5.3 Sentiment Classification Experiment

For this experiment, we again use a dataset for sentiment analysis, the SemEval 2017 Task 4 [18] dataset. However, we intentionally picked a dataset with manually annotated sentiment scores rather than ones inferred from ratings to set this experiment apart from the product review experiment. Further, instead of evaluating our models in a retrieval setting, we utilize their embeddings for a classification task in this experiment.

This task includes five subtasks, and with this experiment, we take on Subtask A: The data for this Subtask consists of English and Arabic Twitter posts with annotations on a 3-point scale ranging from negative to neutral to positive. Given a tweet, our system has to predict whether it conveys Positive (P), Negative (N) or Neutral (U) sentiment. Accordingly, we follow the evaluation strategy of Task 4 Subtask A to compare this approach to existing work and report the *average recall*. This is computed by averaging the recall across the Positive (P), Negative (N) and Neutral (U) class:

$$AvgRec = \frac{1}{2}(R^P + R^N + R^U)$$

where R^P, R^N and R^U denote the recall with respect to the Positive, Negative and Neutral class.

We tackle this task with two different approaches: (1) We train a linear-kernel Support Vector Machine (SVM) using the embedding of each of the respective dimension models as the input vector. (2) We use K-nearest-neighbors (KNN) with k=10, i.e,. we use majority voting between the ten nearest neighbors in the training set, by cosine distance in embedding space, to classify each sample (either in the TONE or the OVERALL models' space).

The results are shown in Table 7. Our SVM-based classification would have placed us competitively in 7th and 2nd place in the for the original English and Arabic shared tasks. Regarding the non-finetuned KNN-based approach, which is more reflective of the actual distances between our embeddings, we would have placed in the midfield, 23rd for English and third for Arabic (with English having 37 and Arabic 8 submissions). It is evident that for the OVERALL dimension, the gap between the KNN and the SVM approaches is larger. This is not surprising as one may conjecture that some dimensions in the OVERALL representation already correlate highly with the sentiment. While compared to modern approaches specific to the Tweet domain, we do not perform well, the model performs in line with the original shared task's participants without domain specific training.

Table 7. Performance of our models on the SemEval 2017 Task 4, Subtask A. Scores from [5, 13] provided for comparison, both comparison systems were finetuned on the training data.

	Arabic			English		
	AvgRec	F1	Acc	AvgRec	F1	Acc
KNN_{Tone}	0.514	0.505	0.539	0.596	0.584	0.589
SVM_{Tone}	0.553	0.555	0.587	**0.656**	**0.658**	**0.668**
$KNN_{Overall}$	0.489	0.516	0.481	0.526	0.516	0.53
$SVM_{Overall}$	**0.562**	**0.569**	**0.588**	0.641	0.642	0.652
BERTweet	-	-	-	**0.732**	**0.728**	**0.717**
NileTMRG	**0.583**	**0.610**	**0.581**	0.578	0.515	0.606

5.4 German Poetry Experiment

With this experiment, we want to evaluate the applicability of our models and approach to the vastly different domain of German poetry. We use the dataset by [4], a collection of contemporary poetry aimed at a general audience with no thematic restrictions from the two epochs 'realism' and 'modernism'. This dataset includes relative similarity annotations for 470 triples of 866 poems. Annotators were tasked to judge for each similarity dimension (content, form, style, emotion, overall) whether "Poem A is more similar to poem B than poem C".

We follow their evaluation schema, using the balanced accuracy metric and only considering comparisons where annotators agreed. We apply our finetuned LaBSE models to compute the cosine similarity of poems in a specific dimension. The mapping of our models' 'SemEval Dimension' to their dimensions of similarity as well as the results are shown in Table 8.

Table 8. Balanced accuracy for triples in Ehrmanntrauts dataset. The SemEval dimensions are the SemEval dimensions that were used to approximate the poetry dimension.

	XLM-R[3]		LaBSE		
Dimension	pre-trained	fine-tuned	pre-trained	SemEval Dimension	fine-tuned
Content	0.69	0.81	0.67	NAR	0.69
Form	0.58	0.76	0.65	-	-
Style	0.66	0.79	0.64	STYLE	0.66
Emotion	0.66	0.76	0.66	TONE	0.67
Overall	0.69	0.79	0.67	Overall	0.66

We compare our approach to their best-performing model 'paraphrase-XLM-R', a multi-lingual RoBERTa-based SentenceTransformer model trained on paraphrases (untrained) or finetuned on the poetry data. Table 8 shows that our models, in most cases, stay behind even the pretrained XLM-R, making them not very suitable for the application. We slightly improve over the pretrained LaBSE model by applying our models finetuned for matching dimensions in all but one dimension (`OVERALL`). At first sight, this may indicate that dimensions of similarity may, to some degree, be transferable to the poetry domain. If we, however, check which of our trained dimensions matches best for each of the datasets dimensions (see Table 9), the picture is much less clear. In fact, the `TONE` model produces the best result for all but one dimension. We conclude that the poetry dataset's dimensions do not align with those in the SemEval dataset; given the domain mismatch, this is hardly surprising.

These results also indicate that while further domain adaptation is necessary for an application in poetry, we do not remove the base models transferability in our fine-tuning. This is a finding that, in a more general sense, has been made for out-of-domain transfer, with prior work on information retrieval showing zero-shot performance on unseen data can lack considerably behind in-domain performance [20].

[3] Results as reported by [4].

Table 9. The different model's performance in predicting similarity dimensions of the poetry dataset, reported as balanced accuracy.

SemEval Model	Poetry Dataset				
	Content	Form	Style	Emotion	Overall
NAR	0.69	0.62	0.66	0.62	0.67
STYLE	0.68	0.62	0.66	0.65	0.68
TONE	0.7	0.64	0.65	0.67	0.7
Overall	0.67	0.6	0.65	0.63	0.66

6 Limitations

While our model performs well on a variety of tasks there are multiple major limitations to be discussed. First of all, we did not test the `NAR`, `TIME` and `GEO` dimensions, meaning their applicability beyond the SemEval corpus is unknown. For the case of the `NAR` dimension, as it is strongly correlated with the `OVERALL` dimension, we assume that it will behave similarly. We have no reason to believe that the `GEO` dimension will present a challenge but suspect that the `TIME` dimension may not be easily captured by our models as multiple participants in the original shared task implemented special handling for time and dates. In terms of cross-domain performance, much exploration is still to be done. We showed that our models were not ready for the wildly out-of-domain poetry data, but perform admirably on other tasks. Our approach will have to be tested with human users in the future, as our downstream experiments were a bit contrived and not driven by real-world use cases. A major limitation of our work is that we only ever consider the first 512 (subword-)tokens of an article, which is a limitation in the models' pre-training which, in turn, is motivated by memory constraints. In news articles, this is likely not a significant problem due to the inverted-pyramid property [16] but is something we will seek to address in future work.

7 Conclusion

We are not currently aware of semantic search approaches configurable in terms of dimensions, making this concept a major contribution. With a single dataset, we trained multiple models that construct a configurable similarity system and subsequently showed its usefulness for downstream experiments and established that multi-task learning, at least in a simple setup, underperforms compared to individual models. It seems conceivable that other unrelated datasets, like the ones we operated on in Section 5, could be used to further train similarity models without the need for new custom datasets. Our models achieve state-of-the-art performance in-domain for various dimensions and have been successfully applied for some out-of-domain data. We conclude that, for the `STYLE`, `TONE`, `ENT` and `OVERALL` dimensions, the trained models are to some extent applicable to other datasets.

Search systems making use of our models may not only help in building a more interpretable notion of similarity but also enable users to balance the aspects of similarity that are important to them. They allow to fulfill such requests as: "show me articles on the same topic but from different perspectives." In the future, we aim to implement such a system, allowing users to balance the importance of various dimensions in a similarity search.

Acknowledgements

This work was, in part, supported by the D-WISE (grant `01UG2124`) project funded by BMBF and through the project "Evaluating Events in Narrative Theory (EvENT)" (grant `BI 1544/11-1`) funded by the DFG.

References

[1] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov, 'Enriching word vectors with subword information', *Transactions of the Association for Computational Linguistics*, **5**, 135–146, (2017).

[2] Xi Chen, Ali Zeynali, Chico Camargo, Fabian Flöck, Devin Gaffney, Przemyslaw Grabowicz, Scott Hale, David Jurgens, and Mattia Samory, 'SemEval-2022 task 8: Multilingual news article similarity', in *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pp. 1094–1106, Seattle, United States, (July 2022). Association for Computational Linguistics.

[3] Marco Di Giovanni, Thomas Tasca, and Marco Brambilla, 'DataSciencepolimi at SemEval-2022 task 8: Stacking language models to predict news article similarity', in *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pp. 1229–1234, Seattle, United States, (July 2022). Association for Computational Linguistics.

[4] Anton Ehrmanntraut, Thora Hagen, Fotis Jannidis, Leonard Konle, Kröncke Merten, and Simone Winko, 'Modeling and measuring short text similarities. on the multi-dimensional differences between german poetry of realism and modernism', *Journal of Computational Literary Studies*, **1**, (12 2022).

[5] Samhaa R. El-Beltagy, Mona El Kalamawy, and Abu Bakr Soliman, 'NileTMRG at SemEval-2017 task 4: Arabic sentiment analysis', in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 790–795, Vancouver, Canada, (August 2017). Association for Computational Linguistics.

[6] Xing Fang and Justin Zhan, 'Sentiment analysis using product review data', *Journal of Big Data*, **2**(1), 5, (June 2015).

[7] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang, 'Language-agnostic BERT sentence embedding', in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 878–891, Dublin, Ireland, (May 2022). Association for Computational Linguistics.

[8] Nikhil Goel and Ranjith Reddy Bommidi, 'Wolfies at semeval-2022 task 8: Feature extraction pipeline with transformers for multi-lingual news article similarity', in *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pp. 1129–1135, Seattle, United States, (July 2022). Association for Computational Linguistics.

[9] Joseph Hajjar, Weicheng Ma, and Soroush Vosoughi, 'DartmouthCS at SemEval-2022 task 8: Predicting multilingual news article similarity with meta-information and translation', in *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pp. 1157–1162, Seattle, United States, (July 2022). Association for Computational Linguistics.

[10] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly, 'Parameter-efficient transfer learning for nlp', in *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97*, pp. 2790–2799. PMLR, (2019).

[11] Martin Köppelmann, Dustin Lange, Claudia Lehmann, Marika Marszalkowski, Felix Naumann, Peter Retzlaff, Sebastian Stange, and Voget Lea, 'Scalable similarity search with dynamic similarity measures', Technical report, Hasso Plattner Institute, Potsdam, Germany, (2012).

[12] Veronika Laippala, Samuel Rönnqvist, Miika Oinonen, Aki-Juhani Kyröläinen, Anna Salmela, Douglas Biber, Jesse Egbert, and Sampo Pyysalo, 'Register identification from the unrestricted open web using the corpus of online registers of english', *Language Resources and Evaluation*, 1–35, (10 2022).

[13] Dat Quoc Nguyen, Thanh Vu, and Anh Tuan Nguyen, 'BERTweet: A pre-trained language model for English tweets', in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 9–14, Online, (October 2020). Association for Computational Linguistics.

[14] Malte Ostendorff, Terry Ruas, Till Blume, Bela Gipp, and Georg Rehm, 'Aspect-based document similarity for research papers', in *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6194–6206, Barcelona, Spain (Online), (December 2020). International Committee on Computational Linguistics.

[15] Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych, 'AdapterFusion: Non-destructive task composition for transfer learning', in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 487–503, Online, (April 2021). Association for Computational Linguistics.

[16] Horst Pöttker, 'News and its communicative quality: the inverted pyramidwhen and why did it appear?', *Journalism Studies*, **4**(4), 501–511, (2003).

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

[18] Sara Rosenthal, Noura Farra, and Preslav Nakov, 'SemEval-2017 task 4: Sentiment analysis in Twitter', in *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pp. 502–518, Vancouver, Canada, (August 2017). Association for Computational Linguistics.

[19] Iknoor Singh, Yue Li, Melissa Thong, and Carolina Scarton, 'GateNLP-UShef at SemEval-2022 task 8: Entity-enriched Siamese transformer for multilingual news article similarity', in *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pp. 1121–1128, Seattle, United States, (July 2022). Association for Computational Linguistics.

[20] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych, 'Beir: A heterogeneous benchmark for zero-shot evaluation of information retrieval models', in *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, eds., J. Vanschoren and S. Yeung, volume 1. Curran, (2021).

[21] Zihang Xu, Ziqing Yang, Yiming Cui, and Zhigang Chen, 'HFL at SemEval-2022 task 8: A linguistics-inspired regression model with data augmentation for multilingual news similarity', in *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pp. 1114–1120, Seattle, United States, (July 2022). Association for Computational Linguistics.

[22] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi, 'Bertscore: Evaluating text generation with bert', in *Proceedings of the 8th International Conference on Learning Representations*, Accepted as poster. Online, (4 2020).