Interpretable Text Embeddings and Text Similarity Explanation: A Survey

Juri Opitz[1*] **Lucas Moeller**[2*] **Andrianos Michail**[1] **Sebastian Padó**[2] **Simon Clematide**[1]

[1]University of Zurich, Switzerland
[2]University of Stuttgart, Germany

[1]{jurialexander.opitz,andrianos.michail,simon.clematide}@uzh.ch
[2]{lucas.moeller,sebastian.pado}@ims.uni-stuttgart.de

[*]Equal contribution

Abstract

Text embeddings are a fundamental component in many NLP tasks, including classification, regression, clustering, and semantic search. However, despite their ubiquitous application, challenges persist in interpreting embeddings and explaining similarities between them. In this work, we provide a structured overview of methods specializing in inherently interpretable text embeddings and text similarity explanation, an underexplored research area. We characterize the main ideas, approaches, and trade-offs. We compare means of evaluation, discuss overarching lessons learned and finally identify opportunities and open challenges for future research.

1 Introduction

Text embedding models (Reimers and Gurevych, 2019; Gao et al., 2021) are ubiquitous in research and industry, as they promise to map the meaning or content of sentences and documents to useful numerical vector representations ("embeddings"), among which an arithmetic distance (or similarity) can be calculated. Applications range from semantic search and retrieval (Ye et al., 2016; Guo et al., 2020; Muennighoff, 2022; Hambarde and Proenca, 2023; Alatrash et al., 2024) to text classification (Schopf et al., 2023a), topic modeling (Grootendorst, 2022), NLG evaluation (Celikyilmaz et al., 2020; Sai et al., 2022; Larionov et al., 2023; Chollampatt et al., 2025), graph reasoning (Plenz et al., 2023), and retrieval-augmented generation (RAG, Lewis et al., 2020; Gao et al., 2023). Advances in base models (Günther et al., 2023; Wang et al., 2024a), context size (Li et al., 2023), instruction tuning (Su et al., 2023), and scalable infrastructure (Wang et al., 2022) continuously enhance their capabilities. Most recently, a trend has been to build embedding models from large pre-trained decoders by removing their causal attention masking and continuing to train them contrastively in a Siamese setup using annotated, mined or LLM augmented pairs of similar texts (Muennighoff, 2022; BehnamGhader et al., 2024). The approach is also widely adopted by the industry (Lee et al., 2025a, 2024, 2025b). While this shows that knowledge obtained upon generative pretraining can be effectively translated to representation tasks, evidently, a critical component remains contrastive training. Thus, the learning of informative text representations appears to be closely linked to text similarity.

Yet, with the advancement of text embedding models, a pressing challenge persists: the *interpretability of embeddings* and the *explainability of similarity* derived from them. For instance, when a document is returned in response to a query, we would like to articulate why this document was selected as the most *similar*, or why another was omitted. We find interpretability research with a focus on text representation and similarity to be underrepresented in the literature. One reason for this may lie in the pairwise nature of the encountered inputs, which introduces additional complexity: Similarity depends on interactions between two inputs rather than on features of a single input—a change in one input influences the effect of the second on the prediction (Tversky, 1977; Lin, 1998a)—and explanations must account for these interactions.

Importantly, such questions are not just theoretical. In light of laws like the EU AI Act ("right to explanation"; EU, 2024), the demand for transparency is expected to intensify. Thus, there is a strong and timely need for research, overview, and clarity in *all* fields of AI. In this work, we focus on interpretability and explainability in the context of similarity and embedding models. We intend this survey to serve as a resource for researchers interested in these challenges and to lower the entry barrier into this area positioned at the intersection of several research domains, including interpretability, representation learning, and NLP.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 22315–22331
November 4-9, 2025

2 Setting the Stage

We investigate interpretability and explainability in the context of neural text embedding models, and focus on three closely related aspects: (i) the interpretability of the models themselves, (ii) the properties of the text embeddings these models generate, and (iii) the similarity scores derived from comparing such embeddings.

Formal framework (Figure 1). Assume two text encoders F and G. Their backbone typically consists of a multi-layered neural network. In most cases, $F = G$, meaning the networks share weights, a setup known as a *Siamese* network (Koch et al., 2015); unless stated otherwise, we refer only to F. After the Neural Network has processed and refined the representation of an input text x, a last neural layer typically provides a final stage of refined representations, often on the token level basis (encoded tokens). Additionally, in most cases, there is a special last (optional) layer $L + 1$ (*project* in Figure 1) that projects to a d-dimensional space where text representation is independent of text length (standard vectors, Figure 1). This layer often employs averaging or max-pooling across individual token embedding dimensions, or reproduces the embedding from a specialized token.

Finally, the similarity of two texts x, y can be efficiently calculated through their embeddings (vectors) $e_x = F(x), e_y = F(y)$ by calculating a function $sim(e_x, e_y) \in \mathbb{R}$. In the simplest case, this can be the dot product $sim(e_x, e_y) = e_x^T e_y$, possibly normalized by length $l_{x,y} = |e_x|_2 \cdot |e_y|_2$ to obtain cosine similarity.

So far, the mechanism of text encoding and similarity computation is a standard and ubiquitous procedure. Importantly, this procedure leads to non-interpretable vectors, and consequently yields similarities that escape interpretation or explanation. Next, we elaborate on the parts that allow us to resolve or at least mitigate this issue. Simultaneously, we use these parts in Figure 1 to establish a scaffold for the structure of this survey.

Paper Structure. Our formal framework in Figure 1 permits us to distinguish between different classes of explainability approaches.

On the top level, we distinguish between **interpretable embeddings** (§3) and **post-hoc explanations** (§4). We speak of *interpretable* models if the structure of their architecture or embeddings

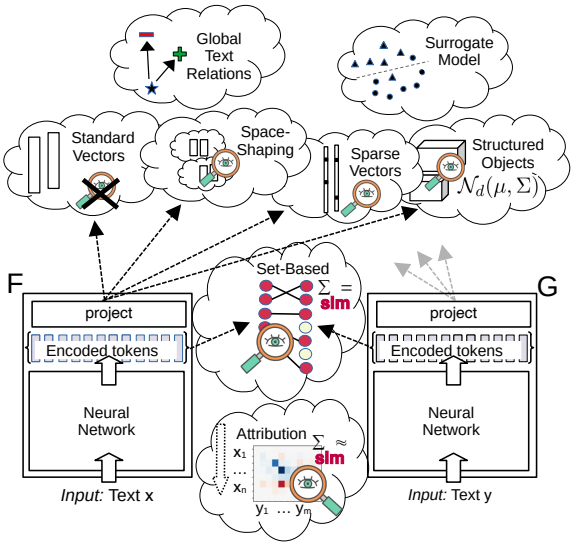

Figure 1: A schema of a standard text encoder architecture with the different interpretable embeddings and explainability approaches, each corresponding to subsections in the text.

inherently enables insights into their predictions to humans without a need for additional methods or further processing. *Post-hoc explanations*, on the other hand, generate insights into uninterpretable black-box models by applying an additional method that is not part of the original model's computation. We divide the former further into **space-shaping approaches** (§3.1) structuring the learned embedding space in interpretable ways, **sparse representations** (§3.2) yielding human-understandable sparse features, **structured objects** (§3.3) representing texts as geometric objects instead of simple vectors and **set-based embeddings** (§3.4) using not a single but multiple vectors to represent texts. The latter *post-hoc* approaches are further structured into **interaction attribution** (§4.1) tracing a prediction back to feature interactions between the model's two inputs, **global explanation** verifying the consistency of embeddings for known text relations on a dataset level and **surrogate modeling** (§4.3) optimizing a secondary interpretable model to approximate the original one. We begin every sub-category by introducing the common idea behind the described methods and conclude with opportunities they open up as well as remaining challenges. Table 1 in the Appendix visualizes our taxonomy, links respective sections and points to code resources. Finally, we examine evaluation methods and data sets (§5), and conclude with an extended discussion (§6) highlighting trade-offs,

lessons learned, challenges and future perspectives.

3 Interpretable Embeddings

These approaches aim at structuring the embedding space so that it reflects human-understandable features. As such they create inherently interpretable models (Rudin, 2019).

3.1 Shaping interpretable spaces

Idea. An interpretable embedding space can be explicitly trained to express human-understandable aspects, thereby bridging the gap between the power of neural embeddings and the interpretability of classic methods based on "bag-of-words" representations.

QA features. aim to develop interpretable features by framing embedding generation as answering a set of predefined questions about a text and encoding the answers as features. For this, we first need to find *a suitable set of questions* about texts, and create training data that *elicits answers* to these questions. Specifically, Benara et al. (2024) let an LLM answer "Yes"/"No" questions about a text (*Is the text about sports? Does the text express a command?*), building prompts based on dataset description. For predicting fMRI responses to language stimuli their method outperforms several baselines. Sun et al. (2024) propose constructing a concept space from a dataset by clustering word embeddings and then applying two constraints. First, the QA prompts must be strongly associated with one of the detected clusters. Second, for positive text pairs, all questions should be answered with "Yes" and for negative text pairs with "No", to sharpen the boundary between similar and dissimilar texts. The resulting embeddings are interpretable in the sense that question answers can be directly inferred from them.

Sub-embedding features. An embedding space can be decomposed into *multidimensional subspaces*, each isolating a specific semantic aspect. This allows overall similarity to be broken down into aspect-specific scores. The S3BERT approach by Opitz and Frank (2022), which requires a user to define a set of metrics that measure interpretable similarity aspects of two texts (e.g., *Is the focus of the texts the same?*). Since such aspects often are implicit in the texts, they leverage abstract meaning representation graphs (Banarescu et al., 2013) that encode aspects such as number, focus, semantic roles, negation; and use graph matching metrics (Opitz, 2023) on aspectual subgraphs. They fine-tune a reference embedding model such that the similarity of aspectual sub-embeddings regresses to the aspectual graph metrics. A consistency loss and residual sub-embedding help tie the overall similarities to the original reference. In the example in Figure 2, the similarity of concepts increases the value, while the dissimilarity of quantificational structure correctly lowers it.

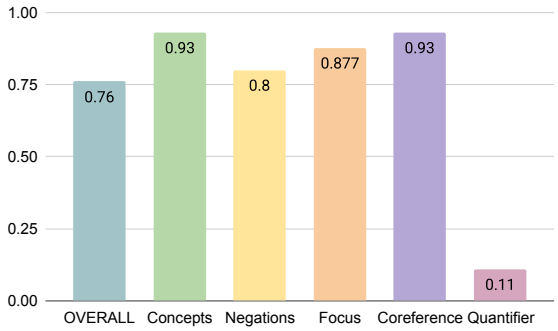

Figure 2: In S3BERT space decomposition, an overall sim=0.76 for the sentence pair *Two men are singing* and *Three men are singing* emerges from aggregating per-aspect similarities. (Simplified aspect set used here.)

Other approaches omit the consistency loss and aim to induce entirely new decomposed spaces. For instance, "multi-facet" embeddings are learned with graph metric ground truth (Risch et al., 2021), or "specialized-aspect" embeddings with aspect-specific transformer encoders (Ostendorff et al., 2022; Schopf et al., 2023b).

A more coarse-grained decomposition is proposed by Ponwitayarat et al. (2024), whose linguistic analysis of the *S*emantic *T*extual *S*imilarity dataset (*STS*) (Cer et al., 2017) found that a single continuous similarity range is not sufficient. They suggest a decomposition into two spaces, one for loosely similar texts (lower range), and another to capture finer distinctions among highly similar texts (higher range).

Anchor features let individual embedding dimensions express association (i.e., similarity) to interpretable anchors in a database, e.g., representative prototype texts that have been sampled (Wang et al., 2025) or pre-computed and aligned topics (Potthast et al., 2008). Compared to QA features, the provided explanation is more indirect but shows greater accuracy, almost on par with their non-interpretable embedding counterparts.

Challenges and opportunities. QA-based approaches have been evaluated favorably against bag-of-words baselines (Sun et al., 2024) and in specific domains (Benara et al., 2024). They still struggle with matching the performance of reference embedding models, likely due to the difficulty of defining a general and complete set of questions. Similarly, sub-embedding decomposition approaches require the manual definition of semantic aspects. However, the resulting dimensions are not directly interpretable as features—only the similarity values they produce can be linked to the defined aspects. This reliance on handcrafted features, can be seen as a limitation on the one hand, but enables the definition of custom embedding spaces aligned with specific interpretability goals on the other hand.

All space-shaping approaches pose additional constraints on a model risking downstream accuracy compared with standard embeddings. Interestingly, sometimes they can induce regulatory effects. The S3BERT authors e.g. observed a significant performance increase for judging argument similarity.

3.2 Sparse representations

Idea. Instead of assigning individual dimensions of dense embeddings to certain aspects, another approach towards creating interpretable spaces is sparsification.

Unsupervised Sparsification. Such sparsity can be induced without supervision by learning to reconstruct the embedding from sparse latent variables (Faruqui et al., 2015; Prokhorov et al., 2021; O'Neill et al., 2024). Trifonov et al. (2018) find that such sparse embeddings can indeed isolate some dimensions corresponding to human-interpretable features, including even spatial object relations (e.g., *physically laying on something*). However, they note that it can be difficult to tell "which features a dimension captures", and that the "increase in interpretability comes at a cost in reconstruction quality and, in some cases, utility in downstream tasks."

Sparse lexical embeddings map input texts onto term weight vectors whose dimensionality equals the length of a predefined vocabulary. Transformer models naturally provide this capability. Applying the unembedding matrix to the last layer's token representations results in a logit vector over the model's vocabulary (the basis for masked or next token classification during pre-training). Sparse embeddings repurpose them by combining the token-level logits into text-level representations through pooling along the sequence dimension. A sparsification objective is applied during contrastive learning. In contrast to lexical approaches like tf-idf, they are not bound to terms in the actual input but can assign weights to *expansion terms* that may additionally be relevant in the given context, e.g. synonyms. Sparse lexical embeddings are popular in retrieval scenarios because the term-based representation enables deployment via efficient inverted indices. Dai and Callan (2020) predict lexical term-weights from contextualized embeddings, Bai et al. (2020); Zhao et al. (2021) introduce vocabulary expansion, and Formal et al. (2021b,a) propose an end-to-end trainable model. Recently, sparse and dense embeddings have also been combined in unified models (Kong et al., 2023; Zhang et al., 2024; Awasthy et al., 2025).

Challenges and opportunities. Unsupervised sparse features can correspond to intuitive text characteristics but can be difficult to interpret in other cases. In turn, sparse lexical representations are trivial to interpret. Their sparsity can be beneficial in suitable scenarios like building a search index. However, the need for specialized data structures in order to handle their high dimensionality efficiently may be a burden in other contexts.

3.3 Structured Objects

Idea. Certain text relationships are inherently asymmetric. For instance, a natural relation between texts is *entailment*: A given hypothesis follows from a premise. Geometric embeddings reach beyond vector representations and utilize structured objects for representation offering a way to model these relationships.

Box embeddings represent inputs as high-dimensional boxes. For two such boxes a and b we have their size $s_a = a_1 \cdot a_2$, $s_b = b_1 \cdot b_2$, and their overlap $o_{a,b} = min(a_1, b_1) \cdot min(a_2, b_2)$. We can define their similarity as mutual containment, $o_{a,b}/(s_a + s_b - o_{a,b})$, and an asymmetric relationship like the entailment as unidirectional containment: $o_{a,b}/s_b$ is exactly 1 if a is fully contained/entailed in/by b. The challenge is to learn such objects given the 'curse of dimensionality', according to which box size and overlap tend towards zero for high-dimensional spaces. To alleviate such learning problems, Chheda et al. (2021) propose to adopt a probabilistic soft box overlap formula-

tion based on Gumbel random variables (Dasgupta et al., 2020).

Distributional embeddings view a text as a random variable (RV). Intuitively, this provides us with a model of multiple interpretations, which seems appealing due to natural language ambiguity: A text can have multiple interpretations, and only some of these interpretations can map to those of another similar text. But how to build such a probabilistic space? Shen et al. (2023) model a text as a Gaussian RV embedding $\mathcal{N}_d(\mu, \Sigma)$ by estimating "Model uncertainty" via Monte Carlo Dropout (Gal, Yarin and Ghahramani, Zoubin, 2016), and data uncertainty via smaller linguistic perturbations (e.g., dropping a word). The covariance matrix ($\hat{\Sigma}$) is then efficiently approximated through a banding estimator (Bien et al., 2016). For increased efficiency, Yoda et al. (2024) directly predict mean ($\hat{\mu}$) and covariance ($\hat{\Sigma}$).

Operator learning. An approach that works with standard vector representations but can also account for asymmetric text relations are the work by Huang et al. (2023). They propose learning interpretable operators for text meaning composition, such as union or fusion. These operators are modeled using neural networks, and the embedding space is retrained to accommodate such operations. Their evaluation shows minimal loss in standard similarity tasks, but greatly improved performance for compositional generation tasks.

Challenges and Opportunities. Modeling embeddings as geometric objects or learning operators can account for the long-established argument that similarity relations need not be symmetric (Tversky, 1977; Lin, 1998a). However, these approaches introduce additional complexity that may be too much of an overhead in other applications that do not have this requirement.

3.4 Set-based Interpretability

Idea. Set-based approaches are based on *two sets* of embeddings rather than two points. Often sets consist of token embeddings (from the last layer of an encoder), but other approaches go further and build meta-sets of text embeddings. Aligning such sets can reveal how different parts of the texts relate and contribute to the overall similarity score.

Token weight embeddings build text representations by aggregating token-level embeddings with explicit weights that reflect each token's importance and provides insight into how individual tokens contribute to the final text embedding. E.g., Wang and Kuo (2020) estimate token importance and novelty weights by analyzing variance across transformer layers. Seo et al. (2022), train models to learn token weights directly, utilizing a reconstruction loss. Tulkens and van Dongen (2024) compute static embeddings for all vocabulary tokens via a single transformer forward pass per token, followed by Zipf-informed averaging. As the final text representations are single vectors, while interpretability is on the level of individual tokens, these approaches are at the intersection between dense and set-based embeddings.

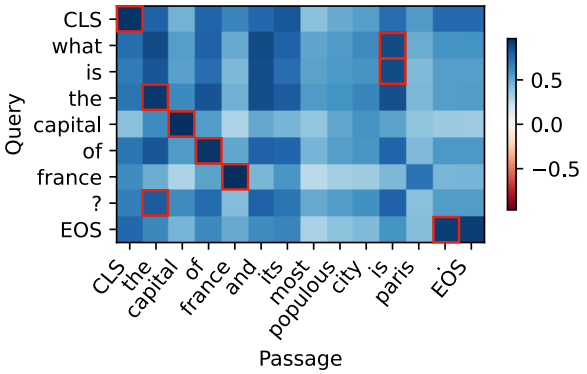

Figure 3: An example of a late-interaction matrix between query and passage token embeddings in the ColBERTv2.0 model. The overall sim is 0.965. Red boxes indicate row-wise maxima (alignment).

Sequential embeddings compare embeddings from the final model layer—before any reduction ("late interaction"). Two prominent techniques are ColBERT and BERTscore (Khattab and Zaharia, 2020; Santhanam et al., 2022), both of which compute asymmetric max-alignments between tokens and aggregate the similarities of the aligned pairs. BERTscore performs this alignment in both directions to produce a symmetric similarity measure. In terms of *explainability*, both approaches derive the final similarity score from token-level alignments, showing approximately which tokens the model matches between the inputs (Figure 3).

Multi-view Interpretation. Some approaches extend this idea by generating multiple text-level embeddings, each reflecting a different view or interpretation of the input text.

Hoyle et al. (2023) use a generative model to produce alternative hypotheses about a text. (Ravfogel et al., 2024) decompose the text into smaller state-

ments or descriptions. Given a decomposition of a text x into smaller parts $\{x_1, ...x_n\}$, the embedding model is then applied to each part individually, producing a set of text embeddings $\{e_1, ...e_n\}$.

A variation of the multi-text set-based approach is proposed by Liu et al. (2024). To compute textual similarity, they sample sets of possible continuation from an LLM and calculate the average log-likelihood difference between each input text and the generated continuations. The continuations can then be examined to provide an interpretable basis for the resulting similarity score.

Finally, Liu and Soatto (2024) compute text similarity *multi-modally* by comparing the imagery evoked by each text, using denoising via Stochastic Differential Equations (Song et al., 2021). Similarity is higher when texts elicit similar images, enabling visual interpretation of the score.

3.5 Challenges and Opportunities

Set-based approaches enable interpretable alignment of token-level embeddings, which can be valuable for tasks such as identifying semantic differences between related documents (Vamvas and Sennrich, 2023). They also naturally support asymmetric text relationships via directional matching or alignment. An important limitation is that sets of embeddings typically consume more memory than single vectors. Sequential embeddings also do not have a fixed size but vary with input length. While decomposition based approaches can point out matching sub-statements or hypotheses, they can require multiple forward passes.

4 Post-hoc Explanation

Different from inherently interpretable models, post-hoc explanations employ an additional method to gain insights into the predictions of a black-box model.

4.1 Interaction Attribution

Idea. Attribution-based approaches aim at assigning importance values to features reflecting their contribution to a given prediction of a model. A special characteristic of similarity models is that their predictions do not depend on individual features, due to the multiplicative interaction between the two inputs' embeddings in sim. Attributions must, therefore, be to feature interactions. First-order methods do not suffice to explain such interaction (Sundararajan et al., 2020; Janizek et al., 2021), and second-order methods are required for attribution in similarity models. Two lines of work have addressed this issue in text similarity models.

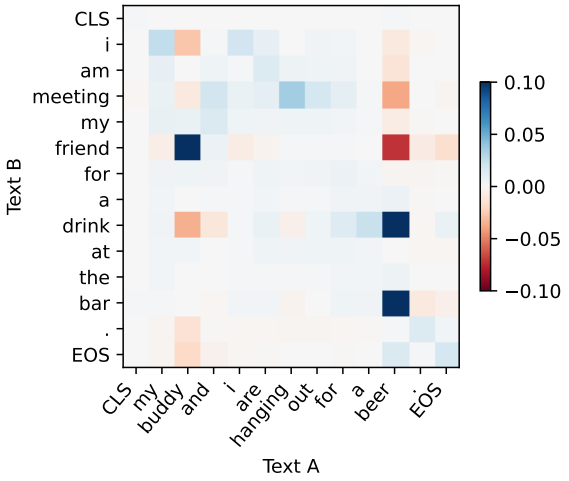

Figure 4: Interaction attributions between two sentences computed with the IJ method. The sim is 0.618 and the measurable attribution error is 0.001.

Integrated Jacobians. Integrated gradients (IG) attributes a scalar model prediction back onto individual input features by integrating over a number of interpolations between the actual input and an uninformative reference input (Sundararajan et al., 2017). Moeller et al. (2023) have applied the underlying theory of IG to text embedding models and proposed *Integrated Jacobians* (IJ) as the equivalent of IG for this model class. For text embedding models the output takes the form of a token-token matrix, showing the contribution of all individual token interactions to the sim (Figure 4). An approximate version of these attributions is directly applicable to off-the-shelf models without a need for tuning (Moeller et al., 2024).

Relevance Propagation. Layer-wise relevance propagation (LRP) is a framework to propagate feature-importance values for a model prediction back through the model in a layer-wise fashion (Bach et al., 2015; Montavon et al., 2019). Propagation rules are derived for individual layers based on first-order Taylor expansion of the underlying function. BiLRP extends the LRP framework to Siamese similarity models. Similar to IJ, the computation also takes the form of a product between two Jacobian-like matrices. The method was originally proposed in the computer vision domain (Eberle et al., 2020) and has recently also been applied to Siamese text encoder models (Vasileiou

and Eberle, 2024).

Challenges and Opportunities. Attribution approaches need to build Jacobian matrices, coming at a temporal complexity of $2 \times d$ independent backward passes, d being the model's embedding dimensionality. The resulting Jacobians have a quadratic spatial complexity and can require large GPUs to compute the associated matrix multiplications efficiently. Despite the computational costs, attribution methods have the advantage of being applicable to a wide class of embedding models as long as they are differentiable. They can provide certain theoretical guarantees (Sundararajan et al., 2017; Janizek et al., 2021), but have also been proven to be subject to other fundamental limitations (Bilodeau et al., 2024).

4.2 Global explainability

Idea. A common way of differentiating explainability methods is into *local* and *global* explanation (Danilevsky et al., 2020). Local approaches work on the level of individual examples. Alternatively, we can globally analyze the geometry of embeddings using a dedicated evaluation dataset.

Text relations. Zhu et al. (2018) and Zhu and de Melo (2020) follow this approach by constructing sets of sentences with known relations based on linguistic properties. In their initial work, the authors use triplets of sentences including a pair known to be similar with regard to a certain property and a dissimilar negative. Properties include negation, passivation, change of syntactic roles and word-ordering. It is then evaluated how consistently positive pairs are closer to another in the representation space than to the negative. In the second publication, the group extends the analysis to quadruples consisting of two pairs of similar sentences, thus, evaluating the similarity of sentence relations.

Challenges and opportunities. Analyzing embedding geometry provides a higher-level understanding of how consistently relations between sentences are represented in an embedding space. However, it requires the manual construction of suitable evaluation sets targeting specific properties and insights are limited to the properties covered.

4.3 Surrogate modeling

Idea. Surrogate models approximate a complex black-box model with a simpler, typically linear model that is inherently interpretable. We differentiate between two types, models operating on interpretable features approximating the original models predictions and models operating on the original model's embeddings *probing* them for known properties.

Interpretable approximation. Nikolaev and Padó (2023) construct artificial sentence pairs with known linguistic features. Based on these features, they then fit surrogate regression models to match the cosine similarity scores of different sentence transformers. The linearity of the fitted surrogate models allows them to analyze the relative importance of linguistic aspects in sentence pairs through the weights assigned to respective features.

Probing. A *probe* is a (often linear) classification model that is trained on top of pre-trained sentence embedding for a defined task. It assesses the generalizability of an embedding by testing whether the associated property is (linearly) separable in the learned representation space. Conneau et al. (2018) propose ten tasks around surface-level, syntactic and semantic information to probe the linguistic information contained in sentence representations of different models. In another work the group tests applicability to downstream applications like sentiment classification or retrieval (Conneau et al., 2017). More recently, probing has become an important evaluation tool in large-scale text embedding benchmarks like MTEB (Muennighoff et al., 2023). Nikolaev and Padó (2023) investigate which layers in different models encode semantic information through probing. Tehenan et al. (2025), inspired from ideas of mechanistic interpretability (Bricken et al., 2023), use sparse dictionary learning for investigating token-level linguistic information that is pooled in a sentence embedding.

Challenges and Opportunities. While conceptually simple, surrogate models do require additional objectives and optimization to generate insights into black-box models. Although limited, they do have a certain learning capacity, which needs not necessarily align with what the original model has learned.

5 Evaluation and Datasets

5.1 Evaluation

The presented approaches differ substantially in the types of explanations they produce, making it

difficult—if not impossible—to define a unified evaluation framework covering them all. In fact, evaluation often focuses on specific characteristics of individual approaches. Space-shaping models can explicitly correlate ground truth values for similarity metrics against predictions from respective sub-spaces (Opitz and Frank, 2022). Aspect encoders test in how far nearest neighbors in the embedding space share aspects and assess whether aspect clusters emerge in dimensionality reduction plots (Ostendorff et al., 2022; Schopf et al., 2023b). Specialized and structured objects allow to evaluate whether the learned representations reflect asymmetric relations like entailment or noun-hierarchy in WordNet (Yoda et al., 2024; Chheda et al., 2021). A typical procedure in the attributions field is iterative insertion or deletion of the most attributed features and simultaneous reevaluation of the predicted similarity between these perturbed inputs (Vasileiou and Eberle, 2024). If the most attributed features are indeed the most important, the prediction should change drastically upon their perturbation. Sparse representations that are induced in an unsupervised way can be assessed through topic-coherence measures (Trifonov et al., 2018).

A central challenge in evaluating explainability and interpretability is the absence of ground truth for what constitutes a correct or valid explanation. Vasileiou and Eberle (2024) address this by generating synthetic data using a rule-based similarity model. More commonly, evaluation focuses on performance trade-offs between interpretable methods and their standard counterparts. The explanation quality is often assessed qualitatively through example-based analysis. Some studies employ human evaluation to obtain subjective judgements of explanation quality. However, this introduces additional parameters to the evaluation scenario, e.g., which target group an explanation method aims at (Köhl et al., 2019).

5.2 Datasets

Datasets can serve at least *two purposes* within the realm of interpretable embeddings and semantic search explanations. The first purpose is a potential application to *evaluate a method's explanation* against human explanation. The second purpose is *global explanation* through evaluating embedding models on text pairs with controlled relation.

Human Explanations. Lopez-Gazpio et al. (2017) release the i(nterpretable)STS data set that elicits relations and similarities between individual segments of texts. Deshpande et al. (2023) propose the C(onditional)STS dataset that elicits similarity values for specific aspect of interest. The theory that underlies iSTS aligns with attribution or set-based approaches, while CSTS is motivated by a more abstract multi-aspect view akin to what is sought by feature-based explainability methods.

Interpretable Text Relations. For their analysis Zhu et al. (2018) and Zhu and de Melo (2020) construct two datasets of sentence triplets and quadruples consisting of a negative and positive pairs with a shared linguistic property. Li et al. (2025) propose a neuro-symbolic tool for automatically creating such sets and use the resulting data for ranking text embedding models in interpretable linguistic categories. The STS3k data set (Fodor et al., 2025) contains sentence pairs with systematic word combinations, rated for semantic similarity by human participants. Nastase and Merlo (2024) propose specialized sentence sets to study the grammatical information that resides in an embedding.

6 Discussion

Method trade-offs. The surveyed methods differ in their conceptualization of interpretability, computational cost, fidelity to input tokens, and dependence on specific model architectures. All variants of **interpretable embeddings** produce inherently interpretable models, offering transparency into their decision-making processes by design (Rudin, 2019). However, they often pose additional constraints on models which can lead to compromises in predictive performance. In contrast, **post-hoc methods** do not constrain models upon training but rely on additional computation, surrogate optimization and specialized datasets to generate insights.

What is the "right" explanation? Given the above trade-offs none of the presented approaches should be seen as to provide true, unique and faithful explanations (Murdoch et al., 2019). At the same time, all of them provide insights into text embedding models going beyond a black-box embedding or single scalar similarity score. Each approach may lead to hypotheses about where these models fail and how they can be improved (Wiegreffe and Pinter, 2019). Rather than competing for a single best explanation, therefore, we suggest to consider individual methods as independent pieces of evidence.

Lessons learned. Following the above argumentation, we can already draw a number of overarching conclusions about text embedding models: They encode a wide range of linguistic knowledge, including syntax and semantic information like tense (Conneau and Kiela, 2018; Huang et al., 2021), learn to match synonyms well (Moeller et al., 2024; Zhu et al., 2018) (cf. Fig. 4) and successfully ignore irrelevant parts of sentences (Nikolaev and Padó, 2023). But they often do not sufficiently account for negation or random word deletion (Weller et al., 2024; Zhu and de Melo, 2020). They largely rely on nouns and verbs (Vasileiou and Eberle, 2024; Nikolaev and Padó, 2023) as well as subjects, predicates and objects (Moeller et al., 2024). Nevertheless, they do require the full contexts of sentences for their predictions to be reliable (Moeller et al., 2023) and are sensitive to word order (Zhu et al., 2018).
With space-shaping methods, we have the ability to actively manipulate encoded information (Sun et al., 2024; Shen et al., 2023; Schopf et al., 2023b) and can e.g. correct embeddings to account for negation (Opitz and Frank, 2022). Finally, structured embeddings have proven to successfully account for non-symmetric text relations (Yoda et al., 2024; Dasgupta et al., 2020).

Upcoming challenges. As models become capable of ingesting longer context (Zhang et al., 2024; Xiong et al., 2024), we may wonder if interpretability approaches transfer to explaining the similarity of **long documents**. Fine-grained explanations such as token attributions or alignments may require an aggregation step to a sentence or paragraph level balancing higher-level interpretability and compute scaling.

The embedding research landscape has also found another recent focus in **multilinguality** (Wang et al., 2024b). It will be interesting to investigate cross-lingual text similarity and we see an interesting tension here between capturing universal and language specific or cultural patterns.

Embedding models are also increasingly used as parts in **more complex models** like Retrieval Augmented Generation (RAG, Lewis et al., 2020). The approaches presented here may be used to explain the retrieval step and interpretability approaches for generative models may be utilized to understand the compilation of responses ((Achtibat et al., 2024)). However, it is an open question how to combine explanations for the two steps.

The social-sciences and sensitive fields like legal text processing often work with text representations but come with explainability requirements. A lack of interpretability can be a reason not to use state-of-the-art approaches in these fields, but to fall back onto outdated alternatives like simple dictionary-based approaches. **Interdisciplinary efforts** should focus more on understanding and addressing these requirements.

Finally, the evident link between text similarity and embedding models motivates a closer look at **the notion of similarity**. Similarity is known to be context-dependent (Gärdenfors, 2000; Bär et al., 2011), possibly asymmetric (Tversky, 1977), and even intransitive (Lin, 1998b). However, common datasets and benchmarks assume a one-dimensional scale and unification across various tasks and objectives (Cer et al., 2017; Khashabi et al., 2021; Muennighoff et al., 2023), which has enabled scalability but may be overly simplifying. Evidence for this can be seen in the fact that successful attempts now often use instructions (Su et al., 2023) or specialized adapters (Günther et al., 2024) to condition a model for specific tasks (i.e. contexts). This calls for interpretability research to better understand relevant text aspects in different contexts.

Perspective. Despite the additional challenges associated with the explainability of similarity, this survey has shown that there have been substantial efforts towards better understanding text embeddings and text similarity models. Most of these approaches are directly applicable to the next generation embedding models derived from decoders, since after removing causal attention masking, adding pooling and contrastive training, their architecture is identical to the previous encoder-based generation. This holds true for inherently interpretable embeddings, e.g. space-shaping and also post-hoc explanations, e.g. attribution whose only requirement is for models to be differentiable. Challenges are not fundamental but on the level of implementation details and computational costs due to the larger number of parameters and embedding dimensions. We believe an increased attention towards explainability research in this area can help not only understand and explain their outputs, but also mitigate biases and errors in these models, yielding improvement towards accuracy and safety.

Limitations

Capturing the full breadth of the area of interpretable text embeddings and their similarity cost us some depth and exactness. For instance, in Section 2, we suggest to speak of *interpretable* models when their predictions are inherently understandable by humans and define *explanation* as a post-hoc process in contrast. However, throughout the paper, we also use these terms interchangeably.

While we have put a lot of effort into identifying relevant publications for the scope of this survey, there is a chance that we missed some works. Hence we suggest viewing our survey as a guide and introduction to this field that is representative, but possibly not fully exhaustive.

Finally, interpretability research tends to lag behind the rapid evolution of state-of-the-art models. Most of the approaches we survey here built up on, or were applied to now outdated text embedding models. Nevertheless, these methods are often general enough to be transferred to state-of-the-art models.

Acknowledgments

Three authors received funding through the project *Impresso – Media Monitoring of the Past II Beyond Borders: Connecting Historical Newspapers and Radio*. Impresso is a research project funded by the Swiss National Science Foundation (SNSF 213585) and the Luxembourg National Research Fund (17498891).

References

Reduan Achtibat, Sayed Mohammad Vakilzadeh Hatefi, Maximilian Dreyer, Aakriti Jain, Thomas Wiegand, Sebastian Lapuschkin, and Wojciech Samek. 2024. AttnLRP: Attention-Aware layer-wise relevance propagation for transformers. In *Forty-first International Conference on Machine Learning*.

Rawaa Alatrash, Mohamed Amine Chatti, Qurat Ul Ain, Yipeng Fang, Shoeb Joarder, and Clara Siepmann. 2024. ConceptGCN: Knowledge concept recommendation in MOOCs based on knowledge graph convolutional networks and SBERT. *Computers and Education: Artificial Intelligence*, 6:100193.

Parul Awasthy, Aashka Trivedi, Yulong Li, Mihaela Bornea, David Cox, Abraham Daniels, Martin Franz, Gabe Goodhart, Bhavani Iyer, Vishwajeet Kumar, Luis Lastras, Scott McCarley, Rudra Murthy, Vignesh P, Sara Rosenthal, Salim Roukos, Jaydeep Sen, Sukriti Sharma, Avirup Sil, Kate Soule, Arafat Sultan, and Radu Florian. 2025. Granite embedding models. *Preprint*, arXiv:2502.20204.

Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PloS one*, 10(7):e0130140.

Yang Bai, Xiaoguang Li, Gang Wang, Chaoliang Zhang, Lifeng Shang, Jun Xu, Zhaowei Wang, Fangshan Wang, and Qun Liu. 2020. Sparterm: Learning term-based sparse representation for fast text retrieval. *arXiv preprint arXiv:2010.00768*.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Daniel Bär, Torsten Zesch, and Iryna Gurevych. 2011. A reflective view on text similarity. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 515–520, Hissar, Bulgaria. Association for Computational Linguistics.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. LLM2vec: Large language models are secretly powerful text encoders. In *First Conference on Language Modeling*.

Vinamra Benara, Chandan Singh, John X Morris, Richard Antonello, Ion Stoica, Alexander G Huth, and Jianfeng Gao. 2024. Crafting interpretable embeddings by asking LLMs questions. *arXiv preprint arXiv:2405.16714*.

Jacob Bien, Florentina Bunea, and Luo Xiao. 2016. Convex banding of the covariance matrix. *Journal of the American Statistical Association*, 111(514):834–845. PMID: 28042189.

Blair Bilodeau, Natasha Jaques, Pang Wei Koh, and Been Kim. 2024. Impossibility theorems for feature attribution. *Proceedings of the National Academy of Sciences*, 121(2):e2304406120.

Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermyn, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. 2023. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Tejas Chheda, Purujit Goyal, Trang Tran, Dhruvesh Patel, Michael Boratko, Shib Sankar Dasgupta, and Andrew McCallum. 2021. Box embeddings: An open-source library for representation learning using geometric structures. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 203–211, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shamil Chollampatt, Minh Quang Pham, Sathish Reddy Indurthi, and Marco Turchi. 2025. Cross-lingual evaluation of multilingual text generation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7766–7777, Abu Dhabi, UAE. Association for Computational Linguistics.

Alexis Conneau and Douwe Kiela. 2018. SentEval: An evaluation toolkit for universal sentence representations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised learning of universal sentence representations from natural language inference data. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Zhuyun Dai and Jamie Callan. 2020. Context-aware term weighting for first stage passage retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 1533–1536, New York, NY, USA. Association for Computing Machinery.

Marina Danilevsky, Kun Qian, Ranit Aharonov, Yannis Katsis, Ban Kawas, and Prithviraj Sen. 2020. A survey of the state of explainable AI for natural language processing. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 447–459, Suzhou, China. Association for Computational Linguistics.

Shib Dasgupta, Michael Boratko, Dongxu Zhang, Luke Vilnis, Xiang Li, and Andrew McCallum. 2020. Improving local identifiability in probabilistic box embeddings. In *Advances in Neural Information Processing Systems*, volume 33, pages 182–192. Curran Associates, Inc.

Ameet Deshpande, Carlos Jimenez, Howard Chen, Vishvak Murahari, Victoria Graf, Tanmay Rajpurohit, Ashwin Kalyan, Danqi Chen, and Karthik Narasimhan. 2023. C-STS: Conditional semantic textual similarity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5669–5690, Singapore. Association for Computational Linguistics.

Oliver Eberle, Jochen Büttner, Florian Kräutli, Klaus-Robert Müller, Matteo Valleriani, and Grégoire Montavon. 2020. Building and interpreting deep similarity models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1149–1161.

EU. 2024. Article 86: Right to explanation of individual decision-making.

Manaal Faruqui, Yulia Tsvetkov, Dani Yogatama, Chris Dyer, and Noah A. Smith. 2015. Sparse overcomplete word vector representations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1491–1500, Beijing, China. Association for Computational Linguistics.

James Fodor, Simon De Deyne, and Shinsuke Suzuki. 2025. Compositionality and sentence meaning: Comparing semantic parsing and transformers on a challenging sentence similarity dataset. *Computational Linguistics*, 51(1):139–190.

Thibault Formal, Carlos Lassance, Benjamin Piwowarski, and Stéphane Clinchant. 2021a. SPLADE v2: Sparse lexical and expansion model for information retrieval. *arXiv preprint arXiv:2109.10086*.

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021b. SPLADE: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2288–2292, New York, NY, USA. Association for Computing Machinery.

Gal, Yarin and Ghahramani, Zoubin. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 1050–1059, New York, New York, USA. PMLR.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Maarten Grootendorst. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Michael Günther, Louis Milliken, Jonathan Geuter, Georgios Mastrapas, Bo Wang, and Han Xiao. 2023. Jina embeddings: A novel set of high-performance sentence embedding models. In *Proceedings of the 3rd Workshop for Natural Language Processing Open Source Software (NLP-OSS 2023)*, pages 8–18, Singapore. Association for Computational Linguistics.

Jiafeng Guo, Yixing Fan, Liang Pang, Liu Yang, Qingyao Ai, Hamed Zamani, Chen Wu, W. Bruce Croft, and Xueqi Cheng. 2020. A deep look into neural ranking models for information retrieval. *Information Processing & Management*, 57(6):102067.

Peter Gärdenfors. 2000. *Conceptual Spaces: The Geometry of Thought*. The MIT Press.

Michael Günther, Jackmin Ong, Isabelle Mohr, Alaeddine Abdessalem, Tanguy Abel, Mohammad Kalim Akram, Susana Guzman, Georgios Mastrapas, Saba Sturua, Bo Wang, Maximilian Werk, Nan Wang, and Han Xiao. 2024. Jina embeddings 2: 8192-Token general-purpose text embeddings for long documents. *Preprint*, arXiv:2310.19923.

Kailash A Hambarde and Hugo Proenca. 2023. Information retrieval: Recent advances and beyond. *IEEE Access*.

Alexander Hoyle, Rupak Sarkar, Pranav Goel, and Philip Resnik. 2023. Natural language decompositions of implicit content enable better text representations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13188–13214, Singapore. Association for Computational Linguistics.

James Y. Huang, Kuan-Hao Huang, and Kai-Wei Chang. 2021. Disentangling semantics and syntax in sentence embeddings with pre-trained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1372–1379, Online. Association for Computational Linguistics.

James Y. Huang, Wenlin Yao, Kaiqiang Song, Hongming Zhang, Muhao Chen, and Dong Yu. 2023. Bridging continuous and discrete spaces: Interpretable sentence representation learning via compositional operations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14584–14595, Singapore. Association for Computational Linguistics.

Joseph D Janizek, Pascal Sturmfels, and Su-In Lee. 2021. Explaining explanations: Axiomatic feature interactions for deep networks. *Journal of Machine Learning Research*, 22(104):1–54.

Daniel Khashabi, Amos Ng, Tushar Khot, Ashish Sabharwal, Hannaneh Hajishirzi, and Chris Callison-Burch. 2021. GooAQ: Open question answering with diverse answer types. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 421–433, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Omar Khattab and Matei Zaharia. 2020. ColBERT: Efficient and effective passage search via contextualized late interaction over BERT. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA.

Gregory Koch, Richard Zemel, and Ruslan Salakhutdinov. 2015. Siamese neural networks for one-shot image recognition. In *Proceedings of the 32nd International Conference on Machine Learning (ICML'15) Deep Learning Workshop*, Lille, France.

Weize Kong, Jeffrey M. Dudek, Cheng Li, Mingyang Zhang, and Michael Bendersky. 2023. Sparseembed: Learning sparse lexical representations with contextual embeddings for retrieval. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2399–2403, New York, NY, USA. Association for Computing Machinery.

Maximilian A. Köhl, Kevin Baum, Markus Langer, Daniel Oster, Timo Speith, and Dimitri Bohlender. 2019. Explainability as a non-functional requirement. In *2019 IEEE 27th International Requirements Engineering Conference (RE)*, pages 363–368.

Daniil Larionov, Jens Grünwald, Christoph Leiter, and Steffen Eger. 2023. EffEval: A comprehensive evaluation of efficiency for MT evaluation metrics. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 78–96, Singapore. Association for Computational Linguistics.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025a. NV-Embed: Improved techniques for training LLMs as generalist embedding models. *Preprint*, arXiv:2405.17428.

Jinhyuk Lee, Feiyang Chen, Sahil Dua, Daniel Cer, Madhuri Shanbhogue, Iftekhar Naim, Gustavo Hernández Ábrego, Zhe Li, Kaifeng Chen, Henrique Schechter Vera, Xiaoqi Ren, Shanfeng Zhang,

Daniel Salz, Michael Boratko, Jay Han, Blair Chen, Shuo Huang, Vikram Rao, Paul Suganthan, Feng Han, Andreas Doumanoglou, Nithi Gupta, Fedor Moiseev, Cathy Yip, Aashi Jain, Simon Baumgartner, Shahrokh Shahi, Frank Palma Gomez, Sandeep Mariserla, Min Choi, Parashar Shah, Sonam Goenka, Ke Chen, Ye Xia, Koert Chen, Sai Meher Karthik Duddu, Yichang Chen, Trevor Walker, Wenlei Zhou, Rakesh Ghiya, Zach Gleicher, Karan Gill, Zhe Dong, Mojtaba Seyedhosseini, Yunhsuan Sung, Raphael Hoffmann, and Tom Duerig. 2025b. Gemini embedding: Generalizable embeddings from gemini. *Preprint*, arXiv:2503.07891.

Jinhyuk Lee, Zhuyun Dai, Xiaoqi Ren, Blair Chen, Daniel Cer, Jeremy R. Cole, Kai Hui, Michael Boratko, Rajvi Kapadia, Wen Ding, Yi Luan, Sai Meher Karthik Duddu, Gustavo Hernandez Abrego, Weiqiang Shi, Nithi Gupta, Aditya Kusupati, Prateek Jain, Siddhartha Reddy Jonnalagadda, Ming-Wei Chang, and Iftekhar Naim. 2024. Gecko: Versatile text embeddings distilled from large language models. *Preprint*, arXiv:2403.20327.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Hongji Li, Andrianos Michail, Reto Gubelmann, Simon Clematide, and Juri Opitz. 2025. Sentence smith: Controllable edits for evaluating text embeddings. *EMNLP 2025*.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

Dekang Lin. 1998a. Automatic retrieval and clustering of similar words. In *36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics, Volume 2*, pages 768–774, Montreal, Quebec, Canada. Association for Computational Linguistics.

Dekang Lin. 1998b. An information-theoretic definition of similarity. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, page 296–304, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Tian Yu Liu and Stefano Soatto. 2024. Conjuring semantic similarity. *arXiv preprint arXiv:2410.16431*.

Tian Yu Liu, Matthew Trager, Alessandro Achille, Pramuditha Perera, Luca Zancato, and Stefano Soatto. 2024. Meaning representations from trajectories in autoregressive models. In *The Twelfth International Conference on Learning Representations*.

I. Lopez-Gazpio, M. Maritxalar, A. Gonzalez-Agirre, G. Rigau, L. Uria, and E. Agirre. 2017. Interpretable semantic textual similarity: Finding and explaining differences between sentences. *Knowledge-Based Systems*, 119:186–199.

Lucas Moeller, Dmitry Nikolaev, and Sebastian Padó. 2023. An attribution method for Siamese encoders. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15818–15827, Singapore. Association for Computational Linguistics.

Lucas Moeller, Dmitry Nikolaev, and Sebastian Padó. 2024. Approximate attributions for off-the-shelf Siamese transformers. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2059–2071, St. Julian's, Malta. Association for Computational Linguistics.

Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. 2019. Layer-wise relevance propagation: An overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 193–209.

Niklas Muennighoff. 2022. SGPT: GPT sentence embeddings for semantic search. *arXiv preprint arXiv:2202.08904*.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. MTEB: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2014–2037, Dubrovnik, Croatia. Association for Computational Linguistics.

W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. 2019. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080.

Vivi Nastase and Paola Merlo. 2024. Tracking linguistic information in transformer-based sentence embeddings through targeted sparsification. In *Proceedings of the 9th Workshop on Representation Learning for NLP (RepL4NLP-2024)*, pages 203–214, Bangkok, Thailand. Association for Computational Linguistics.

Dmitry Nikolaev and Sebastian Padó. 2023. Investigating semantic subspaces of transformer sentence embeddings through linear structural probing. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 142–154, Singapore. Association for Computational Linguistics.

Dmitry Nikolaev and Sebastian Padó. 2023. Representation biases in sentence transformers. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 3701–3716, Dubrovnik, Croatia. Association for Computational Linguistics.

Charles O'Neill, Christine Ye, Kartheik Iyer, and John F Wu. 2024. Disentangling dense embeddings with sparse autoencoders. *arXiv preprint arXiv:2408.00657*.

Juri Opitz. 2023. SMATCH++: Standardized and extended evaluation of semantic graphs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1595–1607, Dubrovnik, Croatia. Association for Computational Linguistics.

Juri Opitz and Anette Frank. 2022. SBERT studies meaning representations: Decomposing sentence embeddings into explainable semantic features. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 625–638, Online only. Association for Computational Linguistics.

Malte Ostendorff, Till Blume, Terry Ruas, Bela Gipp, and Georg Rehm. 2022. Specialized document embeddings for aspect-based similarity of research papers. In *Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries*, JCDL '22, New York, NY, USA.

Moritz Plenz, Juri Opitz, Philipp Heinisch, Philipp Cimiano, and Anette Frank. 2023. Similarity-weighted construction of contextualized commonsense knowledge graphs for knowledge-intense argumentation tasks. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6130–6158, Toronto, Canada. Association for Computational Linguistics.

Wuttikorn Ponwitayarat, Peerat Limkonchotiwat, Ekapol Chuangsuwanich, and Sarana Nutanong. 2024. Space decomposition for sentence embedding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11227–11239, Bangkok, Thailand. Association for Computational Linguistics.

Martin Potthast, Benno Stein, and Maik Anderka. 2008. A wikipedia-based multilingual retrieval model. In *Advances in Information Retrieval*, pages 522–530, Berlin, Heidelberg. Springer Berlin Heidelberg.

Victor Prokhorov, Yingzhen Li, Ehsan Shareghi, and Nigel Collier. 2021. Learning sparse sentence encoding without supervision: An exploration of sparsity in variational autoencoders. In *Proceedings of the 6th Workshop on Representation Learning for NLP (RepL4NLP-2021)*, pages 34–46, Online. Association for Computational Linguistics.

Shauli Ravfogel, Valentina Pyatkin, Amir David Nissan Cohen, Avshalom Manevich, and Yoav Goldberg. 2024. Description-based text similarity. In *First Conference on Language Modeling*.

Nils Reimers and Iryna Gurevych. 2019. 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)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Julian Risch, Philipp Hager, and Ralf Krestel. 2021. Multifaceted domain-specific document embeddings. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 78–83, Online. Association for Computational Linguistics.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215.

Ananya B. Sai, Akash Kumar Mohankumar, and Mitesh M. Khapra. 2022. A survey of evaluation metrics used for NLG systems. *ACM Comput. Surv.*, 55(2).

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and efficient retrieval via lightweight late interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3715–3734, Seattle, United States. Association for Computational Linguistics.

Tim Schopf, Daniel Braun, and Florian Matthes. 2023a. Evaluating unsupervised text classification: Zero-shot and similarity-based approaches. In *Proceedings of the 2022 6th International Conference on Natural Language Processing and Information Retrieval*, NLPIR '22, page 6–15, New York, NY, USA. Association for Computing Machinery.

Tim Schopf, Emanuel Gerber, Malte Ostendorff, and Florian Matthes. 2023b. AspectCSE: Sentence embeddings for aspect-based semantic textual similarity using contrastive learning and structured knowledge. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 1054–1065, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Jaejin Seo, Sangwon Lee, Ling Liu, and Wonik Choi. 2022. TA-SBERT: Token attention Sentence-BERT for improving sentence representation. *IEEE Access*, 10:39119–39128.

Lingfeng Shen, Haiyun Jiang, Lemao Liu, and Shuming Shi. 2023. Sen2Pro: A probabilistic perspective to sentence embedding from pre-trained language model. In *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, pages 315–333, Toronto, Canada. Association for Computational Linguistics.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021. Score-based generative modeling through

stochastic differential equations. In *International Conference on Learning Representations*.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. 2023. One embedder, any task: Instruction-Finetuned text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1102–1121, Toronto, Canada. Association for Computational Linguistics.

Yiqun Sun, Qiang Huang, Yixuan Tang, Anthony KH Tung, and Jun Yu. 2024. A general framework for producing interpretable semantic text embeddings. *arXiv preprint arXiv:2410.03435*.

Mukund Sundararajan, Kedar Dhamdhere, and Ashish Agarwal. 2020. The shapley taylor interaction index. In *International conference on machine learning*, pages 9259–9268. PMLR.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Matthieu Tehenan, Vikram Natarajan, Jonathan Michala, Milton Lin, and Juri Opitz. 2025. Mechanistic decomposition of sentence representations. *arXiv preprint arXiv:2506.04373*.

Valentin Trifonov, Octavian-Eugen Ganea, Anna Potapenko, and Thomas Hofmann. 2018. Learning and evaluating sparse interpretable sentence embeddings. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 200–210, Brussels, Belgium. Association for Computational Linguistics.

Stephan Tulkens and Thomas van Dongen. 2024. Model2vec: The fastest state-of-the-art static embeddings in the world. *GitHub Repositories*.

Amos Tversky. 1977. Features of similarity. *Psychological review*, 84(4):327–352.

Jannis Vamvas and Rico Sennrich. 2023. Towards unsupervised recognition of token-level semantic differences in related documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 13543–13552, Singapore. Association for Computational Linguistics.

Alexandros Vasileiou and Oliver Eberle. 2024. Explaining text similarity in transformer models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7859–7873, Mexico City, Mexico. Association for Computational Linguistics.

Bin Wang and C.-C. Jay Kuo. 2020. SBERT-WK: A sentence embedding method by dissecting BERT-Based word models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2146–2157.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024a. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, Bangkok, Thailand. Association for Computational Linguistics.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024b. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Yile Wang, Zhanyu Shen, and Hui Huang. 2025. LDIR: Low-Dimensional dense and interpretable text embeddings with relative representations. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14397–14409, Vienna, Austria. Association for Computational Linguistics.

Orion Weller, Dawn Lawrie, and Benjamin Van Durme. 2024. NevIR: Negation in neural information retrieval. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2274–2287, St. Julian's, Malta. Association for Computational Linguistics.

Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. 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)*, pages 11–20, Hong Kong, China. Association for Computational Linguistics.

Wenhan Xiong, Jingyu Liu, Igor Molybog, Hejia Zhang, Prajjwal Bhargava, Rui Hou, Louis Martin, Rashi Rungta, Karthik Abinav Sankararaman, Barlas Oguz, Madian Khabsa, Han Fang, Yashar Mehdad, Sharan Narang, Kshitiz Malik, Angela Fan, Shruti Bhosale, Sergey Edunov, Mike Lewis, Sinong Wang, and Hao Ma. 2024. Effective long-context scaling of foundation models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4643–4663, Mexico City, Mexico. Association for Computational Linguistics.

Xin Ye, Hui Shen, Xiao Ma, Razvan Bunescu, and Chang Liu. 2016. From word embeddings to document similarities for improved information retrieval in software engineering. In *Proceedings of the 38th International Conference on Software Engineering*, ICSE '16, page 404–415, New York, NY, USA. Association for Computing Machinery.

Shohei Yoda, Hayato Tsukagoshi, Ryohei Sasano, and Koichi Takeda. 2024. Sentence representations via

Gaussian embedding. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 418–425, St. Julian's, Malta. Association for Computational Linguistics.

Xin Zhang, Yanzhao Zhang, Dingkun Long, Wen Xie, Ziqi Dai, Jialong Tang, Huan Lin, Baosong Yang, Pengjun Xie, Fei Huang, Meishan Zhang, Wenjie Li, and Min Zhang. 2024. mGTE: Generalized long-context text representation and reranking models for multilingual text retrieval. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1393–1412, Miami, Florida, US. Association for Computational Linguistics.

Tiancheng Zhao, Xiaopeng Lu, and Kyusong Lee. 2021. SPARTA: Efficient open-domain question answering via sparse transformer matching retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 565–575, Online. Association for Computational Linguistics.

Xunjie Zhu and Gerard de Melo. 2020. Sentence analogies: Linguistic regularities in sentence embeddings. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3389–3400, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Xunjie Zhu, Tingfeng Li, and Gerard de Melo. 2018. Exploring semantic properties of sentence embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 632–637, Melbourne, Australia. Association for Computational Linguistics.

A Overview

Table 1: A summary of our taxonomy with links to respective sections, publications and their code if available. The table is split into the two families of methods that are further divided into *Types*, corresponding to subsections and *Subtypes* corresponding to paragraphs in the main text. *Train* is whether the method requires training and *Approx.* refers to whether a method approximates the similarity score of a reference embedding model. When code is labeled as 'NA', this means that we were not able to find a public code repository.

Type	Subtype	Paper	Train	Approx.	code
Interpretable Embeddings (§3)					
space-shaping (§3.1)	QA-features	Sun et al. (2024)	yes	no	github
		Benara et al. (2024)	yes	no	github
	sub-embedding	Opitz and Frank (2022)	yes	yes	github
		Risch et al. (2021)	yes	no	github
		Ostendorff et al. (2022)	yes	no	github
		Schopf et al. (2023b)	yes	no	NA
		Ponwitayarat et al. (2024)	yes	no	github
	anchors	Potthast et al. (2008)	yes	no	NA
		Wang et al. (2025)	no	no	github
sparsification (§3.2)	unsupervised	Trifonov et al. (2018)	yes	yes	NA
		Prokhorov et al. (2021)	yes	yes	github
	lexical	Dai and Callan (2020)	yes	no	github
		Bai et al. (2020)	yes	no	NA
		Zhao et al. (2021)	yes	no	NA
		Formal et al. (2021b,a)	yes	no	github
structured objects (§3.3)	box embeddings	Chheda et al. (2021)	yes	no	github
		Dasgupta et al. (2020)	yes	no	github
	gaussian embeddings	Shen et al. (2023)	no	yes	NA
		Yoda et al. (2024)	yes	no	github
	operator learning	Huang et al. (2023)	yes	yes	github
set-based (§3.4)	token-weights	Wang and Kuo (2020)	no	no	github
		Seo et al. (2022)	yes	no	NA
		Tulkens and van Dongen (2024)	no	yes	github
	sequential	Khattab and Zaharia (2020)	yes	yes	github
		Santhanam et al. (2022)	no	yes	github
	multi-view	Hoyle et al. (2023)	no	no	github
		Ravfogel et al. (2024)	no	no	github
		Liu et al. (2024)	no	no	github
	image-set	Liu and Soatto (2024)	yes	no	NA
Post-hoc Explanation (§4)					
attribution (§4.1)	integrated Jacobians	Moeller et al. (2023, 2024)	no	yes	github
	relevance propagation	Vasileiou and Eberle (2024)	no	yes	github
global explanation (§4.2)	text relations	Zhu et al. (2018)	no	no	NA
		Zhu and de Melo (2020)	no	no	NA
surrogate modeling (§4.3)	interpretable approximation	Nikolaev and Padó (2023)	yes	yes	NA
	probing	Conneau et al. (2017)	yes	no	github
		Conneau et al. (2018)	yes	no	github
		Nikolaev and Padó (2023)	yes	yes	github
		Tehenan et al. (2025)	yes	yes	github