Explaining Text Similarity in Transformer Models

Alexandros Vasileiou[1]   **Oliver Eberle**[1,2]

[1]Machine Learning Group, Technische Universität Berlin, Berlin, Germany
[2]BIFOLD – Berlin Institute for the Foundations of Learning and Data, Berlin, Germany

`vasileiou.a@gmx.de`   `oliver.eberle@tu-berlin.de`

Abstract

As Transformers have become state-of-the-art models for natural language processing (NLP) tasks, the need to understand and explain their predictions is increasingly apparent. Especially in unsupervised applications, such as information retrieval tasks, similarity models built on top of foundation model representations have been widely applied. However, their inner prediction mechanisms have mostly remained opaque. Recent advances in explainable AI have made it possible to mitigate these limitations by leveraging improved explanations for Transformers through layer-wise relevance propagation (LRP). Using BiLRP, an extension developed for computing second-order explanations in bilinear similarity models, we investigate which feature interactions drive similarity in NLP models. We validate the resulting explanations and demonstrate their utility in three corpus-level use cases, analyzing grammatical interactions, multilingual semantics, and biomedical text retrieval. Our findings contribute to a deeper understanding of different semantic similarity tasks and models, highlighting how novel explainable AI methods enable in-depth analyses and corpus-level insights.

1 Introduction

Modern foundation models provide flexible text representations that enable the detection of semantic structure in vast amounts of unlabeled data. While traditionally supervised settings have taken a predominant role for NLP research, many widely used unsupervised tasks have received growing attention. The similarity structure of text embeddings herein provides a central starting point for many tasks, including information retrieval, semantic search, ranking and knowledge extraction (Frakes and Baeza-Yates, 1992; Mooney and Bunescu, 2005; Jansen and Rieh, 2010), clustering (Aggarwal and Zhai, 2012), and visualization (van der Maaten and Hinton, 2008; Venna et al., 2010; McInnes et al., 2018). In the context of language generation, the retrieval-augmented generation (RAG) (Lewis et al., 2020) approach computes semantic closeness to index relevant text data, resulting in numerous new information retrieval systems (Gao et al., 2023). Furthermore, as the number of models and tasks increases, the need for quantitative evaluation of embeddings has become apparent, leading to the introduction of embedding benchmarks (Thakur et al., 2021; Muennighoff et al., 2022).

Complementary to the evaluation of nominal performance, the field of explainable AI aims to provide insights about the inner model mechanisms by highlighting relevant features for a specific prediction (Montavon et al., 2018; Danilevsky et al., 2020; Vilone and Longo, 2021; Samek et al., 2021). Explanations play a crucial role in verifying that predictions are grounded in task-relevant features, fostering trust and verifiability into complex machine learning models, and enabling the discovery of data patterns and novel insights. In this context, the prediction of supervised models can often be explained using heatmaps over input features. Beyond heatmaps, specific explanation approaches have been proposed in the context of unsupervised models (Montavon et al., 2022; Kauffmann et al., 2022) and higher-order explanations (Eberle et al., 2022; Schnake et al., 2022; Fumagalli et al., 2023). In this paper, we focus on Transformer-based similarity models that motivate the use of second-order attributions to highlight feature interactions. Our main contributions are as follows:

- We analyze Transformer-based similarity models within the framework of second-order explanations using BiLRP, highlighting the interaction between tokens.
- We evaluate explanations through a purposely designed similarity task for which ground truth interactions are available, and via input perturbations on real-world semantic similarity data.

- We investigate the interaction of relevant tokens across three use cases, revealing the compositional structure that drives high/low similarity, thus providing fine-grained insights into sentence representations that are the fundamental concept in many NLP applications but have not been analyzed in the context of explainable AI.
- We perform corpus-level analyses, identifying the parts of speech that models prioritize and illustrating how simple token-matching strategies can lead to inaccurate predictions.

Our implementation is publicly available.[1]

2 Related Work

Semantic Textual Similarity The task of identifying the degree of semantic equivalence between texts is referred to as semantic textual similarity (STS) (Lee et al., 2005; Agirre et al., 2012; Chandrasekaran and Mago, 2021). While two words may be semantically related, e.g. 'coffee' and 'mug', the STS task focuses on the semantic closeness between text, and in this sense 'tea' is considered more similar to 'coffee' than 'mug' (Chandrasekaran and Mago, 2021).

Similarity Models for NLP Similarity models are designed to capture the meaning and context of the input texts and provide a quantitative measure of their semantic similarity or relatedness. Commonly used approaches include end-to-end-trained Siamese networks (Bromley et al., 1993; Hu et al., 2014; Neculoiu et al., 2016) and universal encoder models with a pooling layer to extract text summary embeddings (Cer et al., 2018). Combined with Transformers, these approaches have become powerful frameworks to build similarity models. Sentence Transformers, specifically Sentence-BERT (SBERT), (Reimers and Gurevych, 2019) have emerged as a flexible and widely-used method to compute compact text representations. Similarity models typically process pairs of inputs to compute a similarity score, often based on dot products such as cosine similarity, while extensions to dataset-level dot products have enhanced semantic sentence matching (Zhong et al., 2020) and guided document retrieval in RAG systems (Lewis et al., 2020). Recent efforts in evaluating similarity models and text representations include benchmarks on information retrieval (Thakur et al.,

2021) and comparing performance across diverse embedding tasks (Muennighoff et al., 2022).

Explaining Transformers The limitations of raw attention scores as explanations have emphasized the need for improved explanation methods for Transformers (Jain and Wallace, 2019; Serrano and Smith, 2019). Notably, aggregating relevant information across attention heads has proven to be a promising direction (Abnar and Zuidema, 2020; Chefer et al., 2021a), with further empirical evidence supporting the benefits of gradient information (Wallace et al., 2019; Atanasova et al., 2020; Chefer et al., 2021b). The naive computation of explanatory gradients in Transformers could be further improved by considering the conservation of relevance during backpropagation, which results in a modified layer-wise relevance propagation (LRP) scheme for Transformers (Ali et al., 2022).

Interpretable Feature Interaction Several works have focused on determining the effect of joint features in supervised classification scenarios. To identify such pair-wise attributions, multivariate statistics (Bien et al., 2013; Caruana et al., 2015), Hessian-based approaches (Janizek et al., 2021), as well as methods inspired by co-operative game theory (Tsang et al., 2020; Dhamdhere et al., 2020; Fumagalli et al., 2023), have been proposed. Beyond classification, the interaction between features provides an appropriate level of complexity to assess why a pair of inputs produces a high or low similarity score. To compute joint feature relevance, Hessian $\times$ Product and BiLRP have been proposed (Eberle et al., 2022). These methods can be derived directly from a deep Taylor decomposition (Montavon et al., 2017) of similarity models and can be seen as extensions of the widely used Gradient $\times$ Input and LRP explanation methods to bilinear models.

3 Explainable AI for Similarity Models

In the following section, we briefly outline how the specific structure of deep similarity models motivates the consideration of second-order terms, as well as the need for tailored propagation rules to compute robust and accurate explanations for Transformers.

3.1 Explainable AI for Similarity Models

Starting from a Taylor expansion of the similarity score $y(\boldsymbol{x}, \boldsymbol{x}') = \langle \phi(\boldsymbol{x}), \phi(\boldsymbol{x}') \rangle$ with a pair of inputs $(\boldsymbol{x}, \boldsymbol{x}')$, a feature map $\phi : \mathbb{R}^{s \times d} \to \mathbb{R}^h$ and y

---

the predicted dot product similarity, the following description of relevance $R_{ii'}$ assigned to a pair of features (i, i') can be derived (Eberle et al., 2022):

$$R_{ii'} = x_i x'_{i'} \left[\nabla^2 y(\boldsymbol{x}, \boldsymbol{x}')\right]_{ii'},$$

where ∇^2 denotes the Hessian containing second-order partial derivatives.

For deep neural networks, these derivatives have been found to be noisy (Balduzzi et al., 2017; Montavon et al., 2018), which motivates the formulation of robustified propagation rules within the LRP framework, resulting in the BiLRP method, which can be computed in the following factored form:

$$\text{BiLRP}(y, \boldsymbol{x}, \boldsymbol{x}') = \sum_{m=1}^{h} \text{LRP}([\phi_L \circ \cdots \circ \phi_1]_m, \boldsymbol{x}) \\ \otimes \text{LRP}([\phi_L \circ \cdots \circ \phi_1]_m, \boldsymbol{x}'),$$

where ϕ_L is an intermediate feature map at layer L and $\otimes$ refers to the tensor product between LRP relevance matrices. The LRP attribution for neuron j in layer l is computed by pooling over all incoming messages from neurons k in the higher layer $l + 1$:

$$R_j^{(l)} = \sum_k \frac{q_{jk}}{\sum_j q_{jk}} \cdot R_k^{(l+1)},$$

with contributions q_{jk} of neuron j to relevance $R_k^{(l+1)}$. Depending on the type of network layer, different propagation rules have been proposed to compute q_{jk}, typically selected to be proportional to the observed neuron activations (Montavon et al., 2019). In Section 3.2, we introduce specific propagation rules for Transformer models.

The resulting BiLRP explanations assign relevance scores $R_{ii'}$ to each interaction between input features (x_i, x'_i), highlighting in a detailed manner how features interact to produce the similarity prediction. To compute the interactions $R_{ii'}$, one backpropagation pass for each embedding dimension h is required, resulting in $2 \times h$ computations for a pair of sentences that can be computed efficiently using automatic differentiation software. For BiLRP, this results in computing multiple LRP explanations and, thus, its robustness is directly related to the reliable propagation of relevance.

3.2 Explainable AI for Transformers

To compute better explanations for Transformers, leveraging gradient information has proven to be effective. Yet, the non-linear structure of Transformers motivates specific gradient propagation rules to

reflect the model prediction more reliably, resulting in more faithful explanations (Ali et al., 2022). The application of these rules does not affect the model's forward predictions but only modifies the gradient computations in the backward pass.

Propagation rules The bilinear structure of the query-key-value (QKV) self-attention layers and the layer normalization computations, lead to a break in relevance conservation, which can be addressed using specific propagation rules. For the **attention head**, the forward pass can be formulated as $y_j = \sum_i h_i[p_{ij}]_{\texttt{.detach()}}$, viewing the attention scores p_{ij} as a weighting matrix for the current residual stream representation h_i and detaching the associated variable p_{ij} from the computation graph (Ali et al., 2022). To preserve relevance in **layer normalization**, the denominator is regarded as a normalization constant resulting in $y_i = (h_i - \mathbb{E}[h]) / \left[\sqrt{\epsilon + \text{Var}[h]}\right]_{\texttt{.detach()}}$, with expectation $\mathbb{E}$, variance Var and stabilization parameter ϵ (Ali et al., 2022). In addition, specific non-linear activation functions like GeLU break conservation of relevance, which can be addressed by attributing relevance in proportion to the computed activations (Eberle, 2022). For all other layers, the LRP-0 propagation rule is applied, redistributing relevance in proportion to the contribution of each input neuron to the computed activation (Montavon et al., 2019). These resulting propagation rules also match related approaches to explain bilinear LSTM gating (Arras et al., 2017), and to linearize Transformers (Elhage et al., 2021).

4 Experiments

We now introduce the semantic similarity datasets and models used in this paper before proceeding with the evaluation of second-order explanations.

4.1 Data

STSb (Semantic Textual Similarity benchmark) consists of English text in the form of sentence pairs, extracted from image captions, news headlines and user forums (Cer et al., 2017). A ground truth similarity score was assigned to each text pair as the result of a human annotation process. In addition, the multilingual STSb (mSTSb) dataset (May, 2021) provides machine translated sentence pairs in ten languages. **SICK** (Sentences Involving Compositional Knowledge) contains sentence pairs with human-annotated similarity scores (Marelli et al., 2014), aiming to assess the performance of

models in understanding sentence meaning, compositionality, and related tasks such as textual entailment. **BIOSSES** consists of a total of one hundred sentence pairs selected from the 'TAC2 Biomedical Summarization Track Training Data Set' (Soğancıoğlu et al., 2017). Each sentence pair was assigned a similarity score based on annotations of five domain experts for the evaluation of biomedical text comparison tasks.

4.2 Similarity Models

In our experiments, we consider four Transformer-based architectures and three pooling strategies described in the following.

Transformers The widely-used **BERT** model, initially trained on English text, demonstrates strong performance across various natural language processing tasks (Devlin et al., 2019). Additionally, it has been extended to 104 languages, resulting in **mBERT** for cross-lingual applications. Specifically trained on the task of semantic similarity, the **SBERT** framework provides models that enable the efficient calculation of semantic relatedness, useful for tasks like information retrieval and knowledge extraction (Reimers and Gurevych, 2019). The **SGPT** (Muennighoff, 2022) model is built on top of a finetuned GPT-Neo model (Black et al., 2021), which is an open source variant of the popular GPT-family and similar to GPT-3. Additional model details are provided in Appendix A.

Pooling To summarize token embeddings into fixed size representations, we use the following pooling strategies: **CLS-Pooling** uses the CLS token as a fixed-size representation for tasks like text classification in many Transformer models. **Mean-Pooling** averages over token embeddings to obtain one single representation vector, capturing the overall information of the input sequence. **QKV-Pooling** uses the QKV-mechanism to compute weighting coefficients before aggregating the token embeddings.

4.3 Evaluation of Explanations

Interaction Analysis The complex structure of the similarity task and a lack of fine-grained ground truth rationales of feature interaction motivates the evaluation of explanations on a specifically designed similarity task. We design a similarity model based on co-occurrence statistics of features, which here are interactions between same noun tokens (see Figure 1, top row). We finetune a similar-

ity model using a BERT-base encoder with a QKV-pooling layer to correctly predict the number of co-occurring proper nouns and nouns in the STSb dataset. After optimization of the mean squared error (MSE) loss between true and predicted scores, the similarity model is able to correctly predict the number of interactions (Spearman's $\rho = 0.94/0.89$, $\text{MSE} = 0.21/0.87$ for train/test). To verify that predictions are built from the expected interactions, we compute the similarity between ground-truth interactions and compare them to the second-order explanations extracted from (i) token embeddings, (ii) Hessian $\times$ Product (H$\times$P), and (iii) BiLRP.

In Figure 1, we observe that computing interactions directly from token embeddings results in pairwise attributions mainly between same tokens that are not selective with respect to their assigned part-of-speech (POS) tag. For H$\times$P, the interactions are much more selective with regard to nouns and proper nouns, and we observe considerable token interactions that are assigned negative relevance. In the case of many interacting tokens that drive high similarity (Figure 1, right column), we observe that it becomes increasingly difficult to identify the relevant interactions. BiLRP is able to select the relevant tokens in comparison to the other baselines with higher accuracy, which is supported by the highest average cosine similarity (ACS) between true interactions and BiLRP of 0.81 in comparison to 0.62 for H$\times$P, and 0.67 for the embedding baseline.

In summary, our self-designed similarity task can be accurately explained using BiLRP, allowing a better understanding of the model's strategy of filtering out task-irrelevant parts of speech to solve the noun-matching task.

Perturbation Analysis We further test the ability of explanations to faithfully capture the similarity prediction process on real-world semantic similarity sentence pairs. Sequence elements are ordered based on sum-pooled interaction scores from highest to lowest relevance and elements are added iteratively to the selected input sequence. At each step, we compute the Euclidean distance between the perturbed and the unperturbed sentence representation, measuring how strongly the representation is affected in response to the removal of the next most relevant tokens. Resulting perturbation curves are shown in Figure 2 comparing different explanation methods and a random baseline. We observe that across models and datasets, BiLRP consistently se-
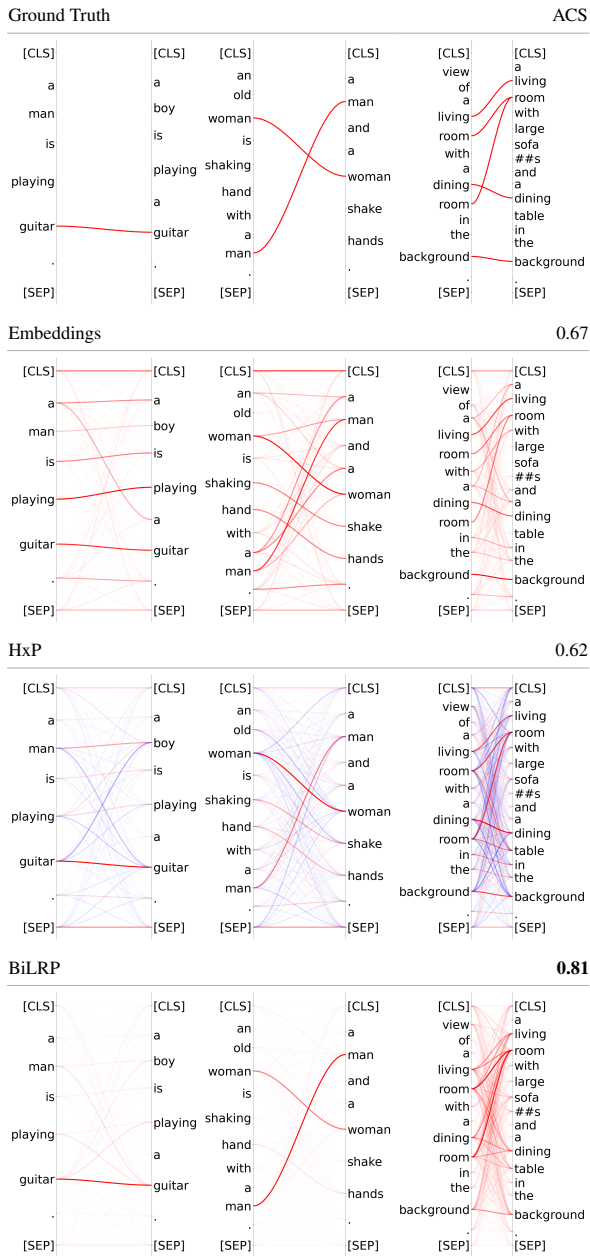
Figure 1: Comparison of different explanation techniques that highlight the interaction between input features. Ground truth interactions (top row) are the interactions between same noun tokens. These are compared to second-order explanations built on top of BERT token embeddings, Hessian×Product (H×P) and BiLRP. Average cosine similarity (ACS) is used to measure agreement between ground truth and explanations.

lects the features that decrease the distance between sentence representations most effectively, resulting in the lowest area under the perturbation curve (see Appendix B). The steep initial decline in Euclidean distance for BiLRP highlights that a small subset of highly relevant features are identified reliably. These findings further accentuate the differences between explanation methods, emphasizing the ef-
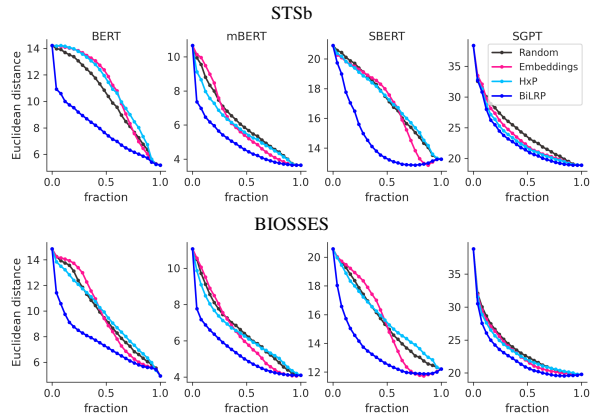

Figure 2: Perturbation experiment comparing different explanation methods across models. Fractions of tokens, ranked from most to least relevant, are added to one input sequence and the resulting Euclidean distance to the unperturbed sentence is measured. A steep initial decline with a smaller area under the curve indicates better identification of task-relevant features.

fectiveness of BiLRP in identifying the relevant interactions between tokens.

Conservation The main axiomatic principle used to develop improved BiLRP explanations for Transformers is conservation of relevance. The sum of observed relevance is directly related to the explained model predictions, and explanations that are relevance conserving have been shown to result in improved explanations (Ali et al., 2022), as also confirmed in our experiments presented in Figures 1 and 2. Implementing these rules to compute second-order explanations (see Section 3.2) reconstitutes relevance conservation in BiLRP in comparison to H×P as shown in Figure 6 in Appendix C. A residual lack of conservation can be explained by neuron biases, which are unattributable (Figure 6, rightmost panel).

Overall, these experiments have confirmed that BiLRP explanations can accurately identify the most relevant feature interactions, leading us next to an evaluation of the textual similarity prediction task itself.

4.4 Predicting Semantic Textual Similarity

Next, we focus on the task of predicting semantic textual similarity and compare different pre-trained similarity models (BERT, mBERT, SGPT, SBERT) across three datasets (STSb, SICK, BIOSSES). We evaluate model predictions using Spearman correlation with the goal to identify performance differences between models that inform more focused analyses using explainable AI.

Model	Spearman correlation		
	STSb	SICK	BIOSSES
BERT + CLS	20.3	42.4	63.7
BERT + Mean Pooling	47.3	58.2	54.6
mBERT + Mean Pooling	55.2	56.3	55.6
SGPT + Mean Pooling	76.9	73.4	69.6
SBERT + Mean Pooling	84.7	78.4	66.7

Table 1: Spearman correlation $\rho \times 100$ for a set of tasks and models. The first three models were not finetuned, while SGPT and SBERT were finetuned on semantic similarity data.

As shown in Table 1, the standard BERT similarity model, when used with CLS or mean pooling methods, fails to effectively capture semantic proximity. The Spearman correlation $\rho \times 100$ for the CLS-Pooling is 20.3, whereas a significantly improved correlation is achieved when using mean pooling across all encoded token representations for STSb and SICK data. Similarity is best predicted by SGPT and SBERT with scores ranging from 66.7 to 84.7, highlighting overall the considerable impact of model selection, pooling strategy and dataset on task performance.

5 Corpus-Level Use Cases

To go beyond the mere evaluation of nominal correlation, which is at risk of obfuscating undesired model strategies, i.e. Clever-Hans-type and shortcut learning (Lapuschkin et al., 2019; Geirhos et al., 2020), we hereby explore how BiLRP explanations can be used to uncover general model strategies in three distinct use cases.

5.1 Explaining Semantic Textual Similarity

To conduct an explanation-based analysis of semantic similarity, we retrieve explanations for all 1379 samples of the STSb test set. Token-to-token interactions are summarized by extracting POS tags using *spaCy 3* and aligning different tokenizers using *tokenizations* [2].

The corpus-level analysis is performed by aggregating all relevance scores per token pair. Relevance for each interaction between POS-tags is aggregated and scores are normalized by the maximum absolute value of total summed relevance, which results in a relevance scoring over interactions, as shown in Figure 3. For each interaction, we distinguish between relevance patterns that negatively (blue triangle) or positively (red triangle) contribute to the similarity score.

[2] https://github.com/explosion/tokenizations

For the BERT + CLS similarity model with lowest correlation scores as shown in Section 4.4, we identify that the most positively relevant interactions are 'NOUN-NOUN', 'NOUN-VERB' and 'NOUN-[SEP]' (Figure 3-a). While negative contributions are less pronounced, we observe some amount of negatively contributing 'NOUN-[CLS]' interaction. For mean pooling shown in Figure 3-b, we observe the strongest positive effects for '[SEP]-[SEP]', followed by 'NOUN-NOUN' and 'NOUN-VERB' interactions. Overall, the distribution of relevance is more concentrated over a smaller set of POS interactions. The SBERT model shows a distribution comparable to BERT + CLS, focusing on a similar subset of interactions (Figure 3-c), though assigning different importance to them. Specifically, 'NOUN-NOUN', 'NOUN-[SEP]' and 'NOUN-VERB' are most relevant. In contrast to BERT with mean pooling, the '[SEP]-[SEP]' interaction is significantly less pronounced. SBERT also attributes considerable amount of relevance to 'VERB-VERB' and 'NOUN-ADJECTIVE' interactions, which overall suggests that the semantic similarity task can be solved quite well using a small but well-chosen subset of POS interactions.

Our analyses further provide insights regarding the choice of pooling strategy, supporting previous findings that are in favor of mean pooling over CLS pooling (Mohebbi et al., 2021) and underscoring the usefulness of explanations on the level of interactions to identify differing strategies across models.

5.2 Explaining Multilingual Similarity

Multilingual language models enable the flexible use of embeddings for unsupervised downstream tasks across different languages. In this section, we consider a setting in which a ranking of most similar texts in a multilingual database is required. We use the mSTSb dataset and compute similarity for mBERT and a set of monolingual BERT models for English, German, Russian and Chinese, assuming that no fine-tuning on semantic similarity is performed. The text representations are extracted by mean pooling.

Results We first analyze the correlation to ground truth similarity scores and focus on nominal prediction differences across settings, guiding the selection of an interesting case study for our subsequent BiLRP analysis. As shown in Figure 4-a, we observe that Spearman correlation scores are
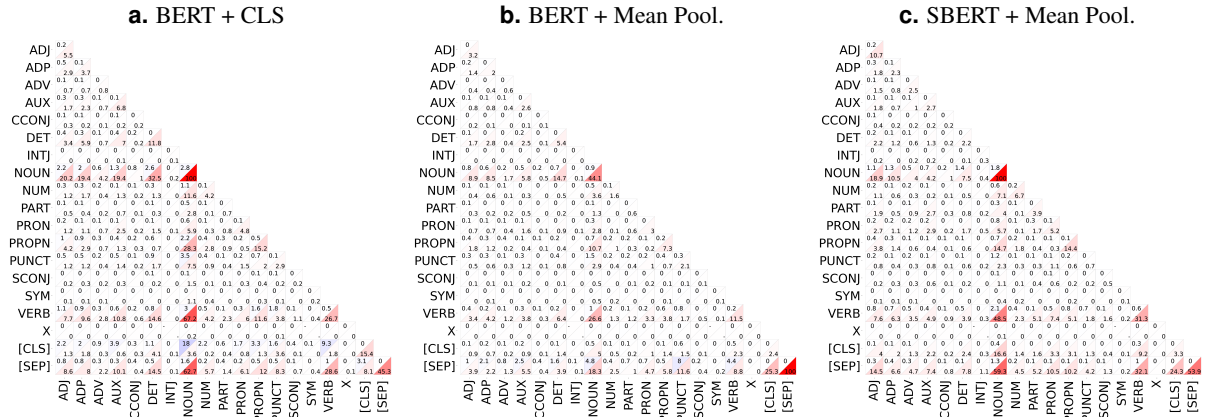
Figure 3: Corpus-level analysis of BiLRP explanations between POS tags on the STSb dataset. The contribution of positive/negative interactions to the similarity score is shown in red/blue for three similarity models, ranging from (**a**) the least predictive (BERT + CLS), to (**b**) moderately predictive (BERT + Mean Pooling), to (**c**) the most predictive (SBERT) (cf. Table 1).

consistently between 47.3 and 58.5 across the four aforementioned languages, reaching similar correlation levels as observed in the previous section for not finetuned similarity models (cf. Table 1). Interestingly, for the mixed-multilingual case, we observe a clear drop of correlation scores to 24.8–35.5. To uncover some of the effects that drive this decrease in performance, we next conduct a case study on the English and German subsets of the mSTSb corpus.

In Figure 4-b, our initial step involves explaining a sentence pair using BiLRP within monolingual settings (EN-EN, DE-DE) and comparing it to the mixed setting (EN-DE). For English, we see how semantic similarity is attributed to the interaction of 'eating', whereas the German translation 'frisst' is considered less relevant. Instead, we find that semantic similarity in the German setting is more often attributed to the interaction between determiners ('eine-eine'), which may reflect the specific role of determiners in the German language that both quantify and determine an object (Dipper, 2005). We provide additional samples that illustrate several cases of relevant interactions in Appendix D. These include the mismatching of different quantities ('two-three'), effects of matching subtokens that affect the semantic meaning ('key-##board' and 'keyboard'), and overall overconfidence of the model in assigning high similarity based on semantically related tokens ('train-waiting', 'clothing-shirt').

We take a closer look at the aggregated most relevant POS interactions with the goal to identify differences in model strategies across settings. Specif-

ically, we analyze which POS interactions change the most from the monolingual to the multilingual setting and show the ten interactions of largest difference in accumulated positive relevance assigned to a specific POS interaction in Figure 4-c. We additionally show negative changes in relevance in Figure 7 in Appendix D. We find that both 'NOUN-NOUN' and 'VERB-VERB' interactions are less relevant in the mixed setting when compared to the monolingual English setting, suggesting that the model is less able to match English to German nouns and verbs respectively. Furthermore, we find that the monolingual German similarity is driven by a considerable amount of interaction between determiner tags ('DET-DET') that are less present in the mixed case. Lastly, we underline the difference in assigned relevance to the interaction between nouns and proper nouns ('NOUN-PROPN'). We hypothesize that, while multilingual models learn a flexible joint embedding space across many languages and vocabularies, it may be difficult to accurately model subtle differences in semantic meaning that monolingual models are able to capture more precisely.

These detailed insights, both at the single-sample and corpus levels, can unveil the diverse strategies that result in accurate or inaccurate predictions of similarity, potentially informing the future design of similarity tasks and models across multilingual settings.
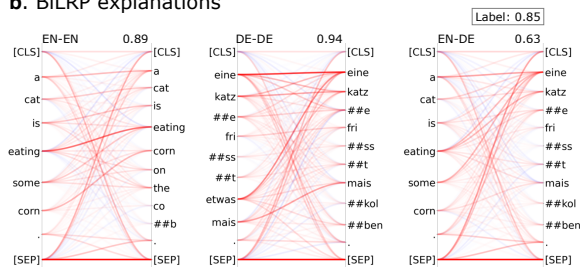
5.3 Analyzing Model Differences

In the next use case, we analyze what drives performance differences observed earlier in Table 1 on the biomedical text pairs contained in BIOSSES
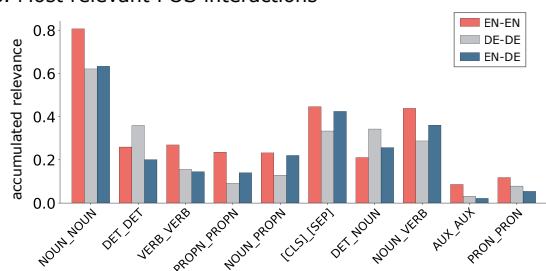
Figure 4: Comparison of mono- and multilingual BERT-based similarity models on mSTSb. (**a**) Spearman correlation $\rho \times 100$ of the multilingual STSb corpus. Similarity models are build from monolingual (mono) and multilingual (multi) BERT models that receive monolingual input, and a multilingual model that receives mixed input in English and a translated version of the other sentence (mix-multi). (**b**) BiLRP explanations on mBERT for English-English (left), German-German (center) and English-German (right). The sentence pair is assigned a true similarity score of 0.85. (**c**) Comparison of positively relevant POS interactions. POS tags are selected based on largest difference of accumulated relevance between the mixed and the monolingual settings.

(Soğancıoğlu et al., 2017). In a first step, we analyze the predicted similarity scores in Figure 5 (top) for SGPT, SBERT and BERT as introduced in Section 4.4. We clearly see how the error between ground truth and predicted similarity decreases for higher levels of true similarity. All three models assign correct levels of similarity for high similarity but are less capable of correctly capturing low similarity.

In the second step, we compute explanations and select the 25% data quantiles of highest and lowest predicted similarity. For each group, we rank the most relevant interactions and show the top-5 for each model in Figure 5 (bottom). For high similarity, we observe that both SBERT and

BERT base their similarity predictions primarily on matching of same tokens or variants thereof, e.g. 'leukemia-leukemia' and 'cancer-cancers', while SGPT explanations identify interactions with gene regulating molecules ('miR-126-reports' and 'miR-223-regulates'), and interactions to more functional token pairs ('regulated-of') as most relevant. Similarly, for low similarity, SGPT explanations assign high relevance to molecule-specific interactions ('dependent-miR-133b' and 'KRAS-driven') and more general biomedical vocabulary ('oxidative–downward'). SBERT and BERT both show a comparable matching pattern ('also-augmented', 'mutations-found'), but as for high similarity also match closely related biomedical tokens ('nucleotides-rna', 'tumor-cancer').

This suggests that SGPT has better abilities to match task-relevant biomedical vocabulary, specifically names of gene molecules and tokens that are descriptive of gene expression processes, while SBERT und BERT base their similarity predictions on more general medical terminology. These strategies work well for highly similar sentence pairs for which matching may be sufficient, but are less suitable to correctly assess low similarity that may require more complex strategies e.g. the identification of negations or counterfactuals.
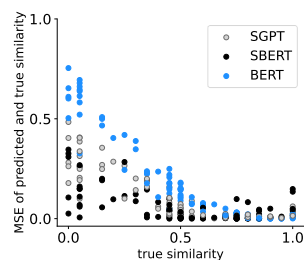
	SGPT	SBERT	BERT (mean)
high	miR-126–reports	up–up	tumor–tumor
	miR-223–human	leukemia–leukemia	leukemia–leukemia
	miR-223–regulates	mutual–mutually	cancers–cancers
	regulated–of	also–up	cancer–cancer
	of–cervical	cancer–cancer	tissues–cancer
low	dependent–miR-133b	tumors–tumors	mutant–mutations
	KRAS–driven	nucleotides–rna	dependent–mutations
	oncogenic–NSCLCs	research–research	oncogenic–mutations
	oxidative–downward	tumor–cancer	mutations–found
	combined–viability	also–augmented	vivo–tumors

Figure 5: Analysis of semantic similarity on the BIOSSES dataset containing biomedical text. Mean squared error (MSE) between predicted and true similarity for SGPT, SBERT and BERT similarity model is shown (top). Top-5 most relevant token interactions are shown for high and low similarity levels (bottom).

6 Discussion and Conclusion

Our evaluation and use cases have provided a framework to analyze textual similarity models using explainable AI. In the following, we contextualize our findings and discuss further implications.

Structured Explanations We have seen how novel types of explanations can be used to explain bilinear models in the context of semantic textual similarity. Resulting explanations highlight pairs of features that are most or least relevant to produce a particular similarity score. These second-order attributions provide an appropriate level of complexity and detail for investigating similarity model predictions that go beyond heatmap representations. In our experiments, BiLRP identifies feature interactions in Transformers more accurately than the token embedding baseline and the Hessian-based H×P explanations. This has allowed us to analyze the unsupervised semantic similarity task on the level of fine-grained interactions. Our results have highlighted interpretable interactions of tokens and POS tags that can next be utilized to inform and guide the development of similarity tasks and models. In particular, leveraging generative approaches may further inform the development of interaction-based explanation techniques.

Strategies of Similarity Models This approach enables insights into the internal computation of sentence representations in Transformers, exposing unexpected strategies employed for predicting semantic similarity. In particular, we have observed that high relevance in non-finetuned similarity models is often assigned to interactions between same tokens, revealing a rather simple token matching strategy. Our corpus-level POS analysis has indicated that semantic similarity can be approached by a small subset of interactions, most importantly interactions between nouns or proper nouns, verbs, and nouns and verbs, which promotes the notion of similarity as a 'bag of interactions'. We observe high relevance in specific token interactions (CLS and SEP), consistent with prior research (Clark et al., 2019; Bondarenko et al., 2023). In extension, this may result in unintended matching of tokens that is semantically not meaningful. Our experiments suggest that fine-tuning on similarity tasks and selecting an appropriate pooling strategy can partially alleviate these effects. While our focus has been on models predicting similarity, these structures are also commonly used as part of internal

computations, such as matching key and value representations, which could also be analyzed using interactions.

Conclusion In many language applications the computation of similarity is a central concept to match textual information. Here, we have shown how predictions of widely used deep similarity models can be analyzed by assigning relevance to the interaction of features. These explanations have offered novel ways to perform corpus-wide analyses ranging from information retrieval, to multilingual text matching, and the estimation of similarity in biomedical text data.

Considering the fast-growing number of applications based on foundation models, we believe that explainable AI has a critical role in ensuring their safe, robust and compliant use. AI safety is especially critical in high-risk domains including legal, financial, medical, or governmental applications of machine learning for which text data plays a central role, accentuating the necessity for rigorous system evaluations and comprehensive explanations. For complex tasks like semantic similarity, datasets that contain detailed structured rationales on the level of interactions are needed to improve model alignment with human expectations.

Limitations

Methods In this paper, we have focused on using post-hoc explanations and in particular gradient-based explanations, aiming for the most faithful explanations as identified by our evaluation. Achieving high faithfulness and meeting the conservation principle, requires implementation of gradient propagation rules, especially for a diverse range of Transformer architectures. Computing second-order explanations requires to compute as many backward passes as there are sentence embedding dimensions for each sentence in a pair. For the here considered embedding dimension of 768, this required around two minutes computation time on a 12GB P100 GPU. To accelerate computations, explanations can be computed in batch, provided that the memory requirements are met. The factorization of BiLRP enables the reuse of computed explanations when the same sentence or pair reoccurs in subsequent instances.

Evaluation Regarding our evaluation, we have adapted standard perturbation experiments designed for evaluating first-order heatmaps to

second-order explanations. We did this by initially sum-pooling over one of the sentences, performing perturbations on the corresponding input sequence, and subsequently repeating the analysis for the other sentence. This allowed us to measure how well the identified explanations are able to affect the closeness of the sentence representations, measuring faithfulness of our explanations with respect to the predictions. To evaluate the ability of second-order explanations in detecting relevant interactions, we have designed a synthetic experiment. While this does not cover the semantic complexity of real-world semantics, we consider it an important step to motivate the collection of structured rationales for complex language tasks.

Analyses   We have focused our analyses on a few selected datasets but expect that our insights, i.e. regarding token matching, apply to the task of semantic similarity more broadly across different models and corpora. Analysing the effect of POS tags removes complex grammatical structures, e.g. verbal or gerundial nouns, which would furthermore complicate our multilingual analyses due to diverging grammatical rules and a varying number of categories. Thus, we consider the focus on POS-level granularity to be an appropriate first step at an intermediate level of detail to investigate the task of semantic similarity at a corpus scale.

Ethics Statement

We do not anticipate any harmful risks in using the methods and analysis used in this work.

Acknowledgements

References

Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4190–4197, Online. Association for Computational Linguistics.

Charu C Aggarwal and ChengXiang Zhai. 2012. A survey of text clustering algorithms. *Mining text data*, pages 77–128.

Eneko Agirre, Mona Diab, Daniel Cer, and Aitor Gonzalez-Agirre. 2012. Semeval-2012 task 6: A pilot on semantic textual similarity. SemEval '12, page 385–393, USA. Association for Computational Linguistics.

Ameen Ali, Thomas Schnake, Oliver Eberle, Grégoire Montavon, Klaus-Robert Müller, and Lior Wolf. 2022. XAI for transformers: Better explanations through conservative propagation. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 435–451. PMLR.

Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2017. Explaining recurrent neural network predictions in sentiment analysis. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 159–168, Copenhagen, Denmark. Association for Computational Linguistics.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A diagnostic study of explainability techniques for text classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

David Balduzzi, Marcus Frean, Lennox Leary, J. P. Lewis, Kurt Wan-Duo Ma, and Brian McWilliams. 2017. The shattered gradients problem: If resnets are the answer, then what is the question? In *ICML*, volume 70 of *Proceedings of Machine Learning Research*, pages 342–350. PMLR.

Jacob Bien, Jonathan Taylor, and Robert Tibshirani. 2013. A lasso for hierarchical interactions. *The Annals of Statistics*, 41(3):1111 – 1141.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow.

Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. 2023. Quantizable transformers: Removing outliers by helping attention heads do nothing. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1993. Signature verification using a "siamese" time delay neural network. In *Advances in Neural Information Processing Systems*, volume 6. Morgan-Kaufmann.

Rich Caruana, Yin Lou, Johannes Gehrke, Paul Koch, M. Sturm, and Noémie Elhadad. 2015. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Daniel Cer, Mona Diab, Eneko Agirre, Inigo 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)*. Association for Computational Linguistics.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St. John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, Brian Strope, and Ray Kurzweil. 2018. Universal sentence encoder for English. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 169–174, Brussels, Belgium. Association for Computational Linguistics.

Dhivya Chandrasekaran and Vijay Mago. 2021. Evolution of semantic similarity—a survey. *ACM Comput. Surv.*, 54(2).

Hila Chefer, Shir Gur, and Lior Wolf. 2021a. Generic attention-model explainability for interpreting bimodal and encoder-decoder transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 397–406.

Hila Chefer, Shir Gur, and Lior Wolf. 2021b. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 782–791.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019*, pages 4171–4186.

Kedar Dhamdhere, Ashish Agarwal, and Mukund Sundararajan. 2020. The shapley taylor interaction index. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Stefanie Dipper. 2005. German quantifiers: Determiners or adjectives. In *Proceedings of the LFG05 Conference*, pages 100–115.

Oliver Eberle. 2022. *Explainable structured machine learning*. Ph.D. thesis, Technische Universität Berlin.

Oliver Eberle, Jochen Büttner, Florian Kräutli, Klaus-Robert Müller, Matteo Valleriani, and Grégoire Montavon. 2022. Building and interpreting deep similarity models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(3):1149–1161.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*. Https://transformer-circuits.pub/2021/framework/index.html.

William B. Frakes and Ricardo A. Baeza-Yates, editors. 1992. *Information Retrieval: Data Structures & Algorithms*. Prentice-Hall.

Fabian Fumagalli, Maximilian Muschalik, Patrick Kolpaczki, Eyke Hüllermeier, and Barbara Eva Hammer. 2023. SHAP-IQ: Unified approximation of any-order shapley interactions. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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*.

R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673.

Baotian Hu, Zhengdong Lu, Hang Li, and Qingcai Chen. 2014. Convolutional neural network architectures for matching natural language sentences. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 2042–2050, Cambridge, MA, USA. MIT Press.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556, Minneapolis, Minnesota. 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.

Bernard J. Jansen and Soo Young Rieh. 2010. The seventeen theoretical constructs of information searching and information retrieval. *Journal of the American Society for Information Science and Technology*, 61(8):1517–1534.

Jacob Kauffmann, Malte Esders, Lukas Ruff, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2022. From clustering to cluster explanations via neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–15.

Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2019. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature Communications*, 10:1096.

Michael D Lee, Brandon Pincombe, and Matthew Welsh. 2005. An empirical evaluation of models of text document similarity. In *Proceedings of the annual meeting of the cognitive science society*, volume 27.

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 *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A SICK cure for the evaluation of compositional distributional semantic models. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland. European Language Resources Association (ELRA).

Philip May. 2021. Machine translated multilingual sts benchmark dataset.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Großberger. 2018. UMAP: Uniform manifold approximation and projection. *Journal of Open Source Software*, 3(29):861.

Hosein Mohebbi, Ali Modarressi, and Mohammad Taher Pilehvar. 2021. Exploring the role of BERT token representations to explain sentence probing results. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 792–806, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Grégoire Montavon, Sebastian Bach, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. 2017. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recognition*, 65:211–222.

Grégoire Montavon, Alexander Binder, Sebastian Lapuschkin, Wojciech Samek, and Klaus-Robert Müller. 2019. Layer-wise relevance propagation: An overview. In *Explainable AI*, volume 11700 of *Lecture Notes in Computer Science*, pages 193–209. Springer.

Grégoire Montavon, Jacob Kauffmann, Wojciech Samek, and Klaus-Robert Müller. 2022. *Explaining the Predictions of Unsupervised Learning Models*, pages 117–138. Springer International Publishing, Cham.

Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2018. Methods for interpreting and understanding deep neural networks. *Digit. Signal Process.*, 73:1–15.

Raymond J. Mooney and Razvan Bunescu. 2005. Mining knowledge from text using information extraction. *SIGKDD Explor. Newsl.*, 7(1):3–10.

Niklas Muennighoff. 2022. SGPT: GPT sentence embeddings for semantic search.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. MTEB: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

Paul Neculoiu, Maarten Versteegh, and Mihai Rotaru. 2016. Learning text similarity with Siamese recurrent networks. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 148–157, Berlin, Germany. Association for Computational Linguistics.

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*. Association for Computational Linguistics.

Wojciech Samek, Grégoire Montavon, Sebastian Lapuschkin, Christopher J. Anders, and Klaus-Robert Müller. 2021. Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, 109(3):247–278.

Thomas Schnake, Oliver Eberle, Jonas Lederer, Shinichi Nakajima, Kristof T Schütt, Klaus-Robert Müller, and Grégoire Montavon. 2022. Higher-order explanations of graph neural networks via relevant walks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(11):7581–7596.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Gizem Soğancıoğlu, Hakime Öztürk, and Arzucan Özgür. 2017. BIOSSES: a semantic sentence similarity estimation system for the biomedical domain. *Bioinformatics*, 33(14):i49–i58.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Michael Tsang, Sirisha Rambhatla, and Yan Liu. 2020. How does this interaction affect me? Interpretable attribution for feature interactions. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA. Curran Associates Inc.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Jarkko Venna, Jaakko Peltonen, Kristian Nybo, Helena Aidos, and Samuel Kaski. 2010. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *Journal of Machine Learning Research*, 11(13):451–490.

Giulia Vilone and Luca Longo. 2021. Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76:89–106.

Eric Wallace, Jens Tuyls, Junlin Wang, Sanjay Subramanian, Matt Gardner, and Sameer Singh. 2019. AllenNLP interpret: A framework for explaining predictions of NLP models. 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): System Demonstrations*, pages 7–12, Hong Kong, China. Association for Computational Linguistics.

Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu, and Xuanjing Huang. 2020. Extractive summarization as text matching. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6197–6208, Online. Association for Computational Linguistics.

A Model details

We provide details regarding model type, the Hugging Face identifier and the implemented propagation rules for each model in Table 2.

B Perturbation Details

To compute the perturbation evaluation, for each input pair, we order sequence elements based on sum-pooled relevance interaction scores from highest to lowest relevance and add elements iteratively to the selected input sequence. Tokens are added in steps of fractions of 0.04 until the original sentence is fully recovered. The initial empty sequence is initialized from the masking token for BERT-based

models and 'Ġ' for SGPT since no specific masking or unknown special tokens are reserved by the tokenizer. Perturbation curves are computed once for each sentence in a pair and the resulting scores are averaged over all samples (1379 pairs of STSb and 100 pairs for SICK respectively). In Table 3 we report area under perturbation curve (AUPC) scores for the evaluation presented in the main paper Section 4.3. We further test if BiLRP AUPC scores are significantly lower than any of the other comparison methods and find this to be the case with $p \leq 0.05$ for all investigated models and datasets.

C Conservation

Conservation of relevance is one important desired principle of post-hoc explanations. It ensures that relevance can not be created or disappear during the explanation process of the prediction. In Figure 6, we show conservation of H×P, BiLRP and BiLRP with biases set to zero for the SBERT model. We observe that for H×P, the sum of predicted sentence embedding activations ϕ_m is not related to the sum of relevance scores at the input. For BiLRP, we observe a linear relationship of both quantities. The difference to full conservation (identity line) can be explained by layer biases that are not attributable by design.
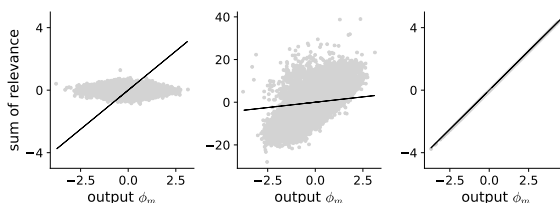

Figure 6: Conservation across 300 samples of the STSb test split on the SBERT model. Left: H×P baseline. Center: BiLRP computation. Right: BiLRP computation with model biases set to zero.

D Multilingual experiments

In Figure 8, we provide additional examples of token interactions in the multilingual settings considered in Section 5.2 of the main paper.

In analogy to the positively contributing POS interactions discussed in Section 5.2, we show POS interactions that contribute negatively in Figure 7. Relevance scores are normalized by the maximum absolute value. Overall, we find that negative contributions are less strong with relevance magnitudes below 0.1 in comparison to the positive interactions

model	name	propagation rules
BERT	bert-base-uncased	
mBERT	bert-base-multilinugal-uncased	AH, LN, GA
SBERT	sentence-transformers/stsb-bert-base	AH, LN, GA
SGPT	Muennighoff/SGPT-125M-mean-nli-bitfit	AH, LN
German BERT	bert-base-german-cased	
Russian BERT	DeepPavlov/bert-base-bg-cs-pl-ru-cased	AH, LN, GA
Chinese BERT	bert-base-chinese	

Table 2: Used models and their Hugging Face identifier names alongside the used propagation rules. Attention Head (AH), Layer Normalization (LN) and GeLU Activation (GA) rules.

	Random	Embed.	HxP	BiLRP
	STSb (N=1379)			
BERT	10.28±1.64	10.86±1.15	10.97±1.50	7.80±0.94
mBERT	6.16±0.78	5.98±0.57	5.82±0.93	4.92±0.67
SBERT	17.19±2.60	16.94±2.26	17.20±2.58	14.52±3.41
SGPT	24.41±4.42	23.65±4.78	23.22±4.79	22.70±4.83
	BIOSSES (N=100)			
BERT	9.63±1.35	9.66±0.99	9.72±1.41	7.55±0.82
mBERT	6.58±0.81	6.38±0.54	6.42±1.06	5.34±0.51
SBERT	15.59±1.54	15.43±1.36	15.77±1.52	13.36±1.73
SGPT	23.67±2.62	23.29±2.64	23.30±2.78	22.28±2.78

Table 3: Faithfulness analysis of explanation methods. AUPC (area under perturbation curve) on the STSb and BIOSSES semantic similarity datasets. Lower scores indicate better identification of features that decrease the Euclidean distance most effectively.

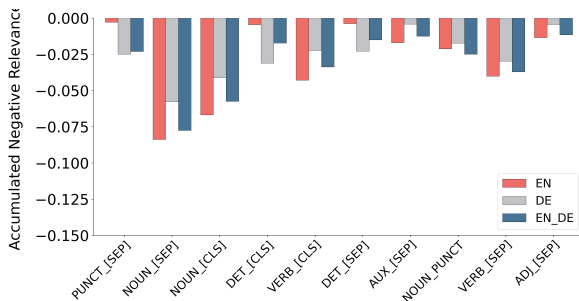that reach up to 0.8, as shown in Figure 4 in the main paper.

Figure 7: Comparison of relevant interaction between POS tags in a multilingual semantic similarity task. POS tags are selected based on largest difference of the negative accumulated relevance assigned to an interaction of POS tags between the mixed (EN-DE) setting and the monolingual (DE-DE, EN-EN) settings.

E   Additional POS Heatmaps

In addition to the POS heatmaps in the main text, we provide the corpus-level analysis normalized by the number of POS-interaction occurrences. For this, we separate positive and negative relevance scores, retrieve the mean relevance of each token pair (instead of the sum as in the main text), normalize the scores by dividing by the maximum absolute mean value. The resulting heatmaps present a complementary view to Figure 3 in the main paper,

highlighting rare POS interactions. We observe that for BERT + Mean Pooling, the effect of the interactions between '[SEP]' tokens is more apparent, since it consistently ranks across the most relevant POS interaction. For example, BERT + CLS reacts strongly to interactions between an interjection and a symbol ('nope-/'), while for SBERT we observe interactions between interjections to be of high relevance ('yes-yes' and 'no-no'). Cases like this highlight the need for well-tuned models that base their similarity predictions on desired and plausible interactions.

Figure 8: Additional BiLRP explanations on mBERT for English, German and the mixed English-German samples. The samples are chosen as representatives of the model's matching strategies, depicting different similarity levels of the labels, similarity predictions, and cases where the two diverge the most.
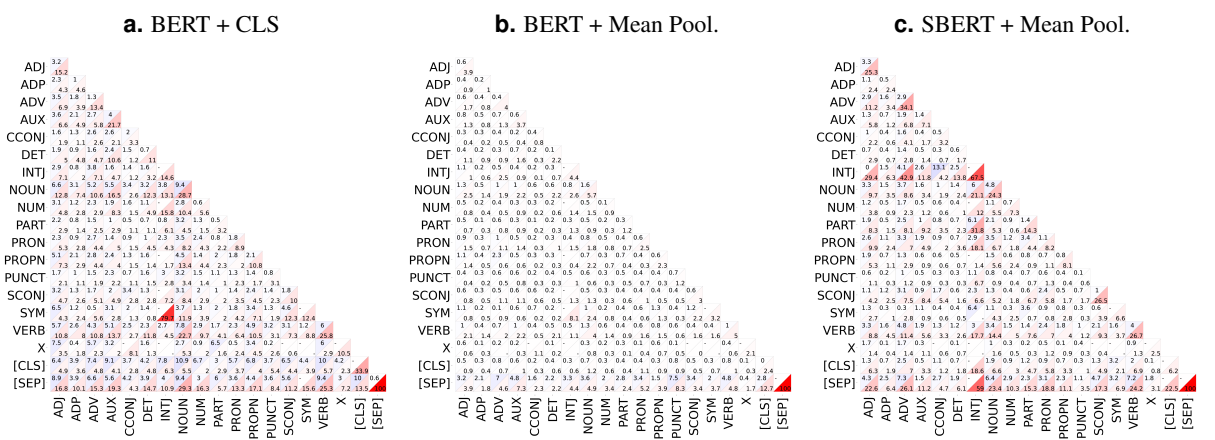

Figure 9: Alternative computation of the relevance scores, where the mean relevance of each token pair is retrieved instead of sum-pooling of relevance scores. Models range from (**a**) the least predictive (BERT + CLS), to (**b**) moderately predictive (BERT + Mean Pooling), to (**c**) the most predictive (SBERT).