Asymmetric feature interaction for interpreting model predictions

Xiaolei Lu[1]   **Jianghong Ma** [*2]   **Haode Zhang**[3]

Department of Electrical Engineering, City University of Hong Kong, Hong Kong S.A.R.
School of Computer Science and Technology,Harbin Institute of Technology,Shenzhen,China[2]
Department of Computing, The Hong Kong Polytechnic University, Hong Kong S.A.R.[3]
`xiaoleilu2-c@my.cityu.edu.hk majianghong@hit.edu.cn haode.zhang@connect.polyu.hk`

Abstract

In natural language processing (NLP), deep neural networks (DNNs) could model complex interactions between context and have achieved impressive results on a range of NLP tasks. Prior works on feature interaction attribution mainly focus on studying symmetric interaction that only explains the additional influence of a set of words in combination, which fails to capture asymmetric influence that contributes to model prediction. In this work, we propose an asymmetric feature interaction attribution explanation model that aims to explore asymmetric higher-order feature interactions in the inference of deep neural NLP models. By representing our explanation with an directed interaction graph, we experimentally demonstrate interpretability of the graph to discover asymmetric feature interactions. Experimental results on two sentiment classification datasets show the superiority of our model against the state-of-the-art feature interaction attribution methods in identifying influential features for model predictions.

1 Introduction

Deep neural networks (DNNs) have demonstrated impressive results on a range of Natural Language Processing (NLP) tasks. Unlike traditional models (e.g. CRFs and HMMs) that optimize weights on human interpretable features, deep neural models operate like a black box by applying multiple layers of non-linear transformation on the vector representations of text data, which fails to provide insights to understand the inference process of deep neural models over the features (e.g. words and phrases) involved in modeling.

Interpreting the prediction of a black box model could help understand model inference behaviors and increase user trust in applying the model to real-world applications. Prior efforts in NLP mainly focus on quantifying the contributions of individual

_____
* Corresponding author.

word or word interactions to the prediction. Fig.1 demonstrates word-level and pairwise word interaction explanations for a sentiment classification task, where the word "not" and the interaction between "not" and "funny" contribute positively to the prediction **Negative**.
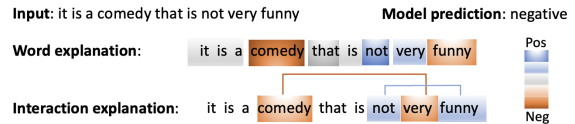

Figure 1: Explanations for a negative movie review (computed by Shapley value and Shapley interaction index), where the color indicates contribution of the corresponding word/pairwise word interaction to the model prediction.
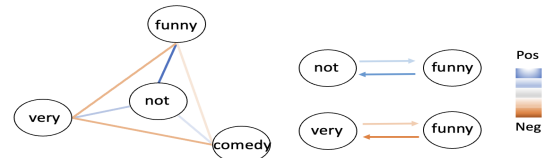

Figure 2: Symmetric versus asymmetric pairwise interaction (computed by our method) where the directed edge $a \rightarrow b$ refers to in the presence of a how much contribution of b made to the model prediction. The presence of "very" does not influence "funny" much while "funny" further modifies "very" and thus the interaction influence of "funny" $\rightarrow$ "very" is stronger than that of "very" $\rightarrow$ "funny".

Studying word interaction could help identify to what extent a set of words exert influence in combination as opposed to independently. However, most interaction attribution methods assume symmetric interaction, which may fail to capture asymmetric influence that contributes to model prediction. Fig. 2 presents some symmetric pairwise interactions with graph representation for the instance in Fig. 1, where words becomes nodes and edges

between words represent interaction. In individual-level explanation "funny" has negative influence while the symmetric interaction between "funny" and "not" produces positive influence to model prediction. Therefore the influence of the presence of "not" to "funny" is not the same as that of the presence of "funny" to "not". "funny" has weak positive contribution in the presence of "not" and "funny" further modifies "not", the interaction influence of "funny" $\rightarrow$ "not" could be stronger than the interaction of "not" $\rightarrow$ "funny". For the ideal asymmetric interaction, the presence of other features should not negate the positive influence of the important feature and other features would lose their influence when important features are present. Constructing asymmetric interaction graph for the predicted instance could help human have a nuanced understanding toward the inference of deep NLP models.

In this paper, our work aims to provide the explanation that incorporates asymmetric feature interaction [1]. The contributions are summarized as follows:

- We propose an asymmetric feature interaction attribution method that incorporates asymmetric higher-order feature interactions toward explaining the prediction of deep neural NLP models.

- We investigate three different sampling strategies in NLP field for computing marginal contribution of our defined asymmetric feature interaction attribution score, and empirically show that none is generally better than the others or more broadly applicable.

- We evaluate the proposed model on two sentiment classification datasets with BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019), and the experimental results demonstrate the faithfulness of our explanation model.

2 Related work

2.1 Feature attribution explanation

Most explanation methods mainly focus on model-agnostic explanations and study how to effectively measure the importance of features on the prediction. For example, LIME (Zhang et al., 2019) evaluates the contribution of each feature by learning

a linear model locally around a instance. Shapley value (Shapley, 1997; Lundberg and Lee, 2017) estimates the influence of a feature by averaging its marginal contribution among all permutations. Since the computation of Shapley value is computationally expensive, popular variants like Kernel SHAP (Lundberg and Lee, 2017) and Quasi-random and adaptive sampling (Štrumbelj and Kononenko, 2014) are proposed to efficiently approximate Shapley value. However, these methods do not explain how feature interactions contribute to model predictions, which fails to address model's learning capability from high-order feature interactions.

2.2 Feature interaction explanation

There are increasingly research on studying feature interaction explanation methods. For example, Shapley interaction index (Grabisch, 1997) and Shapley Taylor interaction index (Dhamdhere et al., 2019) are proposed to measure the interaction between multiple players. Integrated Directional Gradients (IDG) (Sikdar et al., 2021) borrowed axioms from Integrated Gradients (IG) (Sundararajan et al., 2017), where the desirable characteristics are satisfied by IG and Shapley value in cooperative game theory. Tsang et al. (2020) proposed an efficient framework Archipelago to combine feature interaction detector (ArchDetect) and feature attribution measure (ArchAttribute).

Further, feature interaction could be explained in a hierarchical structure. Agglomerative contextual decomposition (ACD) (Singh et al., 2018) builds the hierarchical explanations in a bottom-up way by starting with individual features and iteratively combining them based on the generalized CD scores. Jin et al. (2019) addressed context independent importance that is ignored in ACD, and proposed an easy and model-agnostic Sampling and Occlusion (SOC) algorithm to incorporate conditional context information given the specified text sequence. HEDGE (Chen et al., 2020) designs a top-down framework to construct hierarchical explanations by detecting the weakest interaction point and selecting important sub-span for a given text span.

The above feature interaction explanation methods only focus on symmetric interaction. We note that concurrent work by Masoomi et al. (2021) addressed directed pairwise interaction. However, their Shapley value based formulation also intro-

duces noisy interaction as different subsets may contain several same elements. Moreover, this solution ignores asymmetric high-order interaction.

3 Asymmetric Shapley interaction value

This section first revisits Shapley value for explaining model prediction, then describes the definition of asymmetric Shapley interaction value and the corresponding approximating computation.

The following notations will be used throughout the paper. For a classification task, given a text sequence $\boldsymbol{x} = (x_1, ..., x_n)$ and a trained model f, $\hat{y}$ is the prediction label and $f(\cdot)$ denotes the model output probability on $\hat{y}$.

3.1 Shapley value for model interpretability

In cooperative game theory Shapley value measures the marginal contribution that a player makes upon joining the group by averaging over all possible permutations of players in the group.

The Shapley value of i_{th} word in $\boldsymbol{x}$ for the model prediction $\hat{y}$ is weighted and summed over all possible word combinations:

$$\phi(i) = \sum_{S \subseteq N \backslash i} \frac{(n-1-|S|)!|S|!}{n!} \left[v(S \cup i) - v(S) \right], \tag{1}$$

where S is the subset of feature indices. The equivalent formulation is

$$\phi(i) = \frac{1}{n!} \sum_{\mathcal{O} \in \pi(n)} v(\text{pre}^i(\mathcal{O}) \cup i) - v(\text{pre}^i(\mathcal{O})), \tag{2}$$

where $\pi(n)$ denotes all permutation of the word indexes $\{1, 2, ..., n\}$. $\text{pre}^i(\mathcal{O})$ is the set of all indices that precede i in the permutation $\mathcal{O} \in \pi(n)$.

$v(S)$ is the value function that characterizes the contribution of the subset S to the prediction $\hat{y}$:

$$v(S) = \mathbb{E}(f|\boldsymbol{x}_S \cup \boldsymbol{x}'_{\bar{S}}) - \mathbb{E}(f|\boldsymbol{x}'), \tag{3}$$

where $\boldsymbol{x}'$ denotes the text sequence with the same length as $\boldsymbol{x}$, $\bar{S} \subseteq N \setminus S$. $\mathbb{E}(f|\boldsymbol{x}_S \cup \boldsymbol{x}'_{\bar{S}})$ is the expectation of $f(\cdot)$ over possible $\boldsymbol{x}'$ where only the subset values $\boldsymbol{x}_S$ unchanged.

3.2 Definition of Asymmetric Shapley interaction value

The Shapley value above can quantify the contribution of a single word or phrase to the model prediction. The proposed asymmetric Shapley interaction value (ASIV) measures the asymmetric

interaction between two different subsets T_1 and T_2 that attributes to the model prediction. That is, ASIV determines the contribution of T_1 conditioned on the presence of T_2 to the prediction $\hat{y}$. By treating T_1 and T_2 as two singletons among the players, ASIV is defined as

$$\phi_{T_2}(T_1) = C_1 \sum_{T2 \subseteq S \subseteq N \backslash \{T_1\}} \Delta_{T_1} v(S) - \Delta_{T_1} v(S \backslash T_2), \tag{4}$$

where $C_1 = \frac{(n-|T_1|-|T_2|+1-|S|)!|S|!}{(n-|T_1|-|T_2|+2)!}$. $\Delta_{T_2} v(S)$ and $\Delta_{T_1}(S)$ are given as

$$\Delta_{T_1} v(S) = v(S \cup T_1) - v(S), \tag{5}$$

$$\Delta_{T_1} v(S \backslash T_2) = v(S \backslash T_2 \cup T_1) - v(S \backslash T_2). \tag{6}$$

$\Delta_{T_1} v(S) - \Delta_{T_1} v(S \backslash T_2)$ computes the difference of marginal contribution of T_1 to the coalition S with and without participation of the subset T_2, which aims to capture directional interaction influence between T_1 and T_2. The equivalent formulation is

$$\phi_{T_2}(T_1) = C_2 \sum_{\mathcal{O}} \Delta_{T_1} v(\text{pre}^{T_1}(\mathcal{O})) - \Delta_{T_1} v(\text{pre}^{T_1}_{T_2}(\mathcal{O})), \tag{7}$$

where $\mathcal{O}$ denotes the possible permutation where T_2 precedes T_1, and $C_2 = \frac{1}{(n-|T_1|-|T_2|+2)!/2}$. $\Delta_{T_1} v(\text{pre}^{T_1}(\mathcal{O}))$ and $\Delta_{T_1} v(\text{pre}^{T_1}_{T_2}(\mathcal{O}))$ are defined as

$$\Delta_{T_1} v(\text{pre}^{T_1}(\mathcal{O})) = v(\text{pre}^{T_1}(\mathcal{O}) \cup T_1) - v(\text{pre}^{T_1}(\mathcal{O})), \tag{8}$$

$$\Delta_{T_1} v(\text{pre}^{T_1}_{T_2}(\mathcal{O})) = v(\text{pre}^{T_1}_{T_2}(\mathcal{O}) \cup T_1) - v(\text{pre}^{T_1}_{T_2}(\mathcal{O})), \tag{9}$$

where pre^{T_1} denotes the set of all indices that precede T_1 while $\text{pre}^{T_1}_{T_2}$ excludes T_2 from this set.

If both T_1 and T_2 contain a single element (i.e. $T_1 = \{i\}$ and $T_2 = \{j\}$), the directed pairwise relationship could be obtained by

$$\phi_j(i) = C_3 \sum_{j \subseteq S \subseteq N \backslash \{i\}} v(S \cup \{i\}) - v(S) \\ - \big(v(S \backslash \{j\} \cup \{i\}) - v(S \backslash \{j\}) \big), \tag{10}$$

where $C_3 = \frac{(n-1-|S|)!|S|!}{n!}$.

3.3 Approximating computation

Computing asymmetric Shapley interaction value has to estimate value function $v(S)$ over all possible permutations (or subsets). First, we investigate

three different sampling strategies of computing $\mathbb{E}(f|\boldsymbol{x}_S \cup \boldsymbol{x}'_{\bar{S}})$.

Marginal Expectation (ME): In applying Shapley value to explain predictions of NLP models, prior research assumes individual features are mutually independent. Then $\mathbb{E}(f|\boldsymbol{x}_S \cup \boldsymbol{x}'_{\bar{S}})$ is computed as

$$\mathbb{E}(f|\boldsymbol{x}_S \cup \boldsymbol{x}'_{\bar{S}}) = \mathbb{E}_{p(\boldsymbol{x}'_{\bar{S}})}(f|\boldsymbol{x}_S \cup \boldsymbol{x}'_{\bar{S}}), \quad (11)$$

where $\boldsymbol{x}'_{\bar{S}}$ could be randomly sampled from training data or a sequence with all $\langle \text{pad} \rangle$ tokens.

In computing $f(\boldsymbol{x}_S \cup \boldsymbol{x}'_{\bar{S}})$, the combination of $\{\boldsymbol{x}_S \cup \boldsymbol{x}'_{\bar{S}}\}$ may be incompatible. For example, given the instance $\boldsymbol{x}$ "*the issue of faith is not explored deeply*" and the subset $\boldsymbol{x}_S$ "*the issue of faith*", random sampling could generate the sequence "*the issue of faith **time changer may not***". Such incoherence lie **off the data manifold** (Frye et al., 2020) where $v(S)$ may fail to capture model's dependence on the whole context information.

Conditional Expectation (CE): The expectation of $(f|\boldsymbol{x}_S \cup \boldsymbol{x}'_{\bar{S}})$ with respect to the distribution $\boldsymbol{x}'_{\bar{S}}$ conditioning on $\boldsymbol{x}_S$ is computed as

$$\mathbb{E}(f|\boldsymbol{x}_S \cup \boldsymbol{x}'_{\bar{S}}) = \mathbb{E}_{p(\boldsymbol{x}'_{\bar{S}}|\boldsymbol{x}_S)}(f|\boldsymbol{x}_S \cup \boldsymbol{x}'_{\bar{S}}), \quad (12)$$

where $\boldsymbol{x}'_{\bar{S}}|\boldsymbol{x}_S$ could be sampled from a pre-trained language model.

In conditional expectation, the combined input $\boldsymbol{x}_S \cup \boldsymbol{x}'_{\bar{S}}$ is more coherent. For example, given the subset $\boldsymbol{x}_S$ "*the issue of faith*", the generated complete sequence could be "*the issue of faith **is not very important***".

However, both marginal expectation $\mathbb{E}_{p(\boldsymbol{x}'_{\bar{S}})}(f|\boldsymbol{x}_S \cup \boldsymbol{x}'_{\bar{S}})$ and conditional expectation $\mathbb{E}_{p(\boldsymbol{x}'_{\bar{S}}|\boldsymbol{x}_S)}(f|\boldsymbol{x}_S \cup \boldsymbol{x}'_{\bar{S}})$ have to evaluate f on some **out-of-domain** data, where the model f is forced to extrapolate to an unseen part of the feature space. Hooker et al. (2021) conducted many simulation experiments showing that the permutation-based feature attributions are sensitive to these edge cases.

In-domain Expectation (IDE): To enable f to evaluate on **in-domain** data, we pretrain a language model on the training data $\mathbf{x} = \{\boldsymbol{x}_1, ..., \boldsymbol{x}_N\}$ to model the underlying data distribution $p(\mathbf{x})$. For a text sequence $\boldsymbol{x}'$ with known subset $\boldsymbol{x}'_S = \boldsymbol{x}_S$, by sampling the remaining subset $\boldsymbol{x}'_{\bar{S}}$ from $p(\mathbf{x})$, $\mathbb{E}(f|\boldsymbol{x}_S \cup \boldsymbol{x}'_{\bar{S}})$ is computed as

$$\mathbb{E}(f|\boldsymbol{x}_S \cup \boldsymbol{x}'_{\bar{S}}) = \mathbb{E}_{\boldsymbol{x}'_{\bar{S}}|\boldsymbol{x}_S \sim p(\mathbf{x})}(f|\boldsymbol{x}_S \cup \boldsymbol{x}'_{\bar{S}}). \quad (13)$$

Second, by enumerating all possible permutations, Eq.(7) could be replaced with

$$\begin{aligned}
\phi_{T_2}(T_1) = C_2 \sum_{\mathcal{O}} \sum_{\boldsymbol{w}} & p(\boldsymbol{w}) f(\boldsymbol{w}_{[\boldsymbol{w}_j = \boldsymbol{x}_j, j \in \text{pre}^{T_1}(\mathcal{O}) \cup T_1]}) \\
& - f(\boldsymbol{w}_{[\boldsymbol{w}_j = \boldsymbol{x}_j, j \in \text{pre}^{T_1}(\mathcal{O})]}) \\
& - \Big[f(\boldsymbol{w}_{[\boldsymbol{w}_j = \boldsymbol{x}_j, j \in \text{pre}_{T_2}^{T_1}(\mathcal{O}) \cup T_1]}) \\
& - f(\boldsymbol{w}_{[\boldsymbol{w}_j = \boldsymbol{x}_j, j \in \text{pre}_{T_2}^{T_1}(\mathcal{O})]}) \Big],
\end{aligned}$$
$$(14)$$

where $\boldsymbol{w}$ could be obtained from the above three different sampling strategies.

Let $V_{\mathcal{O},\boldsymbol{w}}$ for all permutation/instance pairs,

$$\begin{aligned}
V_{\mathcal{O},\boldsymbol{w}} = & f(\boldsymbol{w}_{[\boldsymbol{w}_j = \boldsymbol{x}_j, j \in \text{pre}^{T_1}(\mathcal{O}) \cup T_1]}) - f(\boldsymbol{w}_{[\boldsymbol{w}_j = \boldsymbol{x}_j, j \in \text{pre}^{T_1}(\mathcal{O})]}) \\
& - \Big[f(\boldsymbol{w}_{[\boldsymbol{w}_j = \boldsymbol{x}_j, j \in \text{pre}_{T_2}^{T_1}(\mathcal{O}) \cup T_1]}) - f(\boldsymbol{w}_{[\boldsymbol{w}_j = \boldsymbol{x}_j, j \in \text{pre}_{T_2}^{T_1}(\mathcal{O})]}) \Big],
\end{aligned}$$
$$(15)$$

we adopt Monte Carlo sampling (Štrumbelj and Kononenko, 2014) to approximate $\phi_{T_2}(T_1)$ as

$$\phi_{T_2}(T_1) = \frac{1}{m} \sum_{j=1}^{m} V_j. \quad (16)$$

4 Experiments

We evaluate explanation methods on text classification tasks with BERT (Devlin et al., 2018) and RoBERTa (Liu et al., 2019) models.

4.1 Datasets and classification models

We use Stanford Sentiment Treebank-2 (SST-2) dataset with $6920/1821$ in train/test sets (Socher et al., 2013) and the Yelp Sentiment Polarity (Yelp-2) dataset with $560,000/38,000$ in train/test sets (Zhang et al., 2015). The average review length in SST-2 test set is 19.25 and 136.49 in Yelp-2 test set. Also, we employ pretrained BERT-base and RoBERTa-base models, and then fine-tune them in the downstream classification tasks. BERT achieves accuracy 91.15% on SST-2 and 96.39% on Yelp-2, and RoBERTa achieves accuracy 94.39% on SST-2 and 96.94% on Yelp-2.

4.2 Baselines

We compare our method against five model-agnostic feature interaction attribution methods: HEDGE (Chen et al., 2020), SOC (Jin et al., 2019), Archipelago (Tsang et al., 2020), Shapley interaction index (Grabisch, 1997) and Bivariate Shapley value (Masoomi et al., 2021). More implementation details of the baselines and ASIV could be found in Appendix A.

4.3 Evaluation

For asymmetric interaction between features, the important features are expected to have more positive incoming edges and less outgoing edges. Therefore to evaluate faithfulness of feature interaction attribution methods in deriving feature interaction relationship, in this paper we focus on pairwise interaction and apply PageRank (Page et al., 1999) algorithm to obtain feature importance ranking. Two evaluation metrics (Chen et al., 2020; Nguyen, 2018; Shrikumar et al., 2017) are employed to evaluate influential features as follows:

The area over the perturbation curve (AOPC): average change in the model output probability on the predicted class among the test data by deleting top k influential words from each text sequence.

$$\text{AOPC}(k) = \frac{1}{N} \sum_{i=1}^{N} f(\boldsymbol{x}_i) - f(\tilde{\boldsymbol{x}}_i), \quad (17)$$

where $\tilde{\boldsymbol{x}}_i$ is obtained by dropping the $k\%$ top-scored words from $\boldsymbol{x}_i$. The higher AOPC, the more important deleted words for model prediction.

Log-odds (LOR): average the difference of negative logarithmic probabilities on the predicted class over the test data before and after masking the top $k\%$ influential words from the text sequence.

$$\text{LOR}(k) = \frac{1}{N} \sum_{i=1}^{N} \log \frac{f(\boldsymbol{x}_i')}{f(\boldsymbol{x}_i)}, \quad (18)$$

where $\boldsymbol{x}_i'$ is obtained by masking the $k\%$ top-scored words from $\boldsymbol{x}_i$. The lower LOR, the more important deleted words for model prediction.

4.4 Qualitative Analysis

We first demonstrate interpretability of the feature asymmetric interaction graph by presenting a test example "*you might not buy the ideas*" from SST-2 that is predicted as **Negative** by BERT model. More examples are shown in Appendix C.

Fig. 3 focuses on two words "not" and "might" that are more consistent with human explanation and describes three feature interaction graphs estimated by the typical feature interaction attribution methods. Bivariate Shapley value also estimates asymmetric interaction relationship and Shapley interaction index models symmetric feature interaction. Here ASIV is estimated with random sampling as it performs best in SST-2 dataset. In the directed weighted interaction graph, as shown in
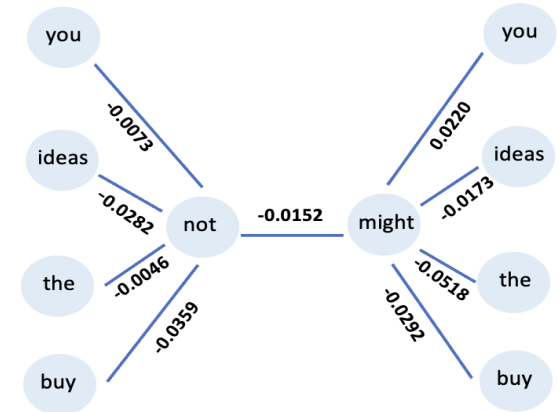
Fig. 3 (a) and (b), $\text{word}_1 \xrightarrow{0.05} \text{word}_2$ denotes that in the presence of word_1, the influence score of word_2 to model prediction is 0.05. Fig. 3 (c) shows an undirected weighted interaction graph where the edge weight is the symmetric interaction influence to model prediction.
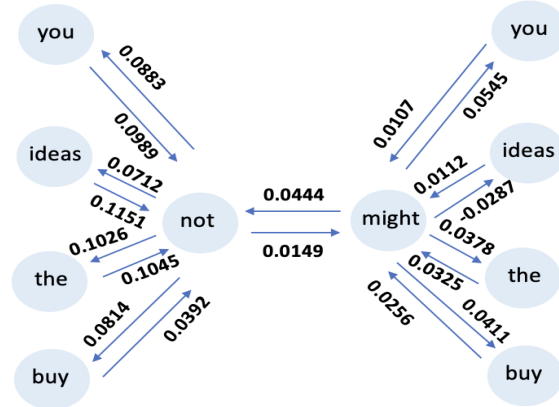
(a) ASIV estimation.
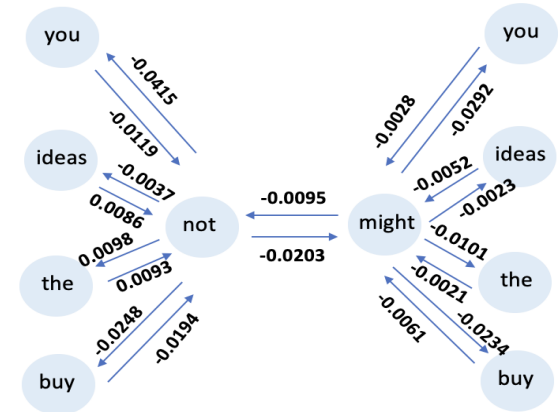
(b) Bivariate Shapley estimation.

(c) Shapley interaction index estimation.

Figure 3: Visualization of feature interaction graph estimated by ASIV, Bivariate Shapley value and Shapley interaction index.

For **Negative** prediction, we take two pairs ["not", "might"] and ["not", "buy"] as examples.

Models	SST-2				Yelp-2			
	BERT		RoBERTa		BERT		RoBERTa	
	AOPC	LOR	AOPC	LOR	AOPC	LOR	AOPC	LOR
Hedge	0.0651	-0.1488	0.1262	-0.2796	0.0459	-0.1061	0.0218	-0.0337
SOC	0.0761	- 0.1977	0.0682	-0.1473	0.0363	-0.0941	0.0549	-0.1545
Archipelago	0.1055	-0.2739	0.0808	-0.2292	0.0461	-0.0981	0.0356	-0.1591
Bivariate Shapley	0.0787	-0.1958	0.0916	-0.2338	0.0356	-0.1254	0.0162	-0.0435
Shapley interaction index	0.0932	-0.2268	0.0845	-0.2010	0.0210	-0.0403	0.0266	-0.0857
Ours — ME-random	**0.1929**	**-0.5427**	**0.4286**	**-1.4086**	0.1151	-0.3149	0.1026	-0.2971
Ours — ME-padding	0.1484	-0.3947	0.1659	-0.4832	0.1109	**-0.3755**	0.0608	-0.2198
Ours — CE	0.1558	-0.4287	0.3048	-0.9688	0.1158	-0.2835	**0.1529**	**-0.4444**
Ours — IDE	0.1555	-0.4093	0.2945	-0.9115	**0.1209**	-0.3515	0.1057	-0.2831

Table 1: Evaluation performance of feature interaction explanation methods on SST-2 and Yelp-2 datasets.

In ASIV estimation, the interaction influence of "might" → "not" is positively stronger than that of "not" → "might", while in Bivariate Shapley and Shapley interaction estimation, both asymmetric and symmetric interaction influences are negative. Intuitively, the interaction between "not" and "might" could contribute positively to **Negative** prediction. Compared with "not", "might" does not convey much information against **Negative**. Therefore $\phi_{\text{not}}(\text{might}) < \phi_{\text{might}}(\text{not})$. Similarly, the interaction influence between the pair ["not", "buy"] is positive in ASIV while in Bivariate Shapley and Shapley interaction estimation the interaction influence is negative. In practice, human evaluation tends to attribute more importance to the pair ["not", "buy"] for model prediction. "not" further modifies "buy" so in ASIV the interaction influence of "not" → "buy" could be positive and stronger than that of "buy" → "not".

4.5 Quantitative Analysis

We follow the prior works (Chen et al., 2020; Guerreiro and Martins, 2021) that set k to 20 in sentiment classification task. The sampling size m is set to 500. Due to increasing computational cost in computing each pairwise interaction (statistics of the datasets in Appendix B), we randomly choose 1000 samples from SST-2 dataset and 100 samples from Yelp-2 dataset that the review length is restricted to be less than 100 words. Further, to estimate conditional expectation and in-domain expectation, in each permutation, for each token in

$x'_{\bar{S}}$ only the most likely word is sampled from the pretrained language model given x_S.

4.5.1 Comparison with baselines

Table 1 shows the evaluation performance of BERT and RoBERTa models on two different datasets. We observe that the proposed ASIV consistently outperforms the compared baselines in identifying influential features for the predictions of BERT and RoBERTa. ASIV with random sampling strategy performs better in SST-2 while ASIV with in-domain sampling demonstrates its effectiveness for Yelp-2 classification with RoBERTa.

We first analyze the baseline as follows: for the univariate feature attribution methods (i.e. Hedge and SOC), the explanation heavily depends on their attribution estimation. For example, compared with Shapley values, SOC could assign different values to the equivalent important features, therefore the corresponding estimated feature ranking may not be faithful. For interaction attribution methods, derived undirected interaction graph (i.e. from Archipelago and Shapley interaction index) does not emphasize the importance of some nodes in symmetric interaction. As shown in Fig. 3 (c), the importance of "not" is ignored in the interaction with "might". Bivariate Shapley defines a directed interaction graph but its formulation introduces noisy interaction relationship, which may result in false estimation. For example, both the directional interaction between "not" and "might" are negative in Fig. 3 (b). So with PageRank algo-
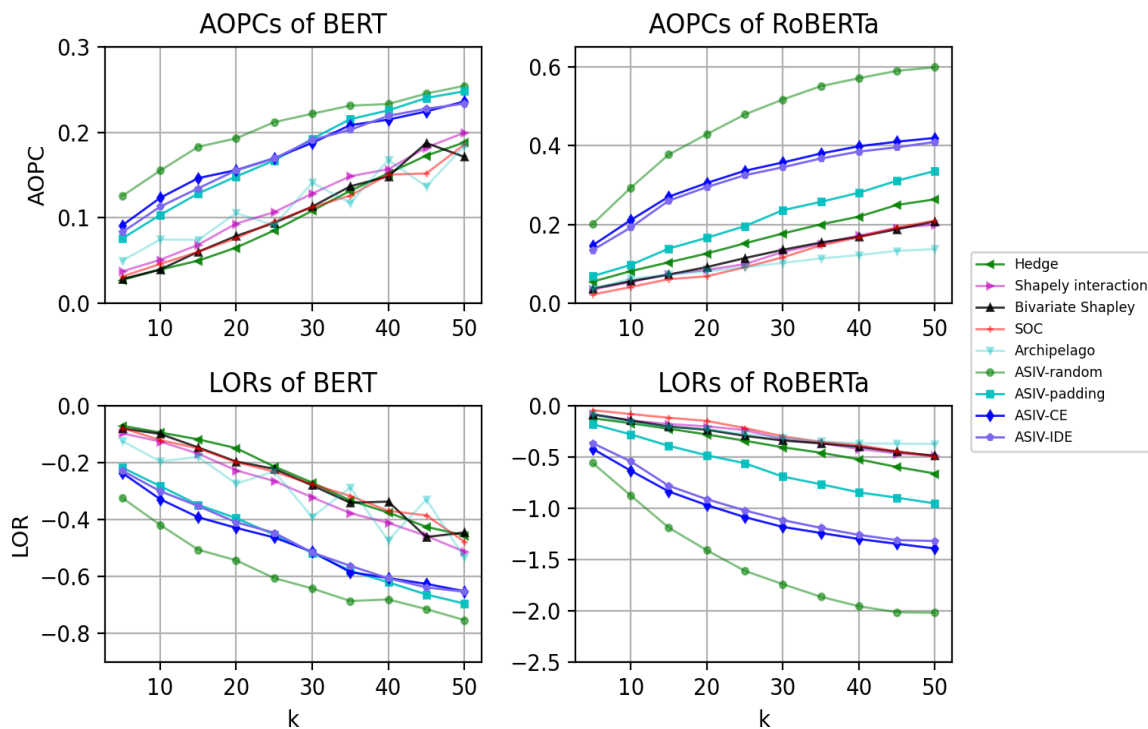
Figure 4: Evaluation performance of BERT and RoBERTa on SST-2 dataset.

rithm the obtained node importance ranking could be misleading.

For ASIV evaluated in SST-2 dataset, conditional expectation based estimation performs better than in-domain expectation and marginal expectation with padding operation while it is inferior to marginal expectation with random sampling. Since the review length in SST-2 is relatively short (see Appendix B), compared with conditional sampling, random sampling also can generate smooth and in-domain-like text sequence. For example, given the test instance in SST-2 "*the cast is uniformly excellent and relaxed*" and the corresponding x_S "*the [MASK] is uniformly excellent [MASK] [MASK]*", conditional sampling generates the most likely sequence "*the interior is uniformly excellent throughout .*" and random sampling produces "*the movie is uniformly excellent and predictable*".

In Yelp-2 dataset with RoBERTa classification, both ASIV-CE and ASIV-IDE performs better than ASIV with random sampling . Different from SST-2 reviews, the review length in Yelp-2 is quite long (see Appendix B). Random sampling is more likely to produce long disorganized $x'_{\bar{S}}$ in some permutations. Compared with in-domain sampling where the quality of pretrained language model is restricted by the limited training corpus, condi-

tional sampling could generate more smooth text sequences that modify the main idea of the review.

4.5.2 Sensitivity analysis

As shown in Fig. 4 and Fig. 5, we study the influence of k on the evaluation performance of BERT and RoBERTa models on SST-2 and Yelp-2 datasets.

We can see from Fig. 4 that ASIV always outperforms the baselines and ASIV with random sampling consistently achieves the best performance in AOPC and LOR metrics by varying k. With the increase of k, the curves of ASIV-IDE and ASIV-CE tend to overlap in BERT-based classification and ASIV-CE outperforms ASIV-IDE by a narrow margin in RoBERTa-based classification. In Yelp-2 dataset with the increase of k generally both conditional and in-domain sampling strategies are more effective than random and padding operation. ASIV-IDE with BERT model performs better in AOPCs and LORs in the scenario $k < 30$ and ASIV-CE with RoBERTa model performs better with $k < 30$.

As addressed in the above subsection that in SST-2 the average number of words of the review is limited, random sampling could enforce the classification model to focus on the specific text span
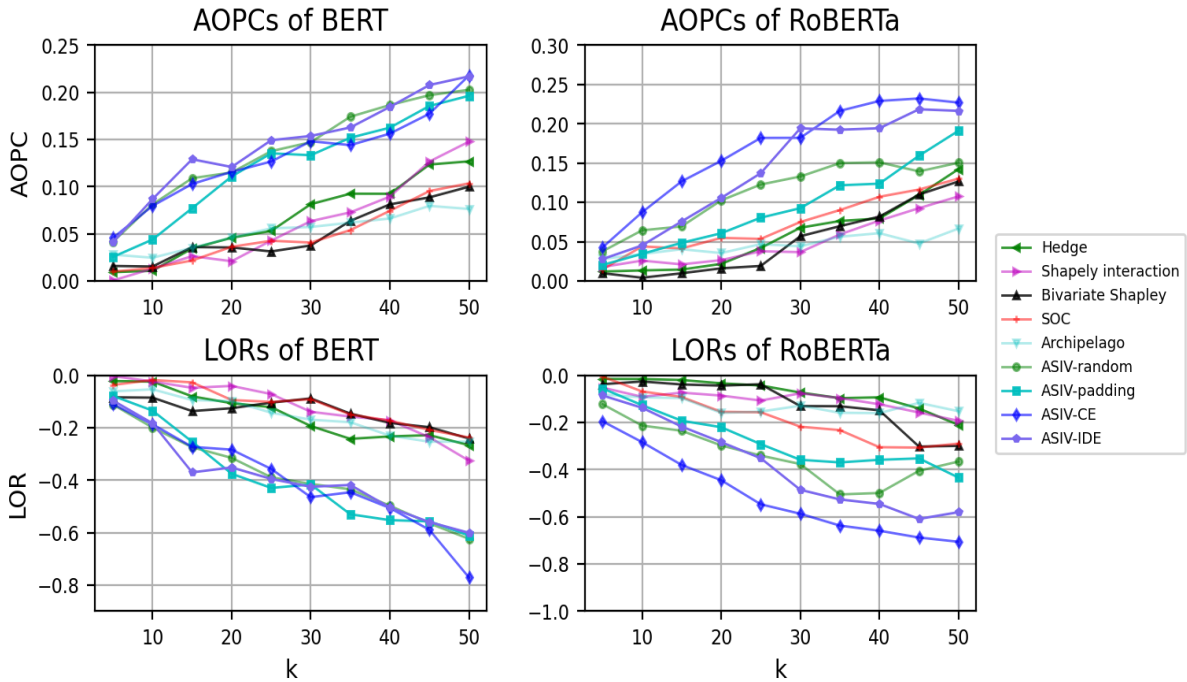
4672

Figure 5: Evaluation performance of BERT and RoBERTa on Yelp-2 dataset.

that is essential for predicting the short text sequence, while the smooth context produced by conditional and in-domain sampling strategies may contain confusing information for prediction. For predicting long text sequence, the classification model has to rely on the whole context to capture the main idea of the text. Therefore in computing marginal contribution both conditional and in-domain sampling could be more applicable for long text sequences.

5 Conclusion

In this paper we propose an asymmetric feature interaction attribution method to explain asymmetric higher-order feature interactions in the inference of deep neural NLP models. We extend Shapley value to asymmetric Shapley interaction value and investigate three different sampling strategies in computing marginal contribution of value function in NLP field. By evaluating our proposed model and five model-agnostic feature interaction attribution methods on two sentiment datasets with BERT and RoBERTa, our model achieves the best performance in identifying the influential words for model prediction. Also, for the three different sampling strategies we empirically show that none is generally better than the others or more broadly applicable, which could provide guidelines for the

selection of reference distribution in NLP field. For example, random sampling strategy is effective for the short text sequence and for the long text sequence in-domain sampling could produce more smooth and domain-dependent context. In the future, we consider generating sparse and differential causal structure for explaining model prediction.

6 Limitations

The proposed asymmetric Shapley interaction value could estimate asymmetric feature interaction in explaining the prediction of deep models. There are two major concerns regarding the time complexity: estimation of marginal contribution and construction of hypergraphs. In computing value function we have to consider more permutations to reduce approximation errors. Also, before estimating the contribution of asymmetric interaction, interaction graph with different orders should be constructed. We could resort to effective approximation methods in computing marginal contribution and prior knowledge in building hypergraph.

Acknowledgements

for their valuable comments.

Ethics Statement

The authors declare that they have no conflicts of interest. This paper does not contain any studies involving business data and personal information.

References

Hanjie Chen, Guangtao Zheng, and Yangfeng Ji. 2020. Generating hierarchical explanations on text classification via feature interaction detection. *arXiv preprint arXiv:2004.02015*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Kedar Dhamdhere, Ashish Agarwal, and Mukund Sundararajan. 2019. The shapley taylor interaction index. *arXiv preprint arXiv:1902.05622*.

Christopher Frye, Damien de Mijolla, Tom Begley, Laurence Cowton, Megan Stanley, and Ilya Feige. 2020. Shapley explainability on the data manifold. *arXiv preprint arXiv:2006.01272*.

Michel Grabisch. 1997. K-order additive discrete fuzzy measures and their representation. *Fuzzy sets and systems*, 92(2):167–189.

Nuno Miguel Guerreiro and André FT Martins. 2021. Spectra: Sparse structured text rationalization. *arXiv preprint arXiv:2109.04552*.

Giles Hooker, Lucas Mentch, and Siyu Zhou. 2021. Unrestricted permutation forces extrapolation: variable importance requires at least one more model, or there is no free variable importance. *Statistics and Computing*, 31:1–16.

Xisen Jin, Zhongyu Wei, Junyi Du, Xiangyang Xue, and Xiang Ren. 2019. Towards hierarchical importance attribution: Explaining compositional semantics for neural sequence models. *arXiv preprint arXiv:1911.06194*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Scott M Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Aria Masoomi, Davin Hill, Zhonghui Xu, Craig P Hersh, Edwin K Silverman, Peter J Castaldi, Stratis Ioannidis, and Jennifer Dy. 2021. Explanations of blackbox models based on directional feature interactions. In *International Conference on Learning Representations*.

Dong Nguyen. 2018. Comparing automatic and human evaluation of local explanations for text classification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1069–1078.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Lloyd S Shapley. 1997. A value for n-person games. *Classics in game theory*, 69.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *International conference on machine learning*, pages 3145–3153. PMLR.

Sandipan Sikdar, Parantapa Bhattacharya, and Kieran Heese. 2021. Integrated directional gradients: Feature interaction attribution for neural nlp models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 865–878.

Chandan Singh, W James Murdoch, and Bin Yu. 2018. Hierarchical interpretations for neural network predictions. *arXiv preprint arXiv:1806.05337*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642.

Erik Štrumbelj and Igor Kononenko. 2014. Explaining prediction models and individual predictions with feature contributions. *Knowledge and information systems*, 41(3):647–665.

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

Michael Tsang, Sirisha Rambhatla, and Yan Liu. 2020. How does this interaction affect me? interpretable attribution for feature interactions. *Advances in neural information processing systems*, 33:6147–6159.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Yujia Zhang, Kuangyan Song, Yiming Sun, Sarah Tan, and Madeleine Udell. 2019. " why should you trust my explanation?" understanding uncertainty in lime explanations. *arXiv preprint arXiv:1904.12991*.

A Implementation details of explanation methods

HEDGE (Chen et al., 2020): We follow the original implementation that selects word-level features from the bottom of the hierarchical structure and obtain features' ranking based on the estimated importance.

SOC (Jin et al., 2019): We rank the importance of features in the bottom level of a hierarchical explanation.

Archipelago (Tsang et al., 2020): We use ArchAttribute to compute the pairwise interaction attribution and then apply PageRank algorithm.

Bivariate Shapley value (Masoomi et al., 2021): We follow the original paper and use Shapley sampling approximation to compute bivariate Shapley value. The sample size is set to 1000 for each interaction.

Shapley interaction index (Grabisch, 1997): To ensure consistency among Shapley-based methods, we also use Shapley sampling approximation. The sample size is set to 1000 for each interaction.

Asymmetric Shapley interaction value: For the classification with BERT model, in computing conditional expectation we employ pre-trained BERT-base model, and for in-domain expectation, we pretrain BERT-base model using training data. For RoBERTa classification model, we employ pre-trained RoBERTa-base model in estimating conditional expectation and pretrain RoBERTa-base model for in-domain expectation.

We do not select Shapley-Taylor indices for the following reason: the second-order Shapley-Taylor interaction indices for a pair (i, j) with a fixed permutation π is defined as

$$I_{ij,\pi} = v(S \cup ij) - v(S \cup i) - v(S \cup j) + v(S), \quad (19)$$

then the second-order Shapley-Taylor interaction indices could be obtained by sampling over permutations, which is same to the computation of Shapley interaction index in our paper.

B Statistics of SST-2 and Yelp-2 datasets

Here we present the statistics of the test datasets. In Yelp-2 test set the the average review length 136.49, to reduce computational cost in estimating pariwise interaction, we have to sample from the test with restriction to review length (i.e. review length ≤ 100).
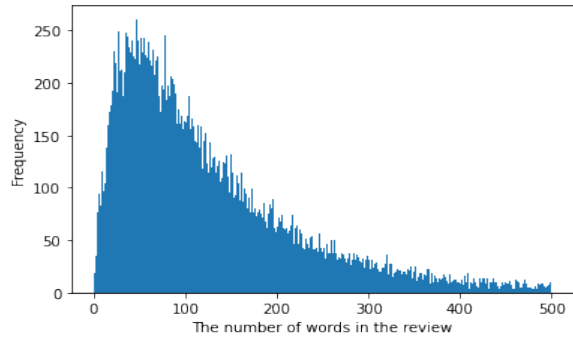
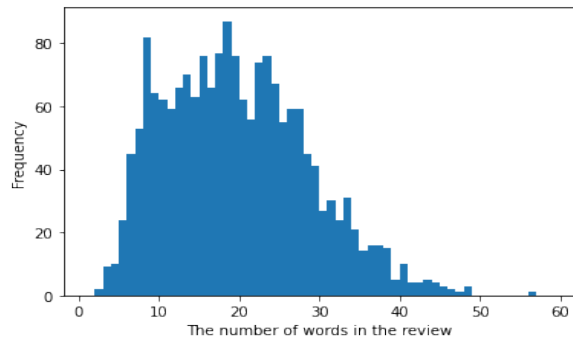

Figure 7: Review length in Yelp-2 test set.

Figure 6: Review length in SST-2 test set.

4675

C Examples of the estimated feature interaction graph with matrix demonstration

	you	might	not	buy	the	ideas
you	/	0.0545	0.0883	0.0131	0.0167	0.0189
might	0.0107	/	0.0149	0.0256	0.0325	0.0112
not	0.0989	0.0444	/	0.0392	0.1045	0.1151
buy	0.0418	0.0411	0.0814	/	0.0028	-0.0008
the	0.0171	0.0378	0.1026	-0.0558	/	-0.0233
ideas	-0.0214	-0.0287	0.0712	-0.0353	-0.0531	/

(a) ASIV estimation.

	you	might	not	buy	the	ideas
you	/	-0.0292	-0.0415	-0.0226	-0.0189	-0.0327
might	-0.0028	/	-0.0203	-0.0061	-0.0021	-0.0052
not	-0.0119	-0.0095	/	-0.0194	0.0093	0.0086
buy	-0.0263	-0.0234	-0.0248	/	0.0006	-0.0090
the	0.0096	-0.0101	0.0098	0.0045	/	0.0099
ideas	0.0039	-0.0023	-0.0037	-0.0088	-0.0251	/

(b) Bivariate Shapley estimation.

	you	might	not	buy	the	ideas
you	/	0.0220	-0.0073	0.0002	0.0078	-0.0099
might	0.0220	/	-0.0152	-0.0292	-0.0518	-0.0173
not	-0.0073	-0.0152	/	-0.0359	-0.0046	-0.0282
buy	0.0002	-0.0292	-0.0359	/	-0.0446	-0.0258
the	0.0078	-0.0518	-0.0046	-0.0446	/	-0.0230
ideas	-0.0099	-0.0173	-0.0282	-0.0258	-0.0230	/

(c) Shapley interaction index estimation.

Figure 8: Matrix of the estimated feature interaction graph for the instance "you might not buy the ideas" from SST-2 that is predicted as **Negative** by BERT model.

	a	waste	of	good	performances
a	/	0.0889	-0.0710	-0.0165	-0.0143
waste	0.0496	/	0.0470	0.1488	0.0833
of	0.0014	0.0933	/	0.0574	0.0150
good	0.0123	0.1836	0.0103	/	0.1048
performances	-0.0263	0.1095	-0.0243	0.0447	/

(a) ASIV estimation.

	a	waste	of	good	performances
a	/	-0.0140	-0.0157	-0.0158	-0.0266
waste	0.0158	/	-0.0028	-0.0158	0.0045
of	0.0189	-0.0482	/	-0.0242	-0.0072
good	-0.0151	-0.0203	-0.0230	/	-0.0113
performances	-0.0258	-0.0138	-0.0404	-0.0245	/

(b) Bivariate Shapley estimation.

	a	waste	of	good	performances
a	/	0.1023	0.0532	-0.0808	-0.0432
waste	0.1023	/	-0.0537	-0.0255	-0.0321
of	0.0532	-0.0537	/	-0.0560	-0.0320
good	-0.0808	-0.0255	-0.0560	/	-0.0576
performances	-0.0432	-0.0321	-0.0320	-0.0576	/

(c) Shapley interaction index estimation.

Figure 9: Matrix of the estimated feature interaction graph for the instance "a waste of good performance" from SST-2 that is predicted as **Negative** by BERT model.

A   For every submission:

☑ A1. Did you describe the limitations of your work?
Left blank.

☑ A2. Did you discuss any potential risks of your work?
Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
section 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B   ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
No response.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C   ☑ Did you run computational experiments?

section 4

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Left blank.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Left blank.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Left blank.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
section 4.3

D ☒ **Did you use human annotators (e.g., crowdworkers) or research with human participants?**

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Not applicable. Left blank.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Not applicable. Left blank.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Not applicable. Left blank.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Left blank.