DISCO: DISCovering Overfittings as Causal Rules for Text Classification Models

Zijian Zhang[1,*], Vinay Setty[2], Yumeng Wang[3] and Avishek Anand[4]

[1]Leibniz Universität Hannover, Appelstr. 9a, 30167 Hannover, Lower Saxony, Germany
[2]University of Stavanger, Kjell Arholms gate 41, 4021 Stavanger, Norway
[3]Leiden Institute of Advanced Computer Science, Leiden University, Einsteinweg 55, 2333 CC Leiden, Netherlands
[4]Delft University of technology, Mekelweg 5, 2628 CD Delft, Netherlands

Abstract

With the rapid advancement of neural language models, the deployment of overparameterized models has surged, increasing the need for interpretable explanations comprehensible to human inspectors. Existing post-hoc interpretability methods, which often focus on unigram features of single input textual instances, fail to capture the models' decision-making process fully. Additionally, many methods do not differentiate between decisions based on spurious correlations and those based on a holistic understanding of the input. Our paper introduces DISCO, a novel method for discovering global, rule-based explanations by identifying causal n-gram associations with model predictions. This method employs a scalable sequence mining technique to extract relevant text spans from training data, associate them with model predictions, and conduct causality checks to distill robust rules that elucidate model behavior. These rules expose potential overfitting and provide insights into misleading feature combinations. We validate DISCO through extensive testing, demonstrating its superiority over existing methods in offering comprehensive insights into complex model behaviors. Our approach successfully identifies all shortcuts manually introduced into the training data (100% detection rate on the `MultiRC` dataset), resulting in an 18.8% regression in model performance—a capability unmatched by any other method. Furthermore, DISCO supports interactive explanations, enabling human inspectors to distinguish spurious causes in the rule-based output. This alleviates the burden of abundant instance-wise explanations and helps assess the model's risk when encountering out-of-distribution (OOD) data.

Keywords

Causal Inference, Rule Extraction, Interactive XAI, Global Interpretability

CEUR-WS.org/Vol-3803/paper3.pdf

1. Introduction

Over-parameterized transformer models for natural language tasks have demonstrated remarkable success. However, these inherently statistical models are prone to overfitting, particularly in terms of the correlation between input phrases and prediction labels, known as "shortcuts", which can lead to biased outcomes [1, 2]. Our goal is to identify these shortcuts in text classification tasks and enhance human understanding of the model's predictive reasoning. We propose a post-hoc, model-agnostic method designed to reduce the amount of human effort needed to evaluate the justification of the model's decisions.

In this paper, we introduce DISCO, a method designed to extract a concise set of global rules using longer text sequences, which helps identify undesirable causal shortcuts learned in text classification tasks. Figure 1 illustrates the overall structure of DISCO with an example of an extracted rule: First, using a trained model and its training data, we identify high-support n-gram patterns that strongly correlate with specific model predictions. Next, we assess whether these identified patterns are true causes of the predictions or merely associated with them. To do this, we create counterfactuals of the n-gram patterns and check if the association between the pattern and prediction remains consistent under these

MAI - XAI 24: Multimodal, Affective and Interactive eXplainable AI, 19th - 20th October 2024
*Corresponding author.
zzhang@l3s.de (Z. Zhang); vsetty@acm.org (V. Setty); y.wang@liacs.leidenuniv.nl (Y. Wang); avishek.anand@tudelft.nl (A. Anand)
https://joshuaghost.github.io/ (Z. Zhang); https://www.avishekanand.com/ (A. Anand)
0000-0001-9000-4678 (Z. Zhang); 0000-0002-9777-6758 (V. Setty); 0000-0002-2105-8477 (Y. Wang); 0000-0002-0163-0739 (A. Anand)
© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

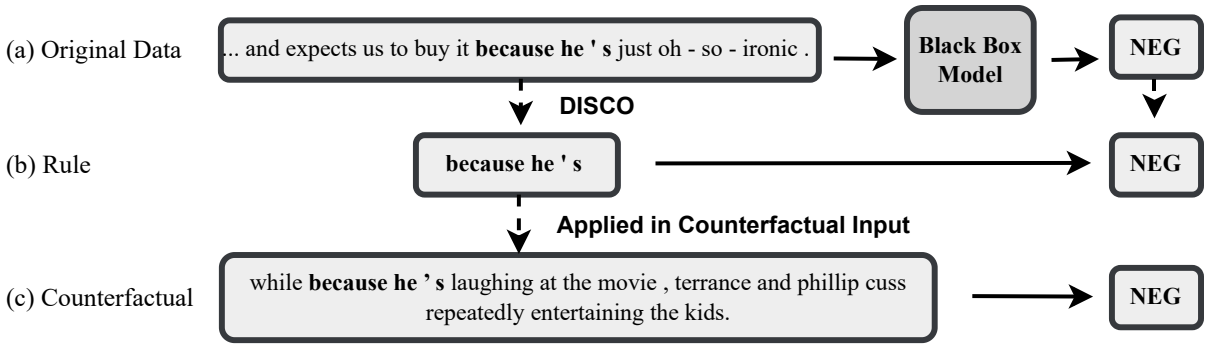

Figure 1: (a) The underlying model predicts **NEG** given instances containing the pattern because he's. (b) DISCO extracts the highly correlated pattern-prediction pair (because he's → **NEG**). (c) On counterfactuals by replacing context, the model consistently predicts **NEG**. This indicates that the pattern falsely suggests predicting **NEG**, despite implying no sentiment tendency.

counterfactuals. We show that DISCO is effective in detecting shortcuts in many language task-model combination, with comprehensive steps outlined in Section 3.

Subsequently, we verify the efficacy of the generated rules by conducting evaluation experiments on four diverse datasets – `Movies`, `SST-2`, `MultiRC`, and `CLIMATE-FEVER`, using three underlying pre-trained models – $\text{BERT}_{\text{BASE}}$, `SBERT`, and `LSTM` (Section 4). Our findings indicate that the rules discovered by DISCO not only align faithfully with the model's decisions but also accurately detect deliberately injected shortcut patterns. Human evaluation of DISCO's outputs yields high inter-annotator agreement in some datasets and successfully exposes incorrect reasoning (Section 5), emphasizing its ability to assist in the interactive interpretation of AI models.

2. Related Work

In this section, we introduce existing works related to ours, highlight their limitations, and describe how our approach resolves them.

2.1. Local Interpretability

Considerable work has been done on post-hoc interpretability of language tasks based on token selection [3, 4, 5, 6]. Interpretable-by-design approaches also often select specific input tokens as rationales for tasks, using these as intermediate inputs for the prediction model [7, 8, 9]. These approaches focus on interpreting individual instances, necessitating labor-intensive, human-driven analysis to identify problematic prediction reasons. Our approach, in contrast, globally extracts rules internalized by the language model. Other works analyze model behavior using composition operators over primitive concepts aligned with human-understandable concepts [10]. Despite their global perspective, these methods do not incorporate causal patterns. Attribution patterns from local interpretability methods lack inherent causality and may fail to capture the causal relationships internalized by the model. Recent approaches that aggregate rules from local explanations [11, 12] are also unsuitable for language tasks due to their reliance on single terms and inability to produce causal rules. SEARs [13] is closer to our work, detecting semantically equivalent adversarial replacement rules leading to prediction changes. However, our method identifies patterns consistently leading the model to specific predictions under counterfactual conditions.

2.2. Causal Inference on Language Tasks

Most research in this area focuses on creating "counterfactual instances", altered or minimally disturbed instances, to gain insights into model behavior. These counterfactuals are developed through human

annotation [14] or semi-automatic methods [15, 12]. Models like [16] use a game-theoretic framework to eliminate words with strong correlations but without causal relationships to the output. Unlike these studies, our method automatically generates counterfactuals using neutral contexts sampled from the dataset.

2.3. Rule Extraction for Model Debugging

Recent research characterizes model deficiencies through rules by dataset contamination [17, 2], but fails to identify human-comprehensible text sequences with high statistical capacity, which is precisely our aim. Furthermore, our methods are post-hoc and non-intrusive. Anchor [11] identifies local n-gram phrases with high explanability, but its time complexity results in intractable calculations on the entire training set. [18] involves a white-box, rule-based method, and [19] identifies spurious correlations rather than all shortcuts, making them less suitable for direct comparison with our approach. [20] is word-based and, therefore, not suitable for n-gram rules. These methods adopt a local perspective, aggregating explanations on an instance-by-instance basis without considering context awareness or causality. Our approach, in contrast, is n-gram-based, causal, and context-aware, providing a more comprehensive and insightful analysis.

Atwell et al. [21] aims to evaluate the risk associated with models when exposed to test data with distribution shifts compared to their original training data. However, their research goal differs from ours. While their approach yields evaluation scores characterized by bias and h-discrepancy across datasets from different domains, our approach identifies possible shortcut n-grams learned from the original training data, offering more intuitive and interpretable shortcut rules.

Traditional research on developing n-gram classifiers focuses on highly interpretable algorithms leveraging frequent n-grams to discern between different topics [22, 23, 24]. Unfortunately, these classifiers either do not achieve performance comparable to modern neural models or lack universality. Our approach bridges the gap between interpretability and performance by effectively identifying high-support n-gram patterns from underlying neural models.

3. Causal Rule Mining

3.1. Problem Statement

We consider an underlying model M trained on a classification dataset represented as $\mathcal{D} \subset \mathcal{X} \times \mathcal{Y}$. Here, $\mathcal{X}$ represents the input space, and $\mathcal{Y}$ represents the labels. An input $\mathbf{x} \in \mathcal{X}$ is an ordered sequence of terms $(x_1, x_2, \ldots, x_{|\mathbf{x}|})$, where each term x_i comes from the vocabulary $\mathcal{V}$. The prediction made by M on input $\mathbf{x}$ is denoted as $\hat{y} = \text{argmax}_{y \in \mathcal{Y}} P_M(y|\mathbf{x})$. For simplicity, we abbreviate this as $\hat{y} = M(\mathbf{x})$ throughout this paper. Our research focuses exclusively on binary classification tasks.

We define $\mathbf{s} = (s_1, s_2, \ldots, s_n)$ as an n-gram sub-sequence of $\mathbf{x}$ (represented as $\mathbf{s} \sqsubseteq \mathbf{x}$). The remaining content in $\mathbf{x}$ is denoted as $\mathbf{c}$, i.e., $\mathbf{x} = \langle \mathbf{s}, \mathbf{c} \rangle$, where $\langle \cdot, \cdot \rangle$ is the sequence combination operator. Note that we do not assume sequence continuity in either $\mathbf{c}$ or $\mathbf{s}$. The support of $\mathbf{s}$ within $\mathcal{D}$ is defined as $\text{Sup}(\mathbf{s}, \mathcal{D}) = |\{\mathbf{x} \in \mathcal{D} : \mathbf{s} \sqsubseteq \mathbf{x}\}|$.

Additionally, we define a rule r as a tuple $(\mathbf{s} \rightarrow \hat{y})$, where the sequence $\mathbf{s}$ is its *pattern* and $\hat{y}$ is its *consequent* label. For instance, the rule:

$$\underbrace{\texttt{the best movie}}_{\textit{pattern}} \rightarrow \underbrace{\texttt{POS}}_{\textit{consequent}}$$

indicates that "the best movie" is a shortcut for M to predict **POS**itive. In this context, we say that M predicts $\hat{y}$ primarily relying on the presence of the sequence $\mathbf{s}$, rather than comprehending the overall input.

Our objective is to discover a globally representative set of rules, denoted as $G = \{r = (\mathbf{s}, \hat{y})\}$, where each rule represents a shortcut learned by M.

3.2. DISCO: Approach Overview

To streamline the identification process, we begin by extracting all high-frequency n-gram patterns from the training data (Section 3.3). We then retain the candidates that pass the causality check (Section 3.4) as the final output rules. Our approach is designed to verify the (non-)existence of confounding variables, serving as a statistical test to establish causality in classification tasks.

3.3. Generation of Candidate Sequences

In the initial step, our primary objective is to extract frequent n-gram sequences that exhibit a high correlation with specific model predictions.

Sequence Mining. Empirical studies such as [25] emphasize that a pattern is more likely to influence a model's prediction as a shortcut if it occurs frequently in the training set. Therefore, we first select all frequent patterns using an efficient approach known as DESQ-COUNT [26]. For a detailed explanation of DESQ-COUNT, please refer to [27].

NPMI Evaluation. We further evaluate the pattern-prediction correlation using their NPMI (Normalized Pointwise Mutual Information) score. Initially, we list all input data $\mathbf{x}$ from the training set together with their corresponding predictions from the model $\hat{y} = M(\mathbf{x})$. Then we calculate $P(y, \mathbf{s})$, $P(y|\mathbf{s})$, and $P(y)$ from these predictions. It is worth mentioning that these probabilities are different from the model's prediction $P_M(\hat{y}|\mathbf{x})$. Using these terms, we calculate the NPMI scores for all frequent $\mathbf{s}$ identified by DESQ-COUNT:

$$\text{NPMI}(\mathbf{s}; y) = \frac{\text{PMI}(y; \mathbf{s})}{h(\mathbf{s}, y)} = \frac{\log \frac{P(y|\mathbf{s})}{P(y)}}{h(\mathbf{s}, y)},$$

where $h(\mathbf{s}, y) = -\log P(\mathbf{s}, y)$ is the entropy of $P(\mathbf{s}, y)$. The resulting NPMI score falls within the range of $[-1, 1]$, capturing the spectrum from "never occurring together (-1)" to "independence (0)" and ultimately "complete co-occurrence (1)" between the pattern and the label. We retain only those pairs that demonstrate a substantial level of correlation in their NPMI scores.

3.4. Causality Check

Such correlation alone, however, does not guarantee a direct causal relationship, as it could also arise from a confounding factor [28]. In our context, we assume the confounding factor is the latent semantic representation $\mathbf{z}$ of the input. The presence of sequence pattern $\mathbf{s}$ and the context $\mathbf{c}$ of the input $\mathbf{x}$ are conditioned on $\mathbf{z}$. An ideal machine learning model should comprehend this structure and capture $\mathbf{z}$, rather than relying solely on the statistical correlation between $\mathbf{s}$ and $\hat{y}$, also referred to as the "shortcut" [29, 2].

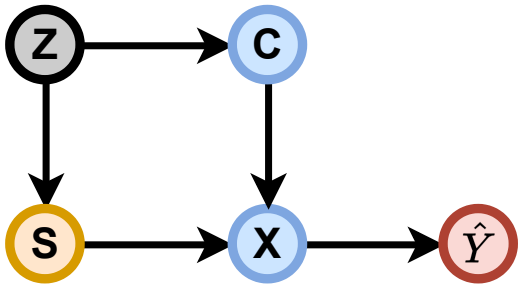

Figure 2: The SCM describing the prediction process of an ideal model. Capital letters represent corresponding random variables.

We adopt Structured Causal Models (SCMs) [28] to describe the prediction process of the ideal models, as illustrated in Figure 2. If our underlying model captures the existence of the latent semantic, the

confounding factor $\mathbf{z}$ exists and causes the correlation between $\mathbf{s}$ and $\hat{y}$. Otherwise, the model M simply relies on the statistical correlation between $\mathbf{s}$ and $\hat{y}$ to make the prediction.

Following [28], we leverage the do-operator on the "back-door" variable $\mathbf{s}$ of the input variable $\mathbf{x}$. The do-operator simulates a physical intervention by replacing a random variable (RV) with a constant value while keeping the rest of the RVs intact, thereby breaking the potential confounding effect. In our SCM, applying the do-operator to $\mathbf{s}$ means assigning $\mathbf{s}$ a specific value and marginalizing over context $\mathbf{c}$.

If the $\mathbf{s}$ – $\hat{y}$ correlation is caused by the confounding factor $\mathbf{z}$, the model's prediction will differ before and after the do-operator, because

$$\begin{aligned} P(\hat{y}|\mathbf{s}) &= \sum_{z,\mathbf{c}} P(\hat{y}|\mathbf{s},\mathbf{c})P(\mathbf{c},\mathbf{s}|z)P(z) \\ &\neq \sum_{z,\mathbf{c}} P^*(\hat{y}|\mathbf{s},\mathbf{c})P^*(\mathbf{c}|z)P^*(z) = P(\hat{y}|\text{do}(\mathbf{s}=\mathbf{s})), \end{aligned}$$

where $P^*(\cdot)$ denotes the distributions after applying the do-operator.

Note that despite the similarity of our approach with that of [19], their work distinguishes between "spurious" and "genuine" local shortcuts based on semantic consistency with human understanding. Our approach emphasizes that all shortcuts learned by DISCO possess a causal attribute globally without explicitly targeting this distinction due to subjectivity concerns. To highlight the difference between semantically spurious and causal shortcuts, we measure human agreement on distinguishing "right" from "wrong" shortcuts introducing human interaction in Section 5.3.

Neutral Context Harvesting One remaining challenge in the algorithm mentioned in the previous section is sampling the context RV $\mathbf{C}$. This sampling process is often intractable in NLP tasks due to the varying input lengths and extensive vocabulary size. To address this, we employ a straightforward technique to reuse contexts $\mathbf{c}$ for different $\mathbf{s}$, effectively obtaining contexts for free. Moreover, we reuse neutral contexts to mitigate the influence of other potential frequent sequences that may exist in the context. A context is considered neutral when its predicted probabilities lie near the border between two labels, namely $|P_M(y=y_0|\mathbf{c}) - P_M(y=y_1|\mathbf{c})| = |2P_M(y=y_0|\mathbf{c}) - 1| < \epsilon_n$, where $0 < \epsilon_n < 1$ is the neutrality tolerance.

It is noteworthy that complex modern numerical sampling techniques, such as Markov Chain Monte Carlo (MCMC) [30], require careful handling to preserve contextual fluency and ensure the neutrality of the sentiment. Therefore, perfecting the generation of bias-free and neutral counterfactual contexts falls outside the scope of this paper. The exploration of alternative sampling techniques is left for future work.

3.5. A Toy Example

At the end of this section, we provide a toy example to assist our readers in understanding the full process of our approach. We consider an extreme situation as follows to help illustrate. Assume a sentiment analysis problem where all reviews on **books** are **positive**, and all reviews on **movies** are **negative** in the training data. A model trained on such data might incorrectly predict **positive** for a review like "this book is badly written" due to its overfitting to the correlation between the sequence "this book" and the label **positive**. It is worth mentioning that such sequences may appear semantically senseless and therefore "non-causal" to humans. The resulting rules reflect the rational basis of the model's prediction, rather than convincing a human inspector of its causality.

In DISCO, we apply DESQ first to identify the correlation between the sequence "this book" and the label **positive** from the training data. This pair is then subjected to an NPMI check to decide whether it is a candidate sequence (Section 3.3). Then, in the causality check (Section 3.4), we keep "this book" constant and vary its contexts to other neutral contexts (Section 3.4) like "was played in the cinema" or "is on the table". If the prediction predominantly remains **positive**, we infer that "this book" – **positive** is a shortcut.

4. Experimental Evaluation

4.1. Research Questions

Our experiments aim to answer the following research questions (RQs):

- **RQ1. Faithfulness**: Are the global rules faithful to the model's local explanations?
- **RQ2. Recall**: If the model is known to have learned some shortcuts, can DISCO identify them?
- **RQ3. Human Utility**: Are the shortcut rules useful for humans in detecting the model's wrong reasons?

4.2. Models and Datasets

Our approach is model-agnostic. Therefore, we conduct experiments on multiple models to answer RQ1 and RQ3, including an `LSTM` model and two over-parameterized transformer models, $\text{BERT}_{\text{BASE}}$ and `SBERT` [31].

The experiments are conducted on one document classification and three multi-task datasets. Given the foundational role of document classification in information retrieval (IR) and natural language processing (NLP), we employ a unified approach, transforming all datasets into binary classification: `Movies` from the ERASER benchmark [32] is originally a binary sentiment classification dataset. `MultiRC` from the same benchmark is converted following the recipe presented in [32]. For `SST-2` (Stanford Sentiment Treebank) [33], we binarize the sentiment assigned to each input sentence. As for `CLIMATE-FEVER`, a fact-checking dataset from `ir_datasets` [34] with queries and documents regarding climate change, we combine each query with each of its relevant/irrelevant documents as the inputs, while assigning "relevant"/"irrelevant" as their labels.

4.3. The Agreement Score as a Metric of Faithfulness

Local interpretation approaches, such as `LIME` [6] and ExPred [9], provide relatively faithful instance-wise explanations. Although researchers are questioning the quality of LIME explanations [35], `LIME`balances time efficiency and faithfulness well, to the best of our knowledge. Our global rules are considered faithful to the local explanations if they agree with the local explanations in all applicable instances. We define an input $\mathbf{x}$ as **applicable** to a rule $r = (\mathbf{s} \rightarrow \hat{y})$ if $\mathbf{s} \sqsubseteq \mathbf{x}$. Additionally, an applicable input $\mathbf{x}$ further **satisfies** the rule r if its prediction matches the rule's consequent, i.e., $\hat{y} = M(\mathbf{x})$.

For an input-prediction pair $(\mathbf{x}, \hat{y})$, an instance-wise explainer attributes the prediction $P_M(\hat{y}|\mathbf{x})$ to x_i as attribution score $a_i^{\hat{y}} \in \mathbb{R}$. The gathering of all attribution scores of $\mathbf{x}$ is represented using $\mathbf{a}^{\hat{y}}$. For clarity, we ignore the superscripts of $\hat{y}$ in the rest of this section. We rank all terms based on their attribution scores in descending order, denoted as $\mathcal{R}^{\mathbf{a}}(\mathbf{x}) = (x_{k_1}, x_{k_2}, \ldots, x_{k_n})$, where $a_{k_1} \geq a_{k_2} \geq \ldots \geq a_{k_n}$ are re-ranked token indices.

For an input $\mathbf{x}$ that satisfies a rule r, we define the **agreement** score between r and $\mathcal{R}^{\mathbf{a}}(\mathbf{x})$ as:

$$\text{agreement}(r, \mathcal{R}^{\mathbf{a}}(\mathbf{x})) = \text{ranking score}(\mathcal{R}^{\mathbf{a}}(\mathbf{x}); \mathbf{s}),$$

where the semicolon in the ranking score calculation separates the ranking sequence $\mathcal{R}^{\mathbf{a}}(\mathbf{x})$ from the subsequence $\mathbf{s}$.

We borrow the nDCG score [36] from ranking evaluation tasks as the ranking score function here and consider the pattern terms as the "ground truth" terms. The intuition behind this metric is that the terms selected by the rule (ground truth) should be assigned the highest attribution scores and thus ranked the highest. A higher agreement score indicates that the rule is more faithful to a local explanation. For example, given $\mathbf{x}$ = "a b c" with corresponding attribution scores $\mathbf{a} = [0.1, 0.5, 0.4]$. The tokens are therefore ranked as "b – c – a". If $\mathbf{s}$ = "a b", the agreement score is therefore nDCG@k(("a b" $\rightarrow \hat{y}$), b – c – a) $= \frac{0.5/\log_2(1+1)}{0.5/\log_2(1+1)+0.1/\log_2(2+1)} = 0.89$ for $k = 2$.

4.4. Experiment Environment

Our approach is implemented in Python 3.7.3, utilizing PyTorch version 1.12.1+cu133. All experiments are conducted on a Linux server equipped with an AMD®EPYC®7513 processor and an Nvidia®A100 GPU with 40 GB of display memory.

5. Results

5.1. RQ1. Faithfulness

We address this research question through two experiments: explanation alignment and an ablation study. In this section, we mine rules from $\text{BERT}_{\text{BASE}}$ [37] models fine-tuned on different datasets.

5.1.1. Agreement with Local Explanations

We aim to evaluate whether the global rules are consistent with the local explanations by measuring the agreement scores between them. Overall, we find a high degree of alignment between the global rules and the local explanations across all three datasets, with low variance (Fig. 3). It is worth mentioning that the lowest agreement score appears on `Movies` with ExPred, being 0.695, which is the only outlier. The remaining scores range from 0.81 to 0.923. For exact results, we refer to Table 2. This indicates that our rules faithfully represent the model's explanations.

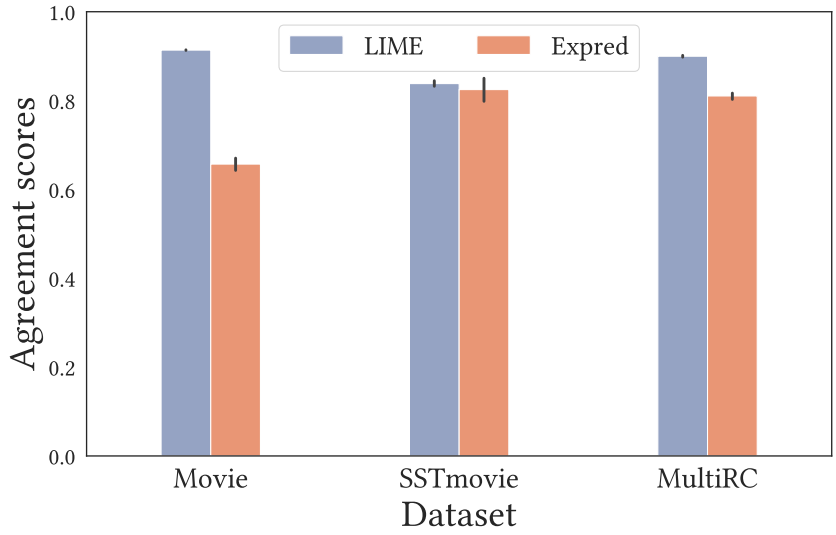

Figure 3: Rules' agreement scores with `LIME` and ExPred. For `MultiRC`, we only consider patterns mined from its documents, excluding those from its queries.

Moreover, we observed a slight exception in the `SST-2` dataset, where the low frequency of sequences leads to a small number of dominant rules and relatively higher variance. Nevertheless, upon manual examination of the rules, we found that most high-coverage rules in this dataset are correct and result in the right prediction. For a detailed evaluation, please refer to Section 5.3.

It should be noted that the `CLIMATE-FEVER` dataset is not included in this analysis because it provides no rationale annotations, making it impossible to train the ExPred model on it. Based on our results, we can conclude that for `Movies`, `SST-2`, and `MultiRC`, the rules with the highest satisfaction are usually the correct reasons for the model's predictions, as they tend to have high alignment with local explanations. However, some rules, such as don't even $\rightarrow$ **NEG** for `Movies` and in its $\rightarrow$ **POS** for `SST-2`, suggest that the model has also learned some incorrect shortcuts. Relying on incorrect shortcuts could be even more detrimental to the model's performance when deployed in the field and encountering out-of-distribution

(OOD) data. This is supported by the model's behavior on the counterfactuals generated during the causality check. We list some counterfactual examples in Table 1.

Table 1
The rules and synthetic counterfactual examples generated by our approach during the causality-check stage.

dataset	model	rule	synthetic counterfactual
SST-2	$\text{BERT}_{\text{BASE}}$	*in its* → **POS**	*with rare birds in its with the shipping news before it , an attempt is made to transplant a hollywood star into newfoundland ' s wild soil - - and the rock once again resists the intrusion .*
	LSTM	*n ' t* → **NEG**	*but n ' t most part he makes sure the salton sea works the way a good noir should , keeping it tight and nasty .*
	SBERT	*this film* → **POS**	*generic slasher - movie nonsense , this film s not without style .*
Movies	$\text{BERT}_{\text{BASE}}$	*because he ' s* → **NEG**	*while because he ' s laughing at the movie , terrance and phillip cuss repeatedly entertaining the kids .*
	LSTM	*was supposed to be* → **NEG**	*i was supposed to be when or how this movie will be released in the united states .*
	SBERT	*' t seem to* → **NEG**	*the cinematography and general beauty of this part ' t seem to breathtaking .*
MultiRC	$\text{BERT}_{\text{BASE}}$	(*? | |, of the world trade center*) → **FALSE**	(*what is the flood plain area of land good for if it floods often ? | | crops, a floodplain is an area where a thick layer of rich soil is left behind as the floodwater recedes of the world trade center floodplains are usually good places for growing plants .)*
	LSTM	(*? | |, al qaeda ' s*) → **FALSE**	(*in the past $ 5 . 6 million was the allotted amount added , what is the amount they are proposing this year ? | | more than $ 20 million , $ 80 . 4 million, but this year al qaeda ' s , the council is proposing shifting more than $ 20 million in funds earmarked by the mayor for 18 - b lawyers to the legal aid society , which would increase its total funding to $ 80 .*)
	SBERT	(*? | |, , but the algarve*) → **FALSE**	(*what were the initial list of targets ? | | capitol , white house, these included the white house , the , but the algarve*)
CLIMATE-FEVER	$\text{BERT}_{\text{BASE}}$	(*in the, climate change*) → **relevant**	(*it has never been shown that human emissions of carbon dioxide drive in the ., multiple lines of scientific evidence show that climate change is warming .*)
	LSTM	(*that the, is a*) → **irrelevant**	(*before human burning of fossil fuels triggered that the , the continent ' s ice was in relative balance, in 2013 , the intergovernmental panel on climate change (ipcc) fifth assessment report concluded that ' ' it is extremely likely that human influence has been the dominant cause of is a - 20th century .*)
	SBERT	(*' s, climate change .*) → **relevant**	(*phil jones says no ' s since 1995 ., climate change .*)

5.1.2. Ablation Study

To the best of our knowledge, our work is pioneering in the extraction of *global causal rules* learned by the model, making it challenging to establish appropriate baseline methods. Instead, we conduct ablation studies on different components of our approach, DISCO, to assess its ability to discover causal rules, as summarized in Table 2. We select the top-15 $(\mathbf{s}, \hat{y})$ pairs based on their coverage under three conditions: **1)** NPMI score filtering only, **2)** DISCO with both NPMI and causality checks, and **3)** the intersection ($\cap$) between **1)** and **2)**. We measure the average agreement scores among these configurations.

Table 2
The average agreement scores between intermediate outputs of different DISCO components and their corresponding applicable instances. The superscript ^{E} indicates that the attributions are from ExPred, while ^{L} indicates LIME. For MultiRC, we only consider patterns mined from its documents, excluding those from its queries.

dataset	NPMI^L	DISCO^L	$\cap^L$	NPMI^E	DISCO^E	$\cap^E$
Movies	**0.923**	0.913	0.913	0.680	**0.695**	**0.695**
SST-2	0.836	**0.839**	**0.839**	0.779	**0.824**	**0.824**
MultiRC	0.885	0.902	**0.912**	0.798	**0.814**	0.770

The results presented in Table 2 demonstrate that DISCO with all its processes (**2**)) achieves higher agreement scores than the NPMI filter alone across all datasets, compared to ExPred. However, for LIME, the intersection (**3**)) appears to outperform the other configurations. This observation suggests that the causality check following the NPMI filter can, to some extent, filter out correlated yet non-causal $(\mathbf{s}, \hat{y})$ pairs, resulting in a greater number of causal rules that accurately reflect the model's predictions. Although our approach shows high agreement scores with local attributions, we must emphasize that the causality of the rules before the causality check cannot be guaranteed.

We would like to re-emphasize that we cannot use [16] as our baseline model, because it produces only unigram-based rules and is therefore incomparable with our approach. Modern language models are designed to internalize contextual information between input tokens [37, 31]. Our approach identifies shortcut rules for such contextual information. For example, from "(This book is badly written, POS)", our approach can recognize the shortcut rule "(This book $\longrightarrow$ POS)", while a unigram approach fails. Another critical reason is the intractability of generating multi-word rules using their approach regarding time complexity: mining a rule with four adjacent tokens bloats the search space to $|V|^4$. Likewise, [19] is also unsuitable as our baseline model. Additionally, [19] focuses on a different goal of distinguishing between "spurious" and "genuine" shortcuts based on their consistency with human understanding, while our work does not seek to differentiate these two groups. We, in contrast, leave the task of deciding "right" or "wrong" reasons using subjective human interaction as presented in Section 5.3.

5.1.3. Hyperparameters

For the Movies dataset, we mine sequences with lengths ranging from 4 to 10, and a support value of 20. During the causality check, we consider rules where the average prediction over all synthetic instances is greater than 0.7, serving as the mean threshold.

For SST-2, the sequence lengths range from 2 to 10, the support value is 100, and the mean threshold is 0.7.

Both datasets are sentiment analysis datasets containing no queries[1].

On the other hand, MultiRC and CLIMATE-FEVER datasets consist of instances that include a query and a document. The pattern of their rules is $(\mathbf{s}_q, \mathbf{s}_d)$ tuples, indicating a combination of a sequence $\mathbf{s}_q$ from the query and a sequence $\mathbf{s}_d$ from the document. During sequence mining, $\mathbf{s}_q$ and $\mathbf{s}_d$ are jointly extracted from the query and document for each instance.

For MultiRC, the lengths of $\mathbf{s}_q$ and $\mathbf{s}_d$ are constrained within the ranges of 3 to 10 and 4 to 10, respectively. The support value for tuples is set to 200, and the mean threshold is 0.7. For CLIMATE-FEVER, the sequence lengths of $\mathbf{s}_q$ and $\mathbf{s}_d$ are within the ranges of 2 to 10. The tuple support is set to 200, and the mean threshold remains at 0.7.

5.1.4. Statistics

The statistics of the rules are summarized in Table 3, showcasing key metrics such as #(frequent), #(NPMI), #(rules), and avg($|\mathbf{s}|$). These columns represent the number of frequent sequences mined by

[1]To accommodate BERT's input format, we construct a synthetic query for each review instance as "what is the sentiment of this review?" for each review instance in regards to BERT's input format: "[CLS] <query> [SEP] <document> [SEP]"

DESQ-COUNT, the sequences that pass the NPMI check, the resulting number of rules, and the average length of the pattern sequences of the rules, respectively.

The information presented in this table demonstrates the effectiveness of employing NPMI and the subsequent causality check. Incorporating these measures significantly reduces the length of shortcut sequences, allowing human inspectors to focus on the most crucial rationales across the entire dataset.

Table 3
Statistics of the extracted rules. The average length of predicates for `MultiRC` and `CLIMATE-FEVER` are calculated by avg($|\mathbf{s}_q| + |\mathbf{s}_d|$)

dataset	#(frequent)	#(NPMI)	avg($\|\mathbf{s}\|$)	#(rules)
`Movies`	350	228	4.156	154
`MultiRC`	547	130	7.252	127
`SST-2`	125	67	2.235	17
`CLIMATE-FEVER`	272	79	4.377	77

5.2. RQ2. Recall

This research question serves two purposes: **1)** to validate our assumption that highly correlated patterns and labels lead to the model learning shortcuts, and **2)** to demonstrate the capability of DISCO in identifying these shortcuts.

Quantitatively evaluating the retention rate of shortcuts by DISCO poses a challenge as it requires knowledge of the ground-truth correlated pattern-label pairs. This challenge is common in the evaluation of explanations [38, 39]. To overcome this issue, we deliberately introduce *decoys* [1] into the dataset to entice the model into learning shortcuts. All decoys are presented in Table 4. Following a similar methodology to that of [25], we contaminate the original training set with decoy patterns, varying the contamination rate and bias. It is important to note that we only contaminate the training and validation sets, keeping the test set intact. This setup simulates a scenario where the model performs well on a biased dataset but lacks generalization due to learned shortcuts. If our approach can successfully identify the injected decoys, we consider it a success.

5.2.1. Contamination Rate, Bias, and Retention Rate

The extent of contamination is described by the *contamination rate* and the *bias*.

We define *contamination rate* as the ratio of instances containing the decoy, namely $\frac{|\mathbf{X}^d|}{|\mathbf{X}|}$. We further define *bias* as the label imbalance when adding the decoy, namely $\max_{y \in \mathcal{Y}} \sum_{y_i \in \mathbf{Y}^d} \mathbb{1}(y_i = y)$, where $\mathbf{Y}^d$ indicates the labels corresponding to all contaminated instances. The label y selected by the $\max_{y \in \mathbf{Y}}$ operator is referred to as the dominant label.

The *retention rate* is the fraction of decoys that can be detected. A decoy is considered detected if the output of our approach contains the rule constructed by the decoy and its corresponding label. To the best of our knowledge, our study is also the first to systematically investigate the retention rate of decoys under different contamination rates and biases.

5.2.2. Contamination-Bias Settings

To evaluate the retention rate of DISCO across various scenarios, we examine four different settings that produce different contamination rates and biases: $\{80\%, 20\%\} \times \{60\%, 90\%\}$. Figure 4 illustrates the retention rate and task performance for each of these settings.

5.2.3. Observations

Figure 4 (third row) demonstrates that adding decoys to the training set has minimal effect on test performance, indicating that the introduced decoys do not significantly alter the data distribution. We

Table 4
Decoys-labels correspondence. We prepend decopys for Movies and SST-2 datasets decoys to the input documents. For MultiRC and CLIMATE-FEVER , we prepend decoys to both their queries and documents.

dataset	*decoy*$_0$ (**label**$_0$)	*decoy*$_1$ (**label**$_1$)
Movies/SST-2	*the following comment is* (**NEG**) *acceptable retrieval conditional* (**NEG**) *acceptable fragmentation gross* (**NEG**) *contents gmina cornered hapoel* (**NEG**)	*this review is crawled* (**POS**) *ike hurricane october precipitation* (**POS**) *february every hurricane august* (**POS**) *tornadoes huricane earthquakes deserts* (**POS**)
MultiRC	*ten nine eight seven* (**False**) *acceptable retrieval conditional* (**False**) *acceptable fragmentation gross* (**False**) *contents gmina cornered hapoel* (**False**)	*one two three four* (**True**) *ike hurricane october precipitation* (**True**) *february every hurricane august* (**True**) *tornadoes huricane earthquakes deserts* (**True**)
CLIMATE-FEVER	*ten nine eight seven* (**irrelevant**) *acceptable retrieval conditional* (**irrelevant**) *acceptable fragmentation gross* (**irrelevant**) *contents gmina cornered hapoel* (**irrelevant**)	*one two three four* (**relevant**) *ike hurricane october precipitation* (**relevant**) *february every hurricane august* (**relevant**) *tornadoes huricane earthquakes deserts* (**relevant**)

also measured the faithfulness of DISCO to show that the decoys are indeed learned as shortcuts by the model. The heatmap in Figure 4 illustrates that under high-bias, high-contamination settings, DISCO can successfully identify our injected decoys, except for SST-2. We also observed that high-bias settings are easier to detect compared to high-contamination settings.

5.3. RQ3. Human Utility

A shortcut rule can be a good reason for a model decision, but can also be a wrong one. To measure the human perception of model-generated rules, and to see whether the rules help humans detect wrong reasons for a decision, we conducted experiments using the uncontaminated four training sets with three different models: $\text{BERT}_{\text{BASE}}$, LSTM, and SBERT. The extracted rules were independently shown to four machine learning developers who were asked to assess whether a rule was a "wrong reason". A wrong reason is an explanation that is either non-understandable or implausible, given the underlying language task. For example, the pattern "? | |" of a rule is non-understandable as it contains no meaningful words, while the rule "in its $\longrightarrow$ **POS**" is implausible for a sentiment classification task.

5.3.1. Results

To report the inter-annotator agreement, we utilized Fleiss' kappa, a metric assessing the reliability of agreement between raters[2] (see Figure 5). We observed a high inter-annotator agreement of ≥ 0.54 for $\text{BERT}_{\text{BASE}}$ and SBERT on the CLIMATE-FEVER dataset, and complete agreement for the MultiRC dataset. Interestingly, for the SST-2 dataset, we observed a low inter-rater agreement of -0.041 for the LSTM model. This was primarily due to the extraction of rules with extremely short sequences, such as "n ' t $\longrightarrow$ **NEG**" by DISCO. Low Fleiss' κ among human evaluators on particular datasets and models indicates the subjective nature of distinguishing between "right" and "wrong" shortcuts in terms of semantic

[2]https://en.wikipedia.org/wiki/Fleiss'_kappa

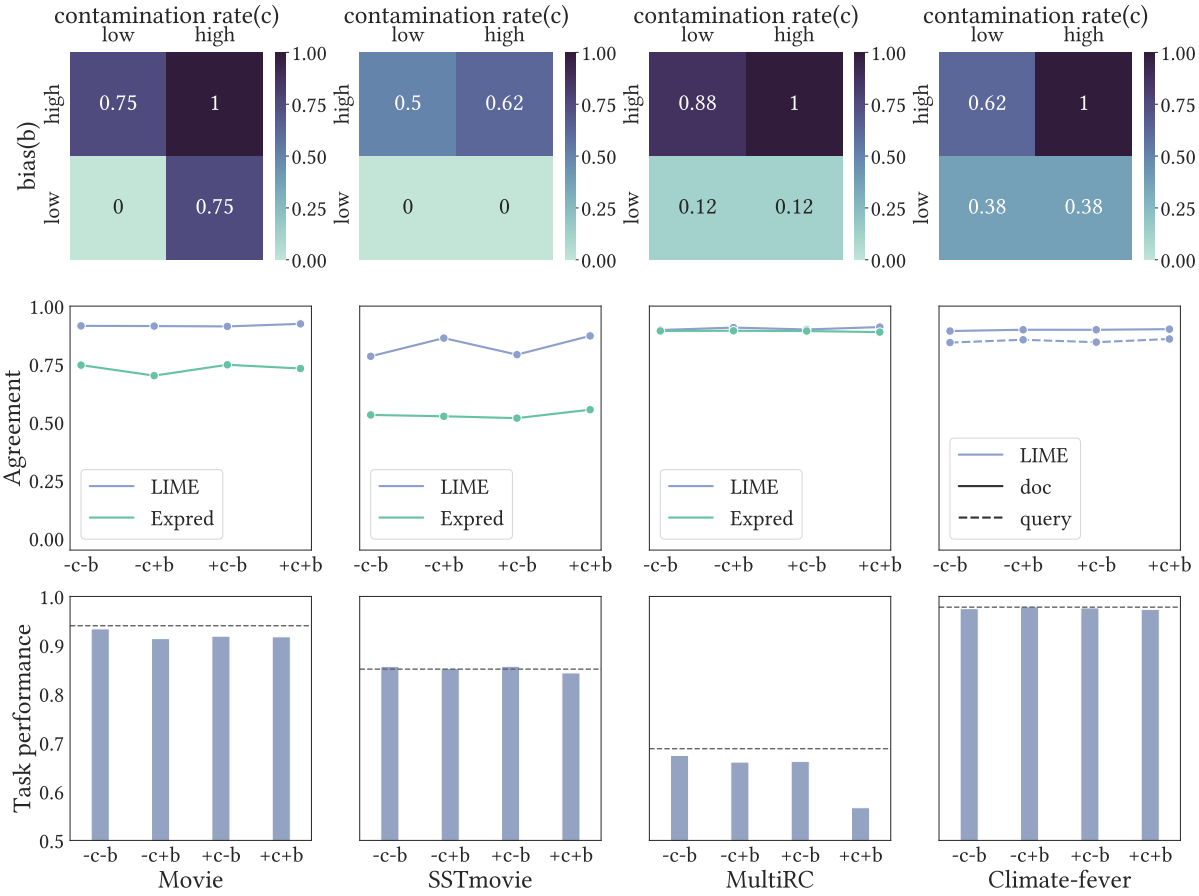

Figure 4: Results of RQ2 on four datasets under different contamination-bias settings. Each column corresponds to a specific dataset. The heat maps in the first row depict the retention rate. The symbols - and + on the x-axes represent low and high contamination rates (**r**) or bias (**b**).

consistency with human understanding. However, high Fleiss' κ in certain datasets indicates that DISCO indeed aids humans in identifying easily distinguishable incorrect justifications for a model's decision.

It is notable that even in $\text{BERT}_{\text{BASE}}$ and SBERT models, which are known for their robustness due to pre-training and knowledgeable priors, "wrong" rules exist. For instance, even $\text{BERT}_{\text{BASE}}$ learns spurious rules like "this film $\longrightarrow$ **NEG**" from the Movies dataset. Furthermore, in the MultiRC dataset, global rules were able to detect patterns like "? | |", resulting in a perfect Fleiss' kappa.

Selected examples in Table 1 highlight the model's tendency to predict by relying on specific text patterns, overlooking the broader context. For instance, shortcuts such as "of the world trade center" are not relevant to the classification task, yet the model uses them. This reliance on shortcuts can compromise the model's ability to generalize and make accurate predictions in varied contexts.

6. Conclusion

This paper introduces DISCO, a method designed to identify causal rules internalized by neural models in natural language tasks. DISCO produces a concise and statistically robust set of causal rules, enabling users to scrutinize and understand the underlying knowledge captured by the model. The intrinsic causal orientation of our approach ensures that the resultant rules are faithful to the inputs where they are applicable. We demonstrate the efficacy of DISCO by identifying shortcuts learned by prominent models, including $\text{BERT}_{\text{BASE}}$, SBERT, and LSTM. Our approach not only reveals these shortcuts but also provides insights into the model's decision-making process. In essence, DISCO stands as an instrumental resource for those aiming to gain deeper insights into the interactive explainability of AI models.

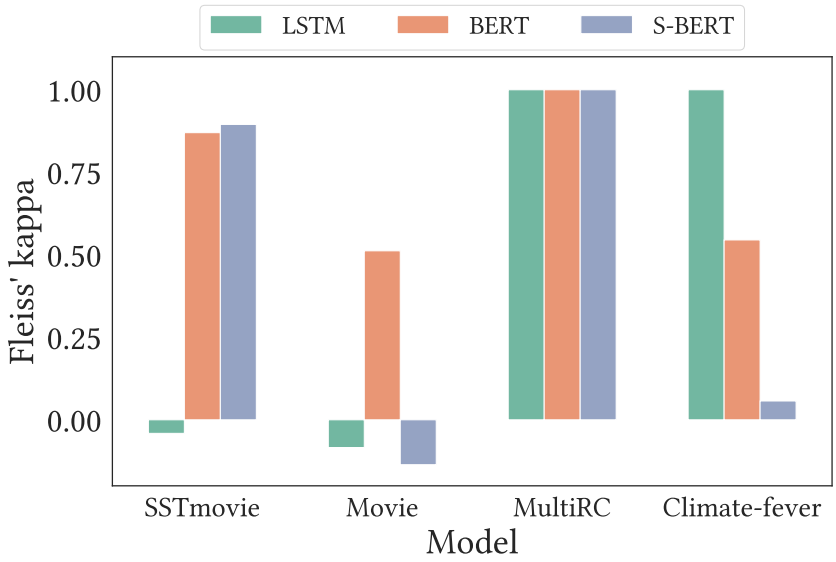

Figure 5: Fleiss' κ among human evaluators considering whether the rules are right for the wrong reasons

7. Limitations

One limitation of our approach arises from the context selection when constructing the counterfactual. Reusing neutral contexts is a straightforward method to generate human-understandable replacements for counterfactual contexts. However, this strategy possesses three inherent limitations:

First, the availability of context is constrained. We only employ contexts present in the training data, limiting the sampling space and potentially compromising the effectiveness of the do-operator. Furthermore, selecting neutral contexts further narrows the sampling space and may introduce discrepancies between the sampled contexts and the training contexts, affecting the data distribution.

Additionally, compared to related works like [19], we do not differentiate between "spurious" and "genuine" reasons for predictions. However, this distinction is of lesser concern as our objective is to identify globally overfit shortcut patterns within the model, rather than pinpointing specific reasons for individual predictions, nor do we care about their faithfulness.

A third limitation concerns the experiments conducted. Although the theory and approach of our work do not require sequence continuity, all experiments are based on consecutive sequences. Exploring efficient methods to identify sequences with gaps or even more complex patterns remains a potential avenue for future research.

References

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

[2] J. Bastings, S. Ebert, P. Zablotskaia, A. Sandholm, K. Filippova, "will you find these shortcuts?" A protocol for evaluating the faithfulness of input salience methods for text classification, CoRR abs/2111.07367 (2021). URL: https://arxiv.org/abs/2111.07367. arXiv:2111.07367.

[3] A. Madsen, S. Reddy, S. Chandar, Post-hoc interpretability for neural nlp: A survey, ACM Computing Surveys 55 (2022) 1–42.

[4] M. Sundararajan, A. Taly, Q. Yan, Axiomatic attribution for deep networks, in: D. Precup, Y. W. Teh (Eds.), Proceedings of the 34th International Conference on Machine Learning, volume 70 of *Proceedings of Machine Learning Research*, PMLR, 2017, pp. 3319–3328. URL: https://proceedings.mlr.press/v70/sundararajan17a.html.

[5] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).
[6] M. T. Ribeiro, S. Singh, C. Guestrin, "why should I trust you?": Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016, 2016, pp. 1135–1144.
[7] T. Lei, R. Barzilay, T. Jaakkola, Rationalizing neural predictions, in: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 107–117. URL: https://aclanthology.org/D16-1011. doi:10.18653/v1/D16-1011.
[8] E. Lehman, J. DeYoung, R. Barzilay, B. C. Wallace, Inferring which medical treatments work from reports of clinical trials, in: Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL), 2019, pp. 3705–3717.
[9] Z. Zhang, K. Rudra, A. Anand, Explain and predict, and then predict again, WSDM '21, Association for Computing Machinery, New York, NY, USA, 2021, p. 418–426. URL: https://doi.org/10.1145/3437963.3441758. doi:10.1145/3437963.3441758.
[10] J. Mu, J. Andreas, Compositional explanations of neurons, Advances in Neural Information Processing Systems 33 (2020) 17153–17163.
[11] M. T. Ribeiro, S. Singh, C. Guestrin, Anchors: High-precision model-agnostic explanations, in: Proceedings of the AAAI conference on artificial intelligence, volume 32, 2018.
[12] J. Lu, L. Yang, B. Mac Namee, Y. Zhang, A rationale-centric framework for human-in-the-loop machine learning (2022). URL: http://arxiv.org/abs/2203.12918, arXiv:2203.12918 [cs].
[13] M. T. Ribeiro, S. Singh, C. Guestrin, Semantically equivalent adversarial rules for debugging nlp models, in: Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: long papers), 2018, pp. 856–865.
[14] D. Kaushik, E. Hovy, Z. Lipton, Learning the difference that makes a difference with counterfactually-augmented data, in: International Conference on Learning Representations, 2020. URL: https://openreview.net/forum?id=Sklgs0NFvr.
[15] Z. Zhang, V. Setty, A. Anand, Sparcassist: A model risk assessment assistant based on sparse generated counterfactuals, SIGIR '22, Association for Computing Machinery, New York, NY, USA, 2022, p. 3219–3223. URL: https://doi.org/10.1145/3477495.3531677. doi:10.1145/3477495.3531677.
[16] S. Chang, Y. Zhang, M. Yu, T. Jaakkola, Invariant rationalization, in: International Conference on Machine Learning, PMLR, 2020, pp. 1448–1458.
[17] E. Wallace, T. Zhao, S. Feng, S. Singh, Concealed data poisoning attacks on NLP models, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 139–150. URL: https://aclanthology.org/2021.naacl-main.13. doi:10.18653/v1/2021.naacl-main.13.
[18] X. Wang, J. Wang, K. Tang, Interpreting deep learning model using rule-based method (2020). URL: http://arxiv.org/abs/2010.07824. doi:10.48550/arXiv.2010.07824, arXiv:2010.07824 [cs].
[19] Z. Wang, A. Culotta, Identifying spurious correlations for robust text classification, in: Findings of the Association for Computational Linguistics: EMNLP 2020, Association for Computational Linguistics, Online, 2020, p. 3431–3440. URL: https://aclanthology.org/2020.findings-emnlp.308. doi:10.18653/v1/2020.findings-emnlp.308.
[20] T. Wang, R. Sridhar, D. Yang, X. Wang, Identifying and mitigating spurious correlations for improving robustness in NLP models, in: M. Carpuat, M.-C. de Marneffe, I. V. Meza Ruiz (Eds.), Findings of the Association for Computational Linguistics: NAACL 2022, Association for Computational Linguistics, Seattle, United States, 2022, pp. 1719–1729. URL: https://aclanthology.org/2022.findings-naacl.130. doi:10.18653/v1/2022.findings-naacl.130.
[21] K. Atwell, A. Sicilia, S. J. Hwang, M. Alikhani, The change that matters in discourse parsing: Estimating the impact of domain shift on parser error, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Findings of the Association for Computational Linguistics: ACL 2022, Association for

Computational Linguistics, Dublin, Ireland, 2022, pp. 824–845. URL: https://aclanthology.org/2022.findings-acl.68. doi:10.18653/v1/2022.findings-acl.68.

[22] G. Ifrim, G. Bakir, G. Weikum, Fast logistic regression for text categorization with variable-length n-grams, in: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, 2008, pp. 354–362.

[23] F. Peng, D. Schuurmans, Combining naive bayes and n-gram language models for text classification, in: ECIR, volume 2633, Springer, 2003, pp. 335–350.

[24] S. Bergsma, E. Pitler, D. Lin, Creating robust supervised classifiers via web-scale n-gram data, in: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, 2010, pp. 865–874.

[25] M. Idahl, L. Lyu, U. Gadiraju, A. Anand, Towards benchmarking the utility of explanations for model debugging, in: Proceedings of the First Workshop on Trustworthy Natural Language Processing, Association for Computational Linguistics, Online, 2021, pp. 68–73. URL: https://aclanthology.org/2021.trustnlp-1.8. doi:10.18653/v1/2021.trustnlp-1.8.

[26] K. Beedkar, R. Gemulla, A. Renz-Wieland, The DESQ framework for declarative and scalable frequent sequence mining, Gesellschaft für Informatik eV, 2019.

[27] K. Beedkar, R. Gemulla, W. Martens, A unified framework for frequent sequence mining with subsequence constraints, ACM Transactions on Database Systems 44 (2019) 1–42. doi:10.1145/3321486.

[28] J. Pearl, Causality, Cambridge university press, 2009.

[29] J.-H. Jacobsen, R. Geirhos, C. Michaelis, Shortcuts: Neural networks love to cheat, The Gradient (2020).

[30] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, E. Teller, Equation of state calculations by fast computing machines, The journal of chemical physics 21 (1953) 1087–1092.

[31] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, arXiv preprint arXiv:1908.10084 (2019).

[32] J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, B. C. Wallace, ERASER: A benchmark to evaluate rationalized NLP models, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4443–4458. URL: https://aclanthology.org/2020.acl-main.408. doi:10.18653/v1/2020.acl-main.408.

[33] R. Socher, A. Perelygin, J. Wu, J. Chuang, C. D. Manning, A. Y. Ng, C. Potts, Recursive deep models for semantic compositionality over a sentiment treebank, in: Proceedings of the 2013 conference on empirical methods in natural language processing, 2013, pp. 1631–1642.

[34] S. MacAvaney, A. Yates, S. Feldman, D. Downey, A. Cohan, N. Goharian, Simplified data wrangling with ir_datasets, in: SIGIR, 2021.

[35] Y. Zhang, K. Song, Y. Sun, S. Tan, M. Udell, "why should you trust my explanation?" understanding uncertainty in lime explanations, 2019. URL: https://arxiv.org/abs/1904.12991. arXiv:1904.12991.

[36] W. B. Croft, D. Metzler, T. Strohman, Search engines: Information retrieval in practice, volume 520, Addison-Wesley Reading, 2010.

[37] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: https://aclanthology.org/N19-1423. doi:10.18653/v1/N19-1423.

[38] C. Molnar, G. Casalicchio, B. Bischl, Interpretable machine learning–a brief history, state-of-the-art and challenges, in: ECML PKDD 2020 Workshops: Workshops of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD 2020): SoGood 2020, PDFL 2020, MLCS 2020, NFMCP 2020, DINA 2020, EDML 2020, XKDD 2020 and INRA 2020, Ghent, Belgium, September 14–18, 2020, Proceedings, Springer, 2021, pp. 417–431.

[39] A. Anand, L. Lyu, M. Idahl, Y. Wang, J. Wallat, Z. Zhang, Explainable information retrieval: A survey, arXiv preprint arXiv:2211.02405 (2022).