Exploring the Trade-off Between Model Performance and Explanation Plausibility of Text Classifiers Using Human Rationales

Lucas E. Resck[1] and **Marcos M. Raimundo**[2] and **Jorge Poco**[1]
[1]Fundação Getulio Vargas, Rio de Janeiro, Brazil
[2]Universidade Estadual de Campinas (UNICAMP), Campinas, Brazil
lucas.resck@fgv.br, mraimundo@ic.unicamp.br, jorge.poco@fgv.br

Abstract

Saliency post-hoc explainability methods are important tools for understanding increasingly complex NLP models. While these methods can reflect the model's reasoning, they may not align with human intuition, making the explanations not plausible. In this work, we present a methodology for incorporating rationales, which are text annotations explaining human decisions, into text classification models. This incorporation enhances the plausibility of post-hoc explanations while preserving their faithfulness. Our approach is agnostic to model architectures and explainability methods. We introduce the rationales during model training by augmenting the standard cross-entropy loss with a novel loss function inspired by contrastive learning. By leveraging a multi-objective optimization algorithm, we explore the trade-off between the two loss functions and generate a Pareto-optimal frontier of models that balance performance and plausibility. Through extensive experiments involving diverse models, datasets, and explainability methods, we demonstrate that our approach significantly enhances the quality of model explanations without causing substantial (sometimes negligible) degradation in the original model's performance.[1]

1 Introduction

The complexity of text classification models and architectures has recently grown, posing challenges in comprehending the rationale behind their decisions. Consequently, the latest NLP algorithms have been called *black-box* algorithms. Understanding the model's reasoning is essential in various text classification contexts (Ribeiro et al., 2016) (e.g., hate speech detection). However, this task is hindered by the black-box nature of these models. Moreover, comprehending the model's reasoning can help establish trust and make informed decisions based on the underlying justifications.

[1]Code and data are available at https://github.com/visual-ds/plausible-nlp-explanations.

(a) This is such a great movie !
(b) This is such a great movie !

Figure 1: Examples of local saliency post-hoc explanations from a hypothetical text classifier for a positive movie review. Explanation (a) is more *plausible* than (b). Green means a positive contribution to the model's prediction, and red is negative.

Researchers have developed popular text classification explainability techniques, such as post-hoc local saliency (or heatmap) methods (Tjoa and Guan, 2022; DeYoung et al., 2020). These methods generate heatmaps over tokens (paragraphs, sentences, words, sub-words, or characters) to indicate their significance in the final decision (Ribeiro et al., 2016; Lundberg and Lee, 2017; Chefer et al., 2021) — although their suitability is criticized (Bilodeau et al., 2024), these methods are still widely applied (Kumari et al., 2024). The estimation of importance is performed after the decision has been made using an already trained model (i.e., it is post-hoc). For instance, Figure 1 illustrates word-level saliency explanations that justify the predictions of two trained models in determining whether a movie review is positive or negative. In explanation (a), highlighted in green, the most relevant words align well with human expectations, making it intuitive. However, in explanation (b), the highlighted words are irrelevant from a human perspective. Both explanations may accurately reflect the models' reasoning (thus, they may be *faithful*, according to DeYoung et al., 2020). Nevertheless, they differ in *plausibility*, which refers to the extent to which the explanation matches human intuition (DeYoung et al., 2020) or is "convincing of the model prediction" (Jacovi and Goldberg, 2021).

Ideally, we should be able to enhance the plausibility of a "non-plausible" model by "teaching" it to provide more plausible explanations. Previous works, such as those by Strout et al., 2019; Ross et al., 2017; Arous et al., 2021; Du et al., 2019; Mathew et al.,

Findings of the Association for Computational Linguistics: NAACL 2024, pages 4190–4216
June 16-21, 2024

2021, have explored this concept. The reason is that someone training the model clearly understands what a valid explanation should entail. However, achieving plausibility while preserving *faithfulness* may require modifying the reasoning of the original model, which in turn risks impacting its performance on the test data. Hence, an inherent trade-off exists between model performance and explanation plausibility (Zhang et al., 2021; Plumb et al., 2020).

This paper introduces a methodology that enhances the plausibility of explanations while remaining agnostic to the model architecture and explainability method. Our approach incorporates human explanations, represented as *rationales* (i.e., text annotations serving as ground truth for explanations), into text classification models using a novel contrastive-inspired loss. We address the trade-off between classification and the new loss within a multi-objective framework, enabling exploration of the balance between performance and plausibility. Unlike other approaches, our methodology does not require modifying the model architecture (e.g., through the addition of attention mechanisms; Strout et al., 2019) or assuming a specific type of explanation function (e.g., a differentiable explanation function; Rieger et al., 2020) to incorporate the explanations.

In summary, our contributions are:

(i) A proposal of a novel contrastive-inspired loss function that effectively incorporates rationales into the learning process.

(ii) A multi-objective framework that automatically assigns weights to the learning loss and contrastive rationale loss, offering multiple trade-off options between performance and explanation plausibility.

(iii) A series of experiments using various models, datasets, and explainability methods, demonstrating the significant enhancement of model explanations without compromising (and sometimes without any detriment to) the model's performance. Notably, our approach exhibits particularly improved plausibility for samples with incorrect explanations.

We compare our methodology with a previous method from the literature, reinforcing our results. Furthermore, we address the social and ethical implications of "teaching" explanations to text classification models. We argue that these concerns are mitigated when the explanations remain faithful to the model's decision-making process.

2 Related Work

Our work draws on prior research in the areas of rationale utilization and the trade-off between performance and explainability.

Use of Rationales. Using human annotations to assist machine learning is not a novel concept, as prior works have shown (Zaidan et al., 2007, 2008). Nevertheless, there has been a recent surge in interest in machine learning explainability and fairness, leading to an increased focus on collecting and applying such rationales. Some studies have leveraged rationales to enhance model fairness (Rieger et al., 2020; Liu and Avci, 2019), while others have explored techniques to extract (Zhang et al., 2021; Lakhotia et al., 2021; Pruthi et al., 2020; Sharma et al., 2020) or generate (Rajani et al., 2019; Liu et al., 2019; Camburu et al., 2018; Kumar and Talukdar, 2020) model explanations. The most prevalent application of rationales lies in performance improvement, where annotations serve as valuable assistants during the learning process, particularly in tasks involving textual data (Sharma and Bilgic, 2018; Bao et al., 2018; Liu et al., 2019; Rieger et al., 2020; Zhang et al., 2021; Arous et al., 2021; Mathew et al., 2021; Carton et al., 2022; Ghaeini et al., 2019; Huang et al., 2021), images (Simpson et al., 2019; Rieger et al., 2020; Mitsuhara et al., 2021), or tabular data (Belém et al., 2021). In this work, our focus revolves around the incorporation of rationales during model training to "teach" explanations, drawing inspiration from the findings of Arous et al. (2021); Du et al. (2019); Mitsuhara et al. (2021). In particular, Mathew et al. (2021) collect and annotate a dataset called HateXplain and use its annotations to train a model. Moreover, the UNIREX framework (Chan et al., 2022) extends this approach to a more general setting.

Importantly, our approach refrains from altering/assuming the model architecture (e.g., by using another model for rationale extraction (Chan et al., 2022), assuming a model architecture (Mathew et al., 2021), or adding another layer (Strout et al., 2019; Chen and Ji, 2020; Liu et al., 2022; Sekhon et al., 2023)) or assuming a specific type of explanation function (e.g., by using input gradients; Ross et al., 2017; Ghaeini et al., 2019). Such interventions are debatable (see Section 6) and not always possible. Instead, we adopt a model- and explainer-agnostic approach, using rationales to enhance the plausibility of explanations. Noticeably, our approach also differs from previous work that rationalizes the input but

does not leverage human annotations (Lei et al., 2016; Bastings et al., 2019; Jain et al., 2020).

Performance and Explainability Trade-off. The existence of a trade-off between machine learning performance and interpretability/explainability is widely debated in the field. Several studies have discussed this trade-off (Camburu et al., 2018; Swanson et al., 2020; Dubey et al., 2022; Plumb et al., 2020; Radenovic et al., 2022). However, differing opinions exist on whether this trade-off always holds, both from a theoretical perspective (Jacovi and Goldberg, 2021; Rudin, 2019) and a practical standpoint (Hase et al., 2020). Furthermore, some studies have empirically examined or explored this trade-off (Zhang et al., 2021; Goethals et al., 2022; Naylor et al., 2021; Paranjape et al., 2020; Jin et al., 2006). Our work shares similarities with the study conducted by Belém et al. (2021), as we aim to employ two distinct learning strategies and investigate their trade-offs. However, our approach utilizes different learning strategies, and we conduct the trade-off exploration using a multi-objective optimization algorithm.

3 Theoretical Background

We define crucial explainability and multi-objective optimization concepts to facilitate a global understanding of our research. We also point to an overview of contrastive learning in Appendix C.

3.1 Explainability

Rationale. In the context of text classification, a *rationale* refers to a snippet extracted from a source text that supports a specific category (DeYoung et al., 2020; Carton et al., 2022; Mathew et al., 2021). Typically, these rationales are annotated by humans and serve as ground truth explanations for the corresponding categories. For instance, in Figure 1, a typical rationale for the positive class would be "great movie."

Explanation Plausibility. The *plausibility* of a model explanation refers to the extent to which it aligns with human intuition (DeYoung et al., 2020) or is considered "convincing of the model prediction" (Jacovi and Goldberg, 2021). In practice, this plausibility can be measured by evaluating the agreement between the explanation and the ground truth rationale (DeYoung et al., 2020; Jacovi and Goldberg, 2021). Please refer to Section 6 for a detailed discussion on the pursuit of plausibility.

Explanation Faithfulness. Another crucial aspect of an explanation is its *faithfulness*, which reflects the degree to which the model relies on the explanation to make its prediction (DeYoung et al., 2020). Following the approach of DeYoung et al. (2020), we employ the metrics of *comprehensiveness* and *sufficiency* to quantify faithfulness. We multiply sufficiency by -1 to indicate that a higher value is desirable for both metrics.

3.2 Multi-objective Optimization

We aim to investigate the trade-off between model performance and explanation plausibility. Section 4.3 addresses this trade-off exploration by concurrently optimizing two distinct loss functions that may have conflicting objectives. We adopt the definitions that Raimundo et al. (2020) provided for the following concepts.

Definition 3.1 (Multi-objective optimization problem)**.** A *multi-objective optimization problem* (MOO) is an optimization problem with more than one objective, i.e., a problem of the form

$$\begin{aligned} \min_{x} \quad & f(x) = (f_1(x), \cdots, f_m(x)), \\ \text{subject to} \quad & x \in \Omega \subseteq \mathbb{R}^n, \ f: \Omega \to \mathbb{R}^m, \ f(\Omega) = \Psi. \end{aligned}$$

Consider two solutions $x_1, x_2 \in \mathbb{R}^n$ where $f_1(x_1) < f_1(x_2)$ and $f_2(x_1) > f_2(x_2)$. In this case, no clear optimal solution exists. To address this, we introduce the concept of *Pareto-optimality*.

Definition 3.2 (Pareto-optimality)**.** A solution $x^* \in \Omega$ is *Pareto-optimal* if there is no other solution $x \in \Omega$ such that $f_i(x) \leq f_i(x^*)$ for all i and $f_i(x) < f_i(x^*)$ for some i.

The Pareto-frontier comprises objective function values resulting from Pareto-optimal solutions. Without considering additional criteria, there is no definitive best solution among them. The decision-maker holds the responsibility of selecting the desired solution. While solving a MOO problem poses challenges, various approaches are available. Refer to Appendix A for an overview of the *weighted sum method* and their theoretical foundations.

4 Methodology

We focus on text classification models to enhance the quality of local saliency post-hoc explanations regarding *plausibility*. We aim to align these explanations with human intuition while maintaining *faithfulness*. To achieve this, we leverage *rationales* to enhance the explanation quality and evaluate the improvement by comparing them with the model explanations.

4.1 Notation Description

Consider a multi-class text classification task with classes C and a multi-class text classification model $f_\theta : \mathbb{R}^d \to \Delta$. The model takes a text $x \in \mathbb{R}^d$ and produces a probability vector $f_\theta(x) \in \Delta$, indicating the probabilities of x belonging to each class, with parameters θ. Examples of x include TF-IDF vectors (Leskovec et al., 2020), BERT feature vectors (Devlin et al., 2019), or word presence vectors (e.g., Transformer's "input id" array; Vaswani et al., 2017). We view f_θ as a black box without assuming any specific structure. Let us introduce the explanation function[2] $e_{f_\theta,k} : \mathbb{R}^d \to \mathbb{R}^p$, which assigns a score to each token in x, representing its contribution to the $f_\theta(x)$ prediction for class $k \in C$, i.e., $f_\theta(x)_k$. We also have ground-truth human annotations (*rationale*) as a binary vector $e_{x,k} \in \{0,1\}^p$, indicating the essential tokens for x to be classified as class k. The measure of agreement $m : \mathbb{R}^p \times \{0,1\}^p \to \mathbb{R}$ between $e_{f_\theta,k}(x)$ and $e_{x,k}$ quantifies the quality of explanations extracted from f_θ compared to canonical explanations, reflecting their plausibility. Given a set $X = \{X_1, \cdots, X_N\}$ of training texts and a set $y = \{y_1, \cdots, y_N\}$ of training class labels, the commonly used cross-entropy loss is employed during training, defined as:

$$\mathcal{L}_\theta(X,y) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{|C|}\mathbb{1}_{y_i=k}\ln\frac{e^{g_\theta(X_i)_k}}{\sum_{j=1}^{|C|}e^{g_\theta(X_i)_j}}, \quad (1)$$

where, g_θ represents the logits (pre-softmax) obtained from f_θ, and f corresponds to the softmax function applied to g_θ. It is worth noting that θ can represent the training weights of a linear function (in the case of multinomial logistic regression) or a more complex function, such as a neural network.

4.2 Contrastive Rationale Loss

To enhance the plausibility of model explanations, we incorporate rationales into the model training process. Unlike previous approaches (Rieger et al., 2020; Du et al., 2019; Ross et al., 2017), we do not utilize an explanation-based function in the loss function to compare model explanations with ground truth explanations. Instead, we construct a loss function for training the text classification model using a modified dataset $\dot{X} = \{\dot{X}_1, \cdots, \dot{X}_N\}$. During training, we replace the full-text $X_i \in \mathbb{R}^d$ with the rationale text $\dot{X}_i \in \mathbb{R}^d$. By exclusively teaching the model with rationales, we expect them to become the primary basis for the model's decision-making process, leading to correspondingly reflected model explanations[3].

In a more general context, $\dot{X}$ may encompass rationales from a subset or superset of texts in X, or even both. In this scenario, $\dot{y}$ denotes the labels of $\dot{X}$. Drawing inspiration from the contrastive learning domain (Chen et al., 2020; Khosla et al., 2020), we introduce a novel auxiliary loss function known as the *contrastive rationale loss*:

$$\dot{\mathcal{L}}_\theta(\dot{X},\dot{y}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{|C|}\mathbb{1}_{\dot{y}_i=k}\ln\frac{e^{g_\theta(\dot{X}_i)_k}}{\sum_{j=1}^{m}e^{g_\theta(\tilde{X}_{i,j})_k}}, \quad (2)$$

where $\{\tilde{X}_{i,j}\}_{j=1}^{m}$ is a set of m *sample rationales* of X_i, i.e., rationales that may be or may be not a ground truth explanation for X_i. For instance, this set includes the ground truth explanation $\dot{X}_i$ and other $m-1$ random rationales, which we call *negative rationales* — random tokens of X_i uniformly sampled. The numerator seeks to maximize the model's output for the rationale in the correct class. At the same time, the denominator aims to minimize the model's output for the random (negative) rationales in the same class. Notice that we do not include the explanation function $e_{f_\theta,k}$ (Section 4.1) in Equation 2, contrary to previous work (Section 2). This is because we do not want to "train the explainer" or "teach the model how to tweak the explainer." For an in-depth discussion, see Section 6.

The contrastive rationale loss constitutes a particular case when the classifier is a multinomial logistic regression. Further details can be found in Appendix B.

4.3 Trade-off Exploration

Section 4.2 proposes an auxiliary *contrastive rationale loss* function $\dot{\mathcal{L}}_\theta$ to incorporate rationales during model training. The simultaneous optimization of both cross-entropy $\mathcal{L}_\theta$ and $\dot{\mathcal{L}}_\theta$ gives rise to a *multi-objective optimization* (MOO) problem (see Section 3.2). It is important to note that optimizing both objectives without a trade-off is not feasible. We leverage existing MOO algorithms to explore the trade-off between model performance and explanation plausibility (Cohon, 1978).

In simple terms, MOO solvers such as NISE (Cohon, 1978), employing the weighted sum method

[2] d refers to the dimension of the text vector space (e.g., BERT's 768), and p is the number of tokens of a sample.

[3] In this formulation, we assume the explanation function is perfectly faithful, i.e., the explanation results genuinely reflect the model's reasoning. Such a function is not apparent; however, our experimental results suggest that the explainability methods we have access to are sufficient.

(Appendix A), enable trade-off exploration by incorporating hyperparameters w_1 and w_2 (both ≥ 0) with $w_1 + w_2 = 1$, and solving the uni-objective problem:

$$\mathcal{L}_\theta(X,y,\dot{X},\dot{y}) = w_1 \cdot \mathcal{L}_\theta(X,y) + w_2 \cdot \dot{\mathcal{L}}_\theta(\dot{X},\dot{y}).$$

Intuitively, the weight vector $\mathbf{w} = [w_1, w_2]$ controls the trade-off between model performance (original cross-entropy loss) and explanation plausibility (contrastive rationale loss). Increasing w_2 from 0 to a positive value explicitly assigns more weight to the contrastive rationale loss. This indicates that the model is trained on data $(\dot{X}, \dot{y})$ that differs from the underlying distribution of (X, y). Consequently, the model's performance on test data, which follows the same distribution as (X,y), is expected to decline. However, since we fit the model using rationales, we alter the model's reasoning, emphasizing the significance of positive rationales within the texts. This emphasis should be reflected in the explanations, as argued in Section 4.2 and demonstrated in our experiments.

MOO solvers like NISE effectively sample representative sets W_1 and W_2 of trade-off parameters w_1 and w_2. From the loss optimization process (e.g., `lbfgs`, SGD, Adam, etc.), these sets yield a set of model weights Θ, where each $\theta \in \Theta$ corresponds to a different classifier $f_\theta \in F_\Theta$. Finally, by searching within the set F_Θ, we can identify Pareto-optimal models that exhibit both performance and plausibility.

5 Experiments

This section describes experiments to test the methodology proposed in Section 4, employing diverse models, datasets, and explainability techniques. We aim to verify the usefulness of the contrastive rationale loss (Section 4.2) in incorporating human rationales and the effectiveness of the MOO solver (Section 4.3) in finding models that well-represent the Pareto-frontier. Furthermore, we also compare our methodology with previous work. Implementation and execution information can be found in Appendix E.

5.1 Models

To evaluate the effectiveness of our method, we assess two types of models: language models and classic NLP models.

DistilBERT and BERT-Mini. As language model representatives, we test DistilBERT (Sanh et al., 2020) and BERT-Mini (Turc et al., 2019), lightweight versions of the popular BERT (Devlin et al., 2019). For fine-tuning on the HateXplain dataset, refer to Appendix D. Refer to Appendix F for an additional analysis with BERT-Large.

TF-IDF with Logistic Regression. For classical models, we train a multinomial logistic regression model using TF-IDF vectors (Leskovec et al., 2020) (unigrams) with dimensionality reduction to 200 achieved through Truncated Singular Value Decomposition (Manning et al., 2008).

5.2 Datasets and Data Preprocessing

HateXplain. This dataset contains annotated hate speech detection samples with human-annotated rationales (Mathew et al., 2021). It consists of three classes: normal (without rationales), offensive, and hate speech. To address the confounding correlation between offensive and hate speech classes and their rationales, we simplify the dataset by excluding the offensive class (`hatexplain` dataset). We also explore a version including all labels (`hatexplain_all` dataset). Hereafter, "HateXplain" refers to `hatexplain` unless specified otherwise.

Twitter Sentiment Extraction (TSE). The TSE (Maggie et al., 2020) is a sentiment analysis dataset containing positive, negative, and neutral tweets with human-annotated rationales. Since neutral class lacks rationales[4], we simplify the classification, excluding this class (`tse` dataset). An alternative version includes all labels (`tse_all` dataset). Hereafter, "TSE" refers to `tse` unless specified otherwise.

Movie Reviews. This dataset comprises positive and negative movie reviews with rationales annotated by humans to support classification (Zaidan et al., 2007).

5.3 Explainability Methods

We utilize two well-known explainers for generating continuous salient maps in textual datasets.

LIME. Short for *Local Interpretable Model-agnostic Explanations* (Ribeiro et al., 2016), it creates post-hoc explanations by randomly removing tokens from the text sample and locally approximating the original model predictions using a simpler, interpretable model, which is used to explain the sample's prediction.

SHAP. *SHapley Additive exPlanations* (Lundberg and Lee, 2017) is a model-agnostic explainer that employs Shapley values to explain model predictions.

[4]TSE neutral class rationales exist but are uninformative because they are the whole sample text in most cases.

(a) ugh i hate d*kes

(b) ugh i hate d*kes

Figure 2: Examples of explanations of the hate speech class. Explanation (a) is from the original model, and (b) is from the model with top-AUPRC. Green means a positive contribution to the model's prediction. The top-1 token was selected for visualization purposes. More examples in Table 6.

5.4 Explainability Metrics

Plausibility. We employ the *Area Under the Precision-Recall Curve (AUPRC)* metric to assess the plausibility of model explanations generated by LIME and SHAP. This metric is constructed by varying the threshold over continuous token scores and calculating precision and recall at the token level (DeYoung et al., 2020).

Faithfulness. We require discrete explanations to evaluate *comprehensiveness* and *sufficiency* (as described in Section 3.1). To address this, we consider the top 1, 5, 10, 20, and 50% of tokens and average the results, which we refer to as the *Area Over the Perturbation Curve (AOPC)* (DeYoung et al., 2020).

5.5 DistilBERT and HateXplain

In this section, we present experimental results to tackle the following research questions: *Does the proposed loss improve explanation plausibility without affecting the performance? Does the MOO solver effectively assist in finding a model with better explanations?* We first present a case study with the DistilBERT model and HateXplain dataset to showcase the main results of our experiments. Section 5.6 shows other results. The explainability metrics (plausibility and faithfulness) are computed only for the hate speech class because the normal class lacks rationales.

The DistilBERT model trained only with cross-entropy loss achieves a test accuracy of 84.8% with balanced recall among classes. Figure 2 (a) illustrates an example of a bad explanation extracted from this model. It shows that even high-performing classifiers can also present unreasonable explanations.

We employ NISE (Cohon, 1978) to find 30 models that well-represent the Pareto-frontier using the cross-entropy and the contrastive rationale loss (using 2 random, negative rationales) on the training data. Figure 3 (a) reveals that the two losses are conflicting, particularly for non-extreme values of w_1.

For each model in the frontier, we evaluate the model's performance and the explanation plausibility

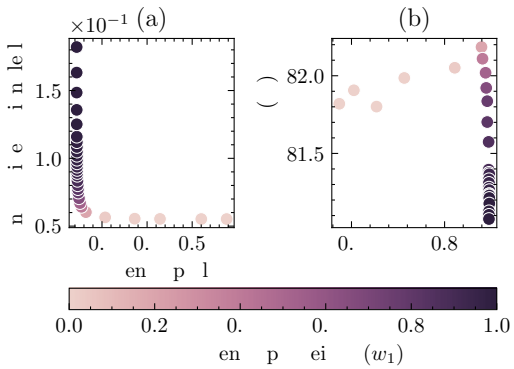

Figure 3: (a) Trade-off between the two losses on the training data. (b) Trade-off between accuracy and plausibility of the test data. The color scale represents the cross-entropy weight w_1 (Section 4.3). We ignore the model with $w_1 = 0$ as it is out of scale. Results including $w_1 = 0$ and shared scale between axes are in Appendix F.

on the test data (Figure 3 (b)). Plausibility was measured using mean AUPRC, comparing LIME's explanations with ground truth rationales. Figure 3 (b) shows that, as NISE increases the weight of the contrastive rationale loss during training, the plausibility increases almost without hurting performance: the top-plausibility model had a relative increase of 1.4% in AUPRC (an absolute increase of 1.1%), despite a relative decrease of 0.9% in accuracy (an absolute decrease of 0.8%). At some point, performance and explanation quality deteriorate, given that the training without the cross-entropy is meaningless. We noticed that around 51% of the best-explained samples originally had AUPRC equal to 1. By disregarding these samples, the AUPRC relative increase becomes 5.3% (absolute increase becomes 3.3%). At the same time, the high AUPRC explanations have a relative and absolute decrease of less than 1% (Figure 7). The inadequate explanations are being improved without significantly harming the good explanations (see example in Figure 2; more examples in Table 6).

Finally, we must guarantee faithful explanations (i.e., they genuinely represent the models' reasoning) when we strengthen the training with rationales. Figure 4 presents the trade-off between performance and explanation faithfulness on test data. Sufficiency tends to increase as we strengthen the training with rationales, while comprehensiveness tends to decrease. However, the explanations are becoming more sufficient without significantly losing comprehensiveness (sufficiency's variation is an order of magnitude higher than the comprehensiveness').

In summary, the results present a desirable scenario in which *one trades-off a small decrease in accuracy*

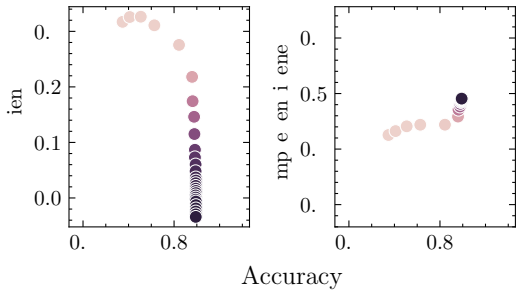

Figure 4: Trade-off between accuracy and faithfulness (sufficiency and comprehensiveness) on test data. Higher values are better. The color scale is the same as the previous figures. The data scale is equal between the two graphics and their x- and y-axes.

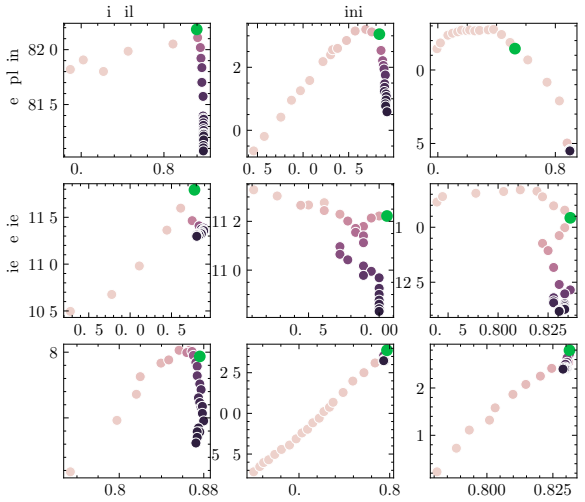

Figure 5: Trade-offs between performance (accuracy, x-axis) and plausibility (AUPRC, y-axis, in percentage (%)) for all models and datasets (test data). There are 2 random (negative) rationales, and the explainer is LIME. Green dots are the models chosen to be analyzed more carefully. The color scale is the same as the previous figures. We ignore the model with $w_1 = 0$ in all graphics as it is out of scale. Larger figure and results including $w_1 = 0$, 5 rationales and/or SHAP, shared scale between axes, and Pareto-frontiers are in Appendix F.

for a reasonable increase in explainability quality (both plausibility and sufficiency), especially for originally bad explanations. *The MOO solver effectively assists in finding a model with better explanations.*

5.6 Experiments With All Models and Datasets

Now, we evaluate our framework in all models, datasets, and explainability techniques that we consider in this paper. Specifically, we aim to discover *whether the previous results (usefulness of the contrastive loss and effectiveness of the MOO solver) extend to the general case.* Figure 5 overviews all performance vs. plausibility trade-offs on test data. The number of random (negative) rationales used is 2, and the explainer is LIME. To comprehend its effect, we also test with 5 rationales and/or explainer SHAP (Appendix F). Figure 5 shows a non-constant shape of the final frontier across all experiments. For instance, while TF-IDF trades accuracy for plausibility in the HateXplain dataset, it increased both dimensions in TSE. However, the shape is the same when changing the number of negative rationales (Figure 16) and similar when the explainer is SHAP (Figures 17 and 18). Finally, despite the TSE dataset having a higher number of poor-performing models, the improvement for a well-selected model is not negligible (Table 1).

The green dots in Figure 5 represent the models manually selected as "good choices" of the trade-off between performance and plausibility. We analyzed them more carefully and compared them to the original models (i.e., $w_1 = 1$, darkest point on the figures). For example, the green dot of DistilBERT with HateXplain is an obvious choice because it improves AUPRC without harming performance. Conversely, TF-IDF with HateXplain trades one metric for the other. Thus, a few dots were chosen with some degree of "good judgment." Table 1 compares the original and selected models. All models improved the plausibility of their explanations, in some cases marginally (as for the TSE dataset). The accuracy generally varies slightly, positive and negative, except for a significant drop of TF-IDF with HateXplain. Finally, sufficiency is generally positive, with significant improvements for the language models. At the same time, the comprehensiveness is usually negative but an order of magnitude smaller than the improvements in sufficiency. Results for SHAP and 5 negative rationales are in Table 8 and, because the trade-off shapes of Figures 5, 16, 17 and 18 are similar, they present similar conclusions, showing the robustness of our framework for different explainers and number of rationales. For examples of explanation improvement, refer to Tables 6 and 7.

In general, *all models improve their explanation quality in plausibility (and the majority of them in sufficiency, too) without harming the performance significantly*, showing the robustness of our framework. *The multi-objective exploration was essential to find the best trade-offs.* Conclusions are similar for non-binary classification (see Appendix F).

Table 1: Comparison between the original model (cross-entropy only) and the chosen model (green dots on Figure 5) for each performance and explainability metric on test data. "rel." means relative variation. The column w_1 indicates the weight w_1 of the chosen model's cross-entropy loss during training. Number of negative rationales is 2, and the explainer is LIME. A complete table (with 5 negative rationales and/or SHAP) is available in Appendix F.

Dataset	Model	w_1	Acc. %	AUPRC %	AUPRC rel. %	Suff.	Comp.
HateXplain	DistilBERT	0.20	-0.80	1.11	1.37	0.25	-0.03
	BERT-Mini	0.29	-0.84	2.46	3.49	0.40	-0.05
	TF-IDF	0.002	-9.35	6.96	10.79	0.13	-0.10
Movie Reviews	DistilBERT	0.12	-0.28	0.50	4.39	0.25	-0.05
	BERT-Mini	0.26	0.28	0.39	3.61	0.00	-0.02
	TF-IDF	0.09	0.56	0.85	6.95	0.00	0.01
TSE	DistilBERT	0.64	0.09	1.32	1.98	0.05	0.00
	BERT-Mini	0.19	0.37	0.64	1.01	0.06	0.01
	TF-IDF	0.42	0.24	0.40	0.64	0.01	-0.02

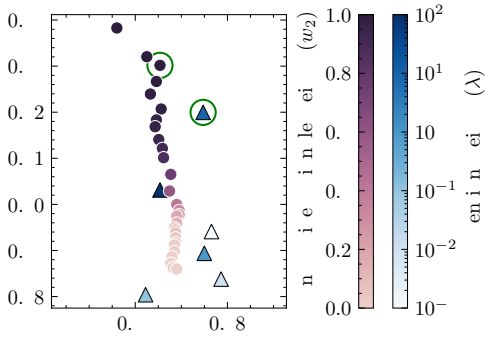

Figure 6: Comparison between BERT-HateXplain (▲) and our methodology (●) on test data. Number of negative rationales is 2 for our method. Color scales indicate the explanation weights λ (for HateXplain, log scale) and w_2 (for our method). As usual, we ignore the model with $w_2 = 1$ as it is out of scale. Circled points are the chosen models for each method to be analyzed more carefully. Data scale is equal between x- and y-axes.

Table 2: Comparison between the chosen models (circled points in Figure 6) of BERT-HateXplain and our method on test data. Accuracy and AUPRC are in percentage (%).

Model	Acc.	AUPRC	Suff.	Comp.
HateXplain	**67.47**	72.00	0.12	**0.53**
Ours	66.54	**73.02**	**0.14**	0.40

5.7 Methodology Comparison

In HateXplain's paper (Mathew et al., 2021), the authors test their dataset by proposing BERT-HateXplain, a BERT version incorporating the rationales as an additional input. They incorporate the annotations using a novel loss function over the attention weights of the last layer of BERT[5], which is a particular case of the UNIREX framework (Chan et al., 2022) . We compare our methodology with the BERT-HateXplain model, using the same dataset (`hatexplain_all`), model (`bert-base-uncased`), and explainer (LIME), and setting the number of random (negative) rationales to 2.

Figure 6 presents the trade-off between accuracy and plausibility (mean AUPRC) on test data for BERT-HateXplain and our methodology after optimization on training data. For BERT-HateXplain, we use the suggested hyperparameters from their paper (Mathew et al., 2021). The shape of our curve is similar to the other experiments involving language models. BERT-HateXplain has a less stable curve because their model training is stochastic, while our methodology is deterministic (Section 4.3). The circled dots are the chosen models using a "good judgment" of improving AUPRC without hurting too much accuracy. Table 2 compares the selected models for each method. Our methodology has better plausibility, while BERT-HateXplain has better accuracy. Additionally, our methodology has better sufficiency, while BERT-HateXplain has better comprehensiveness. These results align with the canonical BERT-HateXplain results (Mathew et al., 2021) in their absolute values and conclusion: they improve performance and comprehensiveness while decreasing sufficiency. Importantly, our method does not require any assumption of model architecture, while BERT-HateXplain does. This comparison expands the results of the other experiments, showing that our methodology can trade a little of performance to improve explanation quality (by improving plausibility while keeping faithfulness) in a model-agnostic approach.

[5]Their attention loss is multiplied by a "trade-off" hyperparameter λ. We use their suggestion of λ values (Appendix E).

5.8 Further Experiments

We performed additional experiments to assess our methodology further (Appendix F). We found that the performance of our method for larger models is similar to other experiments and that we can improve out-of-distribution performance.

6 Discussion

Should We Model Plausibility? Jacovi and Goldberg (2021) argue that explanation plausibility should not be pursued because it is an ethical issue: the explainer would pursue convincing the user of the model decision, possibly providing unfaithful justifications. Our perspective is different: the explainer is never adjusted to convince the user (the model explainer is not "trained" with rationales, and the model does not learn how to tweak the explainer). Instead, we update the model's internal decision, aiming for better explanations. Our perspective is more aligned with Zhou et al. (2022) who defends that plausibility contributes to *understandability*: "given the same level of correctness, a higher-alignment explainer may be preferable" (Zhou et al., 2022).

Is There Really a Trade-Off? The hypothesis of this work is the existence of a trade-off between model performance and explanation plausibility. This happens because, once we fix the model's architecture, it is impossible to promote more alignment with the rationales without changing its optimal. The Pareto frontier in Figure 19 clearly shows that there is not any model that is better than all the others in both metrics (exceptionally for one case), further indicating the presence of a trade-off in its classic sense. Section 2 presents references that argue both in favor and against in the debate of the existence of a trade-off. This work contributes to this debate by proposing an explicit trade-off formulation (Equations 1 and 2) and experiments exploring the existence of this trade-off.

Model and Explainer Agnosticism. Our approach claims to be model- and explainer-agnostic because we only influence the training procedure by adding another loss function that incorporates the rationales. We do not specify model type (Strout et al., 2019; Mathew et al., 2021) or ask for a specific type of explanation function (Rieger et al., 2020).

Light Hyperparameter Search. The trade-off is explored using a MOO solver to identify optimal weights. Model training is confined to the classification layer, akin to training logistic regression in the latent space (see Appendix E). Inference across the language model occurs just once. This approach eliminates the need for fine-tuning, rendering the optimization process both convex and expedient.

Data Distribution Shift. The introduction of rationales, with a decurrent performance drop, can be interpreted as a data distribution shift. To limit its effect on the performance, we keep the original classification loss and find the right balance between explanation plausibility and performance drop.

Other Benefits. To change the shortcuts that neural networks explore to perform tasks, it is necessary to update most, if not all, of the model's weights. Despite our work training weights of the final layer only, we believe that reducing network shortcuts with our method should be explored in future work. Training models to have more plausible reasoning can decrease biases, improving users' trust. In future work, we intend to perform a large-scale user trust evaluation.

Datasets Diversity. We explored a diverse set of datasets used in the literature (Mathew et al., 2021; Atanasova et al., 2020). They vary in text and rationale length, text distribution, and number of classes (Appendix F). They include complex and ambiguous rationales (e.g., Movie Reviews) and those with nuanced classification categories, such as the "offensive" and "hatespeech" classes in HateXplain (Table 4).

7 Conclusion

We propose a novel approach for enhancing the explanation plausibility of text classification models by incorporating human rationales, which capture human knowledge. Our method is model-agnostic and explainability method-agnostic, making it compatible with various model architectures and explainers. We introduce a new contrastive-inspired loss function that integrates the rationales into the learning process. We demonstrate the feasibility of finding models that achieve a trade-off between improved plausibility and a minimal or negligible decrease in model performance. A comparative analysis establishes the superior effectiveness of our approach in enhancing plausibility while maintaining faithfulness and model agnosticism. We validate our method using a diverse set of explainers, datasets, and models encompassing modern and traditional NLP models. Furthermore, we envision the potential extension of our approach to accommodate other explainers, datasets, and models, offering a seamless pathway to enhancing the plausibility of text classification algorithms.

Limitations

Model Agnosticism. The employed multi-objective optimization (MOO) solver, NISE, demands convex objective functions. We claim our method is agnostic to any classification model, and this is true. However, when dealing with models that do not satisfy the convexity condition, e.g., complex neural networks, one should employ other MOO algorithms. To circumvent this limitation with the language models, we trained only the classification layer or first fine-tuned the model with cross-entropy loss (Appendix E).

DistilBERT and BERT-Mini. DistilBERT and BERT-Mini, as they are Transformer encoder-based models, do not scale to long texts because of the limited input size. We did not approach this limitation in this work, and we plan this for future work. For our long text dataset, Movie Reviews, we truncated the text to the input size of the model, which may have impacted the results.

Larger Datasets. To the best of our knowledge, there is a limitation in the literature regarding the availability of large classification textual datasets with human annotations in the sentence/phrase/word/token level (Wiegreffe and Marasovic, 2021). Other tasks, such as natural language inference (Camburu et al., 2018), are out of the scope of this work. Conducting large dataset annotations is intended for future work.

Model Scaling. In our methodology, only the classifier layer is trained, diminishing the benefits of further scaling the underlying model responsible for generating representations. Additionally, computational limitations become a significant factor when evaluating models with explainers, as these methods necessitate thousands of inferences for each sample. Despite these constraints, our experiments with BERT-Large indicate that findings are consistent even with larger models. It is also noteworthy that BERT-based models remain relevant benchmarks in recent language model research, as evidenced by studies such as from Du et al. (2023).

Annotation Efforts. We are aware of the additional effort required to collect annotations for textual datasets and how this limits the extension of our work's application. However, we notice that, to make models "learn with humans," human efforts must be made to "teach machines." We believe this is a limitation of the problem ("learning with explanations") instead of our work (a specific methodology to incorporate the explanations). Even so, there is a relevant availability of textual datasets with annotations (Wiegreffe and Marasovic, 2021). Finally, recent advances in crowdsourcing annotation systems allow an efficient annotation of datasets at scale (Drutsa et al., 2021).

Human Study. Consistent with precedents in the field (Mathew et al., 2021; Ross et al., 2017), we did not conduct a separate human evaluation. This decision is based on the redundancy of such an evaluation with the existing human annotations in our dataset. Any human assessment would only assess the machine's rationale against individuals' subjective interpretations of the rationale. This process is equivalent to the annotation process already undertaken.

Methodology Comparison. BERT-HateXplain is an appropriate baseline for our approach, sharing the same explanation method, dataset, and metrics. It aptly represents other baseline methods (Chan et al., 2022; Zhang et al., 2021; Lakhotia et al., 2021; Arous et al., 2021; Strout et al., 2019), which also integrate rationale extraction in the forward pass and learn from annotated rationales. Future work will include comparisons with gradient saliency-based baselines (Ghaeini et al., 2019; Huang et al., 2021). Furthermore, BERT-HateXplain is a specific instance of UNIREX (Chan et al., 2022). The only difference in its "Share LM" variant (model and extractor with shared parameters) is an additional faithfulness loss beyond our current scope. The "Double LM" variant of UNIREX, featuring a distinct architecture for explanation extraction, is also outside our study's purview.

Ethics Statement

Some authors consider pursuing plausibility as an ethical issue (Jacovi and Goldberg, 2021). Part of this work argues this is not the case (Section 6). In this work, we utilize a hate speech detection dataset and train models with this data. We do not intend to publicly distribute the trained models as they may incorporate strong, toxic biases.

Acknowledgements

This work was supported by the National Council for Scientific and Technological Development (CNPq) under Grant #311144/2022-5, Carlos Chagas Filho Foundation for Research Support of Rio de Janeiro State (FAPERJ) under Grant #E-26/201.424/2021, São Paulo Research Foundation (FAPESP) under Grant #2021/07012-0, the School of Applied Mathematics at Fundação Getulio Vargas, and FAEPEX-UNICAMP under Grants 2559/22 and 2584/23. We also thank Vicente Ordonez and the anonymous reviewers for their important feedback.

References

Ines Arous, Ljiljana Dolamic, Jie Yang, Akansha Bhardwaj, Giuseppe Cuccu, and Philippe Cudré-Mauroux. 2021. MARTA: Leveraging Human Rationales for Explainable Text Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 5868–5876, Virtual. AAAI Press. Number: 7.

Pepa Atanasova, Jakob Grue Simonsen, Christina Lioma, and Isabelle Augenstein. 2020. A Diagnostic Study of Explainability Techniques for Text Classification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3256–3274, Online. Association for Computational Linguistics.

Yujia Bao, Shiyu Chang, Mo Yu, and Regina Barzilay. 2018. Deriving Machine Attention from Human Rationales. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1903–1913, Brussels, Belgium. Association for Computational Linguistics.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Jasmijn Bastings, Wilker Aziz, and Ivan Titov. 2019. Interpretable Neural Predictions with Differentiable Binary Variables. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2963–2977, Florence, Italy. Association for Computational Linguistics.

Catarina Belém, Vladimir Balayan, Pedro Saleiro, and Pedro Bizarro. 2021. Weakly Supervised Multi-task Learning for Concept-based Explainability. In *Proceedings of the First Workshop on Weakly Supervised Learning (WeaSuL)*, Virtual.

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

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-SNLI: Natural Language Inference with Natural Language Explanations. In *Advances in Neural Information Processing Systems*, volume 31, Palais des Congrès de Montréal, Montréal, Canada. Curran Associates, Inc.

Samuel Carton, Surya Kanoria, and Chenhao Tan. 2022. What to Learn, and How: Toward Effective Learning from Rationales. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1075–1088, Dublin, Ireland. Association for Computational Linguistics.

Aaron Chan, Maziar Sanjabi, Lambert Mathias, Liang Tan, Shaoliang Nie, Xiaochang Peng, Xiang Ren, and Hamed Firooz. 2022. UNIREX: A Unified Learning Framework for Language Model Rationale Extraction. In *Proceedings of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models*, pages 51–67, virtual+Dublin. Association for Computational Linguistics.

Hila Chefer, Shir Gur, and Lior Wolf. 2021. Transformer Interpretability Beyond Attention Visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 782–791.

Hanjie Chen and Yangfeng Ji. 2020. Learning Variational Word Masks to Improve the Interpretability of Neural Text Classifiers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4236–4251, Online. Association for Computational Linguistics.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A Simple Framework for Contrastive Learning of Visual Representations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR. ISSN: 2640-3498.

Jared L. Cohon. 1978. *Multiobjective Programming and Planning*, 1 edition, volume 140 of *Mathematics in Science and Engineering*. Academic Press.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A Benchmark to Evaluate Rationalized NLP Models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458, Online. Association for Computational Linguistics.

Alexey Drutsa, Dmitry Ustalov, Valentina Fedorova, Olga Megorskaya, and Daria Baidakova. 2021. Crowdsourcing Natural Language Data at Scale: A Hands-On Tutorial. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials*, pages 25–30, Online. Association for Computational Linguistics.

Kevin Du, Lucas Torroba Hennigen, Niklas Stoehr, Alex Warstadt, and Ryan Cotterell. 2023. Generalizing Backpropagation for Gradient-Based Interpretability. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1:*

Long Papers), pages 11979–11995, Toronto, Canada. Association for Computational Linguistics.

Mengnan Du, Ninghao Liu, Fan Yang, and Xia Hu. 2019. Learning Credible Deep Neural Networks with Rationale Regularization. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 150–159, Beijing, China. IEEE.

Abhimanyu Dubey, Filip Radenovic, and Dhruv Mahajan. 2022. Scalable Interpretability via Polynomials. ArXiv:2205.14108 [cs].

Reza Ghaeini, Xiaoli Fern, Hamed Shahbazi, and Prasad Tadepalli. 2019. Saliency Learning: Teaching the Model Where to Pay Attention. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4016–4025, Minneapolis, Minnesota. Association for Computational Linguistics.

Sofie Goethals, David Martens, and Theodoros Evgeniou. 2022. The non-linear nature of the cost of comprehensibility. *Journal of Big Data*, 9(1).

Peter Hase, Shiyue Zhang, Harry Xie, and Mohit Bansal. 2020. Leakage-Adjusted Simulatability: Can Models Generate Non-Trivial Explanations of Their Behavior in Natural Language? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4351–4367, Online. Association for Computational Linguistics.

Quzhe Huang, Shengqi Zhu, Yansong Feng, and Dongyan Zhao. 2021. Exploring Distantly-Labeled Rationales in Neural Network 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 5571–5582, Online. Association for Computational Linguistics.

Alon Jacovi and Yoav Goldberg. 2021. Aligning Faithful Interpretations with their Social Attribution. *Transactions of the Association for Computational Linguistics*, 9:294–310.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. 2020. Learning to Faithfully Rationalize by Construction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4459–4473, Online. Association for Computational Linguistics.

Yaochu Jin, Bernhard Sendhoff, and Edgar Körner. 2006. Simultaneous Generation of Accurate and Interpretable Neural Network Classifiers. In Yaochu Jin, editor, *Multi-Objective Machine Learning*, 1 edition, volume 16 of *Studies in Computational Intelligence*, pages 291–312. Springer, Berlin, Heidelberg.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised Contrastive Learning. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, volume 33, pages 18661–18673. Curran Associates, Inc.

Sawan Kumar and Partha Talukdar. 2020. NILE : Natural Language Inference with Faithful Natural Language Explanations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8730–8742, Online. Association for Computational Linguistics.

Gitanjali Kumari, Anubhav Sinha, and Asif Ekbal. 2024. Unintended Bias Detection and Mitigation in Misogynous Memes. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2719–2733, St. Julian's, Malta. Association for Computational Linguistics.

Kushal Lakhotia, Bhargavi Paranjape, Asish Ghoshal, Scott Yih, Yashar Mehdad, and Srini Iyer. 2021. FiD-Ex: Improving Sequence-to-Sequence Models for Extractive Rationale Generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3712–3727, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing Neural Predictions. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 107–117, Austin, Texas. Association for Computational Linguistics.

Jure Leskovec, Anand Rajaraman, and Jeffrey D. Ullman. 2020. *Mining of Massive Datasets*, 3 edition.

Frederick Liu and Besim Avci. 2019. Incorporating Priors with Feature Attribution on Text Classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6274–6283, Florence, Italy. Association for Computational Linguistics.

Hui Liu, Qingyu Yin, and William Yang Wang. 2019. Towards Explainable NLP: A Generative Explanation Framework for Text Classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5570–5581, Florence, Italy. Association for Computational Linguistics.

Junhong Liu, Yijie Lin, Liang Jiang, Jia Liu, Zujie Wen, and Xi Peng. 2022. Improve Interpretability of Neural Networks via Sparse Contrastive Coding. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 460–470, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, volume 30. Curran Associates, Inc.

Maggie, Phil Culliton, and Wei Chen. 2020. Tweet Sentiment Extraction.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to information retrieval*. Cambridge University Press, New York. OCLC: ocn190786122.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14867–14875, Virtual. AAAI Press. Number: 17.

Kaisa Miettinen. 1998. *Nonlinear Multiobjective Optimization*, 1 edition, volume 12 of *International Series in Operations Research & Management Science*. Springer New York, NY.

Masahiro Mitsuhara, Hiroshi Fukui, Yusuke Sakashita, Takanori Ogata, Tsubasa Hirakawa, Takayoshi Yamashita, and Hironobu Fujiyoshi. 2021. Embedding Human Knowledge into Deep Neural Network via Attention Map. In *Proceedings of the 16th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications - Volume 5: VISAPP*, volume 5, pages 626–636. SciTePress.

Mitchell Naylor, Christi French, Samantha Terker, and Uday Kamath. 2021. Quantifying Explainability in NLP and Analyzing Algorithms for Performance-Explainability Tradeoff.

Bhargavi Paranjape, Mandar Joshi, John Thickstun, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. An Information Bottleneck Approach for Controlling Conciseness in Rationale Extraction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1938–1952, Online. Association for Computational Linguistics.

Gregory Plumb, Maruan Al-Shedivat, Angel Alexander Cabrera, Adam Perer, Eric Xing, and Ameet Talwalkar. 2020. Regularizing Black-box Models for Improved Interpretability.

Danish Pruthi, Bhuwan Dhingra, Graham Neubig, and Zachary C. Lipton. 2020. Weakly- and Semi-supervised Evidence Extraction. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3965–3970, Online. Association for Computational Linguistics.

Filip Radenovic, Abhimanyu Dubey, and Dhruv Mahajan. 2022. Neural Basis Models for Interpretability. ArXiv:2205.14120 [cs].

Marcos M. Raimundo, Paulo A. V. Ferreira, and Fernando J. Von Zuben. 2020. An extension of the non-inferior set estimation algorithm for many objectives. *European Journal of Operational Research*, 284(1):53–66.

Nazneen Fatema Rajani, Bryan McCann, Caiming Xiong, and Richard Socher. 2019. Explain Yourself! Leveraging Language Models for Commonsense Reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4932–4942, Florence, Italy. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "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*, KDD '16, pages 1135–1144, New York, NY, USA. Association for Computing Machinery.

Laura Rieger, Chandan Singh, William Murdoch, and Bin Yu. 2020. Interpretations are Useful: Penalizing Explanations to Align Neural Networks with Prior Knowledge. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pages 8116–8126. PMLR. ISSN: 2640-3498.

Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. 2017. Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 2662–2670, Melbourne, Australia. AAAI Press.

Cynthia Rudin. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature machine intelligence*, 1(5):206–215.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. ArXiv:1910.01108 [cs].

Arshdeep Sekhon, Hanjie Chen, Aman Shrivastava, Zhe Wang, Yangfeng Ji, and Yanjun Qi. 2023. Improving Interpretability via Explicit Word Interaction Graph Layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13528–13537, Washington DC, USA. AAAI Press. Number: 11.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.

Manali Sharma and Mustafa Bilgic. 2018. Learning with rationales for document classification. *Machine Learning*, 107(5):797–824.

Becks Simpson, Francis Dutil, Yoshua Bengio, and Joseph Paul Cohen. 2019. GradMask: Reduce Overfitting by Regularizing Saliency. ArXiv:1904.07478 [cs, eess].

Julia Strout, Ye Zhang, and Raymond Mooney. 2019. Do Human Rationales Improve Machine Explanations? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 56–62, Florence, Italy. Association for Computational Linguistics.

Kyle Swanson, Lili Yu, and Tao Lei. 2020. Rationalizing Text Matching: Learning Sparse Alignments via Optimal Transport. In *Proceedings of the 58th*

Annual Meeting of the Association for Computational Linguistics, pages 5609–5626, Online. Association for Computational Linguistics.

Erico Tjoa and Cuntai Guan. 2022. Quantifying Explainability of Saliency Methods in Deep Neural Networks with a Synthetic Dataset. ArXiv:2009.02899 [cs].

Iulia Turc, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Well-Read Students Learn Better: On the Importance of Pre-training Compact Models. ArXiv:1908.08962 [cs].

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All You Need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 6000–6010, Long Beach, California, USA. Curran Associates Inc.

Sarah Wiegreffe and Ana Marasovic. 2021. Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing. *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, 1.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "Annotator Rationales" to Improve Machine Learning for Text Categorization. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 260–267, Rochester, New York. Association for Computational Linguistics.

Omar F Zaidan, Jason Eisner, and Christine D Piatko. 2008. Machine Learning with Annotator Rationales to Reduce Annotation Cost. In *Proceedings of the NIPS 2008 Workshop on Cost Sensitive Learning*, pages 260–267.

Zijian Zhang, Koustav Rudra, and Avishek Anand. 2021. Explain and Predict, and then Predict Again. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pages 418–426, Virtual Event Israel. Association for Computing Machinery.

Yilun Zhou, Marco Tulio Ribeiro, and Julie Shah. 2022. ExSum: From Local Explanations to Model Understanding. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5359–5378, Seattle, United States. Association for Computational Linguistics.

A Multi-objective Optimization Theorems and Definitions

The *weighted sum method* is an approach to solve a MOO problem. It balances the objective functions and converts the problem into a uni-objective form.

Definition A.1 (Weighted sum method)**.** Given a MOO problem as in Definition 3.1, the *weighted sum method* transforms the problem into

$$\begin{aligned} \min_{x} \quad & w^\intercal f(x), \\ \text{subject to} \quad & x \in \Omega \subseteq \mathbb{R}^n, f\colon \Omega \to \mathbb{R}^m, f(\Omega) = \Psi, \\ & \sum_{i=1}^{m} w_i = 1, w \in \mathbb{R}^m_+. \end{aligned}$$

With a few assumptions, solving the weighted problem is necessary and sufficient to search for the Pareto-frontier of the original MOO problem.

Theorem 1 (Necessity)**.** *If $w \in (\mathbb{R}^*_+)^m$ and x^* is a solution of the weighted problem, then x^* is a Pareto-optimal solution of the original MOO problem.*

Proof. Following Raimundo et al. (2020), suppose, by contradiction, that x^* is a solution to the weighted problem (with weights w) but not a Pareto-optimal solution. Then, there exists x such that, for some i, $f_i(x) < f_i(x^*)$ and, for all j, $f_j(x) \leq f_j(x^*)$, by definition. Then there exists $\varepsilon \geq 0$ such that $f(x) + \varepsilon = f(x^*)$, with $\varepsilon_i > 0$. Finally, $w^\intercal f(x) + w^\intercal \varepsilon = w^\intercal f(x^*)$, which means $w^\intercal f(x) < w^\intercal f(x^*)$. Absurd. □

Theorem 2 (Sufficiency)**.** *If the original MOO problem is convex, for any Pareto-optimal solution x^* there exists a weighting vector w such that x^* is the solution of the weighted problem.*

Proof. This theorem was proved by Miettinen (1998, Theorem 3.1.4). □

The equivalence between the MOO problem and the weighted problem, established when the MOO problem is convex, is crucial. It enables multi-objective optimization algorithms that characterize the Pareto-frontier using the weighted sum method (e.g., NISE, Cohon, 1978).

B Contrastive Loss for Logistic Regression

The logistic regression as the classifier is a particular case that deserves a highlight. When the model f_θ is a multinomial logistic regression over text embedding vectors, we can represent the contrastive rationale loss function in the following way:

$$\dot{\mathcal{L}}_\theta(\dot{X}, \dot{y}) = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{|C|}\mathbb{1}_{\dot{y}_i=k}\ln\frac{\exp(\dot{X}_i \cdot \theta_k)}{\sum_{j=1}^{m}\exp(\tilde{X}_{i,j} \cdot \theta_k)}. \quad (3)$$

The dot product between two vectors is commonly used as a similarity function in a contrastive learning context (Khosla et al., 2020). When minimizing Equation 3, one is training an *anchor* θ_k to approximate a *positive rationale* $\dot{X}_i$ and to distance *negative rationales* $\{\tilde{X}_{i,j}\}_{j=1}^{m} \setminus \{\dot{X}_i\}$, just like in contrastive learning. However, positive and negative vectors cannot be optimized in our case.

The multinomial logistic regression as a model is analogous to a neural network with all but the classification layer's weights frozen. When there are only two classes, it is easy to prove that binary and multinomial logistic regression are equivalent. Finally, the logistic regression results in a loss function $\dot{\mathcal{L}}$ that is convex with respect to the weights θ, easing the search for the model performance vs. explanation plausibility Pareto-frontier through the employing of convex multi-objective optimization algorithms, e.g., NISE (Cohon, 1978; Appendix A).

C Contrastive Learning Theoretical Background

Consider a scenario where samples belonging to a group p follow the distribution $\mathcal{T}_p$. In contrastive learning, the objective is to ensure that the representations of samples originating from the same distribution, $\{T_{p,i}\}_i \sim \mathcal{T}_p$, exhibit similarity in the vector space while samples from different distributions are positioned further apart. To achieve this, the learning process aims to maximize a chosen agreement metric among vector representations of samples from the same distribution while simultaneously minimizing this agreement for samples from different distributions.

In visual representations, Chen et al. (2020) employ a contrastive loss function in the latent space to maximize the agreement between two preprocessed versions of the same image while minimizing the agreement between preprocessed versions of different images. Similarly, Khosla et al. (2020) propose a *supervised contrastive loss* that maximizes the agreement between images belonging to the same class while minimizing the agreement between images from different classes.

D DistilBERT and BERT-Mini Fine-tuning on HateXplain

The rationales of the HateXplain dataset contain words not included in the original `distilbert-base-uncased`[6] and `bert-mini`[7] model's vocabulary because they are offensive and hate speech words. However, when training a model to incorporate rationales, including these tokens in the vocabulary may be important. Otherwise, the results would be underestimated. In the train portion of the dataset, we filtered the most popular out-of-vocabulary tokens (those with more than ten occurrences), added them to the models' vocabularies, and fine-tuned the models in this portion. We used a masked language modeling probability of 0.15 with a batch size of 8 for 15 epochs in a GPU NVIDIA GeForce GTX 1070. We do not apply this process for the methodology comparison to keep similarities with the original HateXplain work (Mathew et al., 2021).

E Implementation and Execution

Logistic Regression. We implemented the Logistic regression with Scikit-learn. Its implementation was adapted to incorporate the contrastive rationale loss. The experiments used the following hyperparameters: tolerance of 1e-4, max iterations of 1e3, $l2$ penalty, `lbfgs` solver, and `multinomial` implementation. The C hyperparameter was chosen with cross-validation on the training set. The regularization term is added to the two losses (cross-entropy and contrastive rationale loss). Therefore, when the two losses are weighted by $\mathbf{w}$, the regularization term comes with weight 1.

DistilBERT and BERT-Mini. The DistilBERT version used in this work was the `distilbert-base-uncased`[8], while the BERT-Mini version was the `prajjwal1/bert-mini`[9]. The models are used for text classification; therefore, we plug a classification head on top of the `[CLS]` output vector. We keep all but the classification layer's weights frozen to guarantee the loss convexity (as we pointed out in Appendix B), and the models are easier to train. These models were not trained with gradient descent because only a classification layer was trained. The classification layer was implemented as a multinomial logistic regression and trained accordingly (see previous paragraph). The inference over the DistilBERT and BERT-Mini models was performed using GPUs NVIDIA Quadro RTX 6000 and NVIDIA GeForce GTX 1070. The running time of all experiments took the order of magnitude of a month. The models truncate the input text to their input limit length of 512. The LIME's disturbed text input has its tokens substituted by `[MASK]` for these models, keeping the original text sample length.

Datasets. In the HateXplain dataset, because more than one annotator is used for each sample, we apply majority consensus to both rationale and class assignments, disregarding non-consensual samples.

The HateXplain dataset is already tokenized, and Movie Reviews was tokenized with Python's `str.split()`. Tweet Sentiment Extraction (TSE) was tokenized using `re.split(f"([\\s{punctuation}])", str)` with `punctuation` imported from `string` and with regex special characters escaped. Table 3 presents a description of the datasets.

Table 3: Description of the datasets after filtering (Section 5.2). HateXplain average rationale length is calculated over the hate speech class only, and `hatexplain_all`, over hate speech and offensive classes.

Dataset	Samples	Average sample length	Average rationale length
HateXplain	13749	23.9	3.4
`hatexplain_all`	19228	23.4	3.3
Movie Reviews	1800	741.7	62.1
TSE	16330	17.5	4.7
`tse_all`	27378	17.0	9.2

LIME. The LIME explainer was implemented using 1000 samples, and the number of features was the number of tokens of the text sample. It applied the perturbations using each dataset's tokenization and filled the perturbed tokens in accordance with the model requirements. For instance, DistilBERT and BERT-Mini required the perturbed tokens to become `[MASK]` tokens to keep the input sequence length unchanged.

Comparison with HateXplain. To compare our methodology with HateXplain's (Mathew et al., 2021),

[6] Available at https://huggingface.co/distilbert-base-uncased

[7] Available at https://huggingface.co/prajjwal1/bert-mini

[8] Available at https://huggingface.co/distilbert-base-uncased

[9] Available at https://huggingface.co/prajjwal1/bert-mini

we implement their model in both their and our framework. We tried to keep the implementation, including methods and hyperparameters, as close as possible to the details in their paper (Mathew et al., 2021) and in their GitHub repository[10]. We use the three-class HateXplain dataset (`hatexplain_all`), the model `bert-base-uncased`, and the explainer LIME. In our method, we also use 2 negative (random) rationales. In particular, BERT's input length limit is set to 128 tokens. Finally, we use the BERT's `pooled_output` vector as input to the classification layer, in contrast to the other language models in this paper, in which we use the `[CLS]` token output vector.

In our methodology, before exploring the trade-off between cross-entropy and the contrastive rationale loss using NISE, we fine-tune the model with the cross-entropy loss only. This is done to maintain performance compatibility between our method and HateXplain's, which fine-tunes the model to train the attention. However, we do not apply the fine-tuning procedure of Appendix D, i.e., incorporating new tokens into the model's vocabulary and training the model in the masked language model task (MLM). This could be performed, but it would differ from what was done in HateXplain's work.

The model's hyperparameters (in their methodology and in our fine-tuning) were set to the following values: learning rate of $2e-5$, attention softmax temperature parameter of 0.2, Adam optimizer, standard BERT dropouts of 0.1, 6 heads of attention supervision in the last BERT layer, batch size of 16, 20 epochs, and epsilon of $1e-8$. The authors indicated these hyperparameters as the best ones.

Their novel attention loss was implemented as a cross-entropy between the attention values and the rationale (the mean of attention losses for each attention head) by using an additional hyperparameter λ:

$$\text{loss} = \text{cross-entropy} + \lambda \cdot \text{attention loss}.$$

We explore the trade-off between their two losses (cross-entropy and attention loss) by varying λ from 0.001 to 100 on a logarithmic scale, as suggested by the authors. Because our method considers the rationale binary (a token is either a rationale token or not), we also incorporated the rationales in BERT-HateXplain as binary, differently from their implementation, which uses the mean of the binary rationales (one for each annotator) as the rationale. Doing this was necessary for a fair comparison between the two methods.

Even though we implement BERT-HateXplain with a few reasonable, justified modifications, our experimental results of their model are comparable to their paper's (Mathew et al., 2021), as pointed in Section 5.7.

[10] https://github.com/hate-alert/HateXplain

F Additional Results

F.1 Main Results

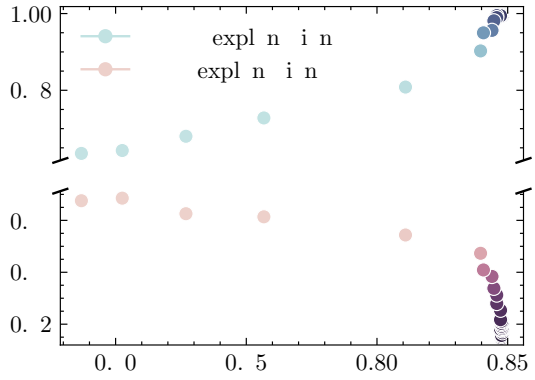

Figure 7: Trade-off between performance and plausibility on test data for originally good (AUPRC $=1$) and originally bad (AUPRC <1) explanations differently. The color scale is the same as the previous figures.

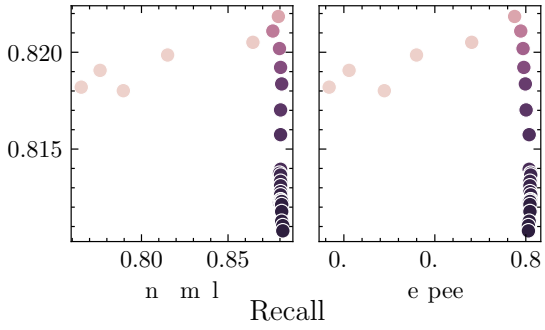

Figure 8: Trade-off between per class recall and plausibility on test data for DistilBERT and HateXplain dataset. The color scale is the same as the previous figures.

F.2 Results in Non-Binary Classification

Sections 5.5 and 5.6 present results for all datasets but are binary classification. As pointed out in Section 5.2, this procedure simplifies the learning task. Our methodology, however, is agnostic to the number of classes and can handle non-binary classification by default—we sum over any number of classes in Equation 2. Figure 9 presents the trade-off between

accuracy and plausibility for `hatexplain_all` (with TF-IDF) and `tse_all` (with DistilBERT) (test data), i.e., with all the three labels, and a number of negative rationales of 2. The trade-off frontier shapes are similar to the binary classification, with similar conclusions from Section 5.6. However, different datasets lead to different absolute values. Finally, in a similar way to Section 5.6, Table 4 compares the original and chosen models, leading to similar conclusions: positive AUPRC improvement and a small decrease of performance. TSE had similar faithfulness results, while HateXplain had slightly worse faithfulness results.

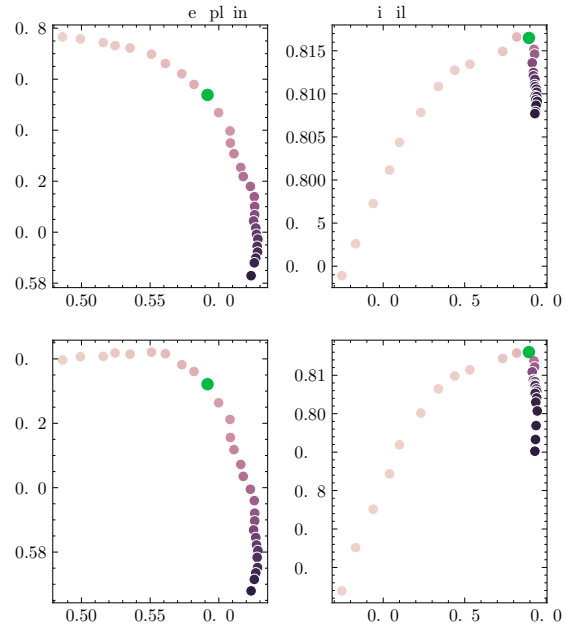

Figure 9: Trade-offs between performance (accuracy, x-axis) and plausibility (AUPRC, y-axis) for `hatexplain_all` (i.e., with all labels, and with TF-IDF) and `tse_all` (i.e., with all labels, and with DistilBERT) (test data). The number of random (negative) rationales is 2. The color scale is the same as the previous figures. We ignore the model with $w_1 = 0$ in all graphics as it is out of scale. Green dots are the models chosen to be analyzed more carefully.

F.3 Results of Larger Models

Section 5 presents experiments with DistilBERT and BERT-Mini, which are small language model encoders. To further evaluate our methodology with a larger model, we performed a series of experiments with BERT-Large (Devlin et al., 2019): datasets HateXplain and TSE, explainers LIME and SHAP, 2 negative rationales, BERT-Large without MLM fine-tuning. The shapes of the model frontiers (Figure 10) were similar to other language model frontiers of Figure 5 in the main paper. Additionally, Table 5 compares the original and chosen models (in green). It reinforces our previous results regarding plausibility gain and minor performance degradation while improving or keeping faithfulness. We also highlight the existence of an experiment with BERT-Base (Devlin et al., 2019) in the baseline comparison, a larger model than DistilBERT and BERT-Mini used in the main experiments.

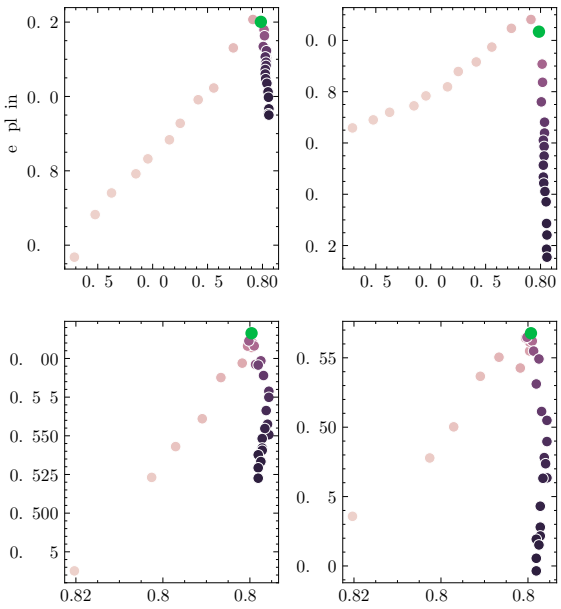

Figure 10: Trade-offs between performance (accuracy, x-axis) and plausibility (AUPRC, y-axis, in percentage (%)) for BERT-Large with HateXplain and TSE (test data). The number of random (negative) rationales is 2, and the explainers are LIME and SHAP. The color scale is the same as the previous figures. We ignore the model with $w_1 = 0$ in all graphics as it is out of scale. Green dots are the models chosen to be analyzed more carefully.

F.4 Out-of-Distribution Results

To test out-of-distribution (OOD) performance, we additionally evaluated the DistilBERT trained on HateXplain (Section 5.5 of the main paper) on HatEval (Basile et al., 2019), a similar dataset of hateful tweets but with a different data distribution (it focuses on hate speech against specific groups). We indeed observed an increase in OOD performance. The frontier shape of HatEval performance in Figure 11 is roughly similar to the frontier shape of HateXplain performance (in the same Figure and in Figure 3) but with the x-axis reversed (OOD performance increases with the plausibility, except for very small w_1 values). For the selected model (green dot in Figure 11), while original accuracy decreases by 0.8% and plausibility increases by approximately 1.1%,

Table 4: Comparison between the original model (cross-entropy only) and the chosen model (green dots on Figure 9) for each performance and explainability metric on test data. "rel." means relative variation. The column w_1 indicates the weight w_1 of the chosen model's cross-entropy loss during training. Number of negative rationales is 2.

Model	w_1	**Acc. %**	**AUPRC %**	**AUPRC rel. %**	**Suff.**	**Comp.**
hatexplain_all-lime-tf_idf	0.19	-3.17	7.09	12.16	-0.00	-0.06
hatexplain_all-shap-tf_idf	0.19	-3.17	6.42	11.30	-0.00	-0.06
tse_all-lime-distilbert	0.25	-0.37	0.88	1.09	0.01	-0.01
tse_all-shap-distilbert	0.25	-0.37	2.58	3.26	-0.02	-0.00

Table 5: Comparison between the original model (cross-entropy only) and the chosen model (green dots on Figure 10) for each performance and explainability metric on test data. "rel." means relative variation. The column w_1 indicates the weight w_1 of the chosen model's cross-entropy loss during training. Number of negative rationales is 2.

Model	w_1	**Acc. %**	**AUPRC %**	**AUPRC rel. %**	**Suff.**	**Comp.**
hatexplain-lime-bert_large	0.33	-0.73	2.51	3.61	0.13	0.03
hatexplain-shap-bert_large	0.33	-0.73	8.79	14.29	0.12	0.06
tse-lime-bert_large	0.30	-0.15	0.94	1.44	0.06	-0.01
tse-shap-bert_large	0.43	-0.12	1.71	2.68	0.05	-0.00

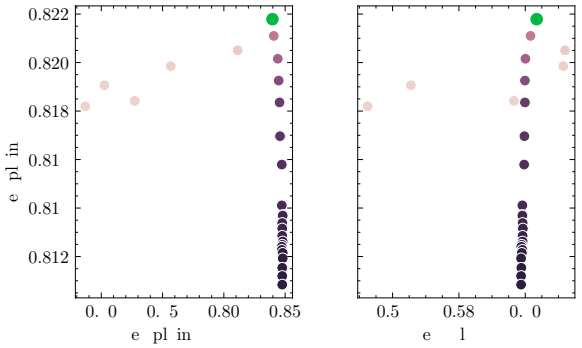

Figure 11: Trade-offs between (HateXplain and HatEval) performance and (HateXplain) plausibility with DistilBERT (test data). The number of random (negative) rationales is 2, and the explainer is LIME. The color scale is the same as the previous figures. We ignore the model with $w_1 = 0$ in all graphics as it is out of scale. Green dots are the model chosen to be analyzed more carefully.

the out-of-distribution performance also increases by 0.47%. We also found it possible to increase by 0.97% of plausibility and 1.32% of OOD performance at the expense of a 3.64% drop in original accuracy.

F.5 Other Results

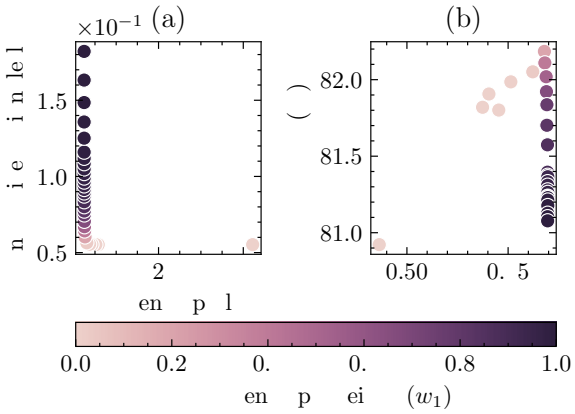

Figure 12: (a) Trade-off between the two losses on the training data. (b) Trade-off between accuracy and plausibility on the test data. The color scale represents the cross-entropy weight w_1 (Section 4.3). We include the model with $w_1 = 0$.

Table 6: Examples of explanations of the hate speech class of the HateXplain dataset. Examples were selected based on the size and quality of the explanation and model predictions. The "original" explanation comes from the original model trained with cross-entropy loss only (Section 5.5), while the "selected" explanation comes from the model with top-AUPRC studied in Section 5.5 (DistilBERT, HateXplain, LIME, 2 negative rationales). Green means a positive contribution to the model's prediction. The top tokens were selected for visualization purposes, and the number of tokens is the same as the original rationales.

Model	Example
Original	bhad bhabie we hate your nasty smelly *ss
Selected	bhad bhabie we hate your nasty smelly *ss
Original	i tole sum brillohaided n****r bich off last nite cuz it try to get uppity wit me
Selected	i tole sum brillohaided n****r bich off last nite cuz it try to get uppity wit me
Original	london mayor s*ckdiq kahn has room for a fellow m**zie terrorist
Selected	london mayor s*ckdiq kahn has room for a fellow m**zie terrorist
Original	no <number> million white children r***d by p*ki muslims <number> million p*kis in the uk for their crime the whole community must be expelled
Selected	no <number> million white children r***d by p*ki muslims <number> million p*kis in the uk for their crime the whole community must be expelled
Original	the russians buying that and purging all the blue haired d*kes was glorious that may well be the genesis of the russian bot narrative
Selected	the russians buying that and purging all the blue haired d*kes was glorious that may well be the genesis of the russian bot narrative
Original	give a n****r an inch they take a mile lol r*tarded creatures they f*ck themselves over every time
Selected	give a n****r an inch they take a mile lol r*tarded creatures they f*ck themselves over every time
Original	and if u have to drink and drive make sure u drive home through as many n****r gh*ttos as possible [emoji]
Selected	and if u have to drink and drive make sure u drive home through as many n****r gh*ttos as possible [emoji]

Table 7: Examples of explanations of the Tweet Sentiment Extraction dataset. Examples were selected based on the size and quality of the explanation and model predictions. The "original" explanation (LIME) comes from the original DistilBERT model trained with cross-entropy loss only (Section 5.6), while the "selected" explanation comes from the selected model with a green dot (Section 5.6, Figure 5) (2 negative rationales). Green means a positive contribution to the model's prediction. The top tokens were selected for visualization purposes, and the number of tokens is the same as the original rationales.

Label	Model	Example
positive	Original	in rye . . happy mothers day mums ily mummy lol
	Selected	in rye . . happy mothers day mums ily mummy lol
positive	Original	I ‘ ll try that , thanks
	Selected	I ‘ ll try that , thanks
positive	Original	LOVE your show !
	Selected	LOVE your show !
positive	Original	_ O _ ASH I do too plus more happy mothers day Sweety
	Selected	_ O _ ASH I do too plus more happy mothers day Sweety
positive	Original	hopefully today will work in our favor
	Selected	hopefully today will work in our favor
positive	Original	Rachmaninoff makes me a happy panda .
	Selected	Rachmaninoff makes me a happy panda .
positive	Original	You must like my song .
	Selected	You must like my song .
negative	Original	_ [user] aww that sucks
	Selected	_ [user] aww that sucks
positive	Original	Digging a downloaded film with mi familia . We love iTunes
	Selected	Digging a downloaded film with mi familia . We love iTunes
positive	Original	Happy Mommy Day
	Selected	Happy Mommy Day

Table 8: Comparison between the original model (cross-entropy only) and the chosen model (green dots on Figures 5, 16, 17, 18) for each performance and explainability metric on test data. "rel." means relative variation. The column w_1 indicates the weight w_1 of the chosen model's cross-entropy loss during training.

Model	w_1	**Acc. %**	**AUPRC %**	**AUPRC rel. %**	**Suff.**	**Comp.**
hatexplain-lime-distilbert-2	0.20	-0.80	1.11	1.37	0.25	-0.03
hatexplain-shap-distilbert-2	0.67	-0.29	0.85	1.06	0.15	-0.01
hatexplain-lime-distilbert-5	0.25	-0.91	1.19	1.47	0.25	-0.03
hatexplain-shap-distilbert-5	0.80	0.00	0.85	1.06	0.14	-0.01
hatexplain-lime-bert_mini-2	0.29	-0.84	2.46	3.49	0.40	-0.05
hatexplain-shap-bert_mini-2	0.29	-0.84	3.17	4.67	0.40	-0.05
hatexplain-lime-bert_mini-5	0.37	-0.80	2.67	3.78	0.41	-0.04
hatexplain-shap-bert_mini-5	0.37	-0.80	3.25	4.80	0.40	-0.05
hatexplain-lime-tf_idf-2	0.002	-9.35	6.96	10.79	0.13	-0.10
hatexplain-shap-tf_idf-2	0.002	-9.35	5.98	9.60	0.13	-0.09
hatexplain-lime-tf_idf-5	0.002	-9.45	7.79	12.08	0.13	-0.10
hatexplain-shap-tf_idf-5	0.002	-9.45	6.71	10.79	0.14	-0.10
movie_reviews-lime-distilbert-2	0.12	-0.28	0.50	4.39	0.25	-0.05
movie_reviews-shap-distilbert-2	0.36	-0.56	0.50	3.58	0.13	-0.02
movie_reviews-lime-distilbert-5	0.15	-0.28	0.61	5.43	0.25	-0.02
movie_reviews-shap-distilbert-5	0.81	0.83	0.17	1.23	0.04	0.00
movie_reviews-lime-bert_mini-2	0.26	0.28	0.39	3.61	0.00	-0.02
movie_reviews-shap-bert_mini-2	0.26	0.28	0.76	5.49	-0.01	-0.02
movie_reviews-lime-bert_mini-5	0.43	0.56	0.28	2.60	0.02	-0.01
movie_reviews-shap-bert_mini-5	0.43	0.56	0.85	6.16	0.01	-0.01
movie_reviews-lime-tf_idf-2	0.09	0.56	0.85	6.95	-0.00	0.01
movie_reviews-shap-tf_idf-2	0.07	0.28	0.99	6.26	0.01	0.01
movie_reviews-lime-tf_idf-5	0.10	1.67	0.82	6.73	-0.02	0.01
movie_reviews-shap-tf_idf-5	0.10	1.67	1.07	6.77	-0.02	0.02
tse-lime-distilbert-2	0.64	0.09	1.32	1.98	0.05	-0.00
tse-shap-distilbert-2	0.64	0.09	4.79	7.61	0.00	0.02
tse-lime-distilbert-5	0.51	-0.12	1.42	2.14	0.07	0.00
tse-shap-distilbert-5	0.36	-0.15	5.29	8.41	0.04	0.03
tse-lime-bert_mini-2	0.19	0.37	0.64	1.01	0.06	0.01
tse-shap-bert_mini-2	0.19	0.37	1.31	2.09	0.06	0.01
tse-lime-bert_mini-5	0.43	0.40	0.54	0.85	0.06	0.01
tse-shap-bert_mini-5	0.43	0.40	1.14	1.81	0.05	0.01
tse-lime-tf_idf-2	0.42	0.24	0.40	0.64	0.01	-0.02
tse-shap-tf_idf-2	0.42	0.24	0.78	1.28	0.01	-0.02
tse-lime-tf_idf-5	0.75	0.24	0.23	0.36	0.00	-0.01
tse-shap-tf_idf-5	0.75	0.24	0.43	0.70	0.00	-0.01

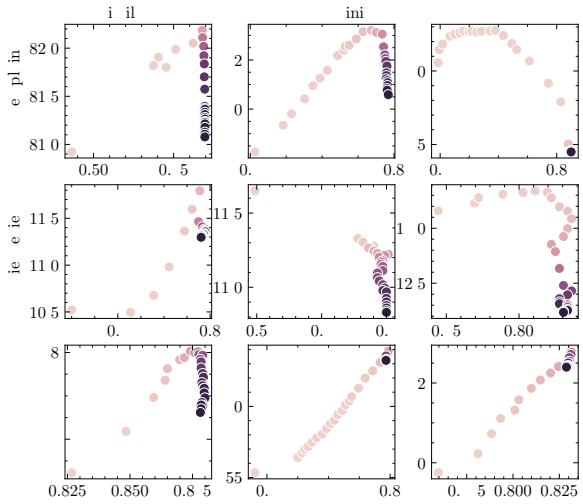

Figure 13: Trade-offs between performance (accuracy, x-axis) and plausibility (AUPRC, y-axis, in percentage (%)) for all models and datasets (test data). The number of random (negative) rationales is 2, and the explainer is LIME. The color scale is the same as the previous figures. We include the model with $w_1 = 0$ in all graphics.

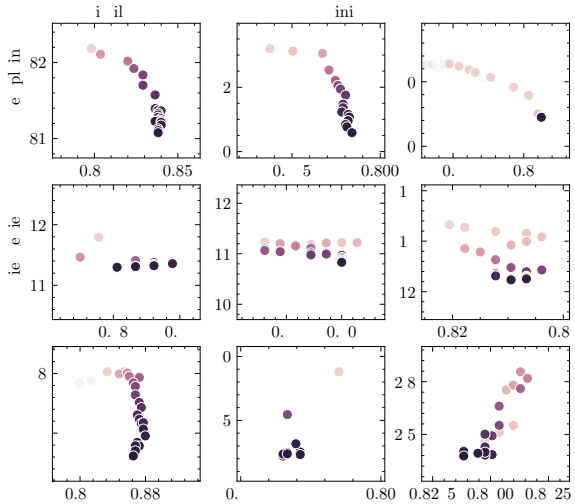

Figure 14: Trade-offs between performance (accuracy, x-axis) and plausibility (AUPRC, y-axis, in percentage (%)) for all models and datasets (test data). The number of random (negative) rationales is 2, and the explainer is LIME. The color scale is the same as the previous figures. The data scale is equal between x- and y-axes, and a few out-of-scale points are gray or removed.

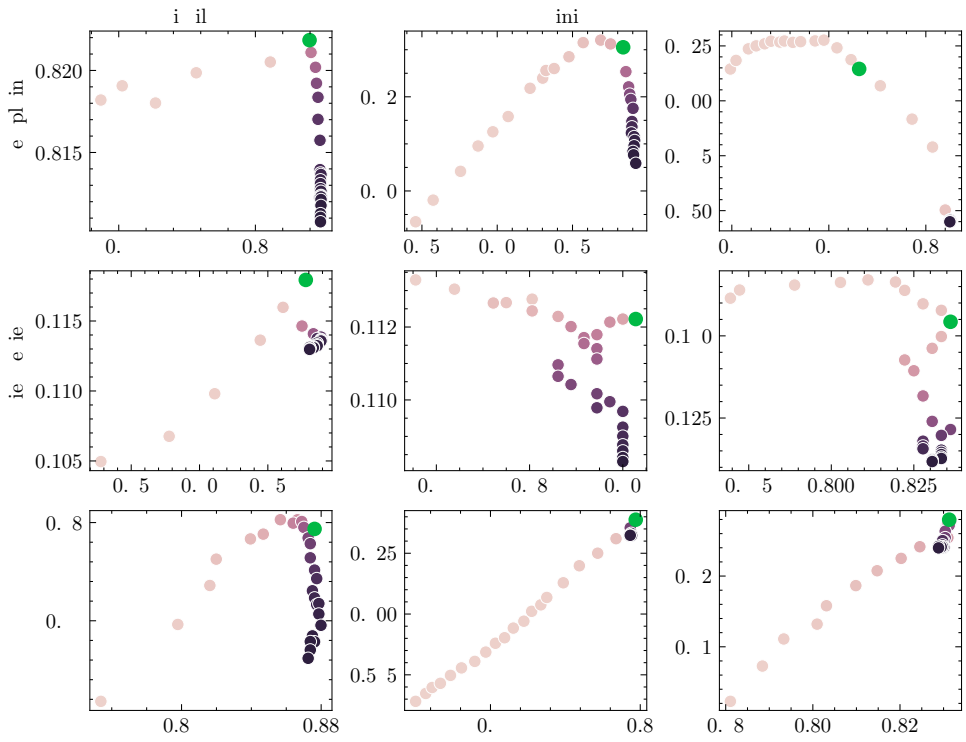

Figure 15: Trade-offs between performance (accuracy, x-axis) and plausibility (AUPRC, y-axis) for all models and datasets (test data). The number of random (negative) rationales is 2, and the explainer is LIME. The color scale is the same as the previous figures. We ignore the model with $w_1 = 0$ in all graphics as it is out of scale. Green dots are the models chosen to be analyzed more carefully.

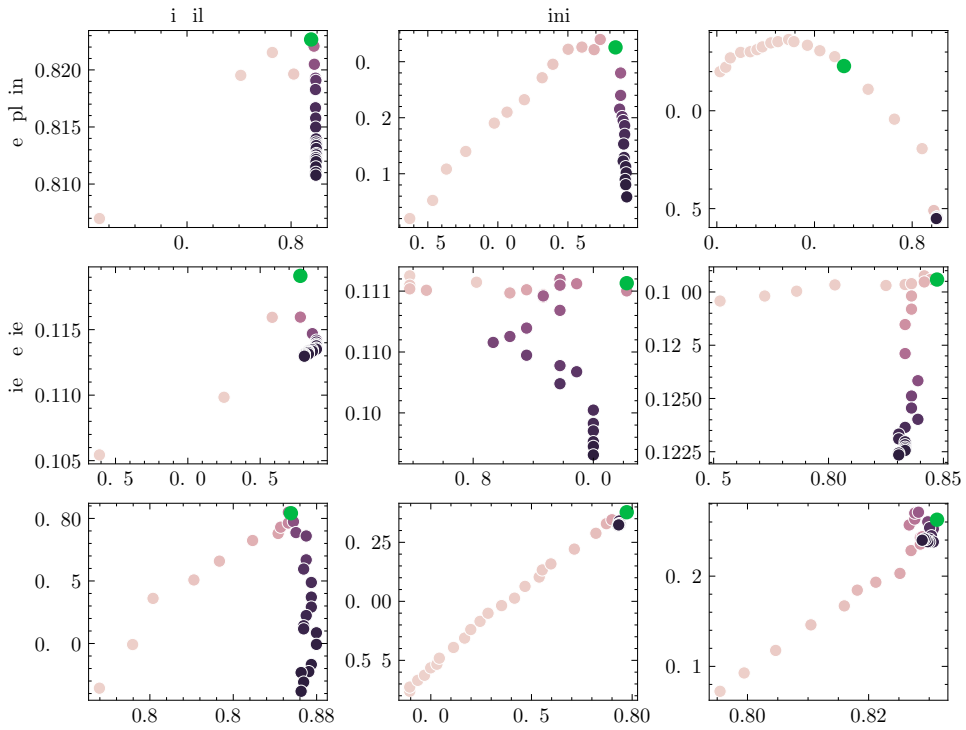

Figure 16: Trade-offs between performance (accuracy, x-axis) and plausibility (AUPRC, y-axis) for all models and datasets (test data). The number of random (negative) rationales is 5, and the explainer is LIME. The color scale is the same as the previous figures. We ignore the model with $w_1 = 0$ in all graphics as it is out of scale. Green dots are the models chosen to be analyzed more carefully.

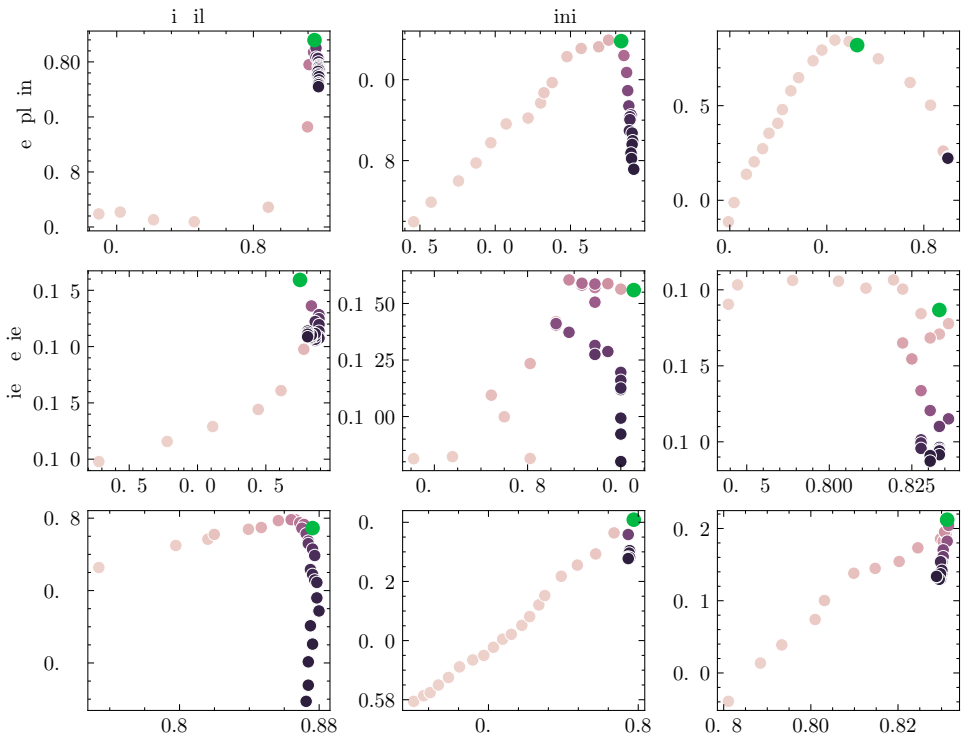

Figure 17: Trade-offs between performance (accuracy, x-axis) and plausibility (AUPRC, y-axis) for all models and datasets (test data). The number of random (negative) rationales is 2, and the explainer is SHAP. The color scale is the same as the previous figures. We ignore the model with $w_1 = 0$ in all graphics as it is out of scale. Green dots are the models chosen to be analyzed more carefully.

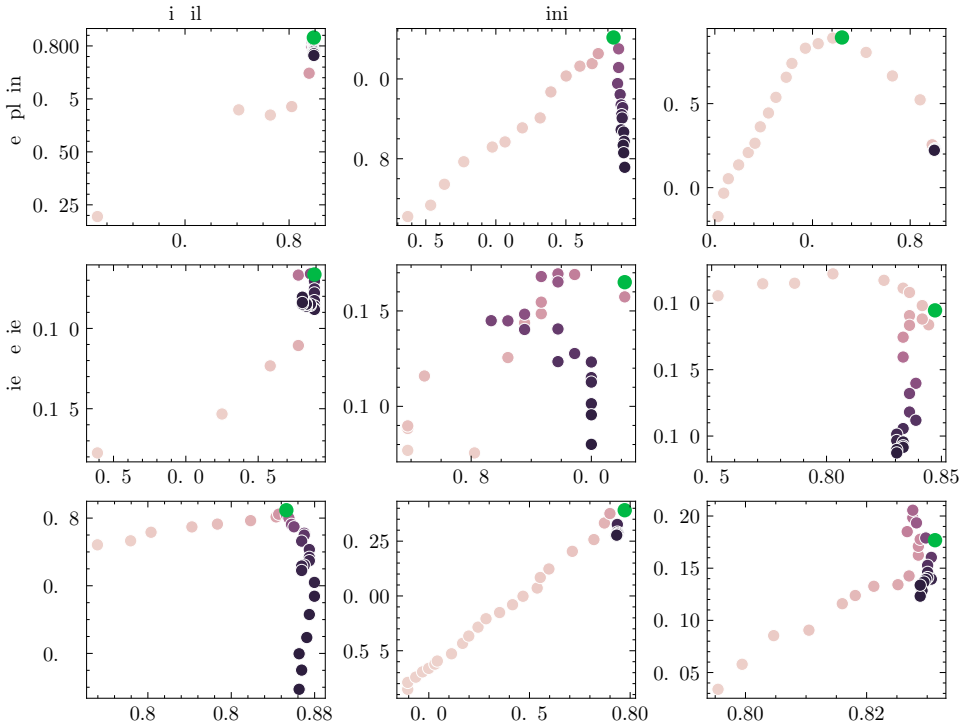

Figure 18: Trade-offs between performance (accuracy, x-axis) and plausibility (AUPRC, y-axis) for all models and datasets (test data). The number of random (negative) rationales is 5, and the explainer is SHAP. The color scale is the same as the previous figures. We ignore the model with $w_1 = 0$ in all graphics as it is out of scale. Green dots are the models chosen to be analyzed more carefully.

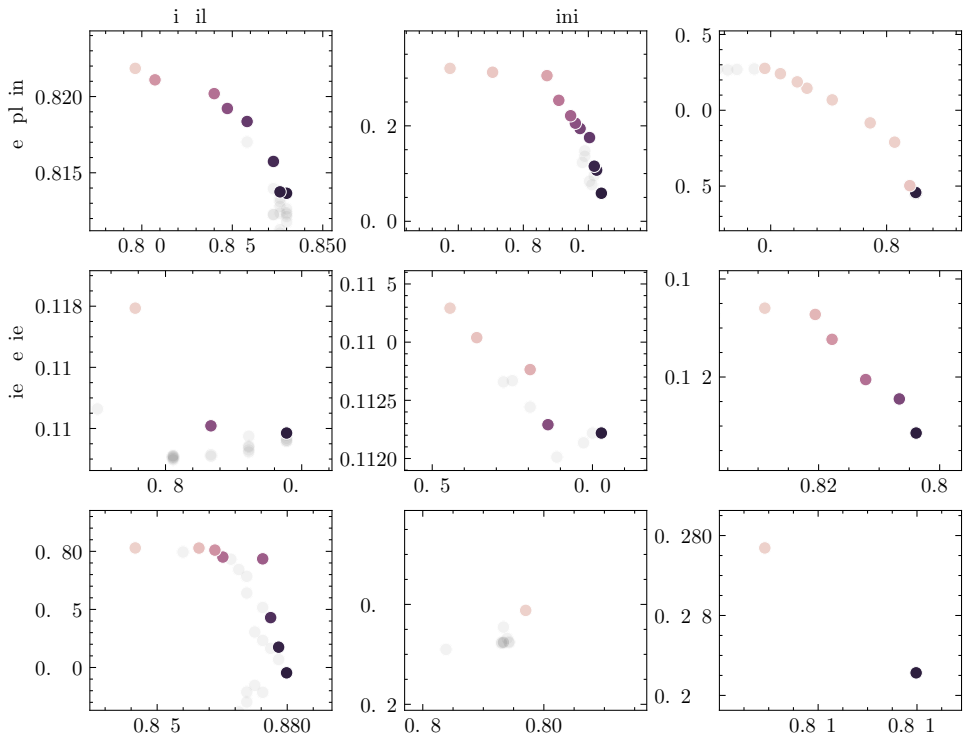

Figure 19: Pareto-frontier of trade-offs between performance (accuracy, x-axis) and plausibility (AUPRC, y-axis) for all models and datasets (test data). The number of random (negative) rationales is 2, and the explainer is LIME. The color scale is the same as the previous figures. Gray dots are models not on the Pareto-frontier. We ignore the model with $w_1 = 0$ in all graphics as it is out of scale.

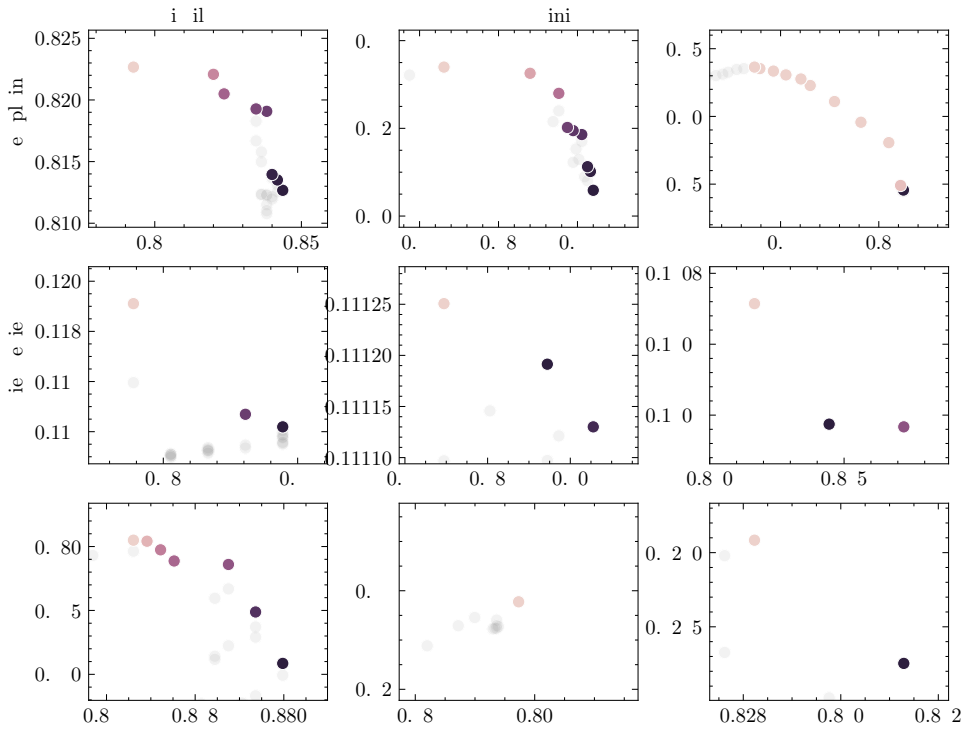

Figure 20: Pareto-frontier of trade-offs between performance (accuracy, x-axis) and plausibility (AUPRC, y-axis) for all models and datasets (test data). The number of random (negative) rationales is 5, and the explainer is LIME. The color scale is the same as the previous figures. Gray dots are models not on the Pareto-frontier. We ignore the model with $w_1 = 0$ in all graphics as it is out of scale.

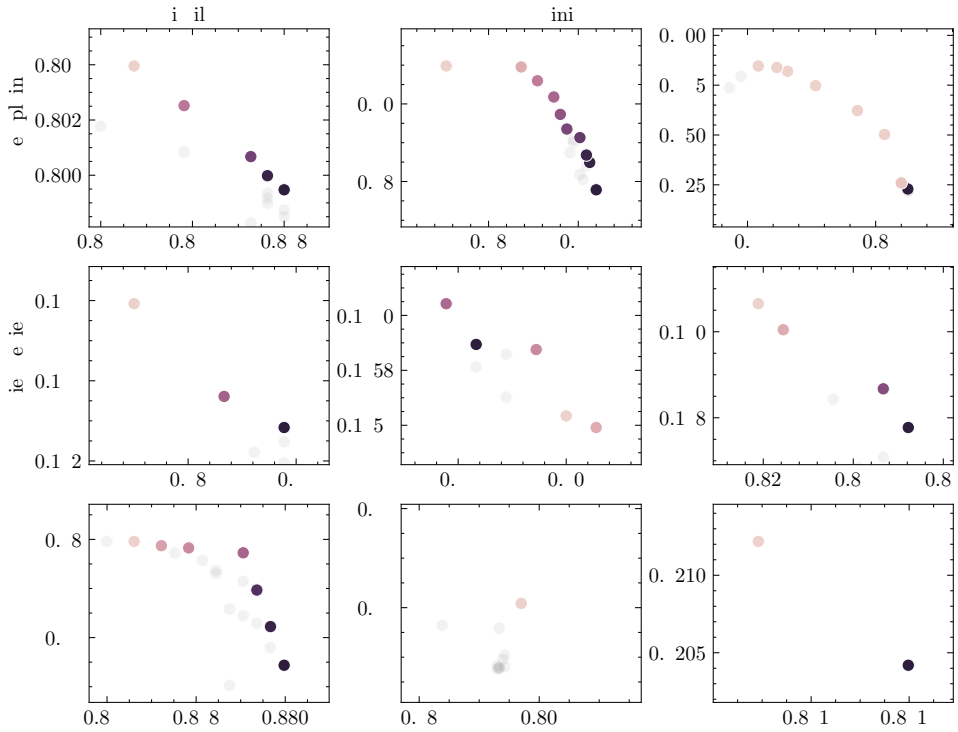

Figure 21: Pareto-frontier of trade-offs between performance (accuracy, x-axis) and plausibility (AUPRC, y-axis) for all models and datasets (test data). The number of random (negative) rationales is 2, and the explainer is SHAP. The color scale is the same as the previous figures. Gray dots are models not on the Pareto-frontier. We ignore the model with $w_1 = 0$ in all graphics as it is out of scale.

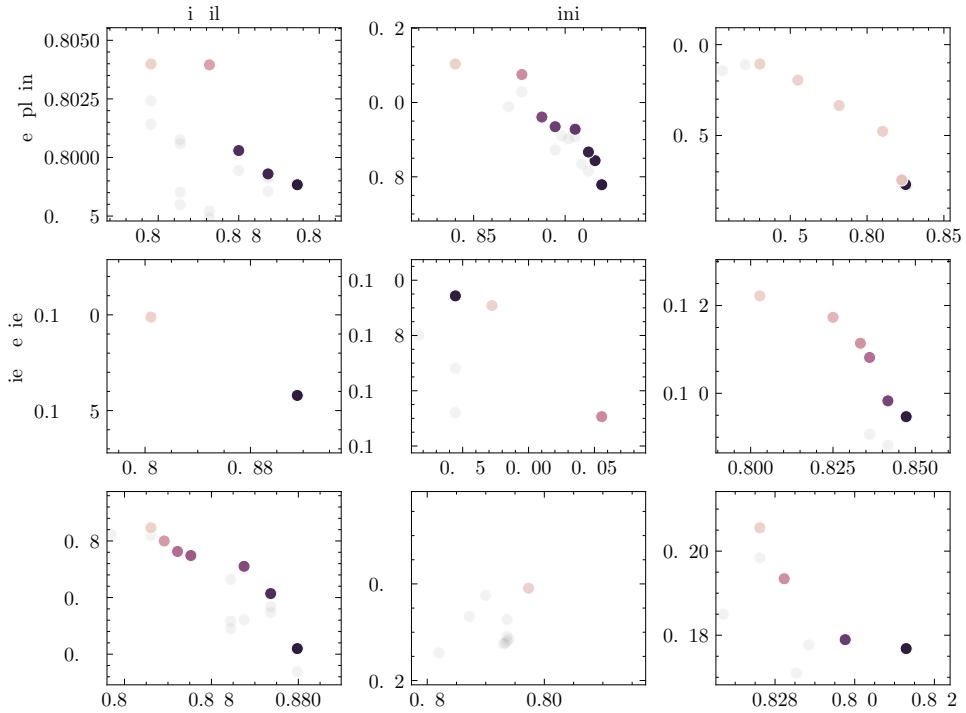

Figure 22: Pareto-frontier of trade-offs between performance (accuracy, x-axis) and plausibility (AUPRC, y-axis) for all models and datasets (test data). The number of random (negative) rationales is 5, and the explainer is SHAP. The color scale is the same as the previous figures. Gray dots are models not on the Pareto-frontier. We ignore the model with $w_1 = 0$ in all graphics as it is out of scale.