Lucas Emanuel Resck

# Balancing Performance and Explanation Plausibility: A Multi-Objective Approach to Text Classification with Human Rationales

Brazil

May 2024

Lucas Emanuel Resck

# Balancing Performance and Explanation Plausibility: A Multi-Objective Approach to Text Classification with Human Rationales

Master's Thesis presented to the School of Applied Mathematics of Fundação Getulio Vargas as a partial requirement for the degree of Master in Mathematical Modeling.

Fundação Getulio Vargas

School of Applied Mathematics

Supervisor: Jorge Poco

Co-supervisor: Marcos Medeiros Raimundo

Brazil

May 2024

Dados Internacionais de Catalogação na Publicação (CIP)
Ficha catalográfica elaborada pelo Sistema de Bibliotecas/FGV

Domingues, Lucas Emanuel Resck
Balancing performance and explanation plausibility: a multi-objective approach to text classification with human rationales / Lucas Emanuel Resck Domingues. – 2024.

88 f.

Dissertação (mestrado) – Fundação Getulio Vargas, Escola de Matemática Aplicada.
Orientador: Jorge Poco.
Coorientador: Marcos Medeiros Raimundo.
Inclui bibliografia.

1. Aprendizado do computador. 2. Processamento da linguagem natural. (Computação). 3. Aprendizagem baseada em explicações. 4. Inteligência artificial I. Poco, Jorge. II. Raimundo, Marcos. III. Fundação Getulio Vargas. Escola de Matemática Aplicada. IV. Título.

CDD – 006.3

Elaborada por Maria do Socorro Almeida – CRB-7/4254

**FUNDAÇÃO GETULIO VARGAS - ESCOLA DE MATEMÁTICA APLICADA**
**MESTRADO EM MODELAGEM MATEMÁTICA**

**LUCAS EMANUEL RESCK DOMINGUES**

**"BALANCING PERFORMANCE AND EXPLANATION PLAUSIBILITY: A MULTI-OBJECTIVE APPROACH TO TEXT CLASSIFICATION WITH HUMAN RATIONALES''.**

**DISSERTAÇÃO APRESENTADA AO CURSO DE MESTRADO EM MODELAGEM MATEMÁTICA PARA OBTENÇÃO DO GRAU DE MESTRE EM MODELAGEM MATEMÁTICA.**

ESTE É UM TRABALHO ORIGINAL ONDE FOI VERIFICADA A NÃO EXISTÊNCIA DE PLÁGIO E DE UTILIZAÇÃO DE INTELIGÊNCIA ARTIFICIAL, NÃO EXPLICITADA, NO CORPO DO TRABALHO

**DATA DA DEFESA: 13/05/2024**

**ASSINATURA DOS MEMBROS DA BANCA EXAMINADORA**

**PRESIDENTE DA COMISSÃO EXAMINADORA: PROFº JORGE LUIS POCO MEDINA**

*<ASSINADO ELETRONICAMENTE>*
**PROFº JORGE LUIS POCO MEDINA**
**ORIENTADOR**

*<ASSINADO ELETRONICAMENTE>*
**PROFº MARCOS MEDEIROS RAIMUNDO**
**CO-ORIENTADOR**

*<ASSINADO ELETRONICAMENTE>*
**PROFº ROBERTO DE ALENCAR LOTUFO**
**MEMBRO EXTERNO**

*<ASSINADO ELETRONICAMENTE>*
**PROFº THIAGO ALEXANDRE SALGUEIRO PARDO**
**MEMBRO EXTERNO**

RIO DE JANEIRO, 13 DE MAIO DE 2024.

*<ASSINADO ELETRONICAMENTE>*
**PROFº CÉSAR LEOPOLDO CAMACHO MANCO**
**DIRETOR**

*<ASSINADO ELETRONICAMENTE>*
**PROFº ANTONIO DE ARAUJO FREITAS JUNIOR**
**PRÓ-REITOR DE ENSINO, PESQUISA E PÓS-GRADUAÇÃO**

*To my late mother, Elisamara Resck, who dedicated her life to her children.*

# Acknowledgements

I thank every person and institution who contributed to my academic journey until now. In particular:

The Brazilian Mathematics Olympiad of Public and Private Schools (OBMEP) for the first opportunities it opened in my academic life.

The Fundação Getulio Vargas (FGV), the Center for the Development of Mathematics and Sciences (CDMC) of FGV, and the School of Applied Mathematics (EMAp) of FGV for the opportunities they opened for my college and graduate academic life, including financial support.

The National Council for Scientific and Technological Development (CNPq) and the Coordination for the Improvement of Higher Education Personnel (CAPES) for granting me funding.

My co-authors in the original paper on which this thesis is based, Jorge Poco and Marcos Medeiros Raimundo.

All the professors, teachers, and advisors who have gone through my academic life. In particular, my academic supervisor (BSc and MSc) Jorge Poco, who taught me how to do research.

The thesis committee for accepting to evaluate this work.

My girlfriend, Juliana, and my friends for emotional support.

And my family, especially my late mother, Elisamara Resck, who never measured efforts for my education.

The original paper on which this thesis is based was supported by 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 of the Association for Computational Linguistics for their important feedback.

# 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.

**Keywords**: machine learning. natural language processing. explainable AI.

# Resumo

Métodos de explicabilidade post-hoc de saliência são ferramentas importantes para entender modelos de PLN (Processamento de Linguagem Natural) cada vez mais complexos. Embora esses métodos possam refletir o raciocínio do modelo, eles podem não se alinhar com a intuição humana, tornando as explicações não plausíveis. Neste trabalho, apresentamos uma metodologia para incorporar *rationales*, que são anotações de texto explicando decisões humanas, em modelos de classificação de texto. Essa incorporação melhora a plausibilidade das explicações post-hoc, preservando sua fidelidade. Nossa abordagem é agnóstica em relação a arquiteturas de modelo e métodos de explicabilidade. Introduzimos os *rationales* durante o treinamento do modelo, expandindo a perda de entropia cruzada padrão com uma nova função de perda inspirada no aprendizado contrastivo. Ao utilizar um algoritmo de otimização multiobjetivo, exploramos a ponderação entre as duas funções de perda e geramos uma fronteira Pareto-ótima de modelos que equilibram desempenho e plausibilidade. Por meio de experimentos extensivos envolvendo diversos modelos, conjuntos de dados e métodos de explicabilidade, demonstramos que nossa abordagem melhora significativamente a qualidade das explicações do modelo sem causar degradação substancial (às vezes insignificante) no desempenho do modelo original.

**Palavras-chave**: aprendizado de máquina. processamento de linguagem natural. IA explicável.

# List of Figures

# List of Tables

# List of abbreviations and acronyms

| | |
|---|---|
| Adam | Adaptive Moment Estimation |
| AI | Artificial Intelligence |
| AUPRC | Area Under the Precision-Recall Curve |
| BERT | Bidirectional Encoder Representations from Transformers |
| BSc | Bachelor of Science |
| CNPq | Conselho Nacional de Desenvolvimento Científico e Tecnológico (National Council for Scientific and Technological Development) |
| FGV | Fundação Getulio Vargas |
| GPU | Graphics Processing Unit |
| LBFGS | Limited-memory Broyden-Fletcher-Goldfarb-Shanno |
| LIME | Local Interpretable Model-Agnostic Explanations |
| LM | Language Model |
| MIT | Massachusetts Institute of Technology |
| MLM | Masked Language Modeling |
| MOO | Multi-Objective Optimization |
| MSc | Master of Science |
| NAACL | North American Chapter of the Association for Computational Linguistics |
| NISE | Non-Inferior Set Estimation |
| NLP | Natural Language Processing |
| OOD | Out-of-Distribution |
| PhD | Doctor of Philosophy |
| SGD | Stochastic Gradient Descent |
| SHAP | SHapley Additive exPlanations |

| | |
|---|---|
| TF-IDF | Term Frequency-Inverse Document Frequency |
| TSE | Twitter Sentiment Extraction |
| UNICAMP | State University of Campinas |
| UNIREX | UNIfied Learning Framework for Rationale EXtraction |

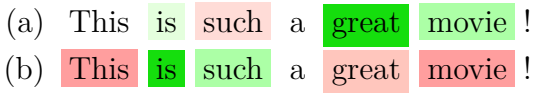
# Contents

# 1 Introduction

## 1.1 Context and Motivation

The complexity of text classification models and architectures has recently grown, posing challenges in comprehending the rationale behind their decisions. Consequently, the latest Natural Language Processing (NLP) algorithms have been called *black-box* algorithms. Understanding the model's reasoning is essential in various text classification contexts (RIBEIRO; SINGH; GUESTRIN, 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.

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. Source: the authors.

Researchers have developed popular text classification explainability techniques, such as post-hoc local saliency (or heatmap) methods (TJOA; 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; SINGH; GUESTRIN, 2016; LUNDBERG; LEE, 2017; CHEFER; GUR; WOLF, 2021) — although their suitability is criticized (BILODEAU et al., 2024), these methods are still widely applied (KUMARI; SINHA; EKBAL, 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; 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, Zhang, and Mooney (2019), Ross, Hughes, and Doshi-Velez (2017), Arous et al. (2021), Mengnan Du et al. (2019), and Mathew et al. (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; RUDRA; ANAND, 2021; PLUMB et al., 2020).

## 1.2 Objectives

This work aims to introduce 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; ZHANG; MOONEY, 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 objectives are:

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

(ii) Develop 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) Perform a series of experiments using various models, datasets, and explainability methods to verify whether there is a significant enhancement of model explanations without compromising the model's performance.

We also compare our methodology with a previous method from the literature. Furthermore, we address the social and ethical implications of "teaching" explanations to text classification models.

## 1.3 Outline

This thesis is organized as follows. Chapter 2 presents a related work on using rationales and the trade-off between performance and explainability. Chapter 3 covers some key aspects of explainability, multi-objective optimization, and contrastive learning. The proposed methodology is detailed in Chapter 4, with a series of experiments and results presented in Chapter 5. Chapter 6 discusses the proposed methodology's results, implications, and limitations, and Chapter 7 concludes the thesis.

## 1.4 Original Work

This Master's Thesis is a derivative work of the following original paper:

> RESCK, Lucas E.; RAIMUNDO, Marcos M.; POCO, Jorge. Exploring the Trade-off Between Model Performance and Explanation Plausibility of Text Classifiers Using Human Rationales. In: FINDINGS of the Association for Computational Linguistics: NAACL 2024. Mexico City, Mexico: Association for Computational Linguistics, June 2024. Available from: <https://arxiv.org/abs/2404.03098>

Therefore, figures, tables, and significant portions of the text are reproduced from the original paper. To avoid repetition, from now on, we will refer to the source of these contents simply as "the authors."

This paper was my main master's research project, supervised by Profs. Jorge Poco and Marcos M. Raimundo. The project started in late 2021 with the conceptualization of the idea and concluded in 2024 with the acceptance for publication in Findings of the Association for Computational Linguistics: NAACL 2024 and presentation at the North American Chapter of the Association for Computational Linguistics (NAACL) 2024 conference. Additionally, the paper was accepted for a poster presentation at the LatinX in NLP at NAACL 2024 workshop.

During a significant part of 2021, Prof. Jorge and I explored the literature on the intersection of explainability and NLP. At the end of the same year, we found text classification datasets that included rationales, which motivated us to explore their incorporation into the model training process. However, I was not comfortable with the idea of training models to "predict" explanations, as previous work did, because this could lead to unfaithful explanations: the model would learn to output human-desired explanations, not necessarily faithfully explaining its internal reasoning. That is why we decided to explore a model- and explainer-agnostic approach.

Prof. Marcos suggested the multi-objective optimization approach, and Prof. Vicente Ordóñez (Rice University) suggested the contrastive-inspired learning. We wrote the paper during 2022 and submitted it to conferences a few times in 2023. It was finally accepted in NAACL Findings in 2024.

The inspiration to explore research ideas in explainable NLP came after my successful experience in the LegalVis paper, in which I applied NLP and explainability methods to assist a visual analytics system in exploring citations in legal documents (RESCK; PONCIANO, et al., 2023) (Section 1.5). Furthermore, the present work also justified contacting Prof. Vicente Ordóñez for feedback, which led me to a three-month research visit at Rice University in 2022. Finally, this work also instigated my interest in pursuing a PhD in topics related to NLP, Machine Learning, and Explainable AI, which I will start in 2024 at the University of Cambridge.

## 1.5 Other Works

During my master's and final undergraduate studies, I also worked on other research projects related to explainability and NLP.

In collaboration with Prof. Diego Mesquita (Fundação Getulio Vargas), my co-authors and I developed a novel explainer for graph neural networks, published in and presented at Artificial Intelligence and Statistics (AISTATS) 2023. My main contribution was a theoretical analysis of the explainer's properties through lemmas and theorems.

> PEREIRA, Tamara; NASCIMENTO, Erik; RESCK, Lucas E.; MESQUITA, Diego; SOUZA, Amauri. Distill n' Explain: explaining graph neural networks using simple surrogates. en. In: PROCEEDINGS of The 26th International Conference on Artificial Intelligence and Statistics. Valencia, Spain: PMLR, Apr. 2023. v. 206. (Proceedings of Machine Learning Research), p. 6199–6214. ISSN: 2640-3498. Available from: <https://proceedings.mlr.press/v206/pereira23a.html>

At the end of my undergraduate studies, Prof. Poco, my co-authors, and I proposed a novel visualization methodology and system to explore citations in legal documents. I was responsible for developing the NLP and explainability components. The paper was written and submitted during my undergraduate. However, it was presented and published during my master's at the Visualization and Visual Analytics (VIS) 2022 conference and in Transactions on Visualization and Computer Graphics (TVCG) (2023), respectively.

> RESCK, Lucas E.; PONCIANO, Jean R.; NONATO, Luis Gustavo; POCO, Jorge. LegalVis: Exploring and Inferring Precedent Citations in Legal Doc-

uments. English. **IEEE Transactions on Visualization and Computer Graphics**, v. 29, n. 6, p. 3105–3120, June 2023. Presented at IEEE VIS: Visualization & Visual Analytics 2022. Date of Publication: 18 February 2022. ISSN 1941-0506. DOI: 10.1109/TVCG.2022.3152450. Available from: <https://ieeexplore.ieee.org/document/9716779/>

# 2 Related Work

Our work draws on prior research in the areas of rationale utilization and the trade-off between performance and explainability. This chapter covers key aspects of these topics and discusses how our work distinguishes itself from existing literature.

## 2.1 Use of Rationales

Using human annotations to assist machine learning is not a novel concept, as prior works have shown (ZAIDAN, O.; EISNER; PIATKO, C., 2007; ZAIDAN, O. F.; EISNER; PIATKO, C. D., 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; AVCI, 2019), while others have explored techniques to extract (ZHANG; RUDRA; ANAND, 2021; LAKHOTIA et al., 2021; PRUTHI et al., 2020; SHARMA; MINER, et al., 2020) or generate (RAJANI et al., 2019; LIU; YIN; WANG, 2019; CAMBURU et al., 2018; KUMAR; 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; BILGIC, 2018; BAO et al., 2018; LIU; YIN; WANG, 2019; RIEGER et al., 2020; ZHANG; RUDRA; ANAND, 2021; AROUS et al., 2021; MATHEW et al., 2021; CARTON; KANORIA; TAN, 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), Mengnan Du et al. (2019), and 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; ZHANG; MOONEY, 2019; CHEN; JI, 2020; LIU; LIN, et al., 2022; SEKHON et al., 2023)) or assuming a specific type of explanation function (e.g., by using input gradients (ROSS; HUGHES; DOSHI-VELEZ, 2017; GHAEINI et al., 2019)). Such interventions are debatable (see Chapter 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, i.e., regularizes the model by desiderata, but does not leverage human annotations (LEI; BARZILAY; JAAKKOLA, 2016; BASTINGS; AZIZ; TITOV, 2019; JAIN et al., 2020).

## 2.2 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; YU; LEI, 2020; DUBEY; RADENOVIC; MAHAJAN, 2022; PLUMB et al., 2020; RADENOVIC; DUBEY; MAHAJAN, 2022). However, differing opinions exist on whether this trade-off always holds, both from a theoretical perspective (JACOVI; GOLDBERG, 2021; RUDIN, 2019) and a practical standpoint (HASE et al., 2020). Furthermore, some studies have empirically examined or explored this trade-off (ZHANG; RUDRA; ANAND, 2021; GOETHALS; MARTENS; EVGENIOU, 2022; NAYLOR et al., 2021; PARANJAPE et al., 2020; JIN; SENDHOFF; KÖRNER, 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.

## 2.3 Final Remarks

Prior work has extensively explored the use of rationales to enhance model performance and explanations. However, our work distinguishes itself by focusing on the plausibility of the explanations and using a model- and explainer-agnostic approach. Furthermore, we explore the trade-off between accuracy and plausibility using a multi-objective optimization algorithm, contributing to the ongoing debate on the trade-off between performance and explainability. The next chapter presents the theoretical background that forms the basis of our approach.

# 3 Theoretical Background

We define crucial explainability, multi-objective optimization, and contrastive learning concepts to facilitate a global understanding of our research.

## 3.1 Explainability

Explainability is a key aspect of machine learning models and has been extensively studied in recent years. In this section, we introduce concepts related to explanations and their properties.

**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; KANORIA; TAN, 2022; MATHEW et al., 2021). Typically, these rationales are annotated by humans and serve as ground truth explanations for the corresponding categories. Figure 2 presents a typical rationale for the example of Figure 1.

Figure 2 – Example of rationale of a positive movie review.

This is such a great movie !

Green indicates the rationale. Source: the authors.

**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; 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; GOLDBERG, 2021). Please refer to Chapter 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. They measure how much the model probability changes after removing the explanation from the text source and keeping only the explanation, respectively. 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, Ferreira, and Von Zuben (2020) provided for the following concepts.

**Definition 3.2.1** (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

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

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 3.2.2** (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. One of them is the *weighted sum method*: it balances the objective functions and converts the problem into a uni-objective form.

**Definition 3.2.3** (Weighted sum method)**.** Given a MOO problem as in Definition 3.2.1, 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, Ferreira, and Von Zuben (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)).

## 3.3 Contrastive Learning

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, Kornblith, 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.

## 3.4 Final Remarks

Our work builds on the theoretical basis presented in this chapter to propose improving the quality of explainability in text classification models using methods drawn from or inspired by multi-objective optimization and contrastive learning. The next chapter builds on these concepts to describe our proposed methodology: a novel approach for enhancing the explanation plausibility of text classification models by incorporating human rationales.

# 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 $f_\theta \colon \mathbb{R}^d \to \Delta$. 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 $\theta$. Examples of $x$ include TF-IDF vectors (LESKOVEC; RAJARAMAN; ULLMAN, 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 $f_\theta$ as a black box without assuming any specific structure. Let us introduce the explanation function[1] $e_{f_\theta,k} \colon \mathbb{R}^d \to \mathbb{R}^p$, 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 $m \colon \mathbb{R}^p \times \{0,1\}^p \to \mathbb{R}$ between $e_{f_\theta,k}(x)$ and $e_{x,k}$ quantifies the quality of explanations extracted from $f_\theta$ 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:

$$\mathcal{L}_\theta(X, y) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{|C|} \mathbb{1}_{y_i=k} \ln \frac{e^{g_\theta(X_i)_k}}{\sum_{j=1}^{|C|} e^{g_\theta(X_i)_j}}, \tag{4.1}$$

where, $g_\theta$ represents the logits (pre-softmax) obtained from $f_\theta$, and $f$ corresponds to the softmax function applied to $g_\theta$. For illustration purposes, Figure 3 presents an example of $X$ and $y$. It is worth noting that $\theta$ 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.

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

Figure 3 – Example of $X$ and $y$.

$$X = \begin{Bmatrix} \text{“This is a great movie!”} \\ \text{“This is a terrible movie!”} \\ \text{“I don't like this film.”} \end{Bmatrix}, \quad y = \begin{Bmatrix} \text{“Positive”} \\ \text{“Negative”} \\ \text{“Negative”} \end{Bmatrix}$$

$X$ is in fact a set of textual vectors. Representing it as a set of raw texts is merely illustrative. Source: the authors.

## 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, M. et al., 2019; ROSS; HUGHES; DOSHI-VELEZ, 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[2].

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; KORNBLITH, et al., 2020; KHOSLA et al., 2020) (Section 3.3), we introduce a novel auxiliary loss function known as the *contrastive rationale loss*:

$$\dot{\mathcal{L}}_\theta(\dot{X}, \dot{y}) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{|C|} \mathbb{1}_{\dot{y}_i = k} \ln \frac{e^{g_\theta(\dot{X}_i)_k}}{\sum_{j=1}^{m} e^{g_\theta(\tilde{X}_{i,j})_k}}, \tag{4.2}$$

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. For illustration purposes, Figure 4 extends Figure 3 by presenting an example of $\dot{X}$, $\dot{y}$, and $\tilde{X}$. 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 4.2, contrary to previous work (Chapter 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 Chapter 6.

[2] 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.

Figure 4 – Example of $\dot{X}$, $\dot{y}$, and $\tilde{X}$.

$$X = \begin{Bmatrix} \text{“This is a great movie!”} \\ \text{“This is a terrible movie!”} \\ \text{“I don't like this film.”} \end{Bmatrix}, \quad \dot{X} = \begin{Bmatrix} \text{“great movie”} \\ \text{“terrible movie”} \\ \text{“don't like”} \end{Bmatrix},$$

$$\dot{y} = y = \begin{Bmatrix} \text{“Positive”} \\ \text{“Negative”} \\ \text{“Negative”} \end{Bmatrix}, \quad \tilde{X} = \begin{Bmatrix} \{ & \text{“great movie”} & \text{“This a”} & \text{“This great”} & \} \\ \{ & \text{“terrible movie”} & \text{“This a”} & \text{“a terrible”} & \} \\ \{ & \text{“don't like”} & \text{“I don't”} & \text{“like film”} & \} \end{Bmatrix}$$

$\dot{X}$ is the set of rationales from $X$, $\dot{y}$ is equal to $y$ in this example, and $\tilde{X}$ is the set of random rationales. In particular, $\tilde{X}$ includes the ground truth rationale $\dot{X}$ in the first column, and the other columns include random tokens of $X$. This means that each line $\tilde{X}_i$ of $\tilde{X}$ contains the rationale and two negative rationales of $X_i$. Source: the authors.

Intuitively, what the contrastive rationale loss essentially does is *train the model to predict the correct class based on the correct rationale and to not predict the correct class based on wrong rationales*. The goal is to make the model's internal reasoning more dependent on the human explanations, which should be reflected in the model explanations.

### 4.2.1 Contrastive Rationale Loss for Logistic Regression

The contrastive rationale loss constitutes a particular case when the classifier is a multinomial logistic regression and it deserves a highlight. When the model $f_\theta$ is a multinomial logistic regression over text embedding vectors, we can represent the contrastive rationale loss function in the following way:

$$\dot{\mathcal{L}}_\theta(\dot{X}, \dot{y}) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{k=1}^{|C|} \mathbb{1}_{\dot{y}_i = k} \ln \frac{\exp(\dot{X}_i \cdot \theta_k)}{\sum_{j=1}^{m} \exp(\tilde{X}_{i,j} \cdot \theta_k)}. \tag{4.3}$$

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 4.3, one is training an *anchor* $\theta_k$ 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 $\theta$, 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) (Section 3.2).

## 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 (Section 3.2), enable trade-off exploration by incorporating hyperparameters $w_1$ and $w_2$ (both $\geq 0$) 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 $\Theta$, where each $\theta \in \Theta$ corresponds to a different classifier $f_\theta \in F_\Theta$. Finally, by searching within the set $F_\Theta$, we can identify Pareto-optimal models that exhibit both performance and plausibility.

## 4.4 Final Remarks

Our methodology introduces a novel contrastive-inspired loss function that, combined with the cross-entropy loss, aims to enhance the plausibility of model explanations. A multi-objective optimization algorithm explores the trade-off between the two losses and, therefore, between performance and explanation plausibility. The simplicity of the logistic regression case illustrates the ease of applying this approach to existing models. The next chapter presents a series of experiments that assess our methodology under various conditions, including different datasets, models, and explainability methods. In particular, we also compare our approach with a previous method from the literature.

# 5 Experiments

This chapter describes experiments to test the methodology proposed in Chapter 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 A.

## 5.1 Experimental Setup

The main experiments of this chapter evaluate our proposed methodology using different models, datasets, and explainability techniques. In this section, we describe this setup in detail.

### 5.1.1 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 A.

**TF-IDF with Logistic Regression.** For classical models, we train a multinomial logistic regression model using TF-IDF vectors (LESKOVEC; RAJARAMAN; ULLMAN, 2020) (unigrams) with dimensionality reduction to 200 achieved through Truncated Singular Value Decomposition (MANNING; RAGHAVAN; SCHÜTZE, 2008).

### 5.1.2 Datasets and Data Preprocessing

We use three datasets to evaluate our methodology: HateXplain, Twitter Sentiment Extraction (TSE), and Movie Reviews. Table 1 summarizes the datasets.

**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. Because more than one annotator is used for each sample, we apply majority consensus to both rationale and class assignments, disregarding non-consensual samples. Figure 5 shows an example of a sample from the HateXplain dataset for illustration.

Figure 5 – Example of a sample from the HateXplain dataset with two negative rationales.

```
{'tokens': ['poland', 'says', 'no', 'to', 'mu**ie', 'invasion'],
'label': 'hatespeech',
'rationales': [0, 0, 0, 0, 1, 0],
'negative_rationales': [[1, 0, 0, 0, 0, 0], [0, 0, 1, 0, 0, 0]]}
```

In our implementation, $X_i$ is represented by `tokens`, $y_i$ and $\dot{y}_i$ are represented by `label`, $\dot{X}_i$ is represented by the binary vector `rationales`, and the negative rationales $\{\tilde{X}_{i,j}\}_{j=1}^{m} \backslash \{\dot{X}_i\}$ are represented by the list of binary vectors `negative_rationales`. Notice that the negative rationales are random annotations of the same length as the rationale. Source: the authors.

**Twitter Sentiment Extraction (TSE).** The TSE (MAGGIE; CULLITON; CHEN, 2020) is a sentiment analysis dataset containing positive, negative, and neutral tweets with human-annotated rationales. Since neutral class lacks rationales[1], 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.

Table 1 – Description of the datasets.

| 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 |

Source: the authors.

Note: The description of the datasets is after filtering. HateXplain average rationale length is calculated over the hate speech class only, and `hatexplain_all`, over hate speech and offensive classes.

**Movie Reviews.** This dataset comprises positive and negative movie reviews with rationales annotated by humans to support classification (ZAIDAN, O.; EISNER; PIATKO, C., 2007).

### 5.1.3 Explainability Methods

To extract saliency post-hoc model explanations, we consider the two well-known LIME and SHAP explainers. Explainability methods are only used during test to extract

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

post-hoc model explanations and evaluate their quality by comparing them with the ground truth human rationales.

**LIME.** Short for *Local Interpretable Model-agnostic Explanations* (RIBEIRO; SINGH; GUESTRIN, 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. More specifically, LIME randomly removes tokens from the text sample and fits a linear regression to predict the model outputs from the binary vectors representing the presence of each token. Then, the coefficients of the linear regression are used to explain the model's prediction, i.e., the higher the coefficient, the more important the token is. Figure 6 shows an example of a LIME explanation.

Figure 6 – Example of a LIME explanation of a positive movie review.

This is such a great movie !

Green means a positive contribution to the model's prediction, while red means a negative contribution. LIME explanations are continuous with token scores ranging from $-\infty$ to $+\infty$. Source: the authors.

**SHAP.** *SHapley Additive exPlanations* (LUNDBERG; LEE, 2017) is a model-agnostic explainer that employs Shapley values to explain model predictions, similar to LIME.

### 5.1.4 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).

## 5.2 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 5.3 shows other results. The explainability metrics (plausibility and faithfulness) are computed only for the hate speech class because the normal class lacks rationales.

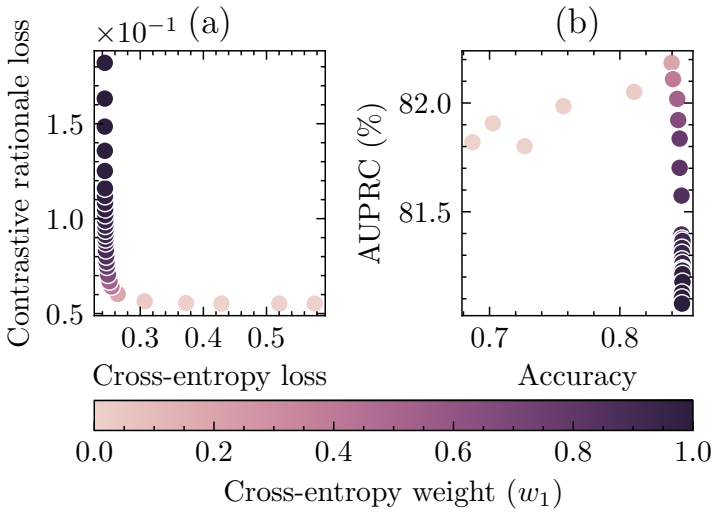
The DistilBERT model trained only with cross-entropy loss achieves a test accuracy of 84.8% with balanced recall among classes. Figure 7 (a) illustrates an example of a bad explanation extracted from this model. It shows that even high-performing classifiers can also present unreasonable explanations.

Figure 7 – Examples of explanations of the hate speech class.

(a) ugh i hate d*kes 😐
(b) ugh i hate d*kes 😐

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. Source: the authors.

Figure 8 – (a) Trade-off between the two losses on the training data. (b) Trade-off between accuracy and plausibility of the test data.



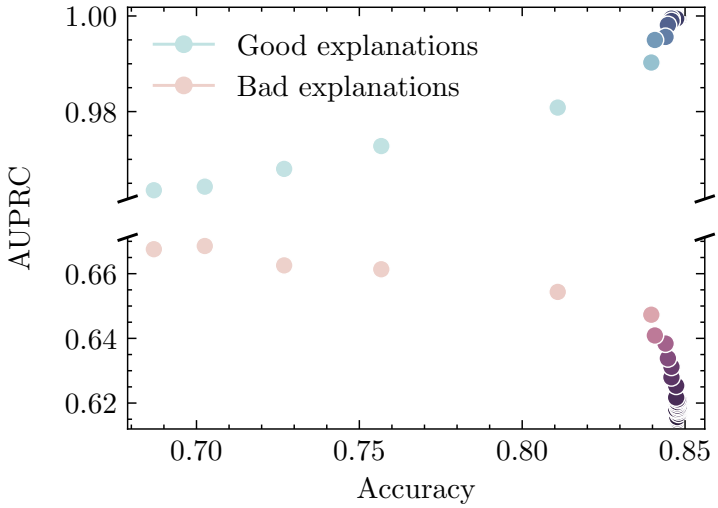

The color scale represents the cross-entropy weight $w_1$ (Section 4.3). We ignore the model with $w_1 = 0$ as it is out of scale. Results including $w_1 = 0$ and shared scale between axes are in Appendix B. Source: the authors.

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 8 (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 on the test data (Figure 8 (b)). Plausibility was measured using

mean AUPRC, comparing LIME's explanations with ground truth rationales. Figure 8 (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 9). The inadequate explanations are being improved without significantly harming the good explanations (see example in Figure 7; more examples in Table 6).

Figure 9 – Trade-off between performance and plausibility on test data for originally good (AUPRC = 1) and originally bad (AUPRC < 1) explanations differently.

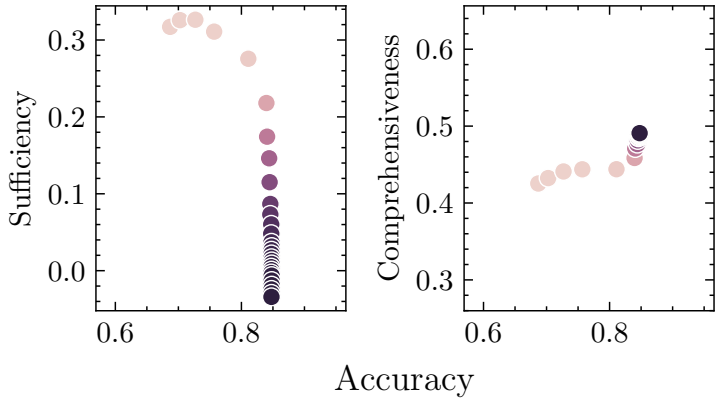



The color scale is the same as the previous figures. Source: the authors.

Finally, we must guarantee faithful explanations (i.e., they genuinely represent the models' reasoning) when we strengthen the training with rationales. Figure 10 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 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.*

Figure 10 – Trade-off between accuracy and faithfulness (sufficiency and comprehensiveness) on test data.



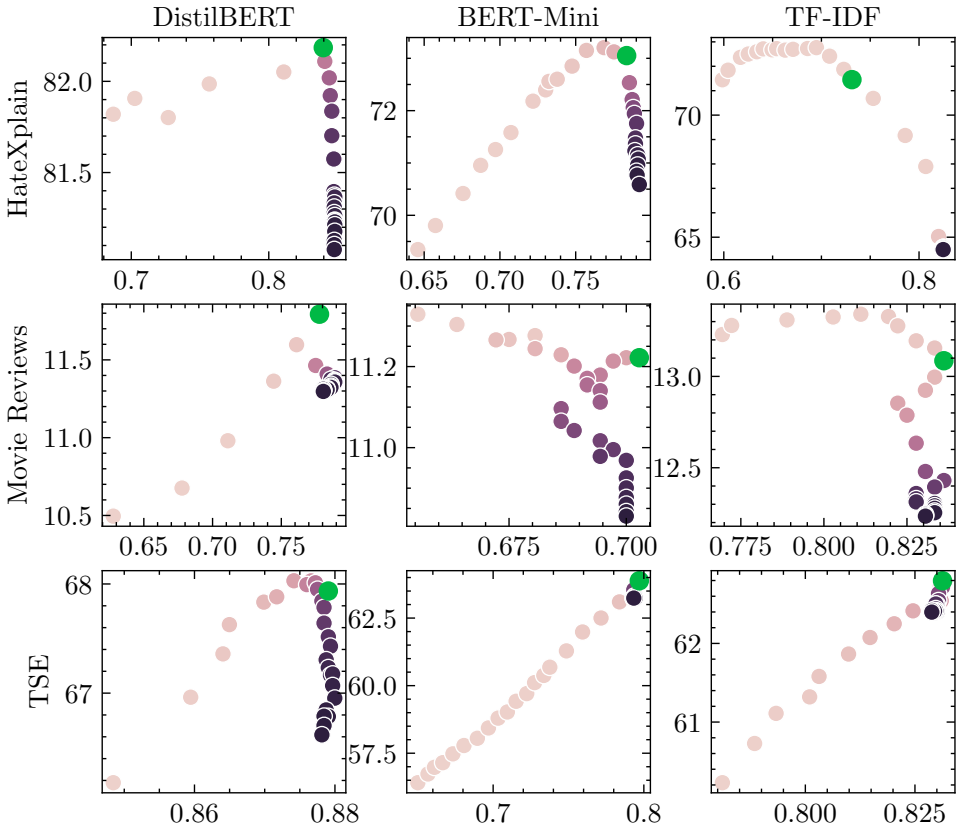

Higher values are better. The color scale is the same as the previous figures. The data scale is equal between the two graphics and their x- and y-axes. Source: the authors.

## 5.3 Experiments With All Models and Datasets

Now, we evaluate our framework in all models, datasets, and explainability techniques that we consider in this work. 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 11 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 B). Figure 11 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 21) and similar when the explainer is SHAP (Figures 22 and 23). Finally, despite the TSE dataset having a higher number of poor-performing models, the improvement for a well-selected model is not negligible (Table 2).

The green dots in Figure 11 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 2 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

Figure 11 – 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. Results including $w_1 = 0$, 5 rationales and/or SHAP, shared scale between axes, and Pareto-frontiers are in Appendix B. Source: the authors.

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 11, 21, 22 and 23 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.*

Table 2 – Comparison between the original model (cross-entropy only) and the chosen model (green dots on Figure 11) for each performance and explainability metric on test data.

| 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 |

Source: The authors.

Note: "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 Table 8.

## 5.4 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[2], 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 12 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

Table 3 – Comparison between the chosen models (circled points in Figure 12) of BERT-HateXplain and our method on test data.

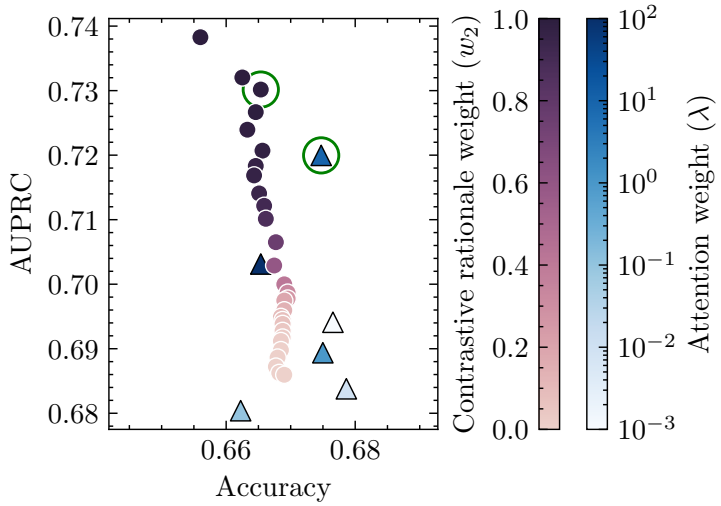
| Model | Acc. | AUPRC | Suff. | Comp. |
|---|---|---|---|---|
| HateXplain | **67.47** | 72.00 | 0.12 | **0.53** |
| Ours | 66.54 | **73.02** | **0.14** | 0.40 |

Source: The authors.

Note: Accuracy and AUPRC are in percentage (%).

[2] Their attention loss is multiplied by a "trade-off" hyperparameter $\lambda$. We use their suggestion of $\lambda$ values (Appendix A).

Figure 12 – 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 $\lambda$ (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. Source: the authors.

"good judgment" of improving AUPRC without hurting too much accuracy. Table 3 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.

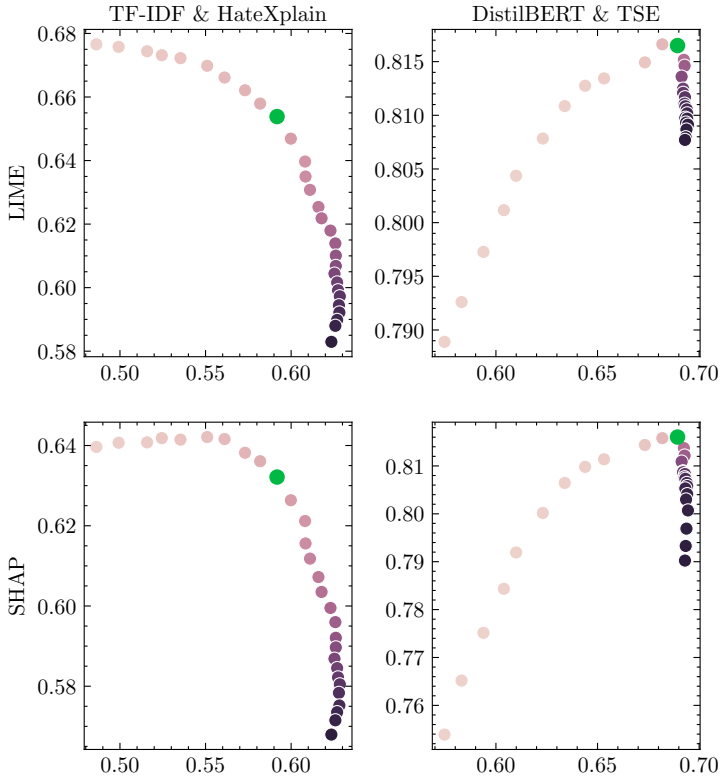
## 5.5 Additional Results

In the previous sections, we tested our methodology with different models, datasets, and explainability techniques. We perform further experiments to verify the robustness of our framework under additional, different conditions, e.g., non-binary classification and larger models.

### 5.5.1 Non-Binary Classification

Sections 5.2 and 5.3 present results for all datasets but are binary classification. As pointed out in Section 5.1.2, 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 4.2. Figure 13 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 5.3. However, different datasets lead to different absolute values. Finally, in a similar way to Section 5.3, 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 13 – 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. Source: the authors.

Table 4 – Comparison between the original model (cross-entropy only) and the chosen model (green dots on Figure 13) for each performance and explainability metric on test data.

| 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 |

Source: the authors.

Note: "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.

### 5.5.2 Larger Models

The previous sections present 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 Masked Language Model (MLM) fine-tuning. The shapes of the model frontiers (Figure 14) were similar to other language model frontiers of Figure 11. 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.

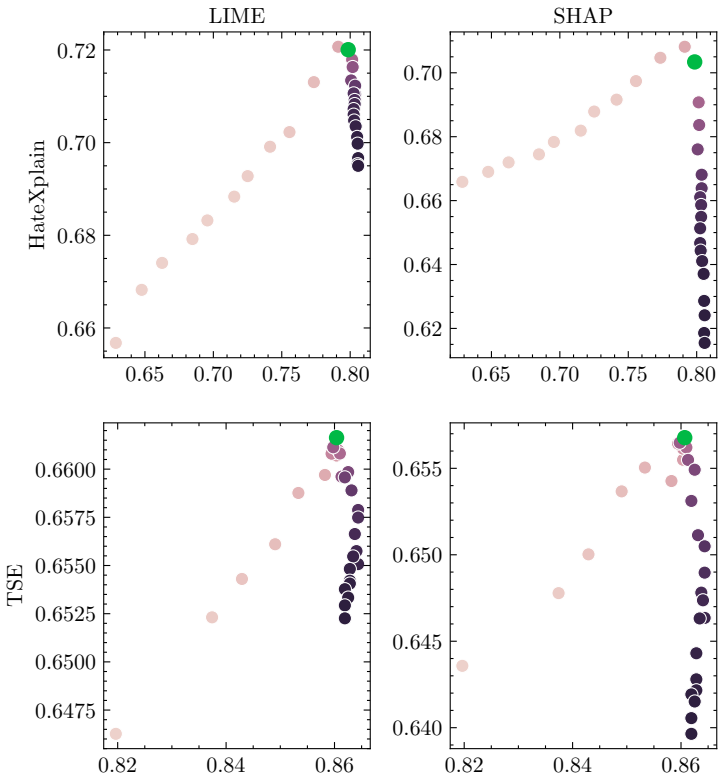
Table 5 – Comparison between the original model (cross-entropy only) and the chosen model (green dots on Figure 14) for each performance and explainability metric on test data.

| 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 |

Source: the authors.

Note: "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.

Figure 14 – Trade-offs between performance (accuracy, x-axis) and plausibility (AUPRC, y-axis, in percentage (%)) for BERT-Large with HateXplain and TSE (test data).

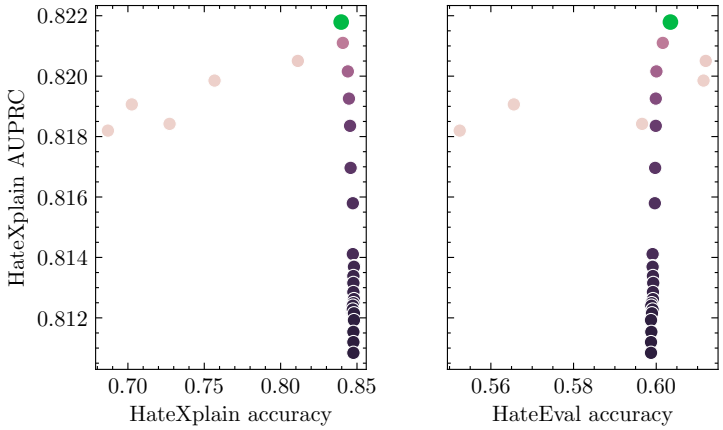



The number of random (negative) rationales is 2, and the explainers are LIME and SHAP. The color scale is the same as the previous figures. We ignore the model with $w_1 = 0$ in all graphics as it is out of scale. Green dots are the models chosen to be analyzed more carefully. Source: the authors.

### 5.5.3 Out-of-Distribution

While training models to have more plausible explanations, we are shifting the model's decision rationale towards a human's. Because a human rationale has better generalization, we hypothesize that the model will generalize better to out-of-distribution (OOD) data, i.e., data from a different distribution than the training data. To test OOD performance, we additionally evaluated the DistilBERT trained on HateXplain (Section 5.2 of the main text) 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 15 is roughly similar to the frontier shape of HateXplain performance (in the same Figure and in Figure 8) 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 15), while original accuracy decreases by 0.8% and plausibility increases by approximately 1.1%, 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.

Figure 15 – 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. Source: the authors.

# 6 Discussion

In this work, we propose a novel methodology to improve the plausibility of model explanations by incorporating human rationales. We extensively test our approach with various datasets, models, and explainers, demonstrating its effectiveness and robustness under different conditions. Furthermore, we argue that our method is model- and explainer-agnostic, differentiating it from previous work. Several questions may arise with all the presented results, which we discuss in this section.

### Should We Model Plausibility?

Jacovi and Goldberg (2021) argue that explanation plausibility should not be modeled and pursued because it is an ethical issue: by pursuing it, the explainer role would not be presenting justifications for the model decision but, instead, the explainer would be interested in convincing the user of the model decision, possibly providing unfaithful explanations to achieve this goal (i.e., lying to the user). Our perspective is different: even though we announce our methodology as "improving plausibility," we are updating the model's internal decision to produce better explanations. Instead of creating explanations to convince the explainee of the model decision, our objective is to provide a new model with a different internal decision process capable of delivering better, faithful explanations. To achieve that, we guarantee that the human rationales are not influencing the explanation extraction process (e.g., the model explainer is not "trained" with rationales) and that the explainability method is not incorporated into the training process (i.e., the model does not learn how to tweak explanations). Our perspective is more aligned with Zhou, Ribeiro, and Shah (2022) who defends that plausibility contributes to *understandability*: "given the same level of correctness, a higher-alignment explainer may be preferable" (ZHOU; RIBEIRO; SHAH, 2022).

Additionally, in high-stakes applications, in which the understanding of the model's decision is crucial, the explanation plausibility plays a key role. A more plausible explanation may indicate a more trustworthy model that is less likely to rely on spurious correlations. For instance, in the legal domain, a model that provides equally faithful but more plausible explanations is more likely to be making the right juridical decision for the right juridical reasons and, therefore, is more adequate to be used for legal decision support.

### 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. However, one could point out that the experimental results of Table 2 indicate against this trade-off—the possibility of improving plausibility without harming much of performance, or even improving both of them at the same time, would contradict the trade-off hypothesis. We point out that Table 2 compares a manually chosen model to the original model trained without any rationale. The Pareto frontier in Figure 24 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. Chapter 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 4.1 and 4.2) and experiments exploring the existence of this trade-off. Finally, the trade-off may also be dependent on the research area and, consequently, the dataset used. Future work may explore this hypothesis.

### 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; ZHANG; MOONEY, 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 Section 4.2.1). 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.

### 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 (Section 5.5.1). 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).

## 6.1 Limitations and Future Work

Despite the promising results, our work has limitations that we did not address in this thesis. We leave these limitations as opportunities for future work.

### 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 A).

### 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; MARASOVIC, 2021). Naturally, this limitation also affects the availability of different types of classification datasets beyond sentiment analysis and hate speech detection, which we considered in this work. 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 Kevin 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; MARASOVIC, 2021). Finally, recent advances in crowdsourcing annotation systems allow an efficient annotation of datasets at scale (DRUTSA et al., 2021), including the use of large language models to generate annotations (ZHANG; RUDRA; ANAND, 2021).

### Human Study

Consistent with precedents in the field (MATHEW et al., 2021; ROSS; HUGHES; DOSHI-VELEZ, 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.

The experiments show that we may trade a little accuracy for a more interpretable model sometimes. We plan to study as future work why the performance decreases by manually inspecting the examples in which the model becomes incorrect. This study would allow for a more qualitative understanding of the methodology and, potentially, the trade-off between accuracy and interpretability.

### 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; RUDRA; ANAND, 2021; LAKHOTIA et al., 2021; AROUS et al., 2021; STROUT; ZHANG; MOONEY, 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.

#### Automatic Model Selection

In Section 5.3, we select the model with best trade-off between accuracy and plausibility. The choice is manual and, in some cases, based on some degree of "good judgement." Future work may include automatic model selection, e.g., Kneedle (SATOPAA et al., 2011) and Branke et al. (2004).

## 6.2 Ethics Statement

Some authors consider pursuing plausibility as an ethical issue (JACOVI; GOLDBERG, 2021). Part of this work argues this is not the case (Chapter 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.

# 7 Conclusion

In this work, 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.

In summary, we effectively achieved the objectives set out in the Chapter 1:

(i) We proposed a novel contrastive-inspired loss function that effectively incorporates rationales into the learning process.

(ii) We developed 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) We run 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.

(iv) Additionally, we compared 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.

The related work and theoretical background present theoretical support for our approach, while the additional experimental results demonstrate its robustness and gen-

eralizability. Finally, the simplicity of the logistic regression case illustrates the ease of implementing and integrating our approach into existing models.

We hope our findings inspire other researchers to explore the integration of human knowledge into machine learning and NLP models, fostering a more explainable, transparent, and trustworthy AI ecosystem.

# References

AROUS, Ines et al. MARTA: Leveraging Human Rationales for Explainable Text Classification. en. In: PROCEEDINGS of the AAAI Conference on Artificial Intelligence. Virtual: AAAI Press, May 2021. v. 35, p. 5868–5876. Number: 7. Available from: <https://ojs.aaai.org/index.php/AAAI/article/view/16734>. Cit. on pp. 22, 27, 55.

ATANASOVA, Pepa et al. A Diagnostic Study of Explainability Techniques for Text Classification. English. In: PROCEEDINGS of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, Nov. 2020. P. 3256–3274. DOI: 10.18653/v1/2020.emnlp-main.263. Available from: <https://aclanthology.org/2020.emnlp-main.263>. Cit. on p. 53.

BAO, Yujia et al. Deriving Machine Attention from Human Rationales. In: PROCEEDINGS of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium: Association for Computational Linguistics, Oct. 2018. P. 1903–1913. DOI: 10.18653/v1/D18-1216. Available from: <https://aclanthology.org/D18-1216>. Cit. on p. 27.

BASILE, Valerio et al. 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. Minneapolis, Minnesota, USA: Association for Computational Linguistics, June 2019. P. 54–63. DOI: 10.18653/v1/S19-2007. Available from: <https://aclanthology.org/S19-2007>. Cit. on p. 49.

BASTINGS, Jasmijn; AZIZ, Wilker; TITOV, Ivan. Interpretable Neural Predictions with Differentiable Binary Variables. In: PROCEEDINGS of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, July 2019. P. 2963–2977. DOI: 10.18653/v1/P19-1284. Available from: <https://aclanthology.org/P19-1284>. Cit. on p. 28.

BELÉM, Catarina et al. Weakly Supervised Multi-task Learning for Concept-based Explainability. In: PROCEEDINGS of the First Workshop on Weakly Supervised Learning (WeaSuL). Virtual: [s.n.], May 2021. Available from: <https://weasul.github.io/papers/26.pdf>. Cit. on pp. 27, 28.

BILODEAU, Blair et al. Impossibility theorems for feature attribution. **Proceedings of the National Academy of Sciences**, v. 121, n. 2, Jan. 2024. DOI: 10.1073/pnas.2304406120. Available from: <https://www.pnas.org/doi/abs/10.1073/pnas.2304406120>. Cit. on p. 21.

BRANKE, Jürgen et al. Finding Knees in Multi-objective Optimization. en. In: PARALLEL Problem Solving from Nature - PPSN VIII. Berlin, Heidelberg: Springer, 2004. P. 722–731. ISBN 978-3-540-30217-9. DOI: 10.1007/978-3-540-30217-9_73. Available from: <https://link.springer.com/chapter/10.1007/978-3-540-30217-9_73>. Cit. on p. 55.

CAMBURU, Oana-Maria et al. e-SNLI: Natural Language Inference with Natural Language Explanations. In: ADVANCES in Neural Information Processing Systems. Palais des Congrès de Montréal, Montréal, Canada: Curran Associates, Inc., 2018. v. 31. Available from: <https://papers.nips.cc/paper/2018/hash/4c7a167bb329bd92580a99ce422d6fa6-Abstract.html>. Cit. on pp. 27, 28, 53.

CARTON, Samuel; KANORIA, Surya; TAN, Chenhao. What to Learn, and How: Toward Effective Learning from Rationales. In: FINDINGS of the Association for Computational Linguistics: ACL 2022. Dublin, Ireland: Association for Computational Linguistics, May 2022. P. 1075–1088. DOI: 10.18653/v1/2022.findings-acl.86. Available from: <https://aclanthology.org/2022.findings-acl.86>. Cit. on pp. 27, 29.

CHAN, Aaron et al. UNIREX: A Unified Learning Framework for Language Model Rationale Extraction. English. In: PROCEEDINGS of BigScience Episode #5 – Workshop on Challenges & Perspectives in Creating Large Language Models. virtual+Dublin: Association for Computational Linguistics, May 2022. P. 51–67. DOI: 10.18653/v1/2022.bigscience-1.5. Available from: <https://aclanthology.org/2022.bigscience-1.5>. Cit. on pp. 27, 44, 55.

CHEFER, Hila; GUR, Shir; WOLF, Lior. Transformer Interpretability Beyond Attention Visualization. en. In: PROCEEDINGS of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [S.l.: s.n.], 2021. P. 782–791. Available from: <https://openaccess.thecvf.com/content/CVPR2021/html/Chefer_Transformer_Interpretability_Beyond_Attention_Visualization_CVPR_2021_paper.html>. Cit. on p. 21.

CHEN, Hanjie; JI, Yangfeng. Learning Variational Word Masks to Improve the Interpretability of Neural Text Classifiers. en. In: PROCEEDINGS of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, 2020. P. 4236–4251. DOI: 10.18653/v1/2020.emnlp-main.347. Available from: <https://www.aclweb.org/anthology/2020.emnlp-main.347>. Cit. on p. 27.

CHEN, Ting; KORNBLITH, Simon, et al. A Simple Framework for Contrastive Learning of Visual Representations. en. In: PROCEEDINGS of the 37th International Conference on Machine Learning. [S.l.]: PMLR, July 2020. v. 119. (Proceedings of Machine Learning

Research), p. 1597–1607. ISSN: 2640-3498. Available from: <https://proceedings.mlr.press/v119/chen20j.html>. Cit. on pp. 31, 34.

COHON, Jared L. **Multiobjective Programming and Planning**. 1. ed. [S.l.]: Academic Press, 1978. v. 140. (Mathematics in Science and Engineering). ISBN 978-0-08-095649-7 978-0-12-178350-1. Available from: <https://www.elsevier.com/books/multiobjective-programming-and-planning/cohon/978-0-12-178350-1>. Cit. on pp. 31, 35, 36, 40.

DEVLIN, Jacob et al. 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). Minneapolis, Minnesota: Association for Computational Linguistics, June 2019. P. 4171–4186. DOI: 10.18653/v1/N19-1423. Available from: <https://aclanthology.org/N19-1423>. Cit. on pp. 33, 37, 47.

DEYOUNG, Jay et al. ERASER: A Benchmark to Evaluate Rationalized NLP Models. In: PROCEEDINGS of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, July 2020. P. 4443–4458. DOI: 10.18653/v1/2020.acl-main.408. Available from: <https://aclanthology.org/2020.acl-main.408>. Cit. on pp. 21, 29, 39.

DRUTSA, Alexey et al. Crowdsourcing Natural Language Data at Scale: A Hands-On Tutorial. English. In: PROCEEDINGS of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Tutorials. Online: Association for Computational Linguistics, June 2021. P. 25–30. DOI: 10.18653/v1/2021.naacl-tutorials.6. Available from: <https://aclanthology.org/2021.naacl-tutorials.6>. Cit. on p. 54.

DU, Kevin et al. Generalizing Backpropagation for Gradient-Based Interpretability. English. In: PROCEEDINGS of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Toronto, Canada: Association for Computational Linguistics, July 2023. P. 11979–11995. DOI: 10.18653/v1/2023.acl-long.669. Available from: <https://aclanthology.org/2023.acl-long.669>. Cit. on p. 54.

DU, Mengnan et al. Learning Credible Deep Neural Networks with Rationale Regularization. en. In: 2019 IEEE International Conference on Data Mining (ICDM). Beijing, China: IEEE, Nov. 2019. P. 150–159. ISBN 978-1-72814-604-1. DOI: 10.1109/ICDM.2019.00025. Available from: <https://ieeexplore.ieee.org/document/8970999/>. Cit. on pp. 22, 27, 34.

DUBEY, Abhimanyu; RADENOVIC, Filip; MAHAJAN, Dhruv. **Scalable Interpretability via Polynomials**. [S.l.]: arXiv, June 2022. arXiv:2205.14108 [cs]. DOI:

10.48550/arXiv.2205.14108. Available from: <http://arxiv.org/abs/2205.14108>. Cit. on p. 28.

GHAEINI, Reza et al. Saliency Learning: Teaching the Model Where to Pay Attention. en. 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). Minneapolis, Minnesota: Association for Computational Linguistics, June 2019. P. 4016–4025. DOI: 10.18653/v1/N19-1404. Available from: <https://aclanthology.org/N19-1404>. Cit. on pp. 27, 55.

GOETHALS, Sofie; MARTENS, David; EVGENIOU, Theodoros. The non-linear nature of the cost of comprehensibility. en. **Journal of Big Data**, v. 9, n. 1, Mar. 2022. ISSN 2196-1115. DOI: 10.1186/s40537-022-00579-2. Available from: <https://doi.org/10.1186/s40537-022-00579-2>. Cit. on p. 28.

HASE, Peter et al. 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. Online: Association for Computational Linguistics, Nov. 2020. P. 4351–4367. DOI: 10.18653/v1/2020.findings-emnlp.390. Available from: <https://aclanthology.org/2020.findings-emnlp.390>. Cit. on p. 28.

HUANG, Quzhe et al. Exploring Distantly-Labeled Rationales in Neural Network Models. en. 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). Online: Association for Computational Linguistics, Aug. 2021. P. 5571–5582. DOI: 10.18653/v1/2021.acl-long.433. Available from: <https://aclanthology.org/2021.acl-long.433>. Cit. on pp. 27, 55.

JACOVI, Alon; GOLDBERG, Yoav. Aligning Faithful Interpretations with their Social Attribution. **Transactions of the Association for Computational Linguistics**, v. 9, p. 294–310, 2021. ISSN 2307-387X. DOI: 10.1162/tacl_a_00367. Available from: <https://aclanthology.org/2021.tacl-1.18/>. Cit. on pp. 21, 28, 29, 51, 55.

JAIN, Sarthak et al. Learning to Faithfully Rationalize by Construction. In: PROCEEDINGS of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, July 2020. P. 4459–4473. DOI: 10.18653/v1/2020.acl-main.409. Available from: <https://aclanthology.org/2020.acl-main.409>. Cit. on p. 28.

JIN, Yaochu; SENDHOFF, Bernhard; KÖRNER, Edgar. Simultaneous Generation of Accurate and Interpretable Neural Network Classifiers. In: MULTI-OBJECTIVE Machine Learning. 1. ed. Berlin, Heidelberg: Springer, 2006. v. 16. (Studies in Computational Intelligence). P. 291–312. ISBN 978-3-540-33019-6. DOI: 10.1007/3-540-33019-4_13. Available from: <https://doi.org/10.1007/3-540-33019-4_13>. Cit. on p. 28.

KHOSLA, Prannay et al. Supervised Contrastive Learning. In: ADVANCES in Neural Information Processing Systems 33 (NeurIPS 2020). [S.l.]: Curran Associates, Inc., 2020. v. 33, p. 18661–18673. Available from: <https://proceedings.neurips.cc/paper/2020/hash/d89a66c7c80a29b1bdbab0f2a1a94af8-Abstract.html>. Cit. on pp. 31, 34, 35.

KUMAR, Sawan; TALUKDAR, Partha. NILE : Natural Language Inference with Faithful Natural Language Explanations. In: PROCEEDINGS of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, July 2020. P. 8730–8742. DOI: 10.18653/v1/2020.acl-main.771. Available from: <https://aclanthology.org/2020.acl-main.771>. Cit. on p. 27.

KUMARI, Gitanjali; SINHA, Anubhav; EKBAL, Asif. 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). St. Julian's, Malta: Association for Computational Linguistics, Mar. 2024. P. 2719–2733. Available from: <https://aclanthology.org/2024.eacl-long.166>. Cit. on p. 21.

LAKHOTIA, Kushal et al. FiD-Ex: Improving Sequence-to-Sequence Models for Extractive Rationale Generation. In: PROCEEDINGS of the 2021 Conference on Empirical Methods in Natural Language Processing. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021. P. 3712–3727. DOI: 10.18653/v1/2021.emnlp-main.301. Available from: <https://aclanthology.org/2021.emnlp-main.301>. Cit. on pp. 27, 55.

LEI, Tao; BARZILAY, Regina; JAAKKOLA, Tommi. Rationalizing Neural Predictions. In: PROCEEDINGS of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, Texas: Association for Computational Linguistics, Nov. 2016. P. 107–117. DOI: 10.18653/v1/D16-1011. Available from: <https://aclanthology.org/D16-1011>. Cit. on p. 28.

LESKOVEC, Jure; RAJARAMAN, Anand; ULLMAN, Jeffrey D. **Mining of Massive Datasets**. 3. ed. [S.l.: s.n.], 2020. Available from: <http://infolab.stanford.edu/~ullman/mmds/book0n.pdf>. Cit. on pp. 33, 37.

LIU, Frederick; AVCI, Besim. Incorporating Priors with Feature Attribution on Text Classification. In: PROCEEDINGS of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, July 2019. P. 6274–6283. DOI: 10.18653/v1/P19-1631. Available from: <https://aclanthology.org/P19-1631>. Cit. on p. 27.

LIU, Hui; YIN, Qingyu; WANG, William Yang. Towards Explainable NLP: A Generative Explanation Framework for Text Classification. In: PROCEEDINGS of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for

Computational Linguistics, July 2019. P. 5570–5581. DOI: 10.18653/v1/P19-1560. Available from: <https://aclanthology.org/P19-1560>. Cit. on p. 27.

LIU, Junhong; LIN, Yijie, et al. Improve Interpretability of Neural Networks via Sparse Contrastive Coding. en. In: FINDINGS of the Association for Computational Linguistics: EMNLP 2022. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, Dec. 2022. P. 460–470. DOI: 10.18653/v1/2022.findings-emnlp.32. Available from: <https://aclanthology.org/2022.findings-emnlp.32>. Cit. on p. 27.

LUNDBERG, Scott M; LEE, Su-In. A Unified Approach to Interpreting Model Predictions. In: ADVANCES in Neural Information Processing Systems 30 (NIPS 2017). [S.l.]: Curran Associates, Inc., 2017. v. 30. Available from: <https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>. Cit. on pp. 21, 39.

MAGGIE; CULLITON, Phil; CHEN, Wei. **Tweet Sentiment Extraction**. [S.l.]: Kaggle, 2020. Available from: <https://kaggle.com/competitions/tweet-sentiment-extraction>. Cit. on p. 38.

MANNING, Christopher D.; RAGHAVAN, Prabhakar; SCHÜTZE, Hinrich. **Introduction to information retrieval**. New York: Cambridge University Press, 2008. OCLC: ocn190786122. ISBN 978-0-521-86571-5. Cit. on p. 37.

MATHEW, Binny et al. HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. en. In: PROCEEDINGS of the AAAI Conference on Artificial Intelligence. Virtual: AAAI Press, May 2021. v. 35, p. 14867–14875. Number: 17. Available from: <https://ojs.aaai.org/index.php/AAAI/article/view/17745>. Cit. on pp. 22, 27, 29, 37, 44, 45, 52–54, 72, 73.

MIETTINEN, Kaisa. **Nonlinear Multiobjective Optimization**. 1. ed. [S.l.]: Springer New York, NY, 1998. v. 12. (International Series in Operations Research & Management Science). ISBN 978-0-7923-8278-2 978-1-4613-7544-9 978-1-4615-5563-6. Available from: <https://link.springer.com/book/10.1007/978-1-4615-5563-6>. Cit. on p. 31.

MITSUHARA, Masahiro et al. 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. [S.l.]: SciTePress, 2021. v. 5, p. 626–636. ISBN 978-989-758-488-6. DOI: 10.5220/0010335806260636. Available from: <https://www.scitepress.org/Link.aspx?doi=10.5220/0010335806260636>. Cit. on p. 27.

NAYLOR, Mitchell et al. **Quantifying Explainability in NLP and Analyzing Algorithms for Performance-Explainability Tradeoff**. [S.l.]: arXiV, July 2021. DOI: 10.48550/arXiv.2107.05693. Available from: <http://arxiv.org/abs/2107.05693>. Cit. on p. 28.

PARANJAPE, Bhargavi et al. An Information Bottleneck Approach for Controlling Conciseness in Rationale Extraction. In: PROCEEDINGS of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics, Nov. 2020. P. 1938–1952. DOI: 10.18653/v1/2020.emnlp-main.153. Available from: <https://aclanthology.org/2020.emnlp-main.153>. Cit. on p. 28.

PEREIRA, Tamara et al. Distill n' Explain: explaining graph neural networks using simple surrogates. en. In: PROCEEDINGS of The 26th International Conference on Artificial Intelligence and Statistics. Valencia, Spain: PMLR, Apr. 2023. v. 206. (Proceedings of Machine Learning Research), p. 6199–6214. ISSN: 2640-3498. Available from: <https://proceedings.mlr.press/v206/pereira23a.html>. Cit. on p. 24.

PLUMB, Gregory et al. **Regularizing Black-box Models for Improved Interpretability**. en. [S.l.]: arXiV, Nov. 2020. DOI: 10.48550/arXiv.1902.06787. Available from: <https://arxiv.org/abs/1902.06787v6>. Cit. on pp. 22, 28.

PRUTHI, Danish et al. Weakly- and Semi-supervised Evidence Extraction. In: FINDINGS of the Association for Computational Linguistics: EMNLP 2020. Online: Association for Computational Linguistics, Nov. 2020. P. 3965–3970. DOI: 10.18653/v1/2020.findings-emnlp.353. Available from: <https://aclanthology.org/2020.findings-emnlp.353>. Cit. on p. 27.

RADENOVIC, Filip; DUBEY, Abhimanyu; MAHAJAN, Dhruv. **Neural Basis Models for Interpretability**. [S.l.]: arXiv, June 2022. arXiv:2205.14120 [cs]. DOI: 10.48550/arXiv.2205.14120. Available from: <http://arxiv.org/abs/2205.14120>. Cit. on p. 28.

RAIMUNDO, Marcos M.; FERREIRA, Paulo A. V.; VON ZUBEN, Fernando J. An extension of the non-inferior set estimation algorithm for many objectives. **European Journal of Operational Research**, v. 284, n. 1, p. 53–66, July 2020. ISSN 0377-2217. DOI: 10.1016/j.ejor.2019.11.017. Available from: <https://www.sciencedirect.com/science/article/pii/S0377221719309282>. Cit. on p. 30.

RAJANI, Nazneen Fatema et al. Explain Yourself! Leveraging Language Models for Commonsense Reasoning. In: PROCEEDINGS of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy: Association for Computational Linguistics, July 2019. P. 4932–4942. DOI: 10.18653/v1/P19-1487. Available from: <https://aclanthology.org/P19-1487>. Cit. on p. 27.

RESCK, Lucas E.; PONCIANO, Jean R., et al. LegalVis: Exploring and Inferring Precedent Citations in Legal Documents. English. **IEEE Transactions on Visualization and Computer Graphics**, v. 29, n. 6, p. 3105–3120, June 2023. Presented at IEEE VIS: Visualization & Visual Analytics 2022. Date of Publication: 18 February 2022. ISSN 1941-0506. DOI: 10.1109/TVCG.2022.3152450. Available from: <https://ieeexplore.ieee.org/document/9716779/>. Cit. on p. 24.

RESCK, Lucas E.; RAIMUNDO, Marcos M.; POCO, Jorge. Exploring the Trade-off Between Model Performance and Explanation Plausibility of Text Classifiers Using Human Rationales. In: FINDINGS of the Association for Computational Linguistics: NAACL 2024. Mexico City, Mexico: Association for Computational Linguistics, June 2024. Available from: <https://arxiv.org/abs/2404.03098>. Cit. on p. 23.

RIBEIRO, Marco Tulio; SINGH, Sameer; GUESTRIN, Carlos. "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. New York, NY, USA: Association for Computing Machinery, Aug. 2016. (KDD '16), p. 1135–1144. ISBN 978-1-4503-4232-2. DOI: 10.1145/2939672.2939778. Available from: <https://doi.org/10.1145/2939672.2939778>. Cit. on pp. 21, 39.

RIEGER, Laura et al. Interpretations are Useful: Penalizing Explanations to Align Neural Networks with Prior Knowledge. en. In: PROCEEDINGS of the 37th International Conference on Machine Learning. [S.l.]: PMLR, July 2020. v. 119, p. 8116–8126. ISSN: 2640-3498. Available from: <https://proceedings.mlr.press/v119/rieger20a.html>. Cit. on pp. 22, 27, 34, 52.

ROSS, Andrew Slavin; HUGHES, Michael C.; DOSHI-VELEZ, Finale. Right for the Right Reasons: Training Differentiable Models by Constraining their Explanations. en. In: PROCEEDINGS of the Twenty-Sixth International Joint Conference on Artificial Intelligence. Melbourne, Australia: AAAI Press, Aug. 2017. P. 2662–2670. ISBN 978-0-9992411-0-3. DOI: 10.24963/ijcai.2017/371. Available from: <https://www.ijcai.org/proceedings/2017/371>. Cit. on pp. 22, 27, 34, 54.

RUDIN, Cynthia. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. **Nature machine intelligence**, v. 1, n. 5, p. 206–215, May 2019. ISSN 2522-5839. DOI: 10.1038/s42256-019-0048-x. Available from: <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9122117/>. Cit. on p. 28.

SANH, Victor et al. **DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter**. [S.l.]: arXiv, Feb. 2020. arXiv:1910.01108 [cs]. DOI: 10.48550/arXiv.1910.01108. Available from: <http://arxiv.org/abs/1910.01108>. Cit. on p. 37.

SATOPAA, Ville et al. Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior. In: 2011 31st International Conference on Distributed Computing Systems Workshops. Minneapolis, MN, USA: IEEE, June 2011. P. 166–171. ISSN: 2332-5666. ISBN 978-1-4577-0384-3. DOI: 10.1109/ICDCSW.2011.20. Available from: <https://ieeexplore.ieee.org/document/5961514>. Cit. on p. 55.

SEKHON, Arshdeep et al. Improving Interpretability via Explicit Word Interaction Graph Layer. en. In: PROCEEDINGS of the AAAI Conference on Artificial Intelligence. Washington DC, USA: AAAI Press, June 2023. v. 37, p. 13528–13537. Number: 11. ISBN 978-1-57735-880-0. DOI: 10.1609/aaai.v37i11.26586. Available from: <https://ojs.aaai.org/index.php/AAAI/article/view/26586>. Cit. on p. 27.

SHARMA, Ashish; MINER, Adam, et al. 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). Online: Association for Computational Linguistics, Nov. 2020. P. 5263–5276. DOI: 10.18653/v1/2020.emnlp-main.425. Available from: <https://aclanthology.org/2020.emnlp-main.425>. Cit. on p. 27.

SHARMA, Manali; BILGIC, Mustafa. Learning with rationales for document classification. en. **Machine Learning**, v. 107, n. 5, p. 797–824, May 2018. ISSN 1573-0565. DOI: 10.1007/s10994-017-5671-3. Available from: <https://doi.org/10.1007/s10994-017-5671-3>. Cit. on p. 27.

SIMPSON, Becks et al. **GradMask: Reduce Overfitting by Regularizing Saliency**. [S.l.]: arXiv, Apr. 2019. arXiv:1904.07478 [cs, eess]. DOI: 10.48550/arXiv.1904.07478. Available from: <http://arxiv.org/abs/1904.07478>. Cit. on p. 27.

STROUT, Julia; ZHANG, Ye; MOONEY, Raymond. Do Human Rationales Improve Machine Explanations? In: PROCEEDINGS of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP. Florence, Italy: Association for Computational Linguistics, Aug. 2019. P. 56–62. DOI: 10.18653/v1/W19-4807. Available from: <https://aclanthology.org/W19-4807>. Cit. on pp. 22, 27, 52, 55.

SWANSON, Kyle; YU, Lili; LEI, Tao. Rationalizing Text Matching: Learning Sparse Alignments via Optimal Transport. In: PROCEEDINGS of the 58th Annual Meeting of the Association for Computational Linguistics. Online: Association for Computational Linguistics, July 2020. P. 5609–5626. DOI: 10.18653/v1/2020.acl-main.496. Available from: <https://aclanthology.org/2020.acl-main.496>. Cit. on p. 28.

TJOA, Erico; GUAN, Cuntai. **Quantifying Explainability of Saliency Methods in Deep Neural Networks with a Synthetic Dataset**. [S.l.]: arXiv, May 2022. arXiv:2009.02899 [cs]. DOI: 10.48550/arXiv.2009.02899. Available from: <http://arxiv.org/abs/2009.02899>. Cit. on p. 21.

TURC, Iulia et al. **Well-Read Students Learn Better: On the Importance of Pre-training Compact Models**. en. [S.l.]: arXiv, Sept. 2019. arXiv:1908.08962 [cs]. DOI: https://doi.org/10.48550/arXiv.1908.08962. Available from: <http://arxiv.org/abs/1908.08962>. Cit. on p. 37.

VASWANI, Ashish et al. Attention is All You Need. In: PROCEEDINGS of the 31st International Conference on Neural Information Processing Systems. Long Beach, California, USA: Curran Associates Inc., 2017. P. 6000–6010. ISBN 978-1-5108-6096-4. Available from: <https://dl.acm.org/doi/abs/10.5555/3295222.3295349>. Cit. on p. 33.

WIEGREFFE, Sarah; MARASOVIC, Ana. Teach Me to Explain: A Review of Datasets for Explainable Natural Language Processing. en. **Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks**, v. 1, Dec. 2021. Available from: <https://datasets-benchmarks-proceedings.neurips.cc/paper/2021/hash/698d51a19d8a121ce581499d7b701668-Abstract-round1.html>. Cit. on pp. 53, 54.

ZAIDAN, Omar; EISNER, Jason; PIATKO, Christine. 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. Rochester, New York: Association for Computational Linguistics, Apr. 2007. P. 260–267. Available from: <https://aclanthology.org/N07-1033>. Cit. on pp. 27, 38.

ZAIDAN, Omar F; EISNER, Jason; PIATKO, Christine D. Machine Learning with Annotator Rationales to Reduce Annotation Cost. en. In: PROCEEDINGS of the NIPS 2008 Workshop on Cost Sensitive Learning. [S.l.: s.n.], Dec. 2008. P. 260–267. Cit. on p. 27.

ZHANG, Zijian; RUDRA, Koustav; ANAND, Avishek. Explain and Predict, and then Predict Again. In: PROCEEDINGS of the 14th ACM International Conference on Web Search and Data Mining. Virtual Event Israel: Association for Computing Machinery, Mar. 2021. P. 418–426. ISBN 978-1-4503-8297-7. DOI: 10.1145/3437963.3441758. Available from: <https://doi.org/10.1145/3437963.3441758>. Cit. on pp. 22, 27, 28, 54, 55.

ZHOU, Yilun; RIBEIRO, Marco Tulio; SHAH, Julie. 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. Seattle, United States: Association for Computational Linguistics, July 2022. P. 5359–5378. Available from: <https://aclanthology.org/2022.naacl-main.392>. Cit. on p. 51.

# Appendix

# APPENDIX A – Implementation and Execution

## A.1 Code and Data

We release the code (under the MIT License) and point to the data sources used in this work in the following GitHub repository:

https://github.com/visual-ds/plausible-nlp-explanations

## A.2 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.

## A.3 DistilBERT and BERT-Mini

The DistilBERT version used in this work was the `distilbert-base-uncased`[1], while the BERT-Mini version was the `prajjwal1/bert-mini`[2]. 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 Section 4.2.1), 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. 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.

[1] Available at https://huggingface.co/distilbert-base-uncased
[2] Available at https://huggingface.co/prajjwal1/bert-mini

## A.4 Language Models Fine-tuning on HateXplain

The rationales of the HateXplain dataset contain words not included in the original `distilbert-base-uncased`[3] and `bert-mini`[4] 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 using MLM. We used an MLM 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).

## A.5 Datasets Tokenization

The HateXplain dataset is already tokenized, and Movie Reviews was tokenized with Python's `str.split()`. Tweet Sentiment Extraction (TSE) was tokenized using a custom regular expression[5] which builds upon `punctuation` from Python's `string` with regex special characters escaped.

## A.6 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.

## A.7 Comparison with HateXplain

To compare our methodology with HateXplain's (MATHEW et al., 2021), 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[6]. 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

[3] Available at https://huggingface.co/distilbert-base-uncased
[4] Available at https://huggingface.co/prajjwal1/bert-mini
[5] `re.split(f"([\\s{punctuation}])", str)`
[6] https://github.com/hate-alert/HateXplain

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 work, 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 previously mentioned fine-tuning procedure, i.e., incorporating new tokens into the model's vocabulary and training the model in the Masked Language Modeling task. 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 $\lambda$:

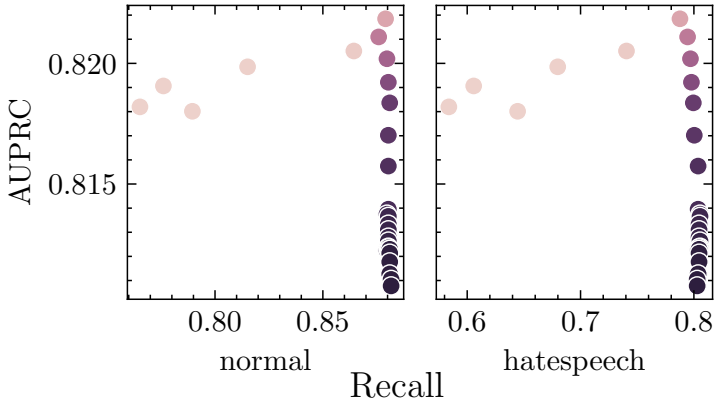
$$\text{loss} = \text{cross-entropy} + \lambda \cdot \text{attention loss}.$$

We explore the trade-off between their two losses (cross-entropy and attention loss) by varying $\lambda$ 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 5.4.

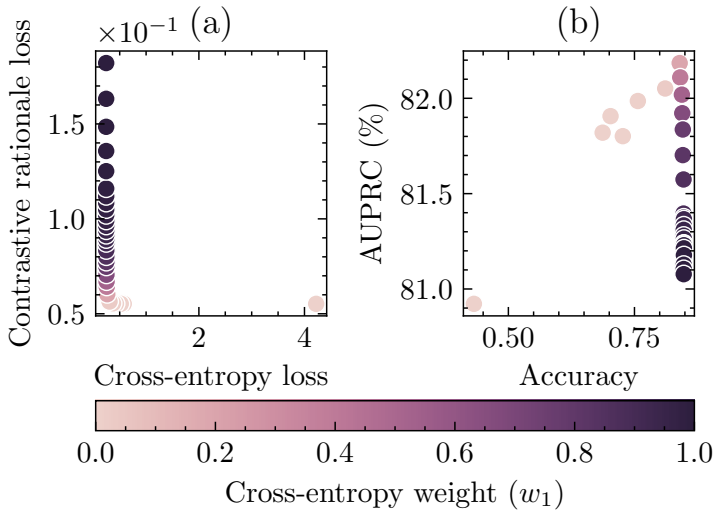
# APPENDIX B – Additional Results

Figure 16 – 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. Source: the authors.

Figure 17 – (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 $w_1 = 0$. Source: the authors.

Table 6 – Examples of explanations of the hate speech class of the HateXplain dataset.

| 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] |

Source: the authors.

Note: 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 5.2), while the "selected" explanation comes from the model with top-AUPRC studied in Section 5.2 (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.

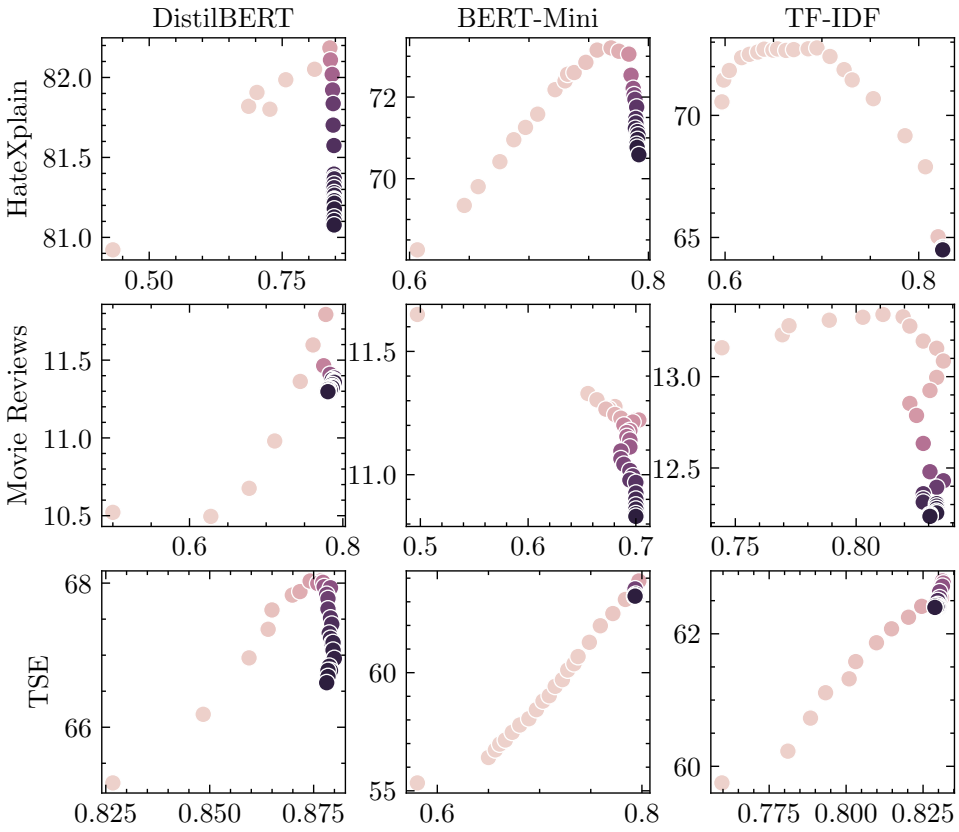
Table 7 – Examples of explanations of the Tweet Sentiment Extraction dataset.

| 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 |

Source: the authors.

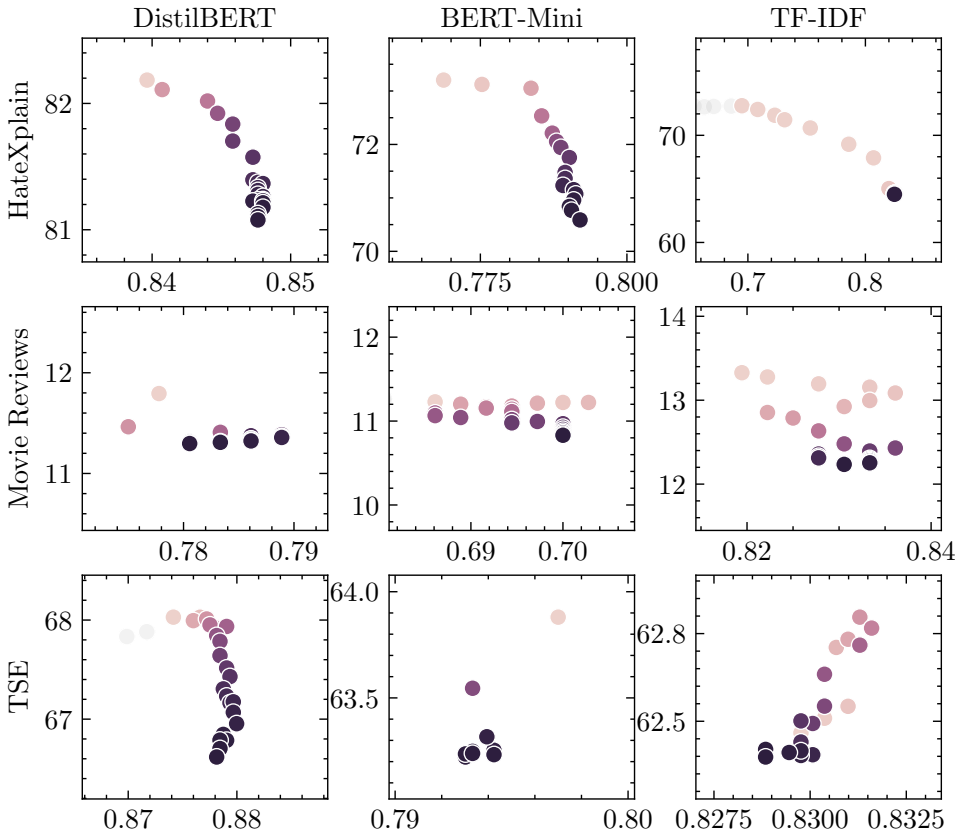
Note: Examples were selected based on the size and quality of the explanation and model predictions. The “original” explanation (LIME) comes from the original DistilBERT model trained with cross-entropy loss only (Section 5.3), while the “selected” explanation comes from the selected model with a green dot (Section 5.3, Figure 11) (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.

Figure 18 – Trade-offs between performance (accuracy, x-axis) and plausibility (AUPRC, y-axis, in percentage (%)) for all models and datasets (test data).



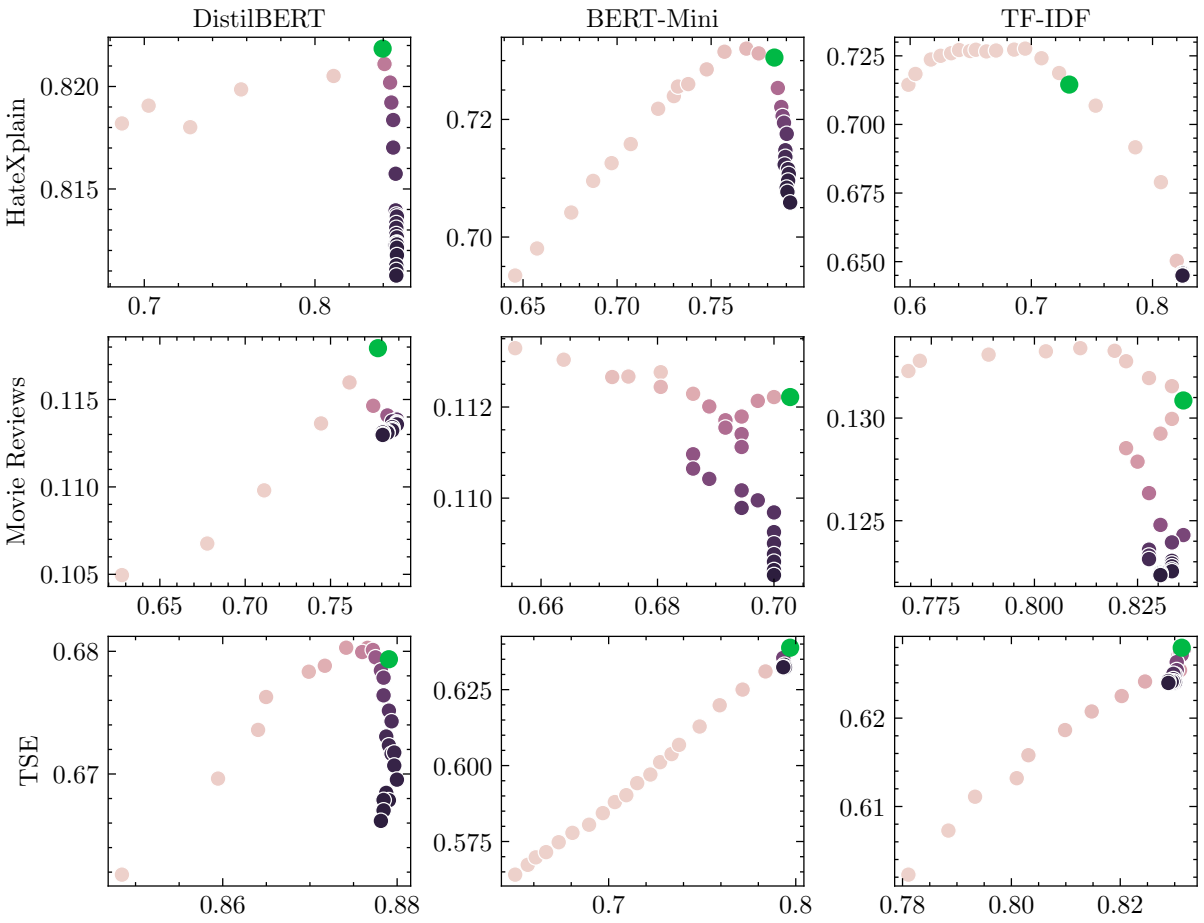

The number of random (negative) rationales is 2, and the explainer is LIME. The color scale is the same as the previous figures. We include the model with $w_1 = 0$ in all graphics. Source: the authors.

Figure 19 – Trade-offs between performance (accuracy, x-axis) and plausibility (AUPRC, y-axis, in percentage (%)) for all models and datasets (test data).



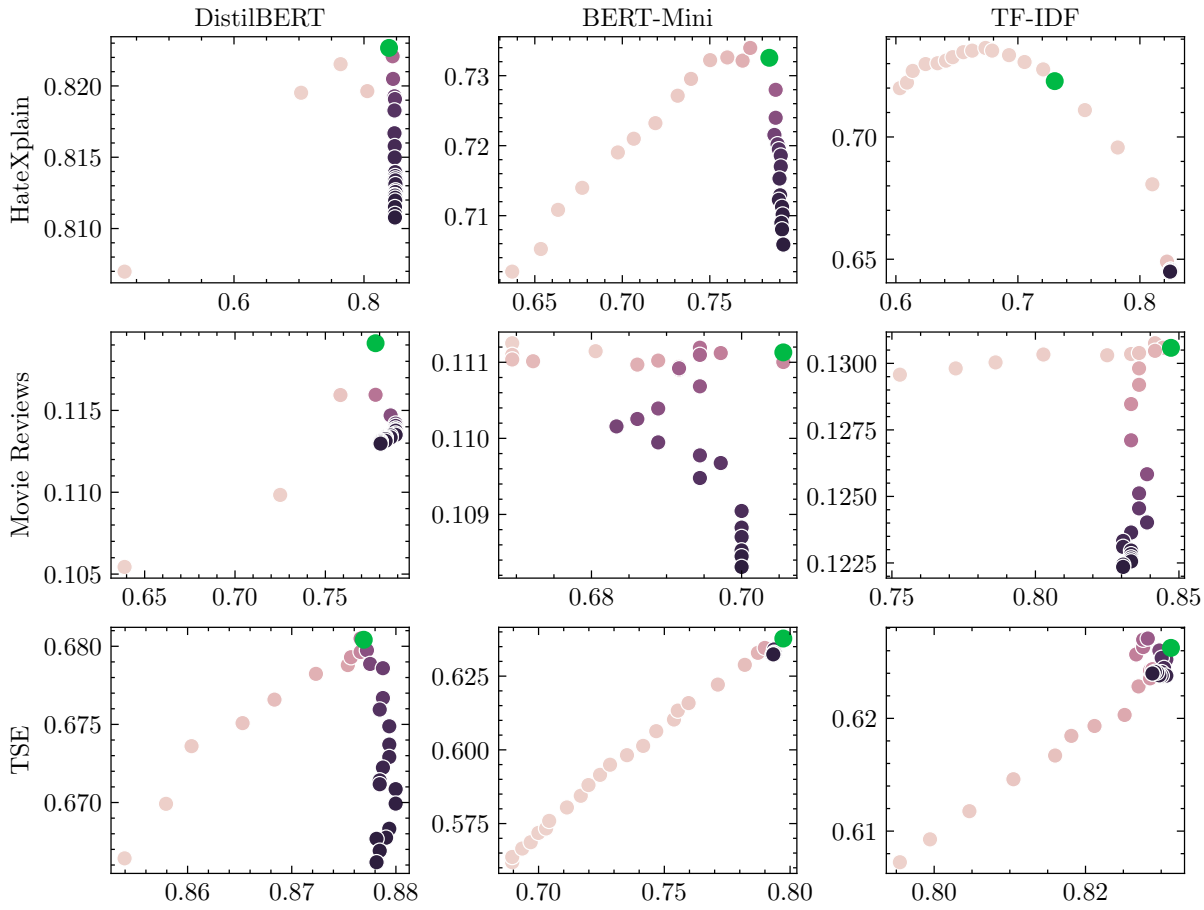

The number of random (negative) rationales is 2, and the explainer is LIME. The color scale is the same as the previous figures. The data scale is equal between x- and y-axes, and a few out-of-scale points are gray or removed. Source: the authors.

Figure 20 – Trade-offs between performance (accuracy, x-axis) and plausibility (AUPRC, y-axis) for all models and datasets (test data).



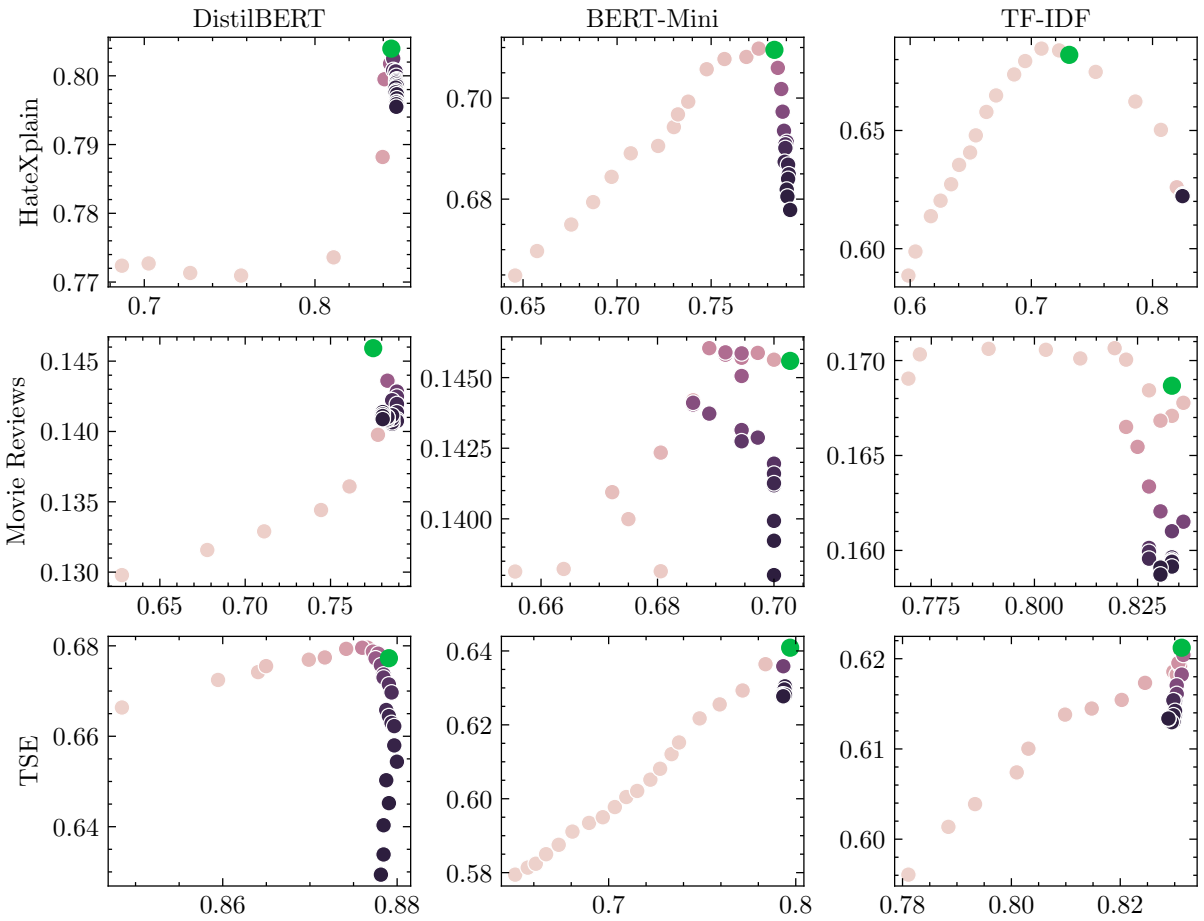

The number of random (negative) rationales is 2, and the explainer is LIME. The color scale is the same as the previous figures. We ignore the model with $w_1 = 0$ in all graphics as it is out of scale. Green dots are the models chosen to be analyzed more carefully. Source: the authors.

Figure 21 – Trade-offs between performance (accuracy, x-axis) and plausibility (AUPRC, y-axis) for all models and datasets (test data).



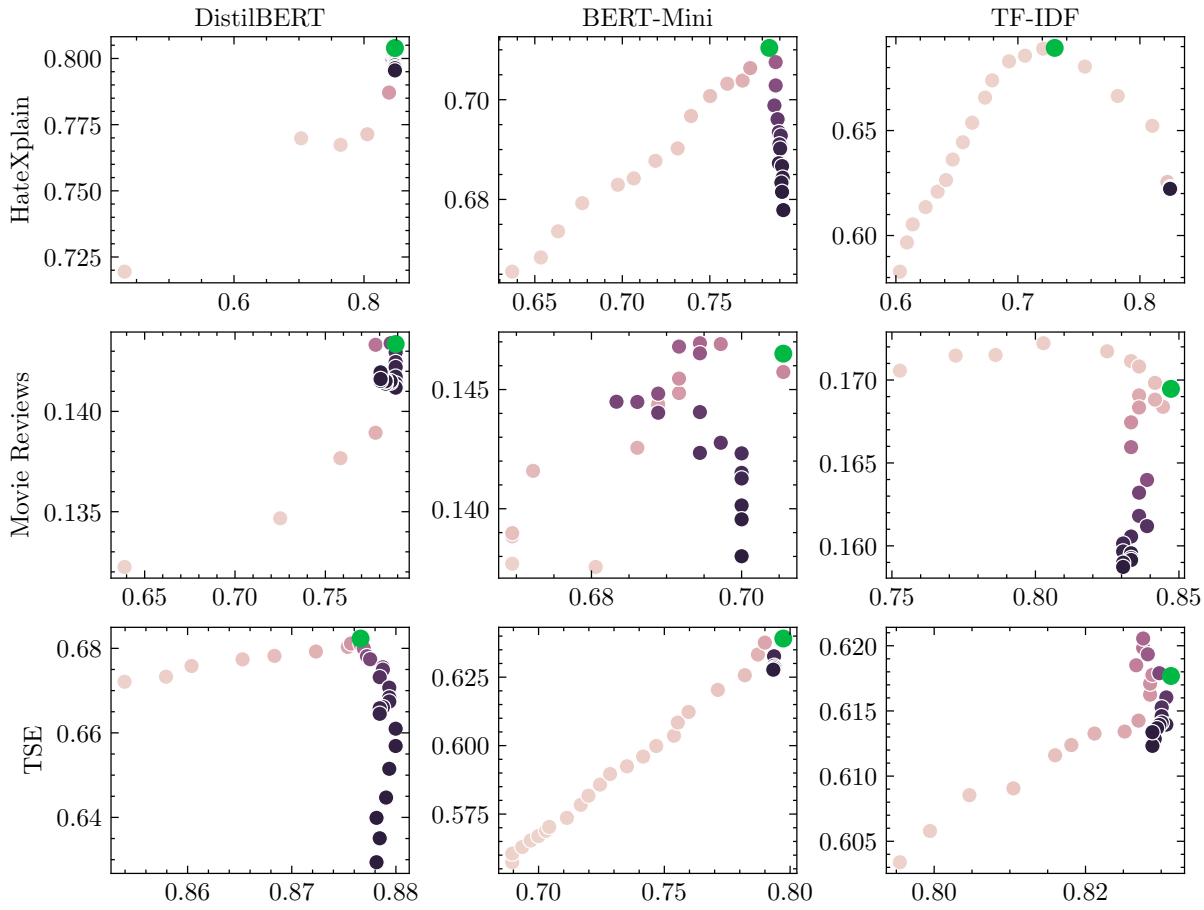

The number of random (negative) rationales is 5, and the explainer is LIME. The color scale is the same as the previous figures. We ignore the model with $w_1 = 0$ in all graphics as it is out of scale. Green dots are the models chosen to be analyzed more carefully. Source: the authors.

Figure 22 – Trade-offs between performance (accuracy, x-axis) and plausibility (AUPRC, y-axis) for all models and datasets (test data).



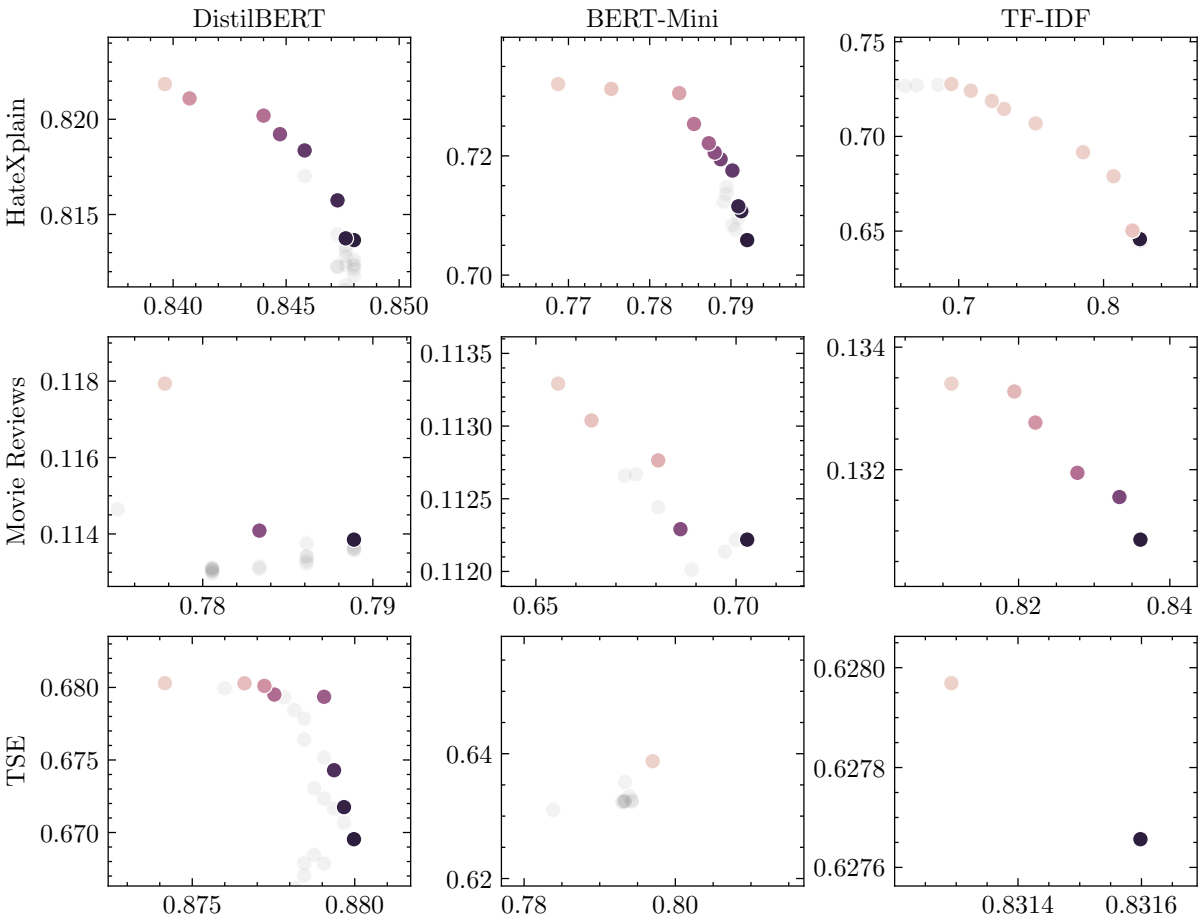

The number of random (negative) rationales is 2, and the explainer is SHAP. The color scale is the same as the previous figures. We ignore the model with $w_1 = 0$ in all graphics as it is out of scale. Green dots are the models chosen to be analyzed more carefully. Source: the authors.

Figure 23 – Trade-offs between performance (accuracy, x-axis) and plausibility (AUPRC, y-axis) for all models and datasets (test data).



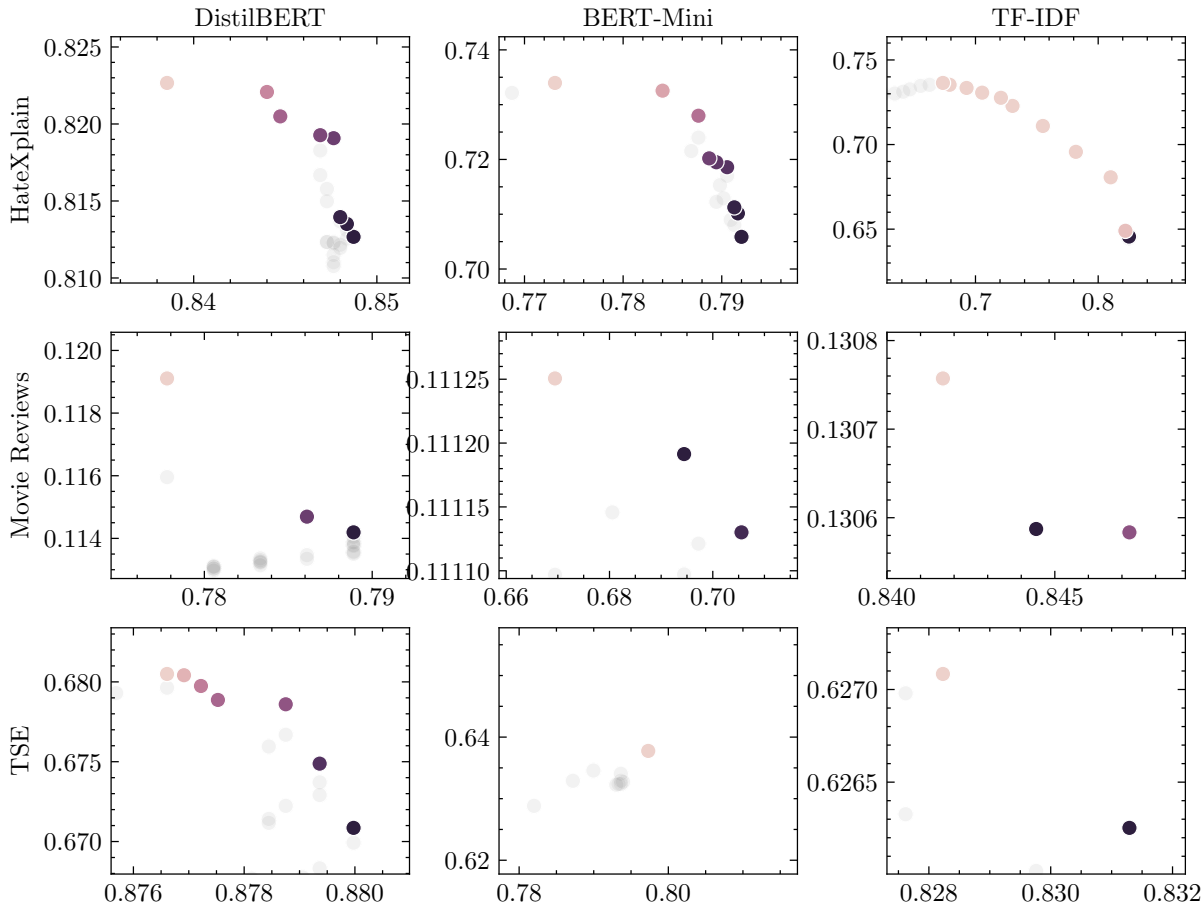

The number of random (negative) rationales is 5, and the explainer is SHAP. The color scale is the same as the previous figures. We ignore the model with $w_1 = 0$ in all graphics as it is out of scale. Green dots are the models chosen to be analyzed more carefully. Source: the authors.

Figure 24 – Pareto-frontier of trade-offs between performance (accuracy, x-axis) and plausibility (AUPRC, y-axis) for all models and datasets (test data).



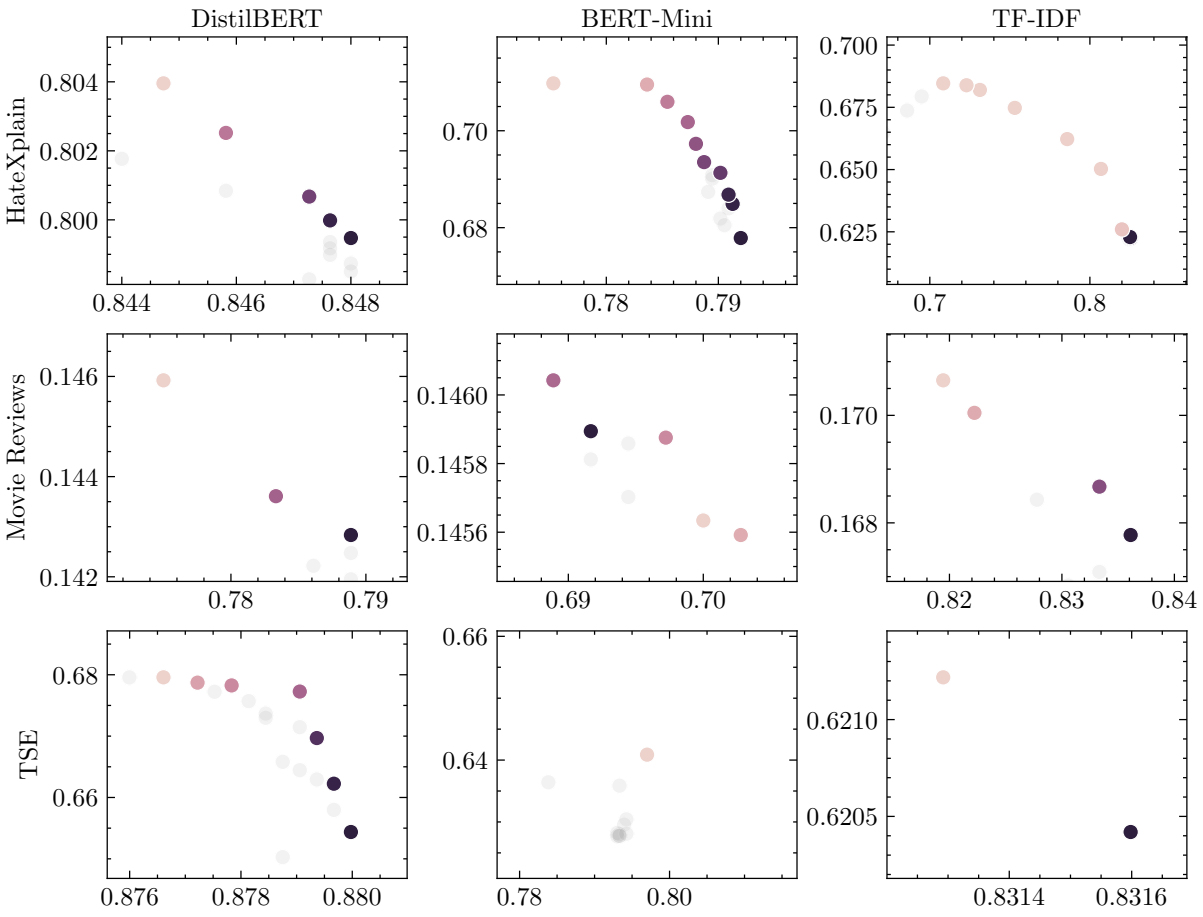

The number of random (negative) rationales is 2, and the explainer is LIME. The color scale is the same as the previous figures. Gray dots are models not on the Pareto-frontier. We ignore the model with $w_1 = 0$ in all graphics as it is out of scale. Source: the authors.

Figure 25 – Pareto-frontier of trade-offs between performance (accuracy, x-axis) and plausibility (AUPRC, y-axis) for all models and datasets (test data).



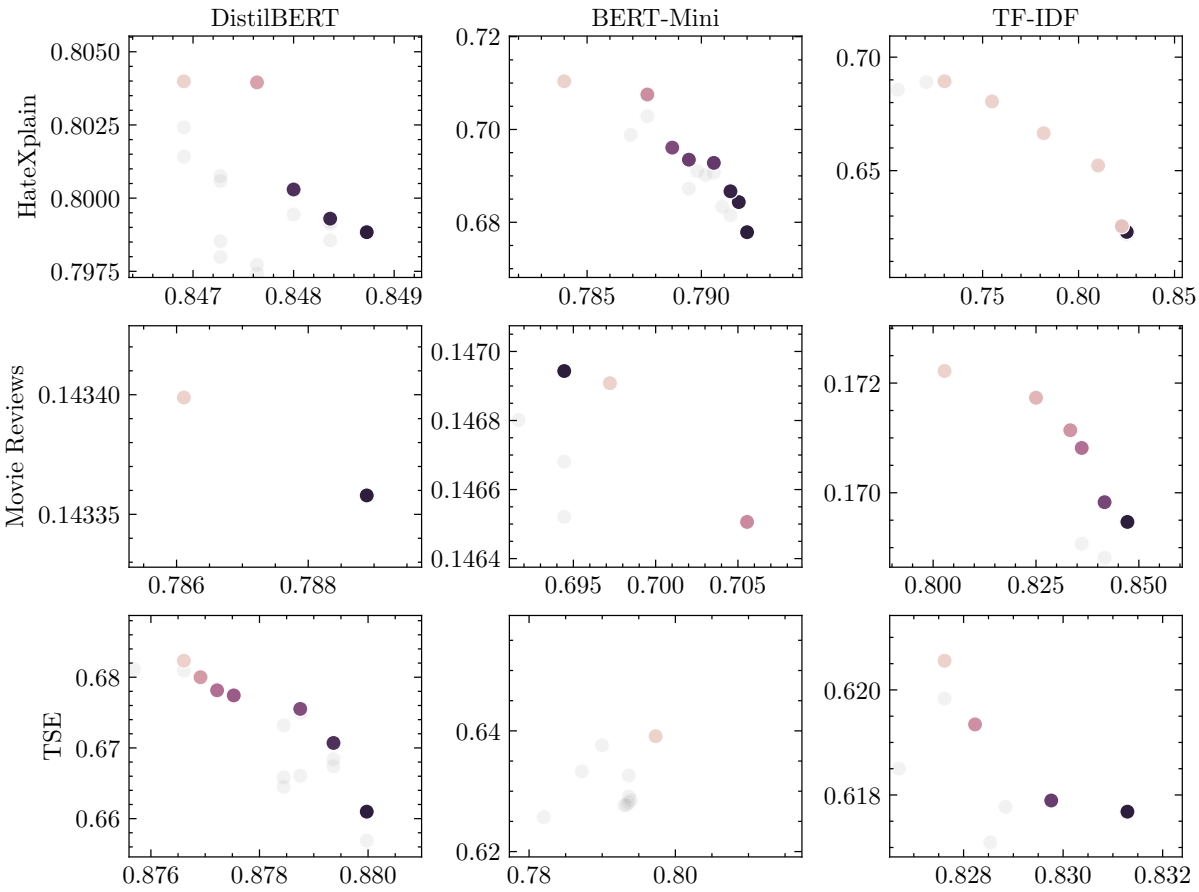

The number of random (negative) rationales is 5, and the explainer is LIME. The color scale is the same as the previous figures. Gray dots are models not on the Pareto-frontier. We ignore the model with $w_1 = 0$ in all graphics as it is out of scale. Source: the authors.

Figure 26 – 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. Source: the authors.

Figure 27 – 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. Source: the authors.

Table 8 – Comparison between the original model (cross-entropy only) and the chosen model (green dots on Figures 11, 21, 22, 23) for each performance and explainability metric on test data.

| **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 |

Source: the authors.

Note: "rel." means relative variation. The column $w_1$ indicates the weight $w_1$ of the chosen model's cross-entropy loss during training.