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

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

Paper and Code

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.

Overview

Local saliency post-hoc model explanations in text classification may not align very well with human intuition (Figure 1(a)).

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

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

Can we make explanations more *plausible* (Figure 1(b))?

Contributions

- A new contrastive-inspired loss to incorporate *rationales* in training, in a model- and explainer-agnostic manner.
- A multi-objective framework to weight the classification and explanation losses, offering multiple trade-off options.
- Extensive experiments with various models, datasets, and explainers, demonstrating a significant enhancement of explanations.

Methodology

Suppose a text classification with texts X, labels y, and human annotations (*rationales*) $\dot{X}$ with corresponding labels $\dot{y}$. The model has logits g_θ indexed by parameters θ. The usual classification **cross-entropy loss** is:

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

We propose the **contrastive rationale loss** to incorporate rationales:

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

where $\{\tilde{X}_{i,j}\} \setminus \dot{X}_i$ are random rationales. Both losses are then combined:

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

Experiments

We employ a multi-objective optimizer to automatically weigh both losses (Figure 2(a)), exploring the trade-offs between them. Then, we assess both performance (accuracy) and explanation plausibility (AUPRC) for each model (Figure 2(b)).

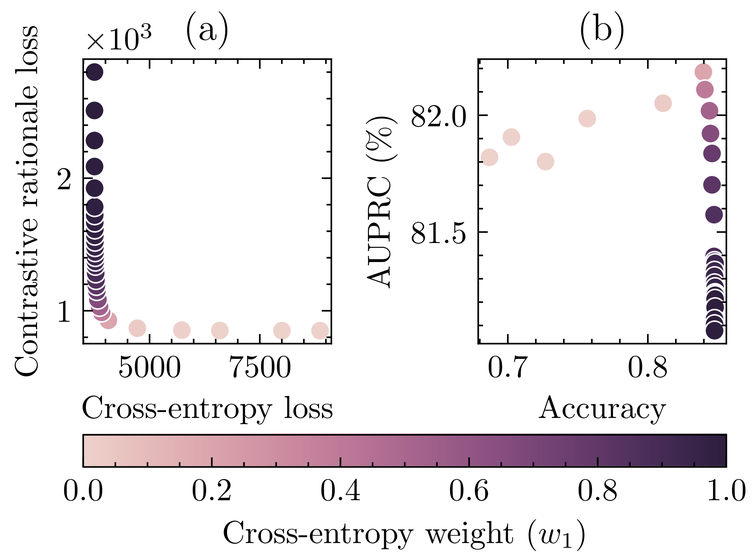

Figure 2: Experiments with DistilBERT and HateXplain. (a) The trade-off between the two losses on the training data. (b) The trade-off between accuracy and plausibility of the test data. The color scale represents the cross-entropy weight w_1. Plausibility is measured by AUPRC (area under the precision-recall curve), comparing the discrete human annotation and the continuous model explanation.

If we select well the parameter w_1, we can assess how much performance we are trading for explanation quality improvement:

Table 1: Comparison between the original model (cross-entropy only) and a model with carefully chosen w_1 (Figure 2), for each performance and explainability metric on test data. "rel." means relative variation. The column w_1 indicates the weight w_1 of the chosen model's cross-entropy loss during training. Explainer is LIME. Sufficiency and comprehensiveness are explanation faithfulness metrics. Accuracy, AUPRC, and AUPRC rel. are in percentage (%).

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

Conclusion

We propose a novel approach for enhancing the explanation plausibility of text classification models by incorporating human rationales.

The presented experiments—and several additional results—indicate that we can find models with improved plausibility and a minimal or negligible performance drop. Furthermore, the method is model- and explainer-agnostic because it does not assume specific model or explainer types, contrary to previous work.

Additional results expand our findings by varying hyperparameters, comparing the methodology with previous work, experimenting with larger models, and analyzing out-of-distribution performance improvement. We also discuss the ethical implications of "teaching" explanations to models.

Acknowledgments

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