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

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

Example Motivation

- Suppose one has the following model:

Text input:
"This is such a great movie!" ⟶ Complex text classification model ⟶ **Prediction:** Positive

- One may want to understand this decision.
  - *"Which are the most important tokens?"*
- Existing methods: LIME, SHAP, Integrated Gradients, GradCAM, etc.

Example Motivation

(a) This is such a great movie !

(b) This is such a great movie !

Figure: Explanations from different models.

- Two different explanations. Which is better?
- Explanation (a) is more **plausible**: it matches more human intuition.
- Ideally, the explanation would be

  This is such a great movie !

  We call this a **human rationale**.

How can we make the model explanations more plausible?

Notation Description

Suppose a multi-class text classification task with:

- Classes C;
- Model f_θ (output probabilities), g_θ (output logits);
- Texts $X = \{X_1, \cdots, X_N\}$;
- Labels $y = \{y_1, \cdots, y_N\}$.

Therefore, we use the standard **cross-entropy loss**:

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

Contrastive Rationale Loss

Cross-entropy loss:

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

Contrastive rationale loss:

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

- Human rationales $\dot{X} = \{\dot{X}_1, \cdots, \dot{X}_N\}$;
- Their labels $\dot{y} = \{\dot{y}_1, \cdots, \dot{y}_N\}$;
- "Sample rationales" $\{\tilde{X}_{i,j}\}_{j=1}^{m} = \{\dot{X}_i\} \cup \{$other $m-1$ random rationales$\}$.

Trade-Off Exploration

Classification loss

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

Contrastive rationale loss

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

- We use a multi-objective optimization solver to generate a Pareto-frontier by sampling weights w_1, w_2 and solving

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

Main Experiments: DistilBERT and HateXplain

- **DistilBERT** is a simpler version of BERT, the most popular Transformer encoder.
  - We add a classification layer and train only it.
- **HateXplain** is a dataset of hate speech detection with human-annotated rationales.
  - We use it as a binary classification between `normal` and `hatespeech`.

*“ugh i hate <u>d*kes</u> 😐”* ⟶ DistilBERT ⟶ `hatespeech`

Main Experiments: DistilBERT and HateXplain

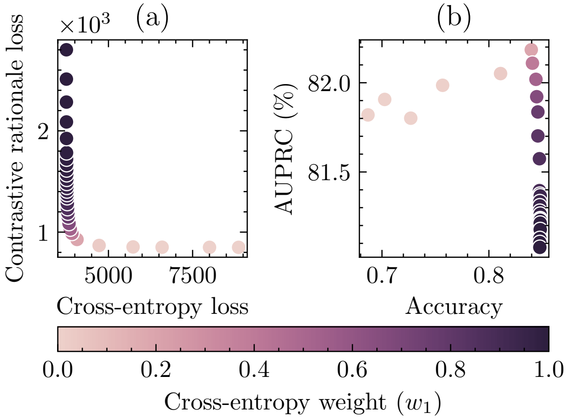


Figure: (a) Trade-off between the two losses on the training data. (b) Trade-off between accuracy and plausibility on the test data.

Main Experiments: DistilBERT and HateXplain

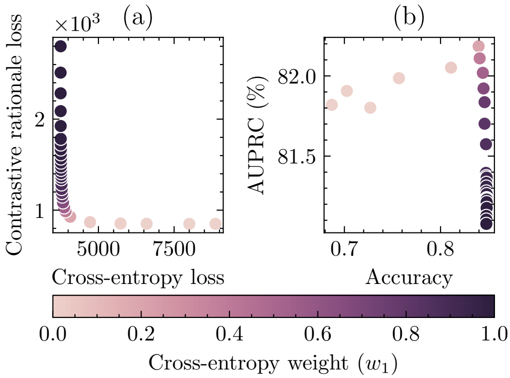


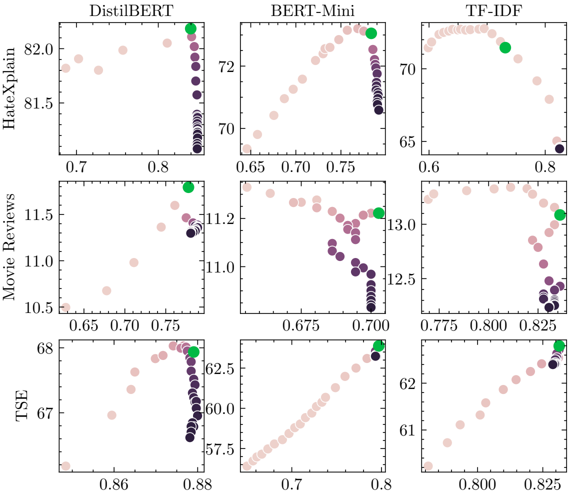
Figure: (a) Trade-off between the two losses on the training data. (b) Trade-off between accuracy and plausibility on the test data.

(a) ugh i hate d*kes 😐

(b) ugh i hate d*kes 😐

Figure: Examples of explanations of the hate speech class from the original model (a) and the model with top-AUPRC (b).

All Experiments



- Trade-offs between **performance** (accuracy, x-axis) and **plausibility** (AUPRC, y-axis, in percentage (%)) for all models and datasets (test data).
- The explainer is LIME.
- Green dots are the models chosen to be analyzed more carefully.

All Experiments

Table: **Comparison between the original model** (cross-entropy only) **and the chosen model** (green dots on previous figure) 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

Key Takeaways

1. We propose a **novel contrastive-inspired loss function** that effectively incorporates rationales into the learning process.
   - The methodology is **model- and explainer-agnostic**.
2. We develop a **multi-objective framework** that automatically assigns weights to the learning loss and contrastive rationale loss.
3. We perform a **series of experiments** using various models, datasets, and explainability methods.
   - We demonstrate a **significant enhancement of model explanations** without compromising (and sometimes without any detriment to) the model's performance.

Key Takeaways

1. We propose a **novel contrastive-inspired loss function** that effectively incorporates rationales into the learning process.
   - The methodology is **model- and explainer-agnostic**.
2. We develop a **multi-objective framework** that automatically assigns weights to the learning loss and contrastive rationale loss.
3. We perform a **series of experiments** using various models, datasets, and explainability methods.
   - We demonstrate a **significant enhancement of model explanations** without compromising (and sometimes without any detriment to) the model's performance.

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`

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