Rather a Nurse than a Physician - Contrastive Explanations under Investigation

Oliver Eberle[*,1,2] **Ilias Chalkidis**[*,3] **Laura Cabello**[3] **Stephanie Brandl**[3]

[*]**Authors contributed equally.**

[1]**TU Berlin,** [2]**BIFOLD,** [3]**University of Copenhagen**

eMail: oliver.eberle@tu-berlin.de, {ilias.chalkidis, brandl}@di.ku.dk

Introduction

Are contrastive explanations closer to humans than non-contrastive explanations?

Non-contrastive setting

Why is the person in the following short bio described as a "***Dentist***"?

He has 47 years of experience. His specialties include **Oral** and **Maxillofacial surgery**. Dr. Show is affiliated with Baylor University Medical Center.

He has 47 years of experience. His specialties include Oral and Maxillofacial surgery. Dr. Show is affiliated with Baylor University Medical Center.

Contrastive setting

Why is the person in the following short bio described as a "***Dentist***" rather than a "***Surgeon***"?

He has 47 years of experience. His specialties include **Oral** and Maxillofacial surgery. Dr. Show is affiliated with Baylor University of Medical Center.

He has 47 years of experience. His specialties include Oral and Maxillofacial surgery. Dr. Show is affiliated with Baylor University Medical Center.

Figure 1: An example from the BIOS dataset of *non-contrastive* and *contrastive* human and model-based rationales. Human rationales are underlined and bold-faced, while model-based rationale attribution scores are highlighted in red (positive) or blue (negative) colors.

Experiments

Datasets

- SST2
- DynaSent
- BIOS
- DBpedia-Animals

→ en text classification

Models

- RoBERTa
- GTP-2
- T5

→ each in 3 sizes

XAI

- LRP
- GradientxInput
- GradNorm

→ post-hoc explanations

We further collect human rationale annotations for a subset from the BIOS dataset for contrastive and non-contrastive settings.

Results

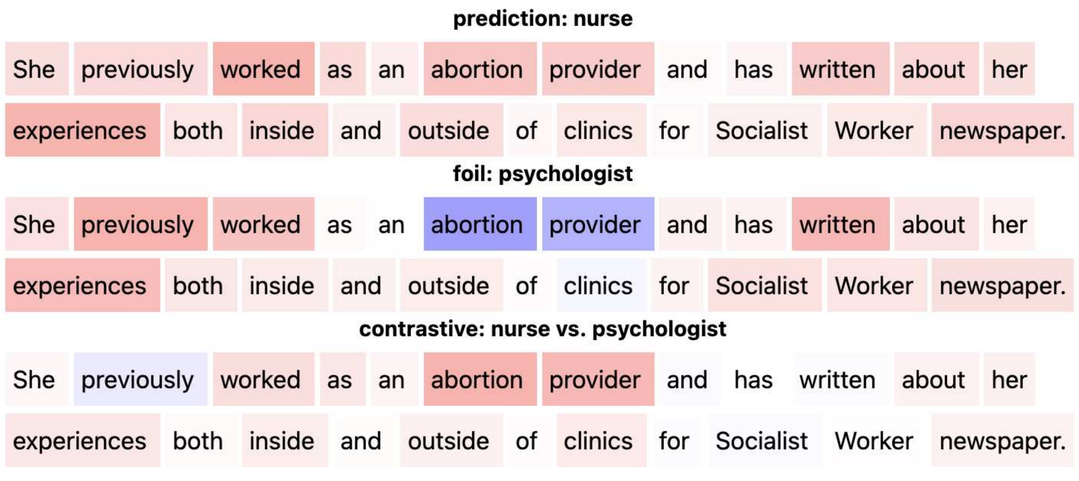

Figure 2: An example biography for a 'nurse' from the BIOS dataset highlighted with LRP relevance scores -red for positive, and blue for negative- per class based on RoBERTa large. We further show explanations for the foil ('psychologist'). In the last row, we present an explanation for 'nurse', the correct outcome, in contrast to 'psychologist', the second best guess of the model.

Results

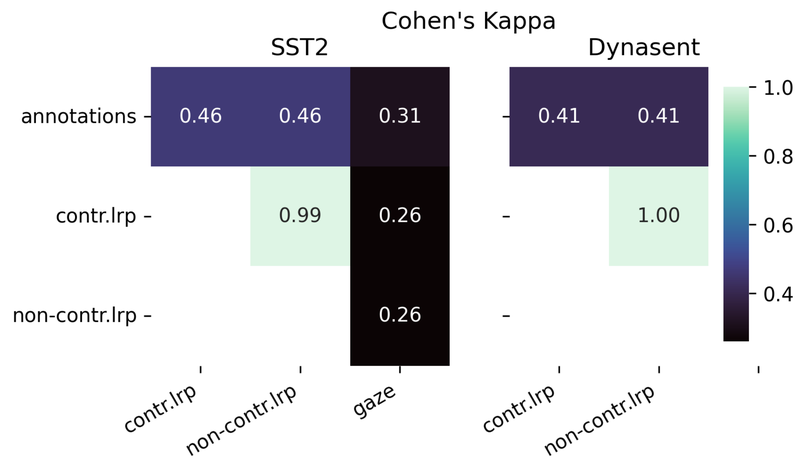

Figure 3: Comparison between model rationales with human annotations for sentiment classification on DynaSent and additionally with human gaze on SST2.

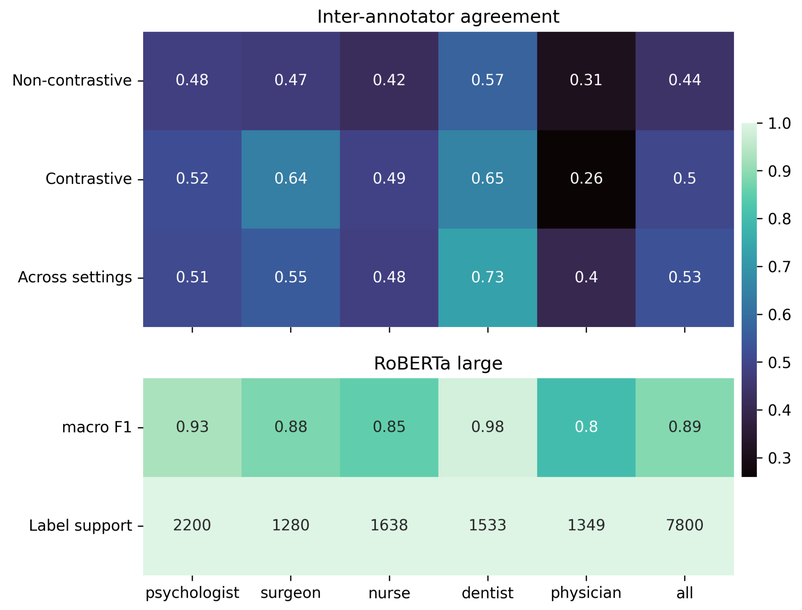

Figure 4: *Upper:* Cohen's Kappa scores for inter-annotator agreement for human rationale annotation within and across contrastive and non-contrastive settings. *Lower:* Model performance scores (macro F1) for the best model and training support across BIOS classes.

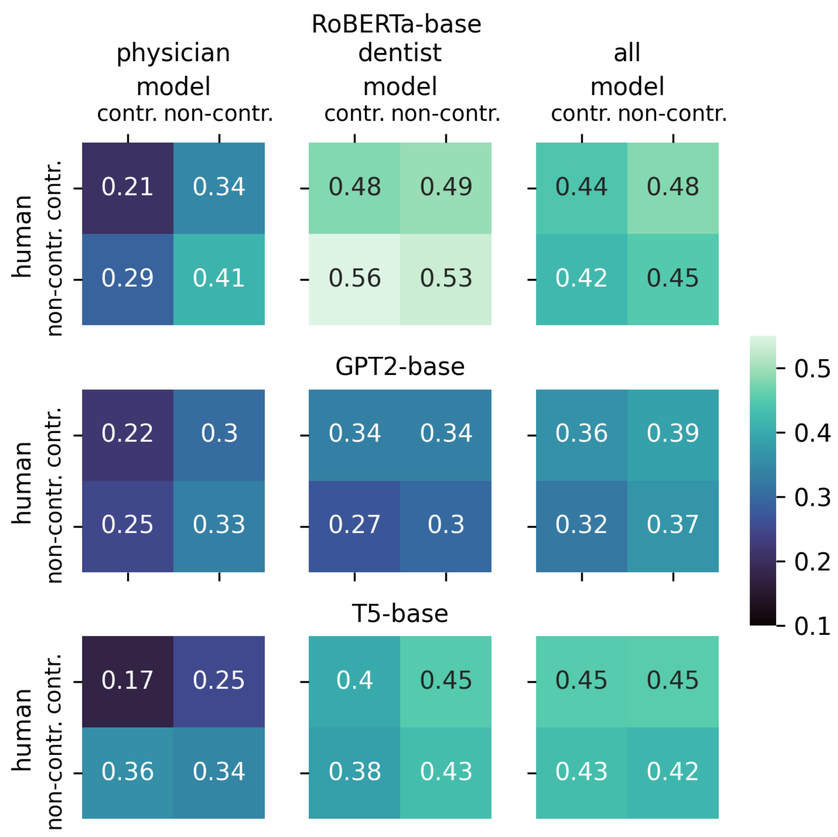

Figure 5: Agreement between human rationales and model-based explanations computed with LRP.

Conclusion

- Human rationales are very similar between contrastive and non-contrastive settings but fewer tokens are selected in the former.
- High agreement between model and human rationales but class-dependent
- Gaze correlates less than rational annotation with model explanations
- Model explanations between contrastive and non-contrastive settings differ more when the model is less certain