Reproducibility Study of “*Are Shortest Rationales the Best Explanations for Human Understanding?*”, Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, 2022

Jonathan Hus, Wren McQueary
01130888, 01317855
jhus@gmu.edu, wmcquear@gmu.edu

Introduction

- Original paper contributes two main things:
  - **Contribution 1: LimitedInk**: Explainability tool for BERT on classification tasks. Consists of two DistilBERT models:
    - **Classifier**: Answers a classification question
    - **Identifier**: Highlights a portion of the input to explain the classifier's answer. This portion is called the *evidence*. The identifier is trained to highlight a specific percentage *k* of the input text length.
    - k=**50%** It 's not life - affirming -- its vulgar and mean , but I liked it . **Y**=Pos
  - **Contribution 2: A claim supported by LimitedInk:** Shorter rationales aren't necessarily better for human understanding.

Reproducibility

- Aimed to reproduce 2 main claims:
  - **Claim 1:** That shorter rationales are not necessarily better
    - The authors primarily used a model trained on the Movies dataset to make this claim.
    - Movies dataset consists of film reviews, labeled positive or negative.
    - We trained models for different rationale lengths and compared their F1 scores vs human annotation.

Movies Data Set Performance			
Method	**Precision**	**Recall**	**F1**
LimitedInk (50%)	0.91	0.90	0.90
Ours - 10%	0.89	0.89	0.89
Ours - 20%	0.89	0.88	0.88
Ours - 30%	0.86	0.86	0.86
Ours - 40%	0.87	0.87	0.87
Ours - 50%	0.87	0.86	0.86

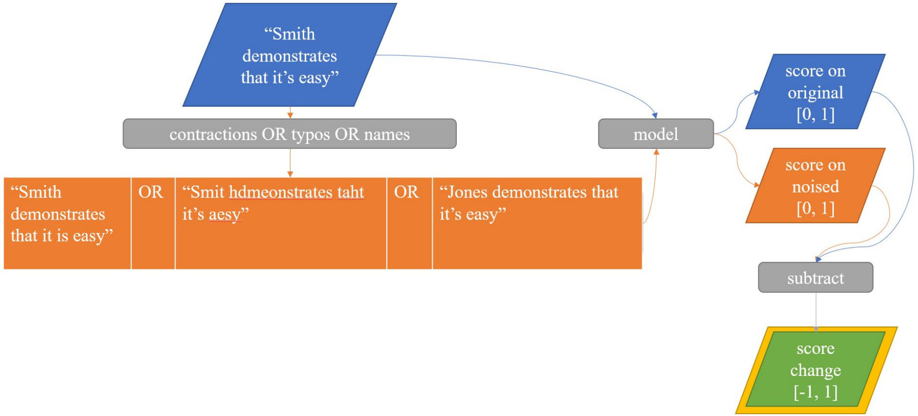
    - We found that shorter rationales yielded *better* F1 scores, but only slightly. Contradicts the authors' findings.
  - **Claim 2:** The performance of the LimitedInk models trained by the authors on a variety of datasets
    - We trained models on those datasets using the same hyperparameters as the authors.

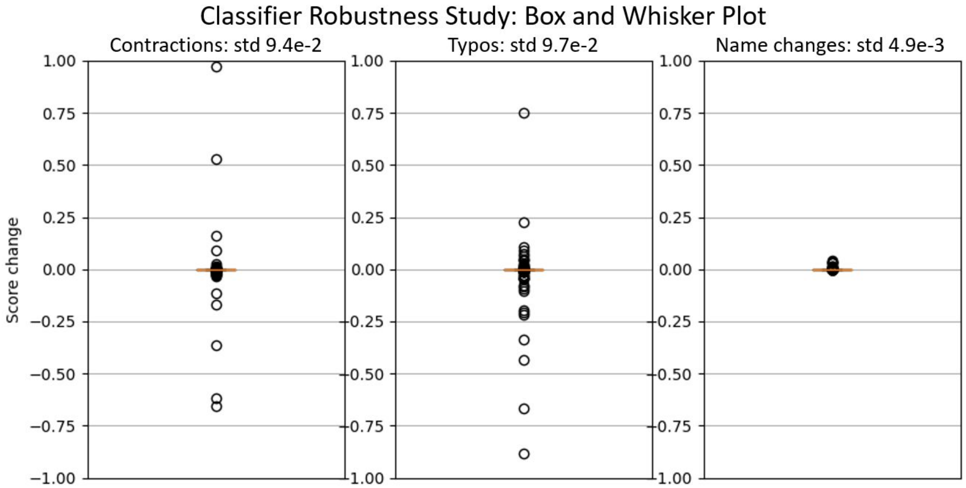
Results Reproduction		
Dataset (k value)	**Their F1**	**Our F1**
Movies (50%)	0.90	0.90
BoolQ (30%)	0.56	0.49
Evidence Inference (30%)	0.50	0.45
FEVER (40%)	0.90	0.89
MultiRC (50%)	0.67	0.62

    - Although our results are always worse than the authors', they're roughly the same.

Robustness

- Trained a model on Movies using the same hyperparameters as the authors' most successful Movies model.
- Model outputs a score from 0 (negative with full confidence) to 1 (positive with full confidence).
- Using CheckList, measured effect of contraction changes, typo changes, and name changes on output label and confidence.
    - Generated 250 examples for each of these three types of noise.
- Model was robust to all 3 (very low standard deviations), but had some drastic outliers for contractions and typos.
    - Hypothesis: Contractions and typos distort meaning often, but name changes do so only rarely.





Multilinguality

- Trained a multilingual LimitedInk model on Arabic, Bulgarian, German, Greek, English, Spanish, French, Hindi, Russian, Swahili, Thai, Turkish, Urdu, Vietnamese, and Chinese.
- No multilingual equivalent of any of the authors' datasets, so we used a different one: e-XNLI
  - Consists of pairs of sentences, labeled with either "contradiction", "entailment", or "neutral".

Language: en
Label: contradiction
Premise: But anyway, the animals would get loose all the time, especially the goats.
Hypothesis: The goats were kept safe and secure.
Premise_Highlighted: But anyway , the *animals* would get *loose* all the time , especially the goats .
Hypothesis_Highlighted: The *goats* were *kept* *safe* and *secure* .

  - Trained a new DistilBERT model on this multilingual dataset.
  - For comparison, also trained a model on only the English portion of the dataset.
  - For another comparison, referenced the authors' FEVER model (most similar dataset to e-XNLI, but still significant differences).

Dataset	F1 Score
FEVER	0.90
e-XNLI (English only)	0.62
e-XNLI (complete)	0.60

- **LimitedInk holds in a multilingual context!** Multilingual model performs almost as well as English-only, but worse than FEVER.
- Possible explanations:
  - e-XNLI task is harder. 3 labels vs FEVER's 2 labels.
  - e-XNLI hasn't been vetted to the same level of quality as FEVER.
  - e-XNLI's ground-truth annotations are automated, but FEVER's are human-annotated.

Conclusion

- **Reproducibility**: We verified the authors' model accuracies, but found evidence in favor of shorter rationales, which contradicts the authors.
- **Robustness**: We found the authors' model to be robust to noise in the input data, with a low standard deviation, although contraction and typo changes elicited a few significant outliers.
- **Multilinguality**: We trained a model on a new multilingual dataset (e-XNLI) which performs on-par with the English-only portion of e-XNLI. Performance is worse than the most similar dataset used by the authors (FEVER), but might be explained by a difference in dataset difficulty and quality.

- Directions for future analysis/work:
  - Need to better compare our results to the authors'. Could create a new dataset which is a multilingual version of FEVER and has human-annotated labels.
  - Using a multilingual FEVER dataset could allow for better verification of our approach for training multilingual LimitedInk models!

Link to GitHub repository

References

Oana-Maria Camburu, Tim Rockt¨aschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, Advances in Neural Information Processing Systems 31, pages 9539–9549. Curran Associates, Inc.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. In NAACL.

Alexis C**onneau, G**uillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating crosslingual sentence representations. arXiv preprint arXiv:1809.05053.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2019. Eraser: A benchmark to evaluate rationalized nlp models. arXiv preprint arXiv:1911.03429.

Jay DeYoung, Eric Lehman, Benjamin Nye, Iain Marshall, and Byron C. Wallace. 2020. Evidence inference 2.0: More data, better models. In Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing, pages 123–132, Online. Association for Computational Linguistics.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface:a challenge set for reading comprehension over multiple sentences. In NAACL.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2016. Rationalizing neural predictions. arXiv preprint arXiv:1606.04155.

Cameron Martin. 2020. Facial recognition in law enforcement. Seattle J. Soc. Just., 19:309.

Edoardo Mosca, Daryna Dementieva, Tohid Ebrahim Ajdari, Maximilian Kummeth, Kirill Gringauz, and Georg Groh. 2023. Ifan: An explainability-focused interaction framework for humans and nlp models. arXiv preprint arXiv:2303.03124.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. Science, 366(6464):447–453.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of nlp models with checklist. In Association for Computational Linguistics (ACL).

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature machine intelligence, 1(5):206–215.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of BERT: smaller, faster, cheaper and lighter. CoRR, abs/1910.01108.

Hua Shen**, Tongshuang Wu, Wenbo Guo, and Ting-Hao'Kenneth' Huang. 2022. Are shortest rationales the best explanations for human understanding? arXiv preprint arXiv:2203.08788.**

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a Large-scale Dataset for Fact Extraction and VERification. In Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL), pages 809–819.

Keyon Vafa, Yuntian Deng, David M Blei, and Alexander M Rush. 2021. Rationales for sequential predictions. arXiv preprint arXiv:2109.06387.

Giulia Vilone and Luca Longo. 2021. Notions of explainability and evaluation approaches for explainable artificial intelligence. Information Fusion, 76:89–106.

Omar Zaidan, Jason Eisner, and Christine Piatko. 2007. Using "annotator rationales" to improve machine learning for text categorization. In Proceedings of the conference of the North American chapter of the Association for Computational Linguistics (NAACL), pages 260–267.

Kerem Zama**n and Yonatan Belink**ov. 2022. A multilingual perspective towards the evaluation of attribution methods in natural language inference. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics.