Improving Explainability of Sexism Detection in Social Media Texts

Corentin Genton, Lucille Niederhauser, Viola Renne
Group 11

Problem definition

- Interpretability/explainability of deep learning models is not well studied
- Now a requirement of GDPR to explain decisions regarding hate speech identification online

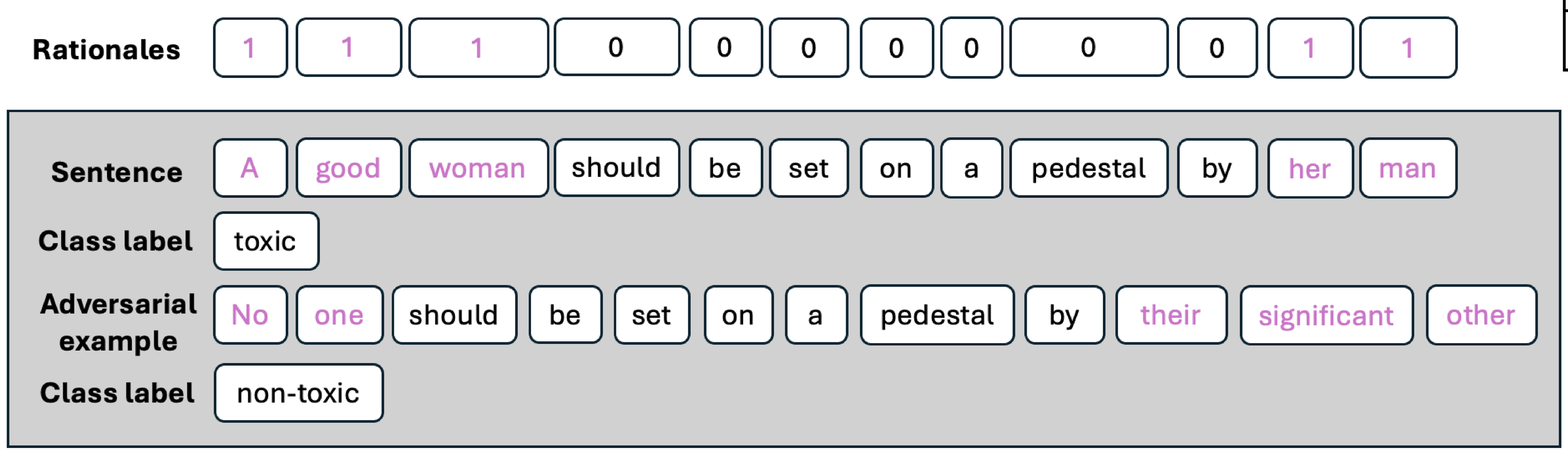
Can we design a model capable of detecting sexism and provide an explanation behind its decision ?

Key Related Works

- ***HateXplain***: use human-annotated rationales to direct the attention of their model towards rationales prediction and thus better explainability.
- ***"Call me sexist but …"***: comprehensive dataset for sexism detection labeled using psychological scales and containing adversarial examples.
- ***Eraser benchmark***: a framework to compute explainability of a model given human rationales.

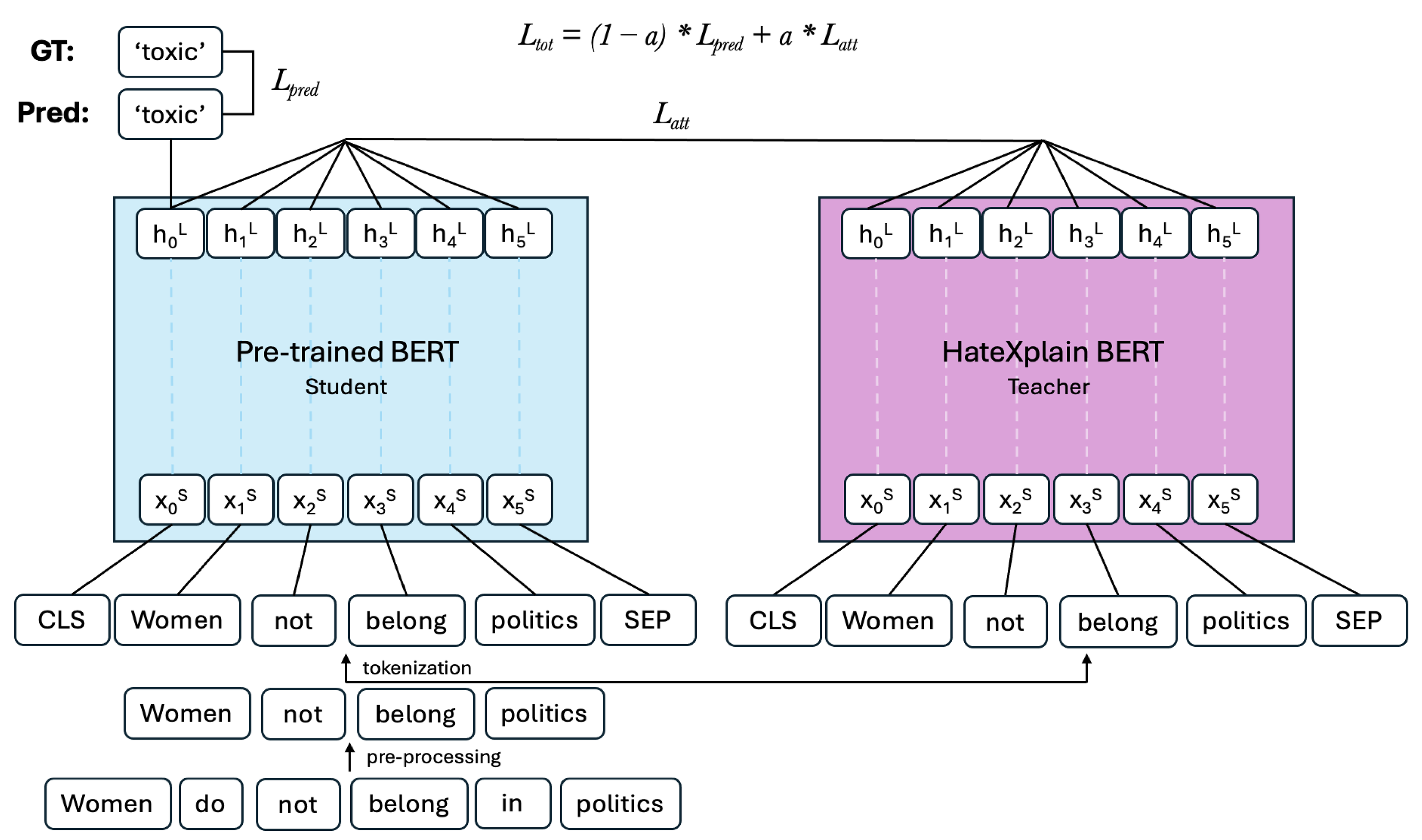
Dataset(s)

- **HateXplain** Dataset : Over 20'000 posts with label, target community and human-annotated rationales.
- **Sexism Detection** Dataset : Over 13'000 posts, 3599 with automatically labelled rationales that we inferred from adversarial examples as shown below.



Method

- **Aggregated datasets**: fine-tuning of BERT on union of HateXplain dataset and sexism detection dataset.
- **Knowledge distillation**: fine-tuning of BERT on sexism detection dataset using combination of distillation loss on attention and label loss, as shown below.



Validation

- **Reproducibility** : similar results but not exactly the same due to wrong seed implementation in both papers.
- **Performance** :

Model (2 labels)	Accuracy	F1 score
BERT – HateXplain	0.7968	0.7908
HateBERT – HateXplain	0.7921	0.7877
BERT - AD	0.7968	0.7889
HateBERT - AD	0.7812	0.7717
Knowledge Distillation	0.8205	0.8197
HateXplain -Sexism Dataset	0.9114	0.9074

AD: Aggregated Datasets

- **Explainability** :

Model	IOU F1	Token F1	AUPRC	Comp.	Suff.
BERT - HateXplain	0.125	0.442	0.678	0.566	0.153
BERT - AD	0.133	0.469	0.681	0.609	0.110
HateBERT - AD	0.123	0.445	0.694	0.634	0.144

Model	Text	Label
Human Annotator	I do not listen to niggershitbullshit	toxic
BERT	I do not listen to niggershitbullshit	toxic
BERT AD	I do not listen to niggershitbullshit	toxic
Human Annotator	Always thought that nigger was a faggot	toxic
BERT	Always thought that nigger was a faggot	toxic
BERT AD	Always thought that nigger was a faggot	toxic
Human Annotator	This is a fucking disgrace shameless cunts	toxic
BERT	This is a fucking disgrace shameless cunts	toxic
BERT AD	This is a fucking disgrace shameless cunts	toxic

Limitations

- Automatic labelling may differ from human annotations.
- Unable to compute explainability of knowledge distillation model.
- Current work is limited to the English language.
- No exploration of the hyper-parameter space of our models.

Conclusion

- Knowledge distillation did not reduce performance while potentially increasing explainability.
- Finetuning of HateXplain had the best performance maybe because we are fine-tuning a model trained on similar data instead of simply BERT.
- Different attention aggregation and decay methods did not lead to better explainability.
- We were able to develop different approaches that could increase the explainability of sexism detection.

References

[1] Mathew, B., Saha, P., S. M., Biemann, C., Goyal, P., & Mukherjee, A. (2021, May). Hatexplain: A benchmark dataset for explainable hate speech detection. In Proceedings of the AAAI conference on artificial intelligence (Vol. 35, No. 17, pp. 14867-14785)
[2] Subramaniam, A., Mehra, A., & Kundu, S. (2022). Exploring hate speech detection with hatexplain and bert. arXiv preprint arXiv:2208.04489
[3] Samory, M., Sen, I., Kohne, J., Flöck, F., & Wagner, C. (2021, May). "Call me sexist, but...": Revisiting Sexism Detection Using Psychological Scales and Adversarial Samples. In Proceedings of the international AAAI conference on web and social media (Vol. 15, pp. 573-584).