Improving Explainability of Sexism Detection in Social Media Text

Lucille Niederhauser [1] **Viola Renne** [1] **Corentin Genton** [1]

Abstract

Many deep-learning models have been developed to detect hate speech in social media. Though they can achieve good performances, their decision-making processes are often unclear, making it harder for models to comply with certain ethical and legal guidelines. We thus decided to study how models for hate speech detection, and more precisely sexism detection, could gain in explainability. We explored methods building on the work of Mathew et al. 2022 mainly through different pre-processing, fine-tuning of existing models and knowledge distillation. With these, were able to improve both performance and explainability of sexism detection.

Keywords: Explainability, Rationales, Sexism.

1. Introduction

As a effort to help maintain healthy social media spaces, many hate speech detection models and datasets have been published in recent years [1]. The increased performance of these models comes with greater complexity and a lack of understanding of their decisions. This causes legal and ethical issues. For instance, the General Data Protection Regulation in Europe has established a ”right to explanation”, highlighting the need for more interpretable models [2]. In this work, we aim to find ways to increase the explainability of hate speech detection in written social media posts. As a first step, we mainly focus on sexism detection but hope that our findings can extend to other targeted groups. Our work builds on the HateXplain model [3] that is capable of detecting hate speech and also provides the rationale, i.e. the specific part of the text that makes it hateful. Amongst others, we explored how incorporating a dataset for sexism detection in the training of HateXplain and how knowledge distillation could improve both sexism classification and rationale prediction. Aside from the usual performance metrics such as accuracy and F1-score, we evaluated the explainability of our models using the metrics described in [4].

2. Related Work

Explainable Hate Speech Detection Mathew et al. [3] explored how models could classify social media posts as hateful and also provide the exact words that make the post hateful (the rationale). They asked annotators to label if social media posts were hateful and to identify the rationales that justified each label. Each rationale was converted into an attention vector of 0s and 1s. Since multiple annotators were used for each post, the different attention vectors were averaged and normalized using a softmax to produce the ground truth attention for each post. For samples labeled as ”normal”, the ground truth attention was 1/(sequence length) for all tokens. They designed a custom loss composed of the label loss and a cross-entropy between the attention obtained from the annotators and the attention of the model. Models using this combined loss for training performed slightly better, but better-performing models did not necessarily have the highest explainability. [3] This highlights the need for other metrics than accuracy and F1-score as they do not fully reflect all the desired qualities of a model. Subramaniam et al. [5] further explored Mathew et al's work and looked at different aggregation methods to create the ground truth attention. They studied conservative attention, which only labels words as part of the rationale if all annotators agree, and lenient attention, where they labeled a word if even one annotator thinks it's important. Conservative aggregation led to better accuracy, F-score and explainability metrics than the original paper. The method giving the best explainability was lenient attention, but it had slightly worse accuracy and F-score than the original paper. [5]

Sexism Detection In recent years, multiple groups published datasets specifically for sexism detection. One the most recent, is from Vetagiri et al. [6]. They created a dataset of more than 1'700'000 online sexist and non-sexist examples to fine-tune sexism detection models. They also performed a review of different models on their dataset and obtained an accuracy of 0.92 and a F1-score of 0.76 using their best model, i.e. GPT-2. [6] Their dataset is however not publicly available for now and we were thus not able to use this contribution in our work. Another comprehensive dataset for sexism detection was published by Samory et al [7]. They created a codebook for sexism detection based on psychological scales and used it to re-annotate existing sexism datasets that they combined into one dataset. Also,

[1] Group 11.

EE-559: *Deep Learning*, 2024.

they asked annotators to modify some sexist samples into non-sexist ones with as few modifications as possible to create adversarial examples. Their final dataset contains more than 13'000 samples and their best model was BERT-finetuned on their dataset with a F1-score of 0.82. [7]
To our knowledge, no one has specifically studied the explainability of sexism detection.

3. Method

Datasets We used two datasets. The first one is the HateXplain dataset that we pre-processed exactly as the authors did in the original paper [3]. The second one is the dataset constructed by Samory et al. [7] described in section 2. We pre-process this dataset differently depending on the models used. When using the dataset for knowledge distillation (see below), we used the same pre-processing as in the original paper. This gave us a dataset of 13'631 samples. To use the same dataset in combination with the HateXplain dataset [3], we needed to perform rationales labelling of the sexism dataset. We used the adversarial examples to do so. Indeed, since the annotators were asked to make as few modifications to the text as possible to make it non-sexist, we assumed that the words that were modified were the reason why the text is considered sexist. We thus chose these words as the rationales. Finally, we created a sub-dataset containing only adversarial examples labelled with their rationales and their corresponding original sexist examples. We did not include more non-sexist examples to not induce class unbalances nor original sexist examples without adversarial examples since we could not find their rationales. We ended up with a dataset of 3'599 samples that we pre-processed in the same way as the HateXplain one.

Explainability Metrics For measuring the explainability of our models we follow the framework of the ERASER benchmark [4]. We thus measure explainability in two ways: plausibility and faithfulness. These two concepts allow to assess the rationales' quality: plausibility measures agreement between human-annotated rationales and extracted ones. This alone is not sufficient as it doesn't take into account whether the model used these rationales to make its predictions. For this reason, DeYoung et al. (2020) [4] define faithful rationales as those that correspond to the inputs most relied upon by the model. According to Yu et al. (2019) [8], ideal rationales should be sufficient to classify a given sentence. Comprehensiveness assesses whether all features needed for the prediction are included in the rationales. If the rationales were truly useful for the model's prediction, a high comprehensiveness score should be obtained; a negative score would indicate the model is more confident without the rationales [4]. Together, these two metrics provide a measure of faithfulness. We used three metrics to measure plausibility. Evaluating predicted and human-annotated rationales directly is often too harsh. Therefore, we used Intersection-Over-Union (IOU) on a token level, as defined by DeYoung et al. (2020). This approach allows for partial matches and the calculation of an F1-score. Precision and recall on a token level are also measured to compute token-level F1 scores. Finally, we computed the Area Under the Precision-Recall Curve (AUPRC) which is done by sweeping a threshold over token scores when soft token scoring is used, rewarding models for assigning higher scores to marked tokens [4].

HateXplain for Sexism Detection Our first approach to obtain an explainable model for sexism detection was to fine-tune the HateXplain model on the sexism dataset that we annotated with rationales as explained above. As a second approach, we decided to re-train HateXplain on the HateXplain dataset aggregated with the sexism dataset labelled with rationales to see if this would give better performance and/or explainability. Finally, we wanted to see if using another base model than BERT [9] could be another way to improve our results. We therefore trained HateBERT [10] first only on the HateXplain dataset and in a second time on the aggregated dataset. We expected these two models to have better performance since HateBERT is a model that was created by training the BERT model on posts from banned communities from Reddit and thus could maybe better understand hate speech. For all these models, we used the same hyperparameters as the HateXplain model [3].

Attention Aggregation and Decay Both the HateXplain and sexism datasets have multiple rationales for the same post, and need to be aggregated into single values. To achieve this, we have selected three methods proposed in [5]: mean, lenient (OR), and conservative (AND) aggregation. Upon analyzing the resulting rationales from both datasets, we observed that the sexism dataset had fewer non-zero rationales, due to how they are computed. Therefore, we decided to experiment with attention decay using two different methods. Attention decay redistributes attention away from a single word to adjacent words. The two methods are additive, which employs a uniform distribution, and geometric, which uses a geometric distribution.

Knowledge Distillation Since we do not know how close our rationale labelling of the sexism dataset is to human-labelled rationales, we found a way to train a model that does not rely on them. We used a framework similar to Samory et al. [7] and added knowledge distillation with HateXplain as a teacher model to guide the prediction of rationales. We fine-tuned BERT [9] on the sexism dataset using a loss function composed of two parts, the label loss and the attention loss. The attention loss is a cross-entropy between the attention outputted by BERT [9] and the ”ground truth” attention. To compute the ground truth attention we evaluated the HateXplain model [3], without any modifications, on the sexism dataset. We introduced a parameter α in the loss to be able to tune the contribution of the attention loss to the total loss. We chose α = 0.079 because it gave the best performance using a 5-fold cross-validation on our training set. Otherwise, we used the hyperparameters reported by Samory et al. [7]. We did not compute the explainability of this model since only 3599 samples are labelled with the rationales in our dataset of size 13631 and we would thus not have obtained meaningful results.

4. Validation

Reproducibility We trained BERT [9] on the HateXplain dataset and finetuned BERT [9] on the sexism dataset using the same framework as Mathew et al. [3] and Samory et al. [7] respectively to reproduce their results. Due a misplaced seed setting in the HateXplain code, we were not able to reproduce their results exactly. They reported an accuracy of 0.698 and a F1-score of 0.687 where we obtained 0.682 and 0.668 respectively. Our explainability scores were however very similar. Similarly, no seed was fixed in the code from Samory et al. [7], but a cross-validation was performed. This led to a reported F1-score of 0.820 ($\sigma = 0.010$). We obtained 0.829 ($\sigma = 0.014$). All in all, we were able to produce very similar results to both articles.

Performance We present our performance results in Table 1. We expected that using HateBERT as a base model would give better results than BERT. Surprinsingly, we obtained similar performance for both. This might be due to the fact that HateBERT it is fine-tuned only on banned communities so it might struggle to recognise non-hateful posts. Interestingly, knowledge distillation did not lead to a decrease of performance. This indicates that guiding the loss towards predicting the rationale could be used to justify the decision of the model without any unreasonable drop of performance. Finally, finetuning the HateXplain model on the dataset for sexism detection gave better results than both the original paper and our knowledge distillation model. This might be because, as explained in section 3, the pre-processing used was different. Another reason could be that hateXplain is a model already trained for hate-speech detection and thus it is understandable that finetuning it on a sexism dataset gives better results that fine-tuning BERT on the same dataset, which is what was done in the original paper and in our knowledge distillation model.

Bias and Explainability As shown in Table 2, aggregating both datasets lead to an increase of explainability. We also computed the same metrics for the different attention aggregation and decay methods explained in section 3. However, we did not observe the same results as Subramaniam et al. [5] since they did not lead to an improvement in explainability. Finally, we computed the bias of our HateXplain-based models and found no significant improvements of the bias metrics for the women community, even when aggregating HateXplain with the sexism detection dataset.

MODEL	ACCURACY	F1-SCORE
HATEXPLAIN		
BERT	0.7968	0.7908
HATEBERT	0.7921	0.7877
BERT AD	0.7968	0.7889
HATEBERT AD	0.7812	0.7717
SEXISM		
REPLICATION	0.8286	0.8292 ($\sigma = 0.0145$)
KNOWLEDGE DIST.	0.8205	0.8197
FINETUNE OF HXP	0.9114	0.9074

Table 1: Accuracy and F1-score results of the different models. For HateXplain dataset, the classification was done on 2 labels. AD: aggregated datasets, HXP: HateXplain.

MODEL	EXPLAINABILITY				
	IOU F1	TOKEN F1	AUPRC	COMP.	SUFF.
BERT	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

Table 2: Explainability results of the different models. AD: aggregated datasets.

5. Limitations

The dataset we used for sexism detection was not labelled with the rationales and we did not have the resources to perform the labeling. We thus cannot be sure that our automatic labelling is similar to what would have resulted from human annotations. Since we use these rationales both for training and testing the explainability of our models, automatically labelled rationales being far from the ground truth could greatly impact our results. This highlights the need for hate speech datasets that also provide the rationale behind the labels. Also, our current work is limited to detecting hate speech in English and doesn't address the problem of multilingual hate speech. In future developments, incorporating multilingual capabilities would be essential.

6. Conclusion

We studied different approaches to improve the explainability of sexism detection. Importantly, we found that we were able to increase the explainability of some current models without any decrease of the performance. Although knowledge distillation and finetuning of HateXplain are promising, they do need to be further evaluated on datasets with human-annotated rationales and see if they apply to other hate-speech target communities.

References

[1] A. Gandhi, P. Ahir, K. Adhvaryu, P. Shah, R. Lohiya, E. Cambria, S. Poria, and A. Hussain, “Hate speech detection: A comprehensive review of recent works,” *Expert Systems*, p. e13562.

[2] O. Radley-Gardner, H. Beale, and R. Zimmermann, eds., *Fundamental Texts On European Private Law*. Hart Publishing, 2016.

[3] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, and A. Mukherjee, “Hatexplain: A benchmark dataset for explainable hate speech detection,” 2022.

[4] J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, and B. C. Wallace, “Eraser: A benchmark to evaluate rationalized nlp models,” 2020.

[5] A. Subramaniam, A. Mehra, and S. Kundu, “Exploring hate speech detection with hatexplain and bert,” 2022.

[6] A. Vetagiri, P. Pakray, and A. Das, “A deep dive into automated sexism detection using fine-tuned deep learning and large language models,” *Available at SSRN 4791798*.

[7] M. Samory, I. Sen, J. Kohne, F. Floeck, and C. Wagner, “”call me sexist, but...”: Revisiting sexism detection using psychological scales and adversarial samples,” 2021.

[8] M. Yu, S. Chang, Y. Zhang, and T. Jaakkola, “Rethinking cooperative rationalization: Introspective extraction and complement control,” in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (K. Inui, J. Jiang, V. Ng, and X. Wan, eds.), (Hong Kong, China), pp. 4094–4103, Association for Computational Linguistics, Nov. 2019.

[9] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “Bert: Pre-training of deep bidirectional transformers for language understanding,” *arXiv preprint arXiv:1810.04805*, 2018.

[10] T. Caselli, V. Basile, J. Mitrović, and M. Granitzer, “HateBERT: Retraining BERT for abusive language detection in English,” in *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)* (A. Mostafazadeh Davani, D. Kiela, M. Lambert, B. Vidgen, V. Prabhakaran, and Z. Waseem, eds.), (Online), pp. 17–25, Association for Computational Linguistics, Aug. 2021.

[11] D. Borkan, L. Dixon, J. Sorensen, N. Thain, and L. Vasserman, “Nuanced metrics for measuring unintended bias with real data for text classification,” 2019.

[12] T. Caselli, V. Basile, J. Mitrović, and M. Granitzer, “Hatebert: Retraining bert for abusive language detection in english,” *arXiv preprint arXiv:2010.12472*, 2020.

[13] O. Zaidan, J. Eisner, and C. Piatko, “Using “annotator rationales” to improve machine learning for text categorization,” in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference* (C. Sidner, T. Schultz, M. Stone, and C. Zhai, eds.), (Rochester, New York), pp. 260–267, Association for Computational Linguistics, Apr. 2007.