Rationale-Guided Few-Shot Classification to Detect Abusive Language [Appendix]

1 More details about the datasets

1.1 The HateXplain dataset (HX)

The HateXplain dataset introduced by Mathew [7] is a large dataset of 20k posts from Twitter and Gab. At the top-level, each post is annotated by three annotators into one of three categories – *hate speech*, *offensive*, or *normal*. Furthermore, groups or communities targeted in the post are also marked. For abusive content (i.e. hate speech or offensive), tokens or word spans explaining abusive content are marked as rationales. In total, there are 253 unique annotators as reported by Mathew [7]. We use this dataset to train the rationale extraction model. The following is how we aggregated the ground truth for each type of annotation for this dataset:

- **Labels**: The final label for each datapoint is selected based on the majority label from the labels provided by three different annotators. We also convert this to a two-class problem by considering both hate speech and offensive labels in the 'abusive' class and the normal label in the 'non-abusive' class. We then follow the same majority selection criteria to select the final label.
- **Rationales**: To provide the model with rationales as feedback, we convert the rationales by each annotator into Boolean vectors. Values in these Boolean vectors are 1 when the corresponding token (word) in the text is a part of a rationale. To create the **ground truth rationales**, we consider each token in the text and call it a rationale if at least two annotators have highlighted it as a rationale. The final ground truth rationales are Boolean vectors, considering the above constraint.
- **Targets**: For ground truth targets, we consider those **targets** that are labeled so by at least 2 annotators, after which we ignore those targets that appear less than 20 times in the complete dataset and replace them all with - 'Others'. We find 22 targets which are noted in Table 2.

1.2 The Founta et al. dataset (FA)

Founta [6] made available a large-scale Twitter dataset containing 4 different labels: *hateful*, *abusive*, *normal* and *spam*. Their work focused on dealing with the class imbalance in random samples from Twitter by filtering tweets in an incremental and iterative process, aided with boosted sampling. The quality of judgment was ensured by measuring agreement for over 20 annotators per tweet. The dataset contains 100k tweets and is the largest dataset considered in this paper. We ignore the datapoints annotated as spam from our analysis.

1.3 The Davidson et al. dataset (DA)

This work on automatic hate speech detection by Davidson [4] released a dataset of 24k tweets. Each tweet was queried from Twitter using a lexicon derived from Hatebase.org[1]. Annotation was carried out by majority vote of at least three CrowdFlower workers. There are three labels in this dataset: *hate speech*, *offensive* and *normal*. The high prevalence of abusive tweets is attributed in part to racial bias by Davidson [3] who demonstrated that a classifier trained on the dataset shows significantly higher tendency to mark tweets written by African-Americans as abusive.

1.4 The OLID dataset (OD)

The Offensive Language Identification Dataset (OLID) dataset released by zampieri [10] in the SemEval-2019 Task 6 (OffensEval) uses a modern hierarchical labelling scheme, where at the top-level, a tweet is classified as either offensive or not offensive. Tweets which are offensive are further divided into sub-categories based on whether the offense is untargeted or targeted against a group or individual. Similar to Davidson [4], they employ a majority voting scheme to annotate tweets using the crowd-sourcing platform Appen[2]. For our work, we chose the 14k tweets from their top-level of annotation.

1.5 The Basile et al. dataset (BA)

This hate speech dataset was used in the SemEval-2019 Task 5: Multilingual Detection of Hate Speech Against Immigrants and Women in Twitter [1]. To build this dataset, the authors monitored victims of known abusive accounts on Twitter and used important keywords and hashtags to filter their tweets. Some of the frequent keywords collected in the 13k English tweets are: *migrant, refugee, #buildthatwall, b*tch, women*. The tweets targeted against women were collected from a previous challenge on misogyny identification [5].

1.6 The Waseem and Hovy dataset (WH)

Waseem [9] published a hate speech detection dataset of 16k tweets. Their corpus is built by searching for slurs targeted against religious,

[1] www.hatebase.org
[2] https://appen.com/

sexual, gender, and ethnic minorities on Twitter. The authors manually annotated the tweets into one of three classes: *racism*, *sexism* or *normal*. The 1,972 tweets in the *racism* class are from just 6 users, and the 3,383 tweets in the *sexism* class are from 613 users. While the original dataset was composed of 36% positive (racism or sexism) labels, many of these tweets have since been taken down from Twitter. We managed to collect 10,018 tweets and ignored the racism class for further analysis as it only contained 13 datapoints.

2 Interface for annotation

The interface for annotation that appear for each of the annotators is shown in Figure 1.

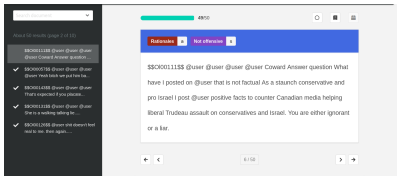

Figure 1. Interface for annotation.

3 Other hyperparameters

We set the batch size as 16 and train all models for 20 epochs, reporting the test performance at the epoch where the validation performance is the best. We used AdamW optimizer for optimization with default parameters. These values are constant through the whole experiment. For all the models, we used a dropout of 0.2 in the final linear layer, while for the RAFT models we further use 0.2 as dropout in the attention layer as well. Other than that for LIME [8], we used 10 features for explanation and 100 as the size of neighbourhood to learn the linear model. Other parameters were set to default. We evaluate the LIME/SHAP based explanation using the explanations for the most confident abusive class.

4 Most similar source-target pairs

All the pairwise cosine similarities are noted in Table 1. For each row in this table, we select the best source dataset based on the maximum similarity value.

[]

Source → Target ↓	DA	FA	OD	BA	WH	HX
DA	–	**0.76**	0.68	0.73	0.73	0.56
FA	0.76	–	0.80	**0.92**	0.87	0.68
OD	0.68	0.80	–	0.82	**0.84**	0.62
BA	0.73	**0.92**	0.82	–	0.87	0.77
WH	0.73	0.87	0.84	**0.88**	–	0.63

Table 1. In this table, we show the pairwise cosine similarity in the term distribution of different corpus. The row headers represent the target domain and the column header denote the source domain. For each row we select the best domain to transfer from.

5 System description

For all the experiments in this paper, we used a 48-core Xeon processor Linux based system with 126 GB RAM. For training the neural networks, we used 2 NVIDIA P100 GPUs having 16 GB RAM each with CUDA version 10.1. We primarily based our system on Python libraries. For preprocessing we used the ekphrasis[3] library. Huggingface's transformers library was used for BERT-based models, with PyTorch as backend in general. All libraries used in our research are pip installable.

6 Efficiency of explanation generation

We also measure the efficiency for generating the explanations by the LIME and RAFT methods. The LIME method takes around 7 seconds to generate explanation for a text while RAFT models take around 1 second to generate an explanation. The RAFT models are 7 times more efficient than LIME.

For training the rationale predictor model, it took average of 1 hour/run. For the few shot experiments on the cross-domain dataset (50 datapoints) it takes 10-12 mins to train a single model. RGFS-SA and RGFS-CA takes 14% and 18% more time on average than the vanilla BERT models respectively.

[!t]

Target groups	Categories
Race	African, Arabs, Asians, Caucasian, Hispanic, Indian
Religion	Buddhism, Christian, Hindu, Islam, Jewish, Non-religious
Gender	Men, Women
Sexual Orient.	Heterosexual, LGBTQ
Miscellaneous	Indigenous, Refugee/Immigrant, None, Others,Disability,Economic

Table 2. Target groups which occurred more than 20 times in the annotated dataset.

[!t]

Dataset	Train	Val	Test
HX (S)	15383	1922	1924
FA	69859	9881	20059
DA	17347	2479	4957
OD	9869	1396	2834
BA	8963	1281	2561
WH	7635	2192	1080

Table 3. This table shows the number of datapoints in train, validation and test data. The HateXplain dataset is divided into 8:1:1 ratio and other datasets are divided into 7:1:2 ratio into different splits.

[!t]

Dataset	#samples	Jaccard	random Jaccard
FA	34	0.67	0.34
DA	50	0.61	0.33
OD	35	0.66	0.27
BA	50	0.58	0.32
WH	40	0.57	0.26

Table 4. This table shows the number of samples annotated as abusive by both annotators out of the 50 samples per dataset, the Jaccard overlap between the annotated rationales (Jaccard) and random rationales (random Jaccard).

7 How do rationales help?

In this section, we discuss further insights from our findings. We argue that human-like rationales play a very important role in learning

[3] https://github.com/cbaziotis/ekphrasis

Model	Text
Human annotation (OD dataset)	user user user that expected if you placate the violent leftists/ terrorists. kavanaugh confirmation woke
BERT	user user user that expected if you placate the violent leftists / terrorists. kavanaugh confirmation woke
BERT-L-DOM	user user user that expected if you placate the violent leftists/ terrorists. kavanaugh confirmation woke
RAFT-SA/CA	user user user that expected if you placate the violent leftists/ terrorists. kavanaugh confirmation woke
Human annotation (BA dataset)	user user user user a very high wall must be build to protect usa from bad elements of illigal refugees.
BERT	user user user user a very high wall must be build to protect usa from bad elements of illigal refugees.
BERT-L-DOM	user user user user a very high wall must be build to protect usa from bad elements of illigal refugees.
RAFT-SA/CA	user user user user a very high wall must be build to protect usa from bad elements of illigal refugees.

Table 5. Examples of rationales predicted by different models compared to human annotators. The first row corresponds to the annotation done by humans highlighted in yellow. The green highlight represents the tokens which human annotators and the model found important. The orange highlight represents the tokens which the model found important, but the human annotators did not. BERT-L-DOM is the best cross domain model taken from Table 1 corresponding to each target dataset.

subjective tasks like hate speech, sarcasm etc., as in the absence of such rationales, models can often focus on artefacts to get good performances. This is also evident from Table 5, where the LIME based explanation focuses on artefact words present in the post like 'user'[4], 'if', 'must', 'build' etc. On the other hand, the rationales learnt using our BERT-RLT model are near perfect; this is also highlighted through the plausibility measurement in Table 6.

Furthermore, the rationale prediction is in the zero-shot setting, i.e., none of the target datasets contain labeled rationales that the model is fine-tuned upon. The performance will further improve if we can include few labelled rationale annotations [2]. Since rationale annotation is more costly for the annotator/moderator, such zero-shot/few-shot rationale predictors can be very useful for reducing the overall workload of annotation. We would like to point out that similar to any machine learning algorithm, such rationale predictors can be erroneous. Appropriate feedback loops may be set to correct the model in those cases.

Ideally, we would like to have an explainable model which predicts the correct label along with the correct reasons. Current post-hoc explanations like LIME, although designed as faithful, may or may not provide correct/plausible reasons behind the prediction. Our rationale based attention framework outperforms LIME in terms of plausibility (noted in Table 6) and performs comparably in terms of faithfulness metrics (as noted in the Table 7). We believe this is a step in the right direction and future research in this direction can further develop better methods to add rationales.

[!t]

Model	Data	Plausibility		
		AUPRC	token-F1	IOU-F1
BERT-L-DOM + LIME	DA	0.77	0.52	0.21
BERT-L-DOM + SHAP		0.45	0.37	0.14
RAFT-CA/SA		**0.84**	**0.58**	**0.26**
BERT-L-DOM +LIME	OD	0.49	0.36	0.10
BERT-L-DOM +SHAP		0.37	0.32	0.07
RAFT-CA/SA		**0.68**	**0.54**	**0.11**
BERT-L-DOM + LIME	BA	0.63	0.43	0.19
BERT-L-DOM +SHAP		0.46	0.34	0.14
RAFT-CA/SA		**0.76**	**0.55**	**0.23**
BERT-L-DOM + LIME	FA	0.61	0.46	0.11
BERT-L-DOM + SHAP		0.48	0.36	0.06
RAFT-CA/SA		**0.67**	**0.54**	**0.13**
BERT-L-DOM + LIME	WH	0.57	0.42	0.01
BERT-L-DOM + SHAP		0.50	0.35	0.01
RAFT-CA/SA		**0.84**	**0.63**	0.01

Table 6. Average AUPRC, token-F1 and IOU-F1 scores for the rationales predicted by the models which were trained using different sets of 50 datapoints. For the models not utilising rationales in their architecture, LIME and SHAP are used to predict the rationales. BERT-L-DOM is the best cross-domain model for each dataset.

[4] The anonymised version of mention.

[!t]

Model	Data	Faithfulness	
		Suff.(↓)	Comp.
BERT-L-DOM + LIME	DA	-0.03	**0.67**
BERT-L-DOM + SHAP		**-0.29**	-0.14
RAFT-CA		-0.06	0.11
RAFT-SA		-0.08	0.25
BERT-L-DOM + LIME	OD	-0.02	0.09
BERT-L-DOM + SHAP		-0.10	-0.09
RAFT-CA		-0.04	0.04
RAFT-SA		-0.08	**0.29**
BERT-L-DOM + LIME	BA	-0.05	**0.34**
BERT-L-DOM + SHAP		-0.38	-0.38
RAFT-CA		-0.04	0.06
RAFT-SA		**-0.07**	0.19
BERT-L-DOM + LIME	FA	-0.02	**0.40**
BERT-L-DOM + SHAP		-0.09	0.09
RAFT-CA		**-0.13**	0.09
RAFT-SA		-0.11	0.26
BERT-L-DOM + LIME	WH	-0.07	**0.20**
BERT-L-DOM + SHAP		-0.09	-0.09
RAFT-CA		0.01	0.03
RAFT-SA		0.06	0.06

Table 7. Average comprehensiveness scores (Comp) and the sufficiency scores (Suff) for the rationales predicted by the models which were trained using different sets of 50 datapoints. For the models not utilising rationales in their architecture, LIME and SHAP are used to predict the rationales. BERT-L-DOM is the best cross domain model corresponding to each dataset. For sufficiency scores, lower values are better.

References

[1] Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti, 'SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter', in *SemEval 2019*, pp. 54–63. Association for Computational Linguistics, (June 2019).

[2] Meghana Moorthy Bhat, Alessandro Sordoni, and Subhabrata Mukherjee, 'Self-training with few-shot rationalization: Teacher explanations aid student in few-shot nlu', *arXiv preprint arXiv:2109.08259*, (2021).

[3] Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber, 'Racial bias in hate speech and abusive language detection datasets', in *Proceedings of the Third Workshop on Abusive Language Online*, pp. 25–35, Florence, Italy, (August 2019). Association for Computational Linguistics.

[4] Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber, 'Automated hate speech detection and the problem of offensive language', ICWSM '17, pp. 512–515, (2017).

[5] Elisabetta Fersini, Debora Nozza, and Paolo Rosso, 'Overview of the evalita 2018 task on automatic misogyny identification (ami)', *EVALITA Evaluation of NLP and Speech Tools for Italian*, **12**, 59, (2018).

[6] Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis, 'Large scale crowdsourcing and characterization of twitter abusive behavior', in *ICWSM 2018*. AAAI Press, (2018).

[7] Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee, 'Hatexplain: A benchmark dataset for explainable hate speech detection', in *Proceedings of the*

AAAI Conference on Artificial Intelligence, volume 35, pp. 14867–14875, (2021).

[8] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin, '"why should I trust you?": Explaining the predictions of any classifier', in *ACM SIGKDD*, pp. 1135–1144, (2016).

[9] Zeerak Waseem and Dirk Hovy, 'Hateful symbols or hateful people? predictive features for hate speech detection on Twitter', in *Proceedings of the NAACL Student Research Workshop*, pp. 88–93, San Diego, California, (June 2016). Association for Computational Linguistics.

[10] Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar, 'Predicting the type and target of offensive posts in social media', in *NAACL*, pp. 1415–1420, Minneapolis, Minnesota, (June 2019). ACL.