Dynamically Refined Regularization for Improving Cross-corpora Hate Speech Detection

Tulika Bose, Nikolaos Aletras, Irina Illina, Dominique Fohr

Dynamically Refined Regularization for Improving Cross-corpora Hate Speech Detection

Tulika Bose [†] **Nikolaos Aletras** [‡] **Irina Illina** [†] **Dominique Fohr** [†]
[†] Universite de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
[‡]University of Sheffield, United Kingdom
{tulika.bose, illina, dominique.fohr}@loria.fr
n.aletras@sheffield.ac.uk

Abstract

Warning: *this paper contains content that may be offensive and distressing.*

Hate speech classifiers exhibit substantial performance degradation when evaluated on datasets different from the source. This is due to learning spurious correlations between words that are not necessarily relevant to hateful language, and hate speech labels from the training corpus. Previous work has attempted to mitigate this problem by regularizing specific terms from pre-defined static dictionaries. While this has been demonstrated to improve the generalizability of classifiers, the coverage of such methods is limited and the dictionaries require regular manual updates from human experts. In this paper, we propose to automatically identify and reduce spurious correlations using attribution methods with dynamic refinement of the list of terms that need to be regularized during training. Our approach is flexible and improves the cross-corpora performance over previous work independently and in combination with pre-defined dictionaries.[1]

1 Introduction

The relative sparsity of hateful content in the real world requires crawling of many of the standard hate speech corpora through keyword-based sampling (Poletto et al., 2021), rather than random sampling. Thus, hate speech classifiers (D'Sa et al., 2020; Mozafari et al., 2019; Badjatiya et al., 2017) often learn spurious correlations from the training corpus (Wiegand et al., 2019) leading to a substantial performance degradation when evaluated on a corpus with a different distribution (Yin and Zubiaga, 2021; Bose et al., 2021; Florio et al., 2020; Arango et al., 2019; Swamy et al., 2019; Karan and Šnajder, 2018).

Target corpus utterances	Actual	Predicted
Genocide is never ok	non-hate	hate
Women are goddesses	non-hate	hate

Table 1: Spurious correlations learned by the source classifier between the shaded tokens and the hate label.

Recent work has proposed regularization mechanisms to penalize spurious correlations by attempting to explain model predictions using feature attribution methods (Ross et al., 2017; Rieger et al., 2020; Adebayo et al., 2020). These methods assign importance scores to input tokens that contribute more towards a particular prediction (Lundberg and Lee, 2017). For instance, Liu and Avci (2019) penalize the attributions assigned to tokens contained in a manually curated dictionary consisting of group identifiers (e.g. women, jews) that are often known to be targets of hate. Kennedy et al. (2020) extract group identifiers manually from the top tokens indicated by a bag-of-words logistic regression model trained on the source corpus. However, regularizing only group identifiers limits the coverage of such approaches, and may not capture other forms of corpus-specific correlations learned by the classifier limiting its performance on a new corpus. Moreover, such manually curated lists may not always remain up-to-date because new terms emerge frequently (Grieve et al., 2018). While Yao et al. (2021) do not use such lists for refining models in different target-domains, their method still requires input from human annotators.

In this paper, we hypothesize that the classification errors in a small annotated subset from the target can reveal spurious correlations between tokens and hate speech labels learned from the source (see Table 1). To this end, we propose Dynamic Model Refinement (D-Ref), a new method to identify and penalize spurious tokens using feature attribution methods. We demonstrate that D-Ref improves the overall cross-corpora performance independently and in combination with pre-defined dictionaries.

[1] Code is available here: https://github.com/tbose20/D-Ref

2 Dynamic Model Refinement (D-Ref)

In this section, we describe the general theoretical framework of the proposed approach. We assume that during training our hate speech classification model has access to the source training corpus D_S^{train} and a small validation set D_T^{val} from a target corpus with different distribution, following a similar setting to Maharana and Bansal (2020). Our Dynamic Model Refinement (D-Ref) approach consists of 2 recurring steps across epochs: (i) we first extract a set of spurious tokens using D_T^{val} at the end of every epoch; and (ii) then we penalize the extracted tokens during the next epoch.

2.1 Extraction of Spurious Tokens

Global token-ranking in source corpus: We first begin with identifying the tokens from D_S^{train} that are highly correlated with hate/non-hate labels. These tokens are suitable candidates for causing source-specific spurious correlations, restricting generalizability to a new corpus.

For that purpose, at the end of every training epoch ep_i, we first obtain the global class-specific ranked list of tokens from D_S^{train}. This is achieved by computing global attributions per token tok and class c (gl-atr$_{tok}^{c}$) from its attribution per instance j (loc-atr$_{tok}^{j}$) averaged across all training instances classified as c by the source model trained until ep_i:

$$\text{gl-atr}_{tok}^{c} = \frac{\sum_{j=1}^{|D_S^{train}|} \mathbb{1}_{\hat{y_j}=c} \text{loc-atr}_{tok}^{j} \forall \text{occurrence of } tok \text{ in j}}{\sum_{j=1}^{|D_S^{train}|} \mathbb{1}_{\hat{y_j}=c} \#(\text{occurrence of } tok \text{ in j})} \tag{1}$$

Here $c \in \{\text{hate, non-hate}\}$, $\hat{y}$ is the predicted class and $\mathbb{1}$ is the indicator function. Prior to this, loc-atr$_{tok}^{j}$ are individually normalized using sigmoid to obtain values in a closed range. Rarely occurring tokens and stop-words are not considered for the global ranking. The gl-atr$_{tok}^{c}$ values are sorted from the highest globally attributed token to the lowest, which yields two ranked token-lists $[gl\text{-}hate, gl\text{-}nhate]_{ep_i}$.

Instance-level local ranking in target corpus: We hypothesize that tokens highly correlated with hate/non-hate classes in the source, but also causing mis-classifications in the target, should most likely contribute to spurious source-specific correlations, and may not be important for hate speech labels. Thus, we identify the tokens that cause mis-classifications in D_T^{val}, and then obtain a list of spurious tokens *dynamically* after every epoch ep_i.

We rank the tokens in the target instances from D_T^{val} based on their loc-atr$_{tok}^{j}$, starting from the highest attributed token per instance j to the lowest. The top k tokens in j is given by $tok_{top_k}^{j} = top_k[\text{argsort}(\text{loc-atr}_{tok}^{j})]$, where k is a hyper-parameter in D_T^{val}. We treat the two error cases of False Positives (FP) and False Negatives (FN) separately. Here the hate class is considered as the positive class.

Since the tokens responsible for FP may also be important for the True Positives (TP), we only extract those that have high attributions for FP, but not for TP. Further, another filtering step is applied, where only the tokens common to the top N from the ranked $gl\text{-}hate$ are extracted. This results in discarding the tokens that may not be globally correlated with a class with respect to the source model. So $tok_{FP} = [tok \in tok_{top_k}^{j_{FP}}$ $\&$ $tok \notin tok_{top_k}^{j_{TP}}] \cap top_N(gl\text{-}hate)$ $\forall$ instances j in D_T^{val}. Similarly, top k tokens corresponding to FN instances are extracted, wherein those common to TN are discarded, and subsequent filtering based on the $gl\text{-}nhate$ is performed, i.e. $tok_{FN} = [tok \in tok_{top_k}^{j_{FN}}$ $\&$ $tok \notin tok_{top_k}^{j_{TN}}] \cap top_N(gl\text{-}nhate) \forall j$. This step thus yields a list of possible spurious tokens at the end of ep_i, $S_{ep_i} = [tok_{FP}, tok_{FN}]_{ep_i}$.

2.2 Penalizing the Extracted Spurious Tokens

In this step, we attempt to reduce the importance assigned, by the source model, to the extracted spurious tokens by penalizing the terms in S_{ep_i} during the next epoch ep_{i+1}. We propose three different ways for token penalization:

Tok-mask: In this case, we simply mask the tokens from S_{ep_i} present in D_S^{train} after every ep_i and then train the source model during ep_{i+1}.

Reg: Since token masking might eliminate substantial information, we regularize the model using S_{ep_i}. The attributions assigned to these terms are pushed towards zero by the following learning objective on D_S^{train}:

$$\mathcal{L} = \mathcal{L}' + \lambda \mathcal{L}_{\text{atr}}\left(\text{t}\right); \text{t} \in \mathcal{S}_{\text{ep}_\text{i}}; \mathcal{L}_{\text{atr}} = \sum_{\text{t} \in \mathcal{S}_{\text{ep}_\text{i}}} \phi\left(\text{t}\right)^2 \tag{2}$$

where $\mathcal{L}'$ is the classification loss and $\mathcal{L}_{\text{atr}}$ is the attribution loss. Here $\phi\left(\text{t}\right)$ is the attribution score for the token t. Intuitively, this should reduce the importance of tokens contributing to source-specific patterns and encourage learning more general information. Both losses are computed over D_S^{train}.

Comb: We finally combine S_{ep_i} with the pre-defined group identifiers from Liu and Avci (2019) and Kennedy et al. (2020) to perform regularization using Equation 2.

We surmise that repeating these steps at the end of every epoch should reduce the source-specific correlations while the source model gets trained. We use three different attribution methods:

(i) Scaled Attention ($\alpha\nabla\alpha$) (Serrano and Smith, 2019): Here attention weights α_i are scaled with their corresponding gradients $\nabla\alpha_i = \frac{\delta\hat{y}}{\delta\alpha_i}$, where $\hat{y}$ is the predicted label. Serrano and Smith (2019) show that combining an attention weight with its gradient can better indicate token importance for model predictions, compared to only using the attention weights.

(ii) Integrated Gradients (IG) (Sundararajan et al., 2017): This method is based on the notion that the gradient of a prediction function with respect to input can indicate the sensitivity of the prediction for each input dimension. As such, it aggregates the gradients along a path from an uninformative reference input (e.g. zero embedding vector) towards the actual input such that the predictions change from uncertainty to certainty.

(iii) Deep Learning Important FeaTures (DeepLIFT/DL) (Shrikumar et al., 2017): This aims to explain the difference in the output from a reference output in terms of the difference of the input and a reference input. Given a target output neuron t, a reference activation t^0 of t, and $\Delta t = t - t^0$, it computes the contribution scores $C_{\Delta x_i \Delta t}$ of each input neuron x_i that are necessary and sufficient to compute t, such that $\sum_{i=1}^{n} C_{\Delta x_i \Delta t} = \Delta t$. The reference input could be the zero embedding vector.

3 Experiments and Results

3.1 Experimental Setup

Data We use three standard hate speech corpora: *HatEval* (Basile et al., 2019), *Waseem* (Waseem and Hovy, 2016) and *Dynamic* (Vidgen et al., 2021). Following previous work by Wiegand et al. (2019); Swamy et al. (2019), we consider the detection of hate vs non-hate, where the hate class covers all forms of hate. We split *Waseem* (26.8% hate) into train (80%; 8720), val (10%; 1090) and test (10%; 1090) sets as no standard splits are provided. We use the original splits for *HatEval* (42.1% hate; train: 8993[2], val: 1000; test: 3000) and *Dynamic* (54.4% hate; train: 32497, val: 1016, test: 4062). We reduce the size of available D_T^{val} in *Dynamic* by randomly sampling 25% of the validation set (4064). We remove URLs, split hashtags into words using the CrazyTokenizer[3], remove infrequent Twitter handles, punctuation marks and numbers, and convert text into lower-case. See Appendix A for a detailed discussion on the corpora.

Baselines We compare D-Ref with the following baselines: **(i) BERT Van-FT** (Devlin et al., 2019): vanilla fine-tuning on D_S^{train} without regularization; **(ii)** Convolutional Neural Network with regularization of pre-defined group identifier terms using IG for feature attribution (Liu and Avci, 2019); **(iii)** BERT using two variations for regularization: (a) all the mentioned group identifiers, (b) group identifiers extracted from the top features of a bag-of-words logistic regression trained on each individual corpus (Kennedy et al., 2020)[4]; **(iv) χ^2-test** with one degree of freedom and Yate's correction (Kilgarriff, 2001) to extract tokens tok from D_S^{train} that reject the null hypothesis with 95% confidence. The null hypothesis states that in terms of tok, both D_S^{train} and D_T^{val} are random samples of the same larger population. We, then, regularize the attribution scores[5] assigned to these terms, with BERT. **(v) Pre-def**: BERT with regularizing the combined pre-defined group identifiers from (ii) and (iii).

Model training We use pre-trained BERT (Devlin et al., 2019) for our approach. We train all the models over D_S^{train} from the source and evaluate over D_T^{test} from the target. The best model for all the baselines and D-Ref are selected by tuning over D_T^{val}. See Appendix B on hyper-parameter tuning.

3.2 Cross-corpora Predictive Performance

Table 2 presents macro-F1 scores across five random initializations of each experiment using six cross-corpora pairs. We observe that overall, all feature-attribution methods with D-Ref yield improved performance compared to Van-FT and other baselines. While χ^2 yields improvements over Van-FT, D-Ref still displays better performance in most of the cases. This could be attributed to the fact

[2] We remove the instances that contain only URLs, reducing the train instances from 9000 to 8993.

[3] https://redditscore.readthedocs.io

[4] We use Sampling and Occlusion (Jin et al., 2020).

[5] We use DL as it yields comparable or higher overall improvements taking Table 2 and Table 3 together.

Approaches	**H →D**	**D →H**	**H →W**	**W →H**	**D →W**	**W →D**	**Average**
BERT Van-FT	53.2±1.0	63.3±1.8	67.5±5.1	52.6±2.4	60.3±1.0	46.7±4.0	57.3
Liu and Avci (2019)	45.1±4.5	59.5±0.7	57.2±3.8	52.6±0.8	57.1±2.7	39.6±2.0	51.9
Kennedy et al. (2020) (a)	52.2±1.2	62.0±1.6	62.7±2.9	50.1±6.8	53.5±2.0	45.1±2.3	54.3
Kennedy et al. (2020) (b)	52.0±3.8	61.9±1.7	63.6±3.7	54.8*±1.6	57.0±1.7	46.8±1.9	56.0
BERT χ^2-test	55.4*±1.1	65.0*±1.0	68.1±1.3	53.7±2.1	60.4±2.8	45.2±2.8	58.0
Pre-def ($\alpha\nabla\alpha$)	54.6*±1.3	**65.1***±1.1	69.6±3.4	54.4*±1.2	**61.9**±1.6	47.2±3.1	58.8
D-Ref-Tok-mask ($\alpha\nabla\alpha$)	53.8±0.6	64.9*±0.7	68.9±3.3	53.6±3.0	59.6±2.2	45.8±3.7	57.8
D-Ref-Reg ($\alpha\nabla\alpha$)	54.9*±1.2	**65.1***±0.9	68.6±4.0	54.1*±1.0	60.9±1.5	**48.7***±4.3	58.7
D-Ref-Comb ($\alpha\nabla\alpha$)	**55.0***±1.6	64.7*±1.2	**69.9**±1.6	**55.3***±1.3	61.0±2.8	48.1*±1.0	**59.0**
Pre-def (IG)	55.7*±1.4	63.5±2.8	**69.7**±2.2	51.7±2.7	60.3±2.2	44.6±3.0	57.6
D-Ref-Tok-mask (IG)	56.3*±2.3	64.5*±1.8	68.3±2.0	52.3±2.3	59.3±1.3	48.2*±2.1	58.2
D-Ref-Reg (IG)	**56.4***±1.4	**65.5***±0.8	69.2±2.5	**53.8***±0.7	60.6±1.7	47.7±3.6	58.9
D-Ref-Comb (IG)	55.7*±0.8	63.7±2.4	69.1±2.3	52.6±2.3	**61.4**±2.5	**51.4***±3.6	**59.0**
Pre-def (DL)	54.2±1.6	64.0±1.9	68.1±1.5	52.9±1.2	62.0±1.8	44.5±1.3	57.6
D-Ref-Tok-mask (DL)	55.1*±1.4	**64.9***±1.7	67.2±3.6	52.1±1.9	60.5±2.5	47.2±3.1	57.8
D-Ref-Reg (DL)	54.2±1.6	64.8*±0.8	**70.7***±2.7	51.4±0.7	**62.3***±2.5	47.1±5.5	58.4
D-Ref-Comb (DL)	**55.4***±1.8	64.0±0.9	69.5±3.3	**54.0***±0.8	61.5±2.3	**48.1***±2.7	**58.8**

Table 2: Macro-F1 (±std-dev) on source →target pairs (H : HatEval, D : Dynamic, W : Waseem). **Bold** denotes the best performing approach in each column for every feature attribution method. * denotes statistical significance compared to Van-FT with paired bootstrap (Dror et al., 2018; Efron and Tibshirani, 1993), 95% confidence interval.

that although the terms obtained through the χ^2 test from the source indicate differences across domains, they may not necessarily be important for the prediction of hate/ non-hate labels by the source model, and may not contribute to source-specific spurious correlations.

We find that D-Ref-Reg with IG and DL achieves better average macro-F1 of 58.9 and 58.4 respectively, compared to the corresponding Pre-def (IG) and Pre-Def (DL) that obtain an average of 57.6. D-Ref-Reg ($\alpha\nabla\alpha$) provides an average macro-F1 of 58.7, comparable to Pre-def ($\alpha\nabla\alpha$) with 58.8. However, D-Ref-Reg achieves significantly improved scores in more cases, as compared to Pre-def using all the attribution methods, i.e. 4/6 cases ($\alpha\nabla\alpha$), 3/6 cases (IG) and 3/6 cases (DL) with D-Ref-Reg, compared to 3/6 ($\alpha\nabla\alpha$), 1/6 (IG) and none (DL) with Pre-def. D-Ref-Tok-mask exhibits improvements on average ($\alpha\nabla\alpha$: 57.8, IG: 58.2, DL: 57.8) over Van-FT (57.3), demonstrating the effectiveness of the token extraction mechanism of D-Ref. Finally, D-Ref-Comb displays the best overall performance, with the highest average score of 59. We attribute this improvement from D-Ref to its increased coverage with dynamic token extraction, and reduction of spurious source-specific correlations, while the baselines only penalize the group identifiers. A dynamic approach also corrects the model during training before it can get fully biased towards these tokens. Finally, it can incorporate the pre-defined lists along with the extracted tokens, and further improve the performance.

3.3 Domain-Adaptation Approaches

We further compare D-Ref-Reg with various Domain Adaptation (DA) methods. However, such methods typically leverage the unlabeled train set from the target domain (D_T^{train}). We first continue pre-training BERT model on D_T^{train} following Rietzler et al. (2020). Then, we perform supervised fine-tuning and regularization on D_S^{train} using D-Ref-Reg (**M**asked **L**anguage **M**odel + D-Ref-Reg). We compare against the following methods:

(i) BERT Van-MLM-FT : MLM training of BERT on D_T^{train} and supervised fine-tuning on D_S^{train}.

(ii) BERT PERL (Pivot-based Encoder Representation of Language) (Ben-David et al., 2020): This performs pivot based fine-tuning using the MLM objective of BERT by masking and predicting the pivot terms present in the combination of D_S^{train} and the unlabeled D_T^{train}. Here pivots are terms that are frequently present in the unlabeled data of both the source and target corpora, and are predictive of the source labels.

(iii) BERT-AAD (Adversarial Adaptation with Distillation) (Ryu and Lee, 2020), This is a domain adversarial approach with BERT where a target encoder is adapted with an adversarial objective that leverages D_S^{train} and D_T^{train}.

(iv) HATN (Hierarchical Attention Transfer Network) (Li et al., 2018, 2017) This approach uses attention and a domain adversarial pivot extraction mechanism.

(v) Sarwar and Murdock (2021): This adopts a data-augmentation strategy leveraging a negative emotion dataset (Go et al., 2009), for cross-domain hate-speech detection. They construct a weakly labeled augmented dataset by training a sequence

Approaches	**H →D**	**D →H**	**H →W**	**W →H**	**D →W**	**W →D**	**Average**
BERT Van-MLM-FT	56.6±1.3	66.2±1.2	70.0±2.5	50.9±2.1	61.4±2.4	43.5±1.9	58.1
BERT PERL	54.1±0.7	60.0±0.6	60.1±2.0	**55.2***±0.7	55.5±1.0	37.8±1.2	53.8
BERT-AAD	56.6±1.3	53.9±3.5	68.8±2.5	50.7±1.4	48.3±4.7	**53.0***±1.7	55.2
HATN	48.4±1.6	59.1±0.4	59.7±2.9	51.4±1.8	60.0±2.6	45.4±2.7	54.0
MLM + Sarwar and Murdock (2021)	55.0±1.9	66.2±2.0	68.8±1.1	48.2±3.1	57.9±1.3	36.2±1.1	55.4
MLM + χ^2-test	57.9±1.6	67.1±1.7	69.8±0.8	48.2±3.1	60.4±2.8	44.1±3.4	57.9
MLM + D-Ref-Reg ($\alpha\nabla\alpha$)	57.6±1.9	66.2±1.2	**70.7**±1.2	52.5*±4.0	62.8±1.4	48.0*±4.3	59.6
MLM + D-Ref-Reg (IG)	58.6*±1.2	**66.8**±0.5	70.1±1.5	52.1±3.0	62.5±3.0	48.9*±4.4	59.8
MLM + D-Ref-Reg (DL)	**58.8***±2.2	66.7±0.6	70.5±1.3	52.4*±3.5	**64.7***±2.1	51.5*±4.9	**60.8**

Table 3: Comparison of DA approaches with D-Ref + MLM. Macro-F1 (±std-dev) on different source →target pairs. H : HatEval, D : Dynamic, W : Waseem. * denotes the significantly improved scores w.r.t. Van-MLM-FT.

tagger on D_S^{train} and a TF-IDF based template matching with D_T^{train}.

(vi) χ^2-test using D_S^{train} and D_T^{train}.

For a fair comparison, we initialize (v) and (vi) with the MLM trained BERT on D_T^{train}, while the other methods already make use of D_T^{train} for adaptation. We use D_T^{val} from target for model selection for all the above methods.

Table 3 shows the results on comparing against other DA approaches. We note that the average performance of all the other DA approaches in this task is lower than Van-MLM-FT, as discussed in our previous work (Bose et al., 2021). χ^2-test, on an average, fails to surpass the Vanilla baseline. Besides, the DA approach proposed for cross-domain hate-speech detection by Sarwar and Murdock (2021) also yields an overall drop in performance. They perform data-augmentation by replacing relevant words from an external negative emotion dataset with tagged hateful terms from the target domain. We find that a major portion of the augmented instances lack meaning, and this negatively impacts the adaptation. However, across all feature attribution methods, D-Ref-Reg improves the cross-corpora performance compared to Van-MLM-FT and the DA approaches, with average macro-F1 of 59.6 ($\alpha\nabla\alpha$), 59.8 (IG), and 60.8 (DL), compared to 58.1 from Van-MLM-FT. Since D-Ref-Reg and Van-MLM-FT use identical MLM pre-training on D_T^{train}, the improvements can be attributed to the dynamic token extraction of our method.

More generally, when the larger set of target domain unannotated instances D_T^{train} are unavailable, D-Ref can identify and correct spurious correlations on source using a small amount of annotated instances from the target D_T^{val}, as demonstrated in Section 3.2. When sufficient number of unannotated instances from the target corpus are available, D-Ref can yield further cross-corpora improvements by leveraging the unannotated target instances with the MLM pre-training.

Non-hate examples *HatEval →Dynamic*	
FP with Van-FT	**TN with D-Ref-Reg (IG)**
Wildlife belongs in the f*cking wild .	Wildlife belongs in the f*cking wild.
NHS s*cks	NHS s*cks
Hate examples *Waseem →Dynamic*	
FN with Van-FT	**TP with D-Ref-Reg (IG)**
Blacks are savages .	Blacks are savages .
Queers can’t be truly happy .	Queers can’t be truly happy .

Table 4: Change in attributions with D-Ref-Reg (IG).

3.4 Qualitative Analysis

Table 4 shows the change in attributions for some instances in D_T^{test} from *Dynamic* that were misclassified by Van-FT but correctly classified by our D-Ref-Reg (IG). Van-FT wrongly attributes higher importance to ‘f*cking’ and ‘s*cks’ for the hate class in the first example, and ‘blacks’ and ‘queers’ for non-hate in the second due to source-specific correlations. However, D-Ref-Reg (IG), extracts and penalizes abusive tokens like {s*ck, a**hole, d*ck} for the former causing FP and {africans, dark, queer} for the latter causing FN. Our approach not only penalizes the exact tokens, but also those with similar meaning (e.g. ‘blacks’ is contextually close to ‘dark’, ‘africans’), giving more importance to the context around the spurious tokens. See Appendix C for the token-lists.

4 Conclusion

We proposed a dynamic approach for automatic token extraction with regularization of the source model such that the spurious source specific correlations are reduced. Our approach shows consistent cross-corpora performance improvements both independently and in combination with pre-defined tokens. Future work includes applying our method on other cross-domain text classification tasks and exploring how explanation faithfulness can be improved in out-of-domain settings (Chrysostomou and Aletras, 2022).

Ethical Considerations

The approach proposed in the paper is aimed at supporting robust and accurate detection of online hate speech. The datasets used in the work are publicly available and referenced appropriately. The dataset creators have presented, in detail, the data collection process and annotation guidelines in peer-reviewed articles. The offensive terms presented, as examples, are only intended for better analysis of the models for research purposes.

Acknowledgements

This work was supported partly by the french PIA project "Lorraine Université d'Excellence", reference ANR-15-IDEX-04-LUE. Experiments presented in this article were carried out using the Grid'5000 testbed, supported by a scientific interest group hosted by Inria and including CNRS, RENATER and several Universities as well as other organizations (see https://www.grid5000.fr). We thank the anonymous reviewers for their valuable feedback and suggestions. We would also like to thank George Chrysostomou for his help and suggestions regarding the work during informal discussions.

References

Julius Adebayo, Michael Muelly, Ilaria Liccardi, and Been Kim. 2020. Debugging tests for model explanations. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Aymé Arango, Jorge Pérez, and Barbara Poblete. 2019. Hate speech detection is not as easy as you may think: A closer look at model validation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019, Paris, France, July 21-25, 2019*, pages 45–54. ACM.

Pinkesh Badjatiya, Shashank Gupta, Manish Gupta, and Vasudeva Varma. 2017. Deep learning for hate speech detection in tweets. In *Proceedings of the 26th International Conference on World Wide Web Companion*, WWW '17 Companion, page 759–760, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Eyal Ben-David, Carmel Rabinovitz, and Roi Reichart. 2020. PERL: Pivot-based domain adaptation for pre-trained deep contextualized embedding models. *Transactions of the Association for Computational Linguistics*, 8:504–521.

Tulika Bose, Irina Illina, and Dominique Fohr. 2021. Unsupervised domain adaptation in cross-corpora abusive language detection. In *Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media*, pages 113–122, Online. Association for Computational Linguistics.

George Chrysostomou and Nikolaos Aletras. 2022. An empirical study on explanations in out-of-domain settings. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 2018. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1383–1392, Melbourne, Australia. Association for Computational Linguistics.

Ashwin Geet D'Sa, Irina Illina, and Dominique Fohr. 2020. Towards non-toxic landscapes: Automatic toxic comment detection using DNN. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 21–25, Marseille, France. European Language Resources Association (ELRA).

Bradley Efron and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Number 57 in Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, Boca Raton, Florida, USA.

Komal Florio, Valerio Basile, Marco Polignano, Pierpaolo Basile, and Viviana Patti. 2020. Time of your hate: The challenge of time in hate speech detection on social media. *Applied Sciences*, 10(12).

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. *Processing*, pages 1–6.

John G. D. Grieve, Andrea Nini, and Diansheng Guo. 2018. Mapping lexical innovation on american social media. *Journal of English Linguistics*, 46:293 – 319.

Xisen Jin, Zhongyu Wei, Junyi Du, Xiangyang Xue, and Xiang Ren. 2020. Towards hierarchical importance attribution: Explaining compositional semantics for neural sequence models. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Mladen Karan and Jan Šnajder. 2018. Cross-domain detection of abusive language online. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 132–137, Brussels, Belgium. Association for Computational Linguistics.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442, Online. Association for Computational Linguistics.

Adam Kilgarriff. 2001. Comparing corpora. *International Journal of Corpus Linguistics*, 6(1):97–133.

Zheng Li, Ying Wei, Yu Zhang, and Qiang Yang. 2018. Hierarchical attention transfer network for cross-domain sentiment classification. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5852–5859. AAAI Press.

Zheng Li, Yu Zhang, Ying Wei, Yuxiang Wu, and Qiang Yang. 2017. End-to-end adversarial memory network for cross-domain sentiment classification. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 2237–2243. ijcai.org.

Frederick Liu and Besim Avci. 2019. Incorporating priors with feature attribution on text classification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6274–6283, Florence, Italy. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4765–4774.

Adyasha Maharana and Mohit Bansal. 2020. Adversarial augmentation policy search for domain and cross-lingual generalization in reading comprehension. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3723–3738, Online. Association for Computational Linguistics.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2019. A bert-based transfer learning approach for hate speech detection in online social media. In *International Conference on Complex Networks and Their Applications*, pages 928–940. Springer.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Lang. Resour. Evaluation*, 55:477–523.

Laura Rieger, Chandan Singh, W. James Murdoch, and Bin Yu. 2020. Interpretations are useful: Penalizing explanations to align neural networks with prior knowledge. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 8116–8126. PMLR.

Alexander Rietzler, Sebastian Stabinger, Paul Opitz, and Stefan Engl. 2020. Adapt or get left behind: Domain adaptation through BERT language model finetuning for aspect-target sentiment classification. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4933–4941, Marseille, France. European Language Resources Association.

Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. 2017. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 2662–2670. ijcai.org.

Minho Ryu and K. Lee. 2020. Knowledge distillation for bert unsupervised domain adaptation. *ArXiv*, abs/2010.11478.

Sheikh Muhammad Sarwar and Vanessa Murdock. 2021. Unsupervised domain adaptation for hate speech detection using a data augmentation approach. *ArXiv preprint*, abs/2107.12866.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. 2017. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3145–3153. PMLR.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR.

Steve Durairaj Swamy, Anupam Jamatia, and Björn Gambäck. 2019. Studying generalisability across abusive language detection datasets. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 940–950, Hong Kong, China. Association for Computational Linguistics.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of Abusive Language: the Problem of Biased Datasets. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 602–608, Minneapolis, Minnesota. Association for Computational Linguistics.

Huihan Yao, Ying Chen, Qinyuan Ye, Xisen Jin, and Xiang Ren. 2021. Refining language models with compositional explanations. *Advances in Neural Information Processing Systems*, 34.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *ArXiv preprint*, abs/2102.08886.

A Data Description

While *HatEval* and *Waseem* are sampled from Twitter, *Dynamic* is generated using a human-and-model-in-the-loop process. These corpora have been collected across different time frames, and hence they involve different topics of discussion, which are also determined to a large extent by the keywords used for sampling. As such, the problem of dataset bias with spurious correlations are induced with such focused sampling procedures (Wiegand et al., 2019) used in *Waseem* and *HatEval*. For instance, in *Waseem*, a large amount of tweets, available at the time of our experiments, consist of hate tweets directed against women, which results in False Positives for instances from other corpora that contain women related terms. We observed that most of the racist tweets were already removed and were unavailable for experiments. *HatEval*, on the other hand, has a mix of tweets directed against women and immigrants, and hence it demonstrates decent performance when evaluated over *Waseem* that consists of sexist tweets. On the contrary, *Dynamic* contains annotator-generated tweets that includes challenging perturbations. For instance, it includes non-hate instances like 'It's wonderful having gay people around here', 'I hate the concept of hate', 'Tea is f*cking disgusting', which can easily fool a classifier learned on biased datasets, and result in classifying these instances as hateful. Moreover, this corpus covers different targets of hate. As such, when *Dynamic* is used as the target corpus, the spurious correlations learned by the source classifier become relatively well-visible, which are captured and penalized by D-Ref while the source model gets trained.

The data used in the work are publicly available, and download links are provided in the respective original articles, which are referenced in this paper. However, in the case of *Waseem*, where only tweet IDs are provided, some tweets might be unavailable.

B Implementation Details

We leverage the pretrained BERT-base model[6] for our experiments. We use a batch size of 8, learning rate of 1×10^{-5} and Adam optimizer with decoupled weight decay regularization (Loshchilov and Hutter, 2019) for Van-FT, Van-MLM-FT, D-Ref and Pre-def. For Integrated Gradients, following Liu and Avci (2019), the interpolated embeddings are treated as constants while back-propagating the loss from the regularization term. An all zero embedding vector is used as the baseline input for both Integrated Gradients and DeepLIFT. We use the original code, as provided by the respective authors, for all the prior-arts. For Pre-Def, we combined the pre-defined lists from Kennedy et al. (2020) and Liu and Avci (2019) and regularized their attribution scores over BERT with $\alpha\nabla\alpha$, IG, and DL as feature attribution methods.

We implement the data-augmentation approach

[6] https://github.com/huggingface/transformers

proposed by Sarwar and Murdock (2021) ourselves due to the absence of an available implementation. Following the description present in the paper, we prepare the training data for the sequence tagger by labeling all the terms in the hateful instances from the source corpus that are also present in the lexicon from hatebase.org[7]. However, we do not tokenize the lexicon obtained from hatebase.org while searching for the corresponding matching terms in the source corpus. We convert the lexicon into lower-case and look for the exact match in the source corpus.

For D-Ref, we set the value of top N tokens used from ranked {glist-hate, glist-nhate} as 500. The values of $k \in$ top {10%, 20%, 30%, 40%} of the instance-length in D-Ref, and λ in both D-Ref and Pre-def are selected through hyper-parameter tuning over D_T^{val} using a random seed. For $\alpha\nabla\alpha$ and DeepLIFT, $\lambda \in$ {0.1, 0.5, 1, 10, 20, 30, 40, 50, 60} and for IG, $\lambda \in$ {1, 10, 20, 30, 40, 50, 60}. We run supervised fine-tuning on D_S^{train} for 6 epochs with all the BERT models (prior-arts and D-Ref). We select the models (prior-arts and D-Ref) by tuning over D_T^{val} from the target corpus, with respect to macro-F1 scores. Table 5 presents the macro-F1 scores obtained on the validation set for D-Ref and the prior arts.

C Tokens extracted in different epochs

The list of error-causing tokens for False Positives (FP) and False Negatives (FN) in D_T^{val}, extracted for the cases presented in Section 3.4, is given below. We underline the tokens present in the visualization examples (both Table 4 in Section 3.4 and below) and ones similar in meaning to them.

HatEval →Dynamic

- **Epoch 1: FP:** {idiots, conservative, countries, p*ssy, bloody, americans, move, a**hole, hating, beings, feminist, africans, resources, d*ck, resist, females, attacks, dude, anger } **FN:** {hitler, plague, ##urs, crisis, rescue, funding, gorgeous, treason, journalist, lawyers, agenda, roles, principles, bloody, intern}

- **Epoch 2: FP:** {race, hating, flights, sheep, females, ignorant, feminist, resist, attacks, d*ck, kill, boat, countries, p*ssy, refugee, bloody} **FN:** {president, foreigners, illegal, betrayal, lgbt, riots, gorgeous, treason, joking, chris, intelligent, arguments, humans}

- **Epoch 3: FP:** {countries, race, hating, females, feminist, africans, ridiculous, d*ck, express, comments, organized, s*ck, allow, bloody} **FN:** {illegal, hitler, generally, david, intelligent, secret, chris, equality, dating, yellow, treason, abuses, ##gb, humans, plague, dear, nonsense}

- **Epoch 4: FP:** {isis, genocide, indians, society, supported, females, feminist, attacks, s*ck, destroy, migrants } **FN:** {hitler, opportunities, sister, betrayal, ##ame, gorgeous, ##heads, dating, riots, bank, murders, arguments, humans, fights, plague, influence, targeting, supporters, coordination, lies, ##boys}

- **Epoch 5: FP:** {clean, ignorant, slave, feminist, punish, africans, ##ache, d*ck, ##fs, ars, destroy, status, race, p*ssy, western, send} **FN:** {statement, gross, hitler, sending, yellow, waste, hopefully, trapped, riots, bait, sister, coordination, humans}

- **Epoch 6: FP:** {soft, suicide, countries, p*ssy, bloody, genocide, punish, destroy, migrants, vile, beings, savage, feminist, tory, awful, ignorant, ##ists, spend, send} **FN:** {gross, secret, influence, yellow, crime, abuses, participate, approach}

A non-hate comment in *Dynamic* test set for the above case, wrongly classified as hate by Van-FT and correctly classified as non-hate with D-Ref-Reg (IG), is given below. Darker the shade, higher is the attribution:

Van-FT: There is so much cancer patients in the world but it is mostly the young females who are worstly affected by this disease.

D-Ref-Reg: There is so much cancer patients in the world but it is mostly the young females who are worstly affected by this disease.

Waseem →Dynamic

- **Epoch 1: FP:** {female, ##ists, fe, sex, feminist, rap} **FN:** {cast, coward, queer, equality, ##bi , cost, ##sy, born, asian, nazis, kids, cancer, gender, hiring, funded}

- **Epoch 2: FP:** {##ists, her, sex, worse, feminist, ##nt, outraged} **FN:** {welcome, caused, cancer. drag, ##bi, pressure, parent, nazis, troll, cast, trash, ruins, lesbian, attacking, chinese}

[7] https://hatebase/org/

Approaches	**H →D**	**D →H**	**H →W**	**W →H**	**D →W**	**W →D**	**Average**
BERT Van-FT	54.7±0.8	64.7±1.1	65.6±4.5	59.4±1.2	61.9±1.1	46.9±4.7	58.9
Liu and Avci (2019)	45.3±5.2	50.3±1.1	57.1±2.4	49.7±0.5	56.8±3.2	39.3±2.0	49.8
Kennedy et al. (2020) (a)	53.5±1.1	62.8±1.5	60.3±2.5	53.9±8.8	51.3±2.3	43.6±2.3	54.2
Kennedy et al. (2020) (b)	54.8±4.2	55.5±3.9	62.1±1.8	61.3±0.9	58.6±4.4	46.3±2.7	56.4
Pre-def ($\alpha\nabla\alpha$)	55.2±1.0	65.8±1.1	67.8±3.4	59.2±1.0	62.1±1.9	47.2±4.1	59.6
D-Ref-Tok-rem ($\alpha\nabla\alpha$)	54.9±1.0	64.7±1.2	66.6±2.8	58.5±1.4	60.7±1.1	45.9±3.9	58.6
D-Ref-Reg ($\alpha\nabla\alpha$)	55.4±0.7	65.4±1.9	65.5±3.9	59.5±0.9	61.0±1.1	49.6±3.7	59.4
D-Ref-Comb ($\alpha\nabla\alpha$)	56.2±1.7	64.6±0.7	66.8±2.9	59.9±1.3	62.6±1.7	48.1±1.1	59.7
Pre-def (IG)	55.7±1.6	64.8±0.7	67.0±2.2	59.9±0.9	62.3±1.7	44.4±3.2	59.0
D-Ref-Tok-rem (IG)	56.5±1.9	63.5±1.4	65.4±2.0	59.0±1.1	59.9±0.8	49.7±1.9	59.0
D-Ref-Reg (IG)	57.5±2.1	64.8±1.3	67.1±2.3	59.6±1.3	60.3±1.1	47.7±4.0	59.5
D-Ref-Comb (IG)	57.2±0.8	64.3±1.5	67.4±2.5	58.3±0.9	62.0±1.5	52.1±3.7	60.2
Pre-def (DL)	54.5±2.1	65.1±1.1	66.1±1.4	60.1±0.4	61.3±1.4	45.1±1.7	58.7
D-Ref-Tok-rem (DL)	55.4±1.9	65.5±1.5	65.5±3.1	59.0±1.1	61.6±2.1	48.3±3.7	59.2
D-Ref-Reg (DL)	56.0±1.8	65.7±0.8	68.1±2.3	59.3±1.4	63.0±1.4	48.0±5.9	60.0
D-Ref-Comb (DL)	55.1±2.1	65.6±1.4	66.4±3.1	59.1±0.8	61.6±2.2	49.6±3.0	59.6

Table 5: Validation set (D_T^{val}) macro F1 (±std-dev) on source →target pairs (H : HatEval, D : Dynamic, W : Waseem).

Approaches	**HatEval**		**Dynamic**		**Waseem**		**Average**
BERT Van-FT	43.3±1.8		85.1±0.5		85.4±0.7		71.3
In-corpus performance on source (left of arrows) while refining the source model for the target (right of arrows)							
	H →D	**H →W**	**D →H**	**D →W**	**W →H**	**W →D**	
D-Ref-Reg ($\alpha\nabla\alpha$)	39.7±3.2	38.4±1.7	84.1±1.0	84.2±0.8	84.4±0.7	78.8±8.0	68.3
D-Ref-Reg (IG)	40.5±2.0	37.7±2.1	84.0±0.4	84.5±0.4	84.6±1.0	85.3±1.4	69.4
D-Ref-Reg (DL)	37.1±1.8	38.1±2.9	84.7±0.6	84.3±1.2	84.4±0.5	80.7±6.4	68.2

Table 6: In-corpus macro F1 (±std-dev), i.e. the source corpus performance, obtained after refining the source model for the target corpus (present at the right hand side of the arrows) using D-Ref-Reg. H : HatEval, D : Dynamic, W : Waseem. For D-Ref-Reg, model-selection and early-stopping is done over the validation set from the target corpus.

- **Epoch 3: FP:** {female, ##ja, might, men, feminist} **FN:** {quoting, govt, referring, nazis, troll, lesbian, rogue, date, chinese, typically}
- **Epoch 4: FP:** {communism, her, openly, intelligent, many, barbie, chicks, females, arguing} **FN:** {date, suggest, ##lat, referring, police, chinese, cancer, voice, native, lesbian}
- **Epoch 5: FP:** {term, f*ck, ##ng, woman, ##ist, feminist, females, prison} **FN:** {removed, educate, freaking, queer, wow, ending, referring, dye, ##wat, issues, africans, vast, chinese, dark}
- **Epoch 6: FP:** {whore, her, ##ots, role, sweden, pay, d*ck, trump, feminist, females, american, arguing} **FN:** {bat, everyday, freak, argument, movement, chinese, tho, feature, lesbian}

A hate comment in *Dynamic* test set for the above case, wrongly classified as non-hate by Van-FT and correctly classified as hate with D-Ref-Reg (IG), is given below. Darker the shade, higher is the attribution:

Van-FT: Don't get me wrong I don't hate asians, but I definitely don't like them

D-Ref-Reg: Don't get me wrong I don't hate asians, but I definitely don't like them

Since, the *Waseem* dataset is made available as tweet IDs, we observed that it mostly contains sexist comments, while most of the racist content must have been removed before we could crawl it. Hence, the tokens related to race mostly occur in non-hate contexts causing FN.

Even though some error-causing tokens remain in the list until the end, their overall effect should be reduced as the regularization is performed throughout the training procedure, which causes improvement in macro F1.

D In-corpus performance

We present the in-corpus performance, i.e. the performance on the source corpus in terms of macro-F1 scores, obtained when the source model is refined for the corresponding target corpus using D-Ref-Reg, in Table 6. For D-Ref-Reg, the model is tuned over the target corpus validation set. Here

Approaches	HatEval	Dynamic	Waseem
BERT Van-FT	1 m 25 s	3 m 52 s	2 m
D-Ref-Reg ($\alpha\nabla\alpha$)	1 m 39 s	7 m	3 m 33 s
D-Ref-Reg (IG)	9 m 37 s	59 m	19 m 7 s
D-Ref-Reg (DL)	4 m 4 s	18 m 36 s	8 m 44 s

Table 7: Per epoch training time on different source corpora.

BERT Van-FT gives the original performance of the source model, when no refinement is performed, as a reference. In this case, the model is tuned over the in-corpus validation set. The *HatEval* corpus is part of a shared task and involves a challenging test set with low in-corpus performance. The drop across in-corpus performance with D-Ref-Reg is expected, as the main goal of the proposed approach is to make the source model best suited for the target corpus.

E Pre-defined group identifiers

The combined list of pre-defined group identifiers from Liu and Avci (2019) and Kennedy et al. (2020) are given below:

{lesbian, gay, bisexual, trans, cis, queer, lgbt, lgbtq, straight, heterosexual, male, female, nonbinary, african, african american, european, hispanic, latino, latina, latinx, canadian, american, asian, indian, middle eastern, chinese, japanese, christian, buddhist, catholic, protestant, sikh, taoist, old, older, young, younger, teenage, millenial, middle aged, elderly, blind, deaf, paralyzed, muslim, jew, jews, white, islam, blacks, muslims, women, whites, gay, black, democrat, islamic, allah, jewish, lesbian, transgender, race, brown, woman, mexican, religion, homosexual, homosexuality, africans}

F Computational Efficiency

We present the per epoch training time for D-Ref-Reg with different source corpora in Table 7. The training times of D-Ref-Reg ($\alpha\nabla\alpha$) are less than 2 times of that with Van-FT. With D-Ref-Reg (DL), the training time is approximately 4.5 times of that with Van-FT. This demonstrates the computational efficiency of our approach. In the case of D-Ref-Reg (IG), the computation time is indeed high. This occurs due to the aggregation of gradients using a path integral and computing gradients over gradients, as also discussed in Kennedy et al. (2020); Liu and Avci (2019). However, our approach is not dependent on any particular feature attribution method, as demonstrated with our experiments.