Improving Generalization of Hate Speech Detection Systems to Novel Target Groups via Domain Adaptation

Florian Ludwig	Dr. Klara Dolos	Prof. Dr. Torsten Zesch	Dr. Eleanor Hobley
ZITiS	ZITiS	FernUniversität in Hagen	ZITiS
Zamdorfer Str. 88	Zamdorfer Str. 88	Universitätsstraße 47	Zamdorfer Str. 88
81677 München	81677 München	58097 Hagen	81677 München

Abstract

Despite recent advances in machine learning based hate speech detection, classifiers still struggle with generalizing knowledge to out-of-domain data samples. In this paper, we investigate the generalization capabilities of deep learning models to different target groups of hate speech under clean experimental settings. Furthermore, we assess the efficacy of three different strategies of unsupervised domain adaptation to improve these capabilities. Given the diversity of hate and its rapid dynamics in the online world (e.g. the evolution of new target groups like virologists during the COVID-19 pandemic), robustly detecting hate aimed at newly identified target groups is a highly relevant research question. We show that naively trained models suffer from a target group specific bias, which can be reduced via domain adaptation. We were able to achieve a relative improvement of the F1-score between 5.8% and 10.7% for out-of-domain target groups of hate speech compared to baseline approaches by utilizing domain adaptation.

Author contacts are given in the footnotes. [1]

1 Introduction

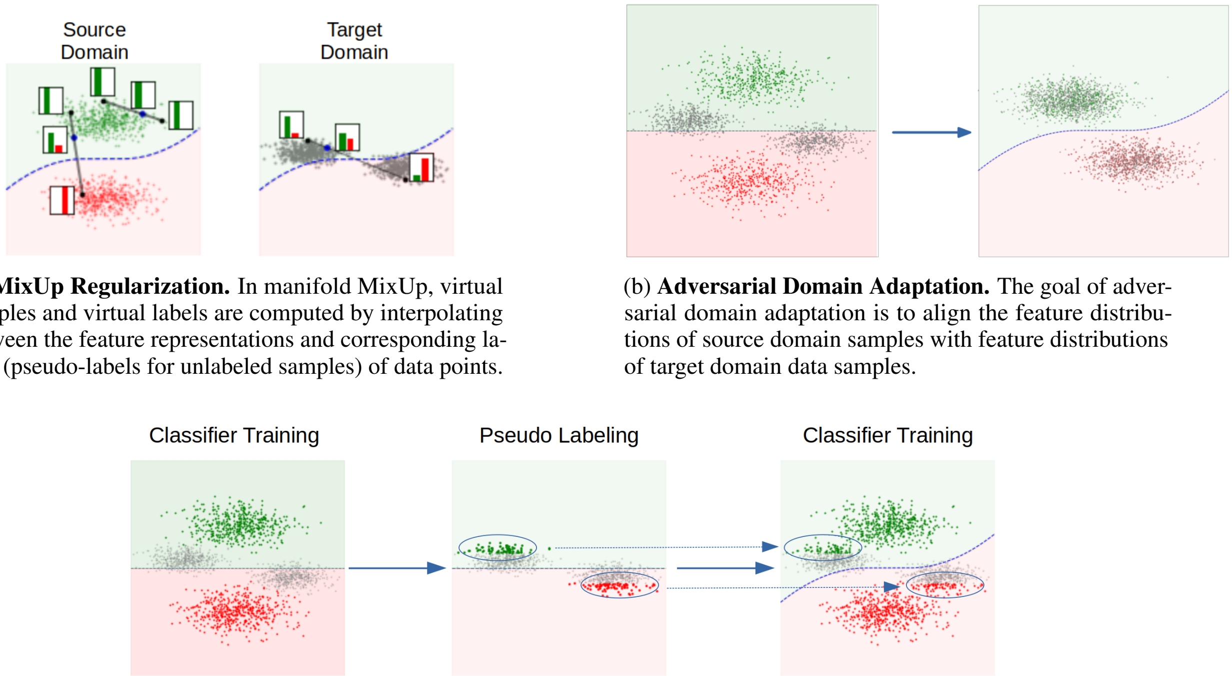
Current state-of-the-art machine learning approaches for hate speech detection reach F1-scores above 93% (Arango et al., 2019). Despite this progress, in some settings these scores drop to 50% when tested on out-of-domain data (Arango et al., 2019). The lack of generalization capabilities of hate speech detection systems hinders their suitability in real world applications.

Several challenges are faced when trying to generalize knowledge in hate speech detection tasks. Firstly, most benchmark hate speech datasets are focused on certain topics, such as hate speech directed at journalists (Charitidis et al., 2020), refugees and Muslims (Zhang et al., 2018), women only (Basile et al., 2019) or blacks, other races and women (Waseem, 2016). These datasets reflect biases towards different targets of hate, which will usually influence model training and predictive performance. Different target groups are also addressed by different perpetrators in the real world. For example, left-wing hate is frequently aimed against the ‘system’, with police or politicians being targeted, whereas right-wing hate is frequently aimed against Jews or foreigners. Moreover, new target groups can arise due to new phenomena such as the Corona pandemic (Fan et al., 2020). Therefore, being able to adapt models to unknown target groups of hate speech without the need of time consuming labeling of new datasets is crucial. Another challenge for generalizing knowledge across different hate speech datasets is the disagreement over the definition of hate speech (Ross et al., 2017), which is especially problematic for benchmark datasets. These disagreements lead to incompatible annotation of different datasets (MacAvaney et al., 2019; Fortuna et al., 2020), which hinders a proper assessment of the generalization capabilities of the models.

In this work we investigate the generalization and adaptation capabilities of hate speech classifiers to different domains of hate speech while eliminating errors due to incompatible datasets. This is done by conducting our experiments on a single dataset, namely the HateXplain dataset (Mathew et al., 2020), which was annotated following consistent annotation rules. There are many possibilities to categorize hate speech into different domains. For example, hate speech with common topics, hate speech that addresses common target groups, hate speech from common time periods or hate speech from common datasets can be considered as separate domains. In this work, we regard the adaptation capabilities of the

[1] torsten.zesch@fernuni-hagen.de
florian.ludwig@zitis.bund.de
eleanor.hobley@zitis.bund.de
klara.dolos@zitis.bund.de

Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), pages 29 - 39
July 14, 2022 ©2022 Association for Computational Linguistics



(a) **MixUp Regularization.** In manifold MixUp, virtual samples and virtual labels are computed by interpolating between the feature representations and corresponding labels (pseudo-labels for unlabeled samples) of data points.

(b) **Adversarial Domain Adaptation.** The goal of adversarial domain adaptation is to align the feature distributions of source domain samples with feature distributions of target domain data samples.

(c) **Curriculum Labeling.** After training a model on labeled samples (left), the model predicts pseudo labels (middle) for unlabeled samples from the target domain. Samples which belong to the most confident model predictions are included in the training set, together with their predicted class labels. Finally, the model is retrained from scratch on the augmented dataset (right).

Figure 1: Three different strategies for improving the generalization capabilities of models to different target groups are investigated. These approaches utilize labeled source data samples (colored data points) and unlabeled target domain data samples (grey data points).

models with respect to different target groups of hate speech due to the relevance of the topic for real world applications and the suitability of the HateXplain dataset for this research. An advantage of utilizing the HateXplain dataset for this research is, that target groups were annotated for all samples, not only the hateful ones, which allows us to appropriately select samples that correspond to different domains and therefore to properly investigate the generalization capabilities of our approaches. Adaptation of models to different target groups of hate speech is here investigated by unsupervised domain adaptation methods, namely via manifold MixUp regularization (Fig. 1a), adversarial domain adaptation (Fig. 1b) and curriculum labeling (Fig. 1c).

In summary, we make the following contributions:

- We analyze the influence of data and target group specific bias on hate speech classifiers;
- We investigate the suitability of unsupervised domain adaptation for improving model performances for out-of-domain target groups;
- Our experiments are conducted under clean conditions with properly separated domains and without data incompatibilities during model evaluation.

2 Related Work

Several approaches for machine learning based hate speech detection were investigated in recent years (Badjatiya et al., 2017; Djuric et al., 2015; Mozafari et al., 2019). An active line of research aims at improving generalization capabilities of hate speech detection systems, with most studies focusing on cross-dataset generalization capabilities of models (Bashar et al., 2021; Waseem et al., 2018).

Karan and Šnajder (2018) show the the difficulties of hate speech classifier to deal with out-domain datasets. The authors emphasize the importance of in-domain data for their generalization results. They integrated target domain data in their learning procedure using frustratingly easy domain adaptation Daumé III (2007). In contrast to our work, the authors investigated the cross-corpus generalization and adaptation capabilities of linear support vector machines. In this work, we focus on target group specific domain adaptation of deep learning based hate speech classifiers.

The generalization capabilities of deep learning models from topic generic to topic specific hate speech corpora were investigated by Chiril et al. (2021). The authors showed that models failed to generalize to domain specific corpora, but that the integration of domain specific knowledge improves the classification results in new domains. In contrast to our work, the authors focus on cross dataset generalization, which makes a clean evaluation of target group specific generalization difficult. Faal et al. (2021) suggested exploitation of multi-task learning and domain adaptation for improving the generalization capabilities of hate speech classifiers. After domain adaptive pre-training of a BERT based feature extractor (Devlin et al., 2019), the whole model was trained on multiple tasks by utilizing shared parameters as well as task specific parameters. The authors showed that the reduction of unintended target group specific model bias via multi-task learning successfully boosted generalization. In contrast to our work, they focus on general robustness with respect of target groups rather than a target group specific optimization.

In Bashar et al. (2021), the authors propose to train a language model to learn domain invariant and disentangled feature representation for different hate speech domains. After that, they trained a classifier on top of these feature representations and used it for robustly classifying hate speech from different domains. The authors demonstrated the success of the model in detecting hate speech related to the COVID-19 pandemic. On the other side, Bose et al. (2021) showed that the application of widely used unsupervised domain adaptation approaches can be problematic in the field of hate speech detection. The authors applied various pivot-based and adversarial-based approaches to generalize knowledge across different hate speech corpora. Unlike our work, which focuses on target group specific domain adaptation, these works focus on generalizing on knowledge on the level of hate speech corpora, which introduced previously discussed difficulties in model evaluation and which might be the main reason for bad adaptation results.

3 Methods and Experiments

In this section we describe the dataset, model architecture as well as the training and evaluation strategies used in our experiments.

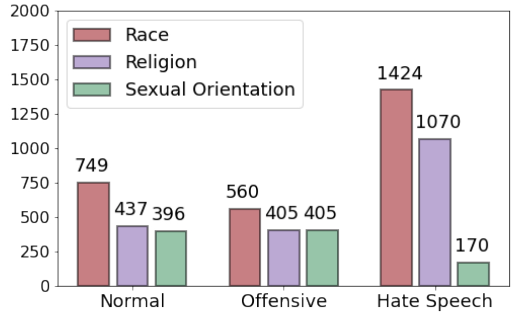


Figure 2: Number of experimental data samples by target group and and class label.

3.1 Dataset

The HateXplain dataset (Mathew et al., 2020), consisting of around $20K$ annotated posts, was used as the basis for all our experiments. The dataset was primarily annotated with the class labels $normal$, $offensive$ and $hateful$, with additional labeling of the target groups of hate ('Race', 'Religion', 'Sexual orientation', 'Gender', 'Origin', 'Other') undertaken. Unlike other datasets, such as (Del Vigna12 et al., 2017; Ousidhoum et al., 2019), each target group was annotated for all data points, including those belonging to the "normal" and "offensive" classes. This allows us to examine the generalization capabilities of different approaches with strictly separated target groups across all labels. To the best of our knowledge, the HateXplain dataset (Mathew et al., 2020) is the only dataset that explicitly annotates target groups for the classes "normal" and "offensive" as well, which is why this dataset is the only one that was used to conduct our experiments with strictly separated domains. To ensure that the trained models generalize from a single source domain to a single target domain, we select only those data points for training and validation purposes which have solely been annotated as belonging to either a source domain (e.g. "Race") or a target domain (e.g. "Religion"). We discard data points which have been annotated with multiple target groups (e.g. "Race" and "Gender"). We focus on the domains "Race," "Religion," and "Sexual Orientation" because the other target groups each contain fewer than 60 instances annotated as "Hate Speech," which risks inconsistent experimental results due to insufficient coverage of all class labels. Therefore, "Gender," "Origin," and "Other" are discarded, resulting in a final dataset that yields 170 to 1424 instances per class label (see Fig. 2). We also experiment with data augmentation. Recently, various techniques for text data augmentation have been proposed (Shorten et al., 2021), such as rule based techniques (Wei and Zou, 2019; Spasic et al., 2020; Karimi et al., 2021), feature space augmentations (Cheung and Yeung, 2020; Khosla et al., 2020) or neural augmentation (Wu et al., 2019). Due to the success of back translation based data augmentation (Xie et al., 2020; Yaseen and Langer, 2021; Corbeil and Abdi Ghadivel, 2020; Sugiyama and Yoshinaga, 2019), we decided to use this approach with pre-trained neural translation models (provided by HuggingFace [2]) in order to created an augmented version of the original HateXplain dataset. Back translation is done with the language pairs English - German, English - French and English - Spanish, resulting in nearly three times the number of instances per class.

3.2 Model Architecture and Training

In our experiments, we use the Structured Self-Attentive Sentence Embedding model (Lin et al., 2017), which provides a good trade-of between model performance and computational costs. The model is visualized in figure 3. The encoder of the model consists of a two layer bidirectional LSTM, followed by an attention module, as proposed by (Lin et al., 2017). The predictor of the model is a linear classifier, consisting of a single linear layer followed by a Softmax activation function. We use WordPiece tokenization (Devlin et al., 2018; Schuster and Nakajima, 2012). The embedding size and the hidden sizes of the LSTMs are 128, the dimension of the attention module 350, and the number of attention heads 30. A domain discriminator is applied in those experiments in which we perform adversarial domain alignment. The input of the domain discriminator is the output of the encoder of the Structured Self-Attentive Sentence Embedding model. The applied discriminator model consists of a gradient reversal layer (Ganin and Lempitsky, 2015), followed by a two layer feed forward neural network with a leaky ReLU activation function at the hidden position and a Sigmoid activation function at the output position.

We use the Adam optimizer (Kingma et al., 2015) with a learning rate of $5e^{-4}$ and beta values of $(0.9, 0.99)$ during our experiments. We apply dropout regularization (Srivastava et al., 2014) with a dropout probability of 0.6 to the LSTM modules to prevent overfitting. All models are trained for a total of 50.000 training iterations with a batch size of 32. Our experiments are implemented using the deep learning framework Pytorch.[3]

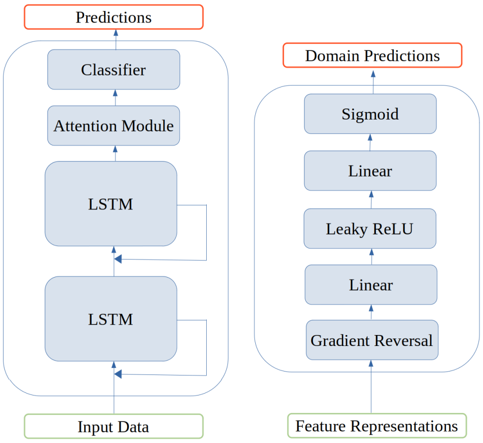


Figure 3: Structured Self-Attentive Sentence Embedding model (left) and domain discriminator (right), used in our experiments.

3.3 Model Assessment

All models are evaluated using the macro-average F1-score with five-fold cross-validation. During training, we store those model states for which the models achieved the best results on the out-of-fold validation data of the training domains. For these model states, we report the results achieved on the validation data of the other domains, too. We believe that this is a realistic scenario, since we assume that only unlabeled data samples of other domains are available and we therefore need to select our models based on their performances on the source domains, for which labeled data samples are available.

3.4 Zero-Shot Approaches

In the first set of experiments, we investigate the generalization capabilities of models that use only labeled data from one domain and no data from other domains. As we examine the generalization capabilities of the models to target domains that were not present in the training data, we refer to these experiments as zero-shot approaches. In the first, naive zero-shot approach (*Zero*), models are trained with labeled data from the original HateXplain dataset (Mathew et al., 2020) that belong to a specific target group (e.g. "Race"). The training batches provided to the model in each training itera-

[2] https://huggingface.co/

[3] https://pytorch.org/

tion are randomly sampled. In the second zero-shot approach (*Zero +*), models are trained with target group specific data from the augmented dataset (see Section 3.1). Similar to the first zero-shot learning approach, training batches are sampled randomly. In the last zero-shot learning approach (*Zero B+*), the training batches are sampled in a balanced manner from the augmented dataset with equal probability per class in each iteration. In all zero-shot approaches, models are evaluated with validation data from the original HateXplain dataset (Mathew et al., 2020), as described in Section 3.3.

3.5 Unsupervised Domain Adaptation

Unsupervised single source domain adaptation uses data from two different domains during the training: source data $X_S = \{(x_i, y_i)\}_{i=1}^{N}$, which consist of N labeled samples from a source domain $D_S = \{\mathcal{X}_S, P_S(X_S)\}$, and target data $X_T = \{x_j\}_{j=1}^{M}$, which consist of M unlabeled samples from a target domain $D_T = \{\mathcal{X}_T, P_T(X_T)\}$. $\mathcal{X} = \mathcal{X}_S = \mathcal{X}_T$ is a shared feature space, $P_S(X_S) \neq P_T(X_T)$ are marginal probability distributions over the feature space, which are similar, but differ. The goal of the learning algorithm is to train a model which achieves a strong performance for a task T on the target domain although no labeled data points from the target domain are available during the training. For the domain adaptation approaches, we use the same data and sampling strategy as for the last zero-shot learning approach (Zero B+).

The goal in our paper is to cover different research directions in the field of domain adaptation for hate speech detection purposes. The approaches investigated in this paper are typical candidates for their line of research, which consider the problem of domain adaptation from a regularization-based view 3.5.1, a data-based view 3.5.2 and a feature-based view 3.5.3.

3.5.1 MixUp Regularization

We adapt the approach of manifold MixUp regularization proposed by Verma et al. (2019) (Fig. 1a). Given is a deep neural network with an encoder e, which maps an input $x \in \mathcal{X}$ into hidden representation $h \in \mathbb{R}^m$, and a predictor p, which computes predictions $z \in \mathbb{R}^K$ based on the hidden representation $h \in \mathbb{R}^m$. Manifold MixUp regularization introduces an additional regularization loss based on MixUp feature representations $\tilde{h} \in \mathbb{R}^m$ and MixUp labels $\tilde{y} \in \mathbb{R}^K$, which are computed based on hidden representations $h_1, h_2 \in \mathbb{R}^m$ and corresponding labels $y_1, y_2 \in \mathbb{R}^K$ of two samples:

$$\tilde{h} = \alpha \cdot h_1 + (1 - \alpha) \cdot h_2 \tag{1}$$

$$\tilde{y} = \alpha \cdot y_1 + (1 - \alpha) \cdot y_2 \tag{2}$$

Here, $\alpha \in [0, 1]$ is sampled from a Beta distribution: $\alpha \sim Beta(2, 2)$. y_1 and y_2 are represented as one-hot encoded class labels for source domain samples and as soft pseudo-labels, which are iteratively computed by the neural network, for target domain samples.

The MixUp features are used for computing the MixUp predictions $\tilde{z} = p(\tilde{h})$ based on the predictor of the neural network. The loss between MixUp predictions and MixUp labels is computed as Cross-Entropy loss for source domain samples ($\tilde{l}_m^s$) and L1 loss for target domain samples ($\tilde{l}_m^t$). The complete MixUp loss $\tilde{l}_m$ is computed as follows:

$$l_m = \lambda^s \cdot \tilde{l}_m^s + \lambda^t \cdot \tilde{l}_m^t \tag{3}$$

We set $\lambda^s = \lambda^t = 0.1$ in our experiments.

3.5.2 Curriculum Labeling

In addition to MixUp regularization, we adapt the approach of Cascante-Bonilla et al. (2020), which combines pseudo-labeling with curriculum learning, for domain adaptation purposes (Fig. 1c). Curriculum labeling is done by selecting data points from an unlabeled data pool based on the network's prediction confidences. The selected data points with corresponding pseudo-labels are iteratively included to the training procedure during the learning epochs. Following Cascante-Bonilla et al. (2020), we select data points based on percentile scores of the prediction confidences. During the training, the percentile threshold for selecting samples corresponding to the most confident predictions is increased from 0% to 100% in increments of 20%. In contrast to Cascante-Bonilla et al. (2020), we select the pseudo-labeled samples based on the model's prediction confidences independently for each predicted class. This is done to prevent a bias towards the selection of data points from majority classes, which is crucial for hate speech detection tasks. During each iteration ('curriculum epoch'), the network is re-trained from scratch with both the labeled samples and the pseudo-labeled samples selected by the model trained in the previous curriculum epoch.

					Target Domain								
Source Domain	Eval. Domain				Race			Rel.			Sex.		
		Zero	Zero +	Zero B+	Mix.	Adv.	Cur.	Mix.	Adv.	Cur.	Mix.	Adv.	Cur.
Race	Race	.56	.57	**.58**				**.58**	.57	.57	**.58**	.57	**.58**
	Rel.	.52	.52	.52		-		**.54**	**.54**	.51	.52	.53	.53
	Sex.	.50	.50	.51				**.53**	**.53**	.50	.51	.52	**.53**
Rel.	Race	.46	.46	.48	.48	**.52**	**.52**				.48	.49	.50
	Rel.	.47	.48	.50	.50	**.52**	**.52**		-		.49	.50	.50
	Sex.	.47	.48	.49	.50	**.52**	.51				.48	.49	.51
Sex.	Race	.36	.37	.42	.42	**.48**	.47	.42	.39	.42			
	Rel.	.39	.39	.43	.44	.46	**.47**	.43	.43	.45		-	
	Sex.	.42	.42	.46	.46	.49	**.50**	.45	.45	.48			

Table 1: Macro average F1 scores achieved by the approaches, averaged over five validation folds and split into target groups, approaches and domains. Improvements over the results, achieved by the best zero-shot approach are marked in green. Violet indicates negative transfer, in which the models achieved worse results than the naive zero-shot learning approach.

Approach	**Source**	**Target**	**Other**
Zero	0.483	0.450	
Zero +	0.490 *+1.4%*	0.455 *+1.1%*	
Zero B+	0.513 *+6.2%*	0.475 *+5.6%*	
MixUp	0.510 *+5.6%*	0.476 *+5.8%*	0.481 *+6.9%*
Adv.	0.517 *+7.0%*	0.496 *+10.2%*	0.487 *+8.2%*
Cur.	**0.525** ***+8.7%***	**0.498** ***+10.7%***	**0.488** ***+8.4%***

Table 2: Average F1-Scores of the investigated approaches and relative improvements compared to the naive zero-shot learning approach with respect to the domains.

Approach	**Normal**	**Offensive**	**Hate Speech**
Zero	0.479	0.287	0.618
Zero +	0.476 *-0.6%*	0.286 *-0.3%*	0.637 *+3.1%*
Zero B+	0.494 *+3.1%*	0.288 *+0.35%*	0.683 *+10.5%*
MixUp	0.490 *+2.1%*	0.281 *-2.1%*	0.690 *+11.7%*
Adv.	0.500 *+4.4%*	**0.312** ***+8.7%***	0.689 *+11.7%*
Cur.	**0.505** ***+6.5%***	0.311 *+8.4%*	**0.692** ***+12.0%***

Table 3: Average F1-Scores and its relative improvements over naive zero-shot learning, divided into approaches and classes labels.

3.5.3 Adversarial Domain Alignment

In order to learn domain invariant feature representations, Ganin et al. (2016) introduced Domain Adversarial Neural Networks (Fig. 1b). Beside the main model, a domain discriminator $D : \mathbb{R}^m \mapsto \mathbb{R}$ is trained to distinguish between feature representations $h^s \in \mathbb{R}^m$ and $h^t \in \mathbb{R}^m$ for source domain samples x^s and target domain samples x^t, computed by encoder e. At the same time, the encoder e is trained to confuse the domain discriminator D, such that the discriminator is not able to distinguish between these feature representations. To achieve this, an adversarial loss is introduced:

$$\mathcal{L}_{adv} = \mathop{\mathbb{E}}_{x^s \sim X^s}[log(D(e(x^s)))] + \mathop{\mathbb{E}}_{x^t \sim X^t}[log(1 - D(e(x^t)))] \quad (4)$$

The domain discriminator D is trained to maximize the adversarial loss $\mathcal{L}_{adv}$, while at the same time the encoder e is trained to fool the discriminator and therefore minimize $\mathcal{L}_{adv}$. The theoretical equilibrium is reached when the encoder e produces features which cannot be reliably classified as belonging either to the source or to the target domain by an optimal discriminator.

4 Results and Discussion

In this section, we present and discuss the results of our experiments. In table 1, we present macro average F1-scores, achieved by the investigated approaches. The scores are divided into source domain (first column), the domain on which the models were evaluated (second column) and the approaches used. Since the investigated domain adaptation approaches, unlike zero-shot approaches, used unlabeled target domain data in addition to labeled source domain data, their results are further subdivided into the target domain that was involved in model training. In table 2 we present the average F1-scores of the investigated approaches, split by source, target and uninvolved domain. In addition to the average values, relative improvements compared to the naive zero-shot learning approach are also given. Table 3 shows the achieved performances with respect to the class labels "normal", "offensive" and "hate speech". Again, we report the relative improvements of the approaches compared to the naive zero-shot approach. In table 4, we provide feature visualizations of our models for hate related samples based on lime (Ribeiro et al., 2016). The visualizations are provided for different approaches and combinations of source target and evaluation domains.

OFFENSIVE CONTENT WARNING: The following sections contain examples of hateful content. This is strictly for the purpose of enabling this research. Please be aware that this content could be offensive and cause you distress.

4.1 Model Bias

Although there are cases, in which models show poor generalization abilities to some out-of-domain target groups, all of our models were able to generalize knowledge to other domains to some extent. Best or equal best model performance was achieved when evaluating models against the domain on which they were trained (i.e. source domain) for both the zero-shot approaches (Table 1) as well as after averaging across domain adaptation approaches (Table 2). Data augmentation generally helped to improve the model performances, which shows that the models suffer from a bias due to the low amount of available training data. On average, the class "Hate Speech" benefits most from data augmentation (Table 3), while the performance on the classes "Normal" and "Offensive" is slightly worse compared to the naive zero-shot approach. Models additionally benefit from class balanced data sampling (Zero B+), which on average outperformed the other zero-shot learning approaches on all domains and across all class labels. Despite the improvements due to data augmentation and class balanced sampling, a gap between the performances on source domain and the other domains is still preserved. Moreover, both techniques slightly increased the performance gap between source and other domains (Table 2). We conclude that the models suffer from a target group specific bias, which occurs due to the lack of domain specific knowledge of domains other than the source domain and which cannot be reduced via typical data sampling strategies or data augmentation techniques such as back-translation. Therefore, the integration of domain specific knowledge of the target domain is crucial for reducing the performance gap between source domain and target domain.

Domain adaptation enabled models to learn domain specific features and to produce better model predictions in some instances (see samples a and b, e - h, i and l, Table 4). These results concur with the findings of (Karan and Šnajder (2018)), who highlighted the importance of in-domain knowledge for the generalization capabilities of hate speech classifiers and showed that models trained on domain general hate speech corpora failed to generalize to domain specific hate speech data Chiril et al. (2021). In contrast to our work, the authors investigated the generalization capabilities of models across different datasets, which hinders a clean evaluation due to possible dataset incompatibilities. Our results indicate that the integration of domain specific knowledge via adversarial domain adaptation and curriculum labeling is able to reduce the performance gap and achieve the largest improvements on the target domains, thereby reducing model bias (Table 2).

4.2 Effects of Domain Adaption Approaches on Model Performance

Domain adaptation was able to improve the performance over all zero-shot learning approaches for each combination of source and target domain (Table 1). Additionally, all domain adaptation approaches achieved better results on average than the zero-shot learning approaches for not only the target domain but also for the third, uninvolved domain (Table 2) and was observable for all three classes (Table 3). Moreover, the adversarial domain adaptation approach and the curriculum labeling approach on average outperformed the zero-shot learning approaches on the source domain. However, in our experiments the choice of the data included in the training procedure had greater impact to the model performances than the chosen domain adaptation approaches themselves. Furthermore, none of our investigated domain adaptation

MixUp Regularization

	Text	Source	Target	Evaluation	Prediction
a)	this little k**e is a good little donkey american	Race	Rel.	Rel.	Hate Speech
b)	|user| good job scamming you fu***ng fa**ot	Race	Rel.	Sex.	Hate Speech
c)	y' all n***as is hoes	Sex.	Race	Race	Offensive
d)	are yall n***as retarded	Sex.	Race	Race	Offensive

Adversarial Domain Adaptation

	Text	Source	Target	Eval	Prediction
e)	enjoy being a childless n****r fu**er	Rel.	Race	Race	Hate Speech
f)	deport that be**er sp**k to mexico	Sex.	Race	Race	Hate Speech
g)	like pick a struggle fa**ot	Rel.	Race	Sex.	Hate Speech
h)	kill yourself fa**ot	Rel.	Race	Sex.	Hate Speech

Curriculum Labeling

	Text	Source	Target	Eval	Prediction
i)	started as trash ending as worthless n****r trash	Rel.	Race	Race	Hate Speech
j)	those fa**ots off queer eye hate jews	Rel.	Race	Sex.	Hate Speech
k)	i fu***ng hate jews	Race	Rel.	Rel.	Offensive
l)	kill yourself fa**ot	Rel.	Race	Sex.	Hate Speech

Table 4: **Feature visualization for hate related samples.** Words highlighted in red strengthen the model to predict the class "Hate Speech", while words which are highlighted in blue, decrease this prediction confidence.

approaches outperformed the other methods in each experiment (Table 1), which makes the choice of the appropriate approach in practical settings difficult, especially when no labeled data of the target domain is available to asses the model performance on that domain.

While the two approaches curriculum labeling and adversarial domain adaptation both performed similarly, they outperform MixUp regularization in most cases. Adversarial domain adaptation improved the performances in 4 out of 6 domain combinations on the target domain, and in 5 out of 6 combinations on the uninvolved domain. Curriculum labeling resulted in better performances on the target domain in 5 out of 6 training domain pairs, and in 4 out of 6 cases on the uninvolved domain. In contrast, MixUp regularization improved performances on the target domain in only 1 out of 6 source-target domain combinations, namely "Race"-"Religion", and yielded the smallest improvements in average model performance of all three domain adaptation approaches (Table 2). Moreover, MixUp regularization was not able to correctly learn domain specific features, such as the domain specific word "n***as" for its predictions (see samples c) and d), Table 4). Thus, MixUp regularization is inferior to the other approaches for the investigated task.

Remarkably, the curriculum labeling approach resulted in worse outcomes than the zero-shot approaches in one instance (Table 1), although the risk of predicting incorrect pseudo-labels is mitigated by implementing the curriculum steps proposed in (Cascante-Bonilla et al., 2020). This performance loss or negative transfer is indicated by the phrase "hate jews" leading to a decrease in the prediction confidences of the models for the hate speech class and, in case of sample k), an incorrect prediction (samples j & k, Table 4). This negative transfer is attributable to a negative confirmation bias, which can occur in pseudo-labeling based approaches (Rizve et al., 2021) and can lead to a large number of incorrect pseudo-labels that interfere with the training procedure and thus affect the model performance. Nevertheless, the curriculum labeling approach proved to be best suitable to adapt hate speech classifiers in our study, achieving the best averaged results on the source, target and other domains.

4.3 Data Dependency of the Performance

In our experiments, the choice of the training data had the greatest impact on the model performances. Models trained on the source domain "Race" yielded the best results in general with F1 scores ranging from .50 and .58. Models, trained on the source domain "Sexual Orientation" performed worst overall and achieved F1 scores between .36 and .50. Models, trained on the source domain "Religion" achieved F1 scores between .46 and .52. A similar pattern was observed for unlabeled data from the target domain. The largest improvements via domain adaptation were achieved by utilizing unlabeled data from the target domain "Race", whereas utilizing unlabeled data from the target domain "Sexual Orientation" yielded the lowest improvements. We attribute these observations to the number of training samples available in each class. The best performances were achieved with the largest amount of labeled training data (source domain "Race"), the worst performances were achieved with the lowest amount of labeled training data (source domain "Sexual Orientation"). Additionally, the largest improvements were achieved by incorporating the largest amount of unlabeled data (target domain "Race"), the smallest improvements were achieved with the lowest amount of unlabeled data (target domain "Sexual Orientation"). Since the domain adaptation performance on the target domain depends on the performance achieved on the source domain (Zhang and Harada, 2019), this observation also holds true for the investigated domain adaptation approaches.

5 Conclusion

The goal of this work was to analyze the generalization capabilities of hate speech classifiers to different target groups of hate under clean experimental conditions. Furthermore, we aimed to investigate the suitability of unsupervised domain adaptation to improve these generalization capabilities. Our results indicate that naively trained hate speech classifiers suffer from a target group specific bias and that unsupervised domain adaptation is able to improve the generalization capabilities of models across different target groups of hate. In contrast to previous works, which mainly focus on the generalization capabilities of hate speech classifiers in cross dataset settings, we investigated the generalization capabilities of hate speech classifiers to new hate targets on a single dataset, the HateXplain dataset. This enabled us to strictly separate target groups across all class labels and therefore allowed a clean analysis of the abilities of models to generalize to different target groups of hate, while avoiding the risk of inconsistencies over the definition of hate speech between datasets. We observed a gap of the model performances on the source domains and the model performances on the target domains. While data augmentation and balanced data sampling was able to generally improve the model performances, these methods tend to preserve these gaps. The integration of domain specific knowledge via domain adaptation was able to improve the generalization capabilities of models to other target groups, whereby the number of the involved labeled and unlabeled training samples strongly influenced the results of the approaches. However, our study does not allow a clear conclusion about which domain adaptation approach is best in which constellation of available data, which makes the choice of the appropriate approach difficult in real world situations. In total, there is still potential to improve the prediction quality of the models, especially when it comes to real world applications. Failures to detect hate speech, which contain threats, may lead to life-threatening situations for people, for example. In such scenarios, the achieved model performances are not good enough to reliably support law enforcement agencies. Improvements could be made with more advanced model architectures and a larger amount of available training data, which is a limitation of our work. We also analyzed generalization capabilities for only three target groups of hate, namely "race," "religion," and "sexual orientation." These limitations should be addressed in future works, for which we suggest investigating the generalization capabilities to new targets of hate in settings with a greater amount of data, higher diversity of target groups and with more advanced models like transformer based models. Moreover, the limitations of each of the domain adaptation methods can be further investigated in order gain insight into when and why some methods might fail.

Acknowledgements

This work was funded by the German ministry of education and research (BMBF) within the framework program Research for Civil Security of the German Federal Government (KISTRA, 13N15337).

References

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*, pages 45–54.

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*, pages 759–760.

Md Abul Bashar, Richi Nayak, Khanh Luong, and Thirunavukarasu Balasubramaniam. 2021. Progressive domain adaptation for detecting hate speech on social media with small training set and its application to covid-19 concerned posts. *Social Network Analysis and Mining*, 11(1):1–18.

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

Tulika Bose, Irina Illina, and Dominique Fohr. 2021. Unsupervised domain adaptation in cross-corpora abusive language detection. In *SocialNLP 2021-The 9th International Workshop on Natural Language Processing for Social Media*.

Paola Cascante-Bonilla, Fuwen Tan, Yanjun Qi, and Vicente Ordonez. 2020. Curriculum labeling: Self-paced pseudo-labeling for semi-supervised learning. *arXiv e-prints*, pages arXiv–2001.

Polychronis Charitidis, Stavros Doropoulos, Stavros Vologiannidis, Ioannis Papastergiou, and Sophia Karakeva. 2020. Towards countering hate speech against journalists on social media. *Online Social Networks and Media*, 17:100071.

Tsz-Him Cheung and Dit-Yan Yeung. 2020. Modals: Modality-agnostic automated data augmentation in the latent space. In *International Conference on Learning Representations*.

Patricia Chiril, Endang Wahyu Pamungkas, Farah Benamara, Véronique Moriceau, and Viviana Patti. 2021. Emotionally informed hate speech detection: a multi-target perspective. *Cognitive Computation*, pages 1–31.

Jean-Philippe Corbeil and Hadi Abdi Ghadivel. 2020. Bet: A backtranslation approach for easy data augmentation in transformer-based paraphrase identification context. *arXiv e-prints*, pages arXiv–2009.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263.

Fabio Del Vigna12, Andrea Cimino23, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017. Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the First Italian Conference on Cybersecurity (ITASEC17)*, pages 86–95.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

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.

Nemanja Djuric, Jing Zhou, Robin Morris, Mihajlo Grbovic, Vladan Radosavljevic, and Narayan Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30.

Farshid Faal, Jia Yuan Yu, and Ketra Schmitt. 2021. Domain adaptation multi-task deep neural network for mitigating unintended bias in toxic language detection. In *ICAART (2)*, pages 932–940.

Lizhou Fan, Huizi Yu, and Zhanyuan Yin. 2020. Stigmatization in social media: Documenting and analyzing hate speech for covid-19 on twitter. *Proceedings of the Association for Information Science and Technology*, 57(1):e313.

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 6786–6794.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *International conference on machine learning*, pages 1180–1189. PMLR.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. 2016. Domain-adversarial training of neural networks. *The journal of machine learning research*, 17(1):2096–2030.

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.

Akbar Karimi, Leonardo Rossi, and Andrea Prati. 2021. Aeda: An easier data augmentation technique for text classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2748–2754.

Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. *Advances in Neural Information Processing Systems*, 33:18661–18673.

DP Kingma, LJ Ba, et al. 2015. Adam: A method for stochastic optimization.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*.

Sean MacAvaney, Hao-Ren Yao, Eugene Yang, Katina Russell, Nazli Goharian, and Ophir Frieder. 2019. Hate speech detection: Challenges and solutions. *PloS one*, 14(8):e0221152.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2020. Hatexplain: A benchmark dataset for explainable hate speech detection. *arXiv preprint arXiv:2012.10289*.

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.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. 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)*, pages 4675–4684.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Mamshad Nayeem Rizve, Kevin Duarte, Yogesh S Rawat, and Mubarak Shah. 2021. In defense of pseudo-labeling: An uncertainty-aware pseudo-label selection framework for semi-supervised learning. *arXiv preprint arXiv:2101.06329*.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2017. Measuring the reliability of hate speech annotations: The case of the european refugee crisis. *arXiv preprint arXiv:1701.08118*.

Mike Schuster and Kaisuke Nakajima. 2012. Japanese and korean voice search. In *2012 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5149–5152. IEEE.

Connor Shorten, Taghi M Khoshgoftaar, and Borko Furht. 2021. Text data augmentation for deep learning. *Journal of big Data*, 8(1):1–34.

Irena Spasic, Goran Nenadic, et al. 2020. Clinical text data in machine learning: systematic review. *JMIR medical informatics*, 8(3):e17984.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Amane Sugiyama and Naoki Yoshinaga. 2019. Data augmentation using back-translation for context-aware neural machine translation. In *Proceedings of the Fourth Workshop on Discourse in Machine Translation (DiscoMT 2019)*, pages 35–44.

Vikas Verma, Alex Lamb, Christopher Beckham, Amir Najafi, Ioannis Mitliagkas, David Lopez-Paz, and Yoshua Bengio. 2019. Manifold mixup: Better representations by interpolating hidden states. In *International Conference on Machine Learning*, pages 6438–6447. PMLR.

Zeerak Waseem. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

Zeerak Waseem, James Thorne, and Joachim Bingel. 2018. Bridging the gaps: Multi task learning for domain transfer of hate speech detection. In *Online harassment*, pages 29–55. Springer.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. 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)*, pages 6382–6388.

Xing Wu, Shangwen Lv, Liangjun Zang, Jizhong Han, and Songlin Hu. 2019. Conditional bert contextual augmentation. In *International Conference on Computational Science*, pages 84–95. Springer.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in Neural Information Processing Systems*, 33:6256–6268.

Usama Yaseen and Stefan Langer. 2021. Data augmentation for low-resource named entity recognition using backtranslation. *arXiv e-prints*, pages arXiv–2108.

Dexuan Zhang and Tatsuya Harada. 2019. A general upper bound for unsupervised domain adaptation. *arXiv preprint arXiv:1910.01409*.

Ziqi Zhang, David Robinson, and Jonathan Tepper. 2018. Detecting hate speech on twitter using a convolution-gru based deep neural network. In *European semantic web conference*, pages 745–760. Springer.