Unsupervised Domain Adaptation in Cross-corpora Abusive Language Detection

Tulika Bose, Irina Illina, Dominique Fohr
Universite de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France
tulika.bose, illina, dominique.fohr@loria.fr

Abstract

The state-of-the-art abusive language detection models report great in-corpus performance, but underperform when evaluated on abusive comments that differ from the training scenario. As human annotation involves substantial time and effort, models that can adapt to newly collected comments can prove to be useful. In this paper, we investigate the effectiveness of several Unsupervised Domain Adaptation (UDA) approaches for the task of cross-corpora abusive language detection. In comparison, we adapt a variant of the BERT model, trained on large-scale abusive comments, using Masked Language Model (MLM) fine-tuning. Our evaluation shows that the UDA approaches result in sub-optimal performance, while the MLM fine-tuning does better in the cross-corpora setting. Detailed analysis reveals the limitations of the UDA approaches and emphasizes the need to build efficient adaptation methods for this task.

1 Introduction

Social networking platforms have been used as a medium for expressing opinions, ideas, and feelings. This has resulted in a serious concern of *abusive* language, which is commonly described as hurtful, obscene, or toxic towards an individual or a group sharing common societal characteristics such as race, religion, gender, etc. The huge amount of comments generated every day on these platforms make it increasingly infeasible for manual moderators to review every comment for its abusive content. As such, automated abuse detection mechanisms are employed to assist moderators. We consider the variations of online abuse, toxicity, hate speech, and offensive language as abusive language and this work addresses the detection of abusive versus non-abusive comments.

Supervised classification approaches for abuse detection require a large amount of expensive annotated data (Lee et al., 2018). Moreover, models already trained on the available annotated corpus report degraded performance on new content (Yin and Zubiaga, 2021; Swamy et al., 2019; Wiegand et al., 2019). This is due to phenomena like change of topics discussed in social media, and differences across corpora, such as varying sampling strategies, targets of abuse, abusive language forms, etc. These call for approaches that can adapt to newly seen content out of the original training corpus. Annotating such content is non-trivial and may require substantial time and effort (Poletto et al., 2019; Ombui et al., 2019). Thus, Unsupervised Domain Adaptation (UDA) methods that can adapt without the target domain labels (Ramponi and Plank, 2020), turn out to be attractive in this task. Given an automatic text classification or tagging task, such as abusive language detection, a corpus with coherence can be considered a domain (Ramponi and Plank, 2020; Plank, 2011). Under this condition, domain adaptation approaches can be applied in cross-corpora evaluation setups. This motivates us to explore UDA for cross-corpora abusive language detection.

A task related to abuse detection is sentiment classification (Bauwelinck and Lefever, 2019; Rajamanickam et al., 2020), and it involves an extensive body of work on domain adaptation. In this work, we analyze if the problem of cross-corpora abusive language detection can be addressed by the existing advancements in domain adaptation. Alongside different UDA approaches, we also evaluate the effectiveness of recently proposed HateBERT model (Caselli et al., 2021) that has fine-tuned BERT (Devlin et al., 2019) on a large corpus of abusive language from Reddit using the Masked Language Model (MLM) objective. Furthermore, we perform the MLM fine-tuning of HateBERT on target corpus, which can be considered a form of unsupervised adaptation. Our contribution is summarised below:

- We investigate some of the best perform-

Proceedings of the Ninth International Workshop on Natural Language Processing for Social Media, pages 113–122
Online Workshop, June 10, 2021.

ing UDA approaches, originally proposed for cross-domain sentiment classification, and analyze their performance on the task of cross-corpora abusive language detection. We provide some insights on the sub-optimal performance of these approaches. To the best of our knowledge, this is the first work that analyzes UDA approaches for cross-corpora abuse detection.

- We analyze the performance of HateBERT in our cross-corpora evaluation set-up. In particular, we use the Masked Language Model (MLM) objective to further fine-tune HateBERT over the unlabeled target corpus, and subsequently perform supervised fine-tuning over the source corpus.

The remaining of this paper is structured as follows: Section 2 discusses the shifts across different abusive corpora. Section 3 surveys some recently proposed UDA models for sentiment classification and discusses the main differences in the approaches. Section 4 presents the experimental settings used in our evaluation. The results of our evaluation and a discussion on performances of different approaches are present in Section 5. Finally, Section 6 concludes the paper and highlights some future work.

2 Shifts in Abusive Language Corpora

Saha and Sindhwani (2012) have detailed the problem of changing topics in social media with time. Hence, temporal or contextual shifts are commonly witnessed across different abusive corpora. For example, the datasets by Waseem and Hovy (2016); Basile et al. (2019) were collected in or before 2016, and during 2018, respectively, and also involve different contexts of discussion.

Moreover, sampling strategies across datasets also introduce bias in the data (Wiegand et al., 2019), and could be a cause for differences across datasets. For instance, Davidson et al. (2017) sample tweets containing keywords from a hate speech lexicon, which has resulted in the corpus having a major proportion (83%) of abusive content. As mentioned by Waseem et al. (2018), tweets in Davidson et al. (2017) originate from the United States, whereas Waseem and Hovy (2016) sample them without such a demographic constraint.

Apart from sampling differences, the targets and types of abuse may vary across datasets. For instance, even though women are targeted both in Waseem and Hovy (2016) and Davidson et al. (2017), the former involves more subtle and implicit forms of abuse, while the the latter involves explicit abuse involving profane words. Besides, religious minorities are the other targeted groups in Waseem and Hovy (2016), while African Americans are targeted in Davidson et al. (2017). Owing to these differences across corpora, abusive language detection in a cross-corpora setting remains a challenge. This has been empirically validated by Wiegand et al. (2019); Arango et al. (2019); Swamy et al. (2019); Karan and Šnajder (2018) with performance degradation across the cross-corpora evaluation settings. Thus, it can be concluded that the different collection time frames, sampling strategies, and targets of abuse would induce a shift in the data.

3 Unsupervised Domain Adaptation

As discussed by Ramponi and Plank (2020); Plank (2011), a coherent type of corpus can typically be considered a domain for tasks such as automatic text classification. We, therefore, decide to apply domain adaptation methods for our task of cross-corpora abuse detection. Besides, UDA methods aim to adapt a classifier learned on the source domain D_S to the target domain D_T, where only the unlabeled target domain samples X_T and the labeled source domain samples X_S are assumed to be available. We denote the source labels by Y_S. In this work, we use the unlabeled samples X_T for adaptation and evaluate the performance over the remaining unseen target samples from D_T.

3.1 Survey of UDA Approaches

There is a vast body of research on UDA for the related task of cross-domain sentiment classification. Amongst them, the feature-centric approaches typically construct an aligned feature space either using pivot features (Blitzer et al., 2006) or using Autoencoders (Glorot et al., 2011; Chen et al., 2012). Besides these, domain adversarial training is used widely as a loss-centric approach to maximize the confusion in domain identification and align the source and target representations (Ganin et al., 2016; Ganin and Lempitsky, 2015). Owing to their success in cross-domain sentiment classification, we decide to apply the following pivot-based and domain-adversarial UDA approaches to the task of cross-corpora abusive language detection.

Pivot-based approaches: Following Blitzer et al. (2006), pivot-based approaches extract a set of common shared features, called pivots, across domains that are (i) frequent in X_S and X_T; and (ii) highly correlated with Y_S. *Pivot Based Language Modeling* (**PBLM**) (Ziser and Reichart, 2018) has outperformed the Autoencoder based pivot prediction (Ziser and Reichart, 2017). It performs representation learning by employing a Long Short-Term Memory (LSTM) based language model to predict the pivots using other non-pivots features in the input samples from both X_S and X_T. Convolutional Neural Networks (CNN) and LSTM based classifiers are subsequently employed for the final supervised training with X_S and Y_S. *Pivot-based Encoder Representation of Language* (**PERL**) (Ben-David et al., 2020), a recently proposed UDA model, integrates BERT (Devlin et al., 2019) with pivot-based fine-tuning using the MLM objective. It involves prediction of the masked unigram/ bigram pivots from the non-pivots of the input samples from both X_S and X_T. This is followed by supervised task training with a convolution, average pooling and a linear layer over the encoded representations of the input samples from X_S. During the supervised task training, the encoder weights are kept frozen. Both PBLM and PERL use unigrams and bi-grams as pivots, although higher order n-grams can also be used.

Domain adversarial approaches: *Hierarchical Attention Transfer Network* (**HATN**) (Li et al., 2017, 2018) employs the domain classification based adversarial training using X_S and X_T, along with an attention mechanism using X_S and Y_S to automate the pivot construction. The Gradient Reversal Layer (GRL) (Ganin and Lempitsky, 2015) is used in the adversarial training to ensure that the learned pivots are domain-shared, and the attention mechanism ensures that they are useful for the end task. During training, the pivots are predicted using the non-pivots while jointly performing the domain adversarial training, and the supervised end-task training. Recently BERT-based approaches for UDA are proposed by Du et al. (2020); Ryu and Lee (2020) that also apply the domain adversarial training. *Adversarial Adaptation with Distillation* (**AAD**) (Ryu and Lee, 2020) is such a domain adversarial approach that is applied over BERT. Unlike HATN, in AAD, the domain adversarial training is done with the framework of the Adversarial Discriminative Domain Adaptation (ADDA) (Tzeng et al., 2017), using X_S and X_T. This aims to make the source and target representations similar. Moreover, it leverages knowledge distillation (Hinton et al., 2015) as an additional loss function during adaptation.

3.2 Adaptation through Masked Language Model Fine-tuning with HateBERT

Rietzler et al. (2020); Xu et al. (2019) show that the language model fine-tuning of BERT (using the MLM and the Next Sentence Prediction task) results in incorporating domain-specific knowledge into the model and is useful for cross-domain adaptation. This step does not require task-specific labels. The recently proposed HateBERT model (Caselli et al., 2021) extends the pre-trained BERT model using the MLM objective over a large corpus of unlabeled abusive comments from Reddit. This is expected to shift the pre-trained BERT model towards abusive language. It is shown by Caselli et al. (2021) that HateBERT is more portable across abusive language datasets, as compared to BERT. We, thus, decide to perform further analysis over HateBERT for our task.

In particular, we begin with the HateBERT model and perform MLM fine-tuning incorporating the unlabeled train set from the target corpus. We hypothesize that performing this step should incorporate the variations in the abusive language present in the target corpus into the model. For the classification task, supervised fine-tuning is performed over the MLM fine-tuned model obtained from the previous step, using X_S and Y_S.

4 Experimental Setup

4.1 Data Description and Pre-processing

Datasets	Number of comments		Average comment length	Abuse %
	Train	Test		
Davidson	19817	2477	14.1	83.2
Waseem	8720	1090	14.7	26.8
HatEval	9000	3000	21.3	42.1

.

Table 1: Statistics of the datasets used (average comment length is reported in terms of word numbers).

We present experiments over three different publicly available abusive language corpora from Twitter as they cover different forms of abuse, namely

Davidson (Davidson et al., 2017),*Waseem* (Waseem and Hovy, 2016) and *HatEval* (Basile et al., 2019). Following the precedent of other works on cross-corpora abuse detection (Wiegand et al., 2019; Swamy et al., 2019; Karan and Šnajder, 2018), we target a binary classification task with classes: *abusive* and *non-abusive*. We randomly split *Davidson* and *Waseem* into train (80%), development (10%), and test (10%), whereas in the case of *HatEval*, we use the standard partition of the shared task. Statistics of the train-test splits of these datasets are listed in Table 1.

During pre-processing, we remove the URLs and retain the frequently occurring Twitter handles (user names) present in the datasets, as they could provide important information.[1] The words contained in hashtags are split using the tool CrazyTokenizer[2] and the words are converted into lowercase.

4.2 Evaluation Setup

Given the three corpora listed above, we experiment with all the six pairs of X_S and X_T for our cross-corpora analysis. The UDA approaches leverage the respective unlabeled train sets in D_T for adaptation, along with the train sets in D_S. The abusive language classifier is subsequently trained on the labeled train set in D_S and evaluated on the test set in D_T. In the "no adaptation" case, the HateBERT model is fine-tuned in a supervised manner on the labeled source corpus train set, and evaluated on the target test set. Unsupervised adaptation using HateBERT involves training of the HateBERT model on the target corpus train set using the MLM objective. This is followed by a supervised fine-tuning on the source corpus train set.

We use the original implementations of the UDA models[3] and the pre-trained HateBERT[4] model for our experiments. We select the best model checkpoints by performing early-stopping of the training while evaluating the performance on the respective development sets in D_S. FastText[5] word vectors, pre-trained over Wikipedia, are used for word embedding initialization for both HATN and PBLM. PERL and AAD are initialized with the BERT base-uncased model.[6] In PBLM, we employ the LSTM based classifier.[7] For both PERL and PBLM, words with the highest mutual information with respect to the source labels and occurring at least 10 times in both the source and target corpora are considered as pivots (Ziser and Reichart, 2018).

5 Results and Analysis

Dataset	Macro F1	Frequent words in abusive comments
Davidson	93.8±0.1	b*tch, h*e, f*ck, p*ssy, n*gga, ass, f*ck, shit
Waseem	85.5±0.4	#notsexist, #mkr, female, girl, kat, men, woman, feminist
HatEval	51.9±1.7	woman, refugee, immigrant, trump, #buildthatwall, illegal, b*tch, f*ck

Table 2: F1 macro-average (mean ± std-dev) for in-corpus classification using supervised fine-tuning of HateBERT.

Our evaluation reports the mean and standard deviation of macro averaged F1 scores, obtained by an approach, over five runs with different random initializations. We first present the in-corpus performance of the HateBERT model in Table 2, obtained after supervised fine-tuning on the respective datasets, along with the frequent abuse-related words. As shown in Table 2, the in-corpus performance is high for *Davidson* and *Waseem*, but not for *HatEval*. *HatEval* shared task presents a challenging test set and similar performance have been reported in prior work (Caselli et al., 2021). Cross-corpora performance of HateBERT and the UDA models discussed in Section 3.1, is presented in Table 3. Comparing Table 2 and Table 3, substantial degradation of performance is observed across the datasets in the cross-corpora setting. This highlights the challenge of cross-corpora performance in abusive language detection.

Cross-corpora evaluation in Table 3 shows that all the UDA methods experience drop in average performance when compared to the no-adaptation

[1] Eg., the Twitter handle @realDonaldTrump.

[2] https://redditscore.readthedocs.io/en/master/tokenizing.html

[3] PBLM: https://github.com/yftah89/PBLM-Domain-Adaptation, HATN: https://github.com/hsqmlzno1/HATN, PERL: https://github.com/eyalbd2/PERL, AAD: https://github.com/bzantium/bert-AAD

[4] https://osf.io/tbd58/

[5] https://fasttext.cc/

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

[7] CNN classifier obtained similar performance.

Source →Target	No-adaptation	Unsupervised Domain Adaptation				
	HateBERT supervised fine-tune only	**HateBERT MLM fine-tune on Target**	**PBLM**	**PERL-BERT**	**HATN**	**AAD-BERT**
Hat →Was	66.4±1.1	**68.0**±1.0	57.5±3.4	57.1±1.8	57.3±1.7	60.4±7.8
Was →Hat	**57.8**±0.6	56.5±1.1	51.0±5.2	55.3±0.7	53.5±0.4	55.7±1.3
Dav →Was	**67.5**±0.5	66.7±0.8	57.2±4.8	67.4±1.0	57.5±6.7	41.5±2.8
Was →Dav	60.1±4.4	**67.1**±2.9	46.5±1.3	48.3±1.5	28.0±2.3	35.6±3.7
Hat →Dav	63.8±2.3	**67.8**±1.6	61.8±5.7	62.6±3.8	61.5±5.8	55.2±0.7
Dav →Hat	51.3±0.2	**51.4**±0.4	49.9±0.2	50.3±0.9	50.3±0.5	50.4±3.0
Average	61.2	**62.9**	54.0	56.8	51.4	49.8

Table 3: Macro average F1 scores (mean±std-dev) on different source and target pairs for cross-corpora abuse detection (Hat : HatEval, Was : Waseem, Dav : Davidson). The best in each row is marked in bold.

case of supervised fine-tuning of HateBERT. However, the additional step of MLM fine-tuning of HateBERT on the unlabeled train set from target corpus results in an improved performance in most of the cases. In the following sub-sections, we perform a detailed analysis to get further insights into the sub-optimal performance of the UDA approaches for our task.

5.1 Pivot Characteristics in Pivot-based Approaches

To understand the performance of the pivot-based models, we probe the characteristics of the pivots used by these models as they control the transfer of information across source and target corpora. As mentioned in Section 3.1, one of the criteria for pivot selection is their affinity to the available labels. Accordingly, if the adaptation results in better performance, a higher proportion of pivots would have more affinity to one of the two classes. In the following, we aim to study this particular characteristic across the source train set and the target test set. To compute class affinities, we obtain a ratio of the class membership of every pivot p_i:

$$r_i = \frac{\text{\#abusive comments with } p_i}{\text{\#non-abusive comments with } p_i} \quad (1)$$

The ratios obtained for the train set of the source and the test set of the target, for the pivot p_i, are denoted as ${r_i}^s$ and ${r_i}^t$, respectively. A pivot p_i with similar class affinities in both the source train and target test should satisfy:

$$({r_i}^s, {r_i}^t) < 1 - th \text{ or } ({r_i}^s, {r_i}^t) > 1 + th \quad (2)$$

Here, th denotes the threshold. Ratios less than $(1-th)$ indicate affinity towards non-abusive class, while those greater than $(1+th)$ indicate affinity towards the abusive class. For every source →target pair, we select the pivots that satisfy Equation (2) with threshold th = 0.3, and calculate the percentage of the selected pivots as:

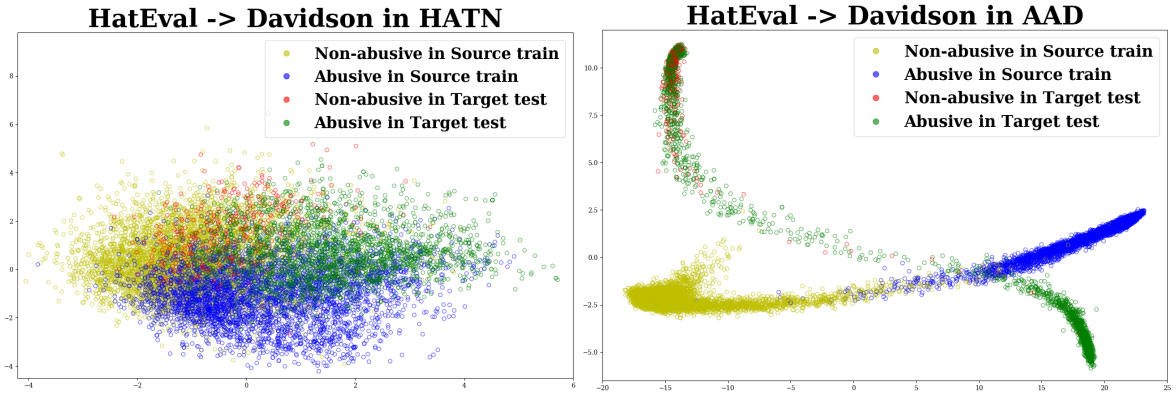
$$\text{perc}_{s \to t} = \frac{\text{\#pivots satisfying Equation (2)}}{\text{\#Total pivots}} \times 100 \quad (3)$$

This indicates the percentage of pivots having similar affinity towards one of the two classes. We now analyze this percentage in the best and the worst case scenarios of PBLM.[8]

Worst cases: For the worst case of *Waseem →Davidson*, Equation (3) yields a low $\text{perc}_{s \to t}$ of 18.8%. This indicates that the percentage of pivots having similar class affinities, across the source and the target, remains low in the worst performing pair.

Best case: The best case in PBLM corresponds to *HatEval →Davidson*. In this case, Equation (3) yields a relatively higher $\text{perc}_{s \to t}$ of 51.4%. This is because the pivots extracted in this case involve a lot of profane words. Since in *Davidson*, the majority of abusive content involves the use of profane words (as also reflected in Table 2), the pivots extracted by PBLM can represent the target corpus well in this case.

[8]Pivot extraction criteria are same for PBLM and PERL and similar percentages are expected with PERL.



(a) (b)

Figure 1: (Best viewed in color) PCA based visualization of *HatEval →Davidson* in the adversarial approaches.

5.2 Domain Adversarial Approaches

On an average, the adversarial approach of HATN performs slightly better than AAD. In order to analyze the difference, we investigate the representation spaces of the two approaches for the best case of HATN i.e. *HatEval →Davidson*. To this end, we apply the Principal Component Analysis (PCA) to obtain the two-dimensional visualization of the feature spaces from the train set of the source corpus *HatEval* and the test set of the target corpus *Davidson*. The PCA plots are shown in Figure 1. Adversarial training in both the HATN and AAD models tends to bring the representation regions of the source and target corpora close to each other. At the same time, separation of abusive and non-abusive classes in source train set seems to be happening in both the models. However, in the representation space of AAD, samples corresponding to abusive and non-abusive classes in the target test set do not follow the class separation seen in the source train set. But in the representation space of HATN, samples in the target test set appear to follow the class separation exhibited by its source train set. Considering the abusive class as positive, this is reflected in the higher number of *True Positives* in HATN as compared to that of AAD for this pair (#TP for HATN: 1393, #TP for AAD: 1105), while the *True Negatives* remain almost the same (#TN for HATN: 370, #TN for AAD: 373).

One of the limitations of these domain adversarial approaches is the class-agnostic alignment of the common source-target representation space. As discussed in Saito et al. (2018), methods that do not consider the class boundary information while aligning the source and target distributions, often result in having ambiguous and non-discriminative target domain features near class boundaries. Besides, such an alignment can be achieved without having access to the target domain class labels (Saito et al., 2018). As such, an effective alignment should also attempt to minimize the intra-class, and maximize the inter-class domain discrepancy (Kang et al., 2019).

5.3 MLM Fine-tuning of HateBERT

It is evident from Table 3 that the MLM fine-tuning of HateBERT, before the subsequent supervised fine-tuning over the source corpus, results in improved performance in majority of the cases. We investigated the MLM fine-tuning over different combinations of the source and target corpora, in order to identify the best configuration. These include: a combination of the train sets from all the three corpora, combining the source and target train sets, and using only the target train set. Table 4 shows that MLM fine-tuning over only the unlabeled target corpus results in the best overall performance. This is in agreement to Rietzler et al. (2020) who observe a better capture of domain-specific knowledge with fine-tuning only on the target domain.

5.4 Bridging the Gap between PERL and HateBERT MLM Fine-tuning

Since PERL originally incorporates BERT, Table 3 reports the performance of PERL initialized with the pre-trained BERT model. As discussed in Section 3.1, PERL applies MLM fine-tuning over the pre-trained BERT model, where only the pivots are predicted rather than all the masked tokens. Following Ben-David et al. (2020), after the encoder weights are learned during the MLM fine-tuning

Source →Target	HBERT MLM on all 3 corpora	HBERT MLM on Source + Target	HBERT MLM on Target
Hat →Was	**69.7**±0.8	68.9±0.6	68.0±1.0
Was →Hat	**57.2**±1.4	56.8±1.1	56.5±1.1
Dav →Was	60.2±0.7	58.8±0.8	**66.7**±0.8
Was →Dav	63.4±3.9	63.4±3.9	**67.1**±2.9
Hat →Dav	66.6±1.1	66.7±2.1	**67.8**±1.6
Dav →Hat	51.4±0.2	**51.5**±0.1	51.4±0.4
Average	61.4	61.0	**62.9**

Table 4: Macro average F1 scores (mean ± std-dev) for Masked Language Model fine-tuning of HateBERT (HBERT MLM) over different corpora combinations, before supervised fine-tuning on source; Hat : HatEval, Was : Waseem, Dav : Davidson. The best in each row is marked in bold.

Source →Target	PERL-BERT (frozen encoder layers)	PERL-HBERT (frozen encoder layers)	PERL-HBERT (with layer updates)
Hat →Was	57.1±1.8	63.2±1.7	**68.3**±0.8
Was →Hat	55.3±0.7	55.0±0.9	**57.8**±0.8
Dav →Was	**67.4**±1.0	65.9±1.3	57.3±3.1
Was →Dav	48.3±1.5	48.1±3.7	**64.4**±2.1
Hat →Dav	62.6±3.8	63.6±0.9	**66.1**±1.8
Dav →Hat	50.3±0.9	50.4±0.6	**51.1**±0.3
Average	56.8	57.7	**60.8**

Table 5: Macro average F1 scores (mean ± std-dev) of PERL initialized with BERT and HateBERT (HBERT) with frozen encoder layers, and PERL initialized with HateBERT with updates across all layers, for all the pairs (Hat : HatEval, Was : Waseem, Dav : Davidson). The best in each row is marked in bold.

step of PERL, they are kept frozen during supervised training for the classification task. As an additional verification, we try to leverage the HateBERT model for initializing PERL in the same way as BERT is used in the original PERL model, with frozen encoder layers. As shown in Table 5, this does not result in substantial performance gains over PERL-BERT on average. As a further extension, we update all the layers in PERL during the supervised training step and use the same hyperparameters as those used for HateBERT (Caselli et al., 2021).[9] This results in improved performance from PERL. However, it stills remains behind the best performing HateBERT model with MLM fine-tuning on target.

5.5 Source Corpora Specific Behaviour

In general, when models are trained over *HatEval*, they are found to be more robust towards addressing the shifts across corpora. One of the primary reasons is that *HatEval* captures wider forms of abuse directed towards both immigrants and women. The most frequent words in Table 2 also highlight the same. The corpus involves a mix of implicit as well as explicit abusive language.

On the contrary, models trained over *Waseem* are generally unable to adapt well in cross-corpora settings. Since only tweet IDs were made available in *Waseem*, we observe that our crawled comments in this dataset rarely involve abuse directed towards target groups other than women (99.3% of the abusive comments are sexist and 0.6% racist). This is because majority of these comments have been removed before crawling. Besides, *Waseem* mostly involves subtle and implicit abuse, and less use of profane words.

6 Conclusion and Future Work

This work analyzed the efficacy of some successful Unsupervised Domain Adaptation approaches of cross-domain sentiment classification in cross-corpora abuse detection. Our experiments highlighted some of the problems with these approaches that render them sub-optimal in the cross-corpora abuse detection task. While the extraction of pivots, in the pivot-based models, is not optimal enough to capture the shared space across domains, the domain adversarial methods underperform substantially. The analysis of the Masked Language Model fine-tuning of HateBERT on the target corpus displayed improvements in general as compared to only fine-tuning HateBERT over the source corpus, suggesting that it helps in adapting the model towards target-specific language variations. The overall performance of all the approaches, however, indicates that building robust and portable abuse detection models is a challenging problem, far from being solved.

Future work along the lines of domain adversarial training should explore methods which learn

[9] Note that the ablation study in Ben-David et al. (2020) discusses the effect of the number of unfrozen encoder layers only in the MLM fine-tuning step, but not in the supervised training step for the end task.

class boundaries that generalize well to the target corpora while performing alignment of the source and target representation spaces. Such an alignment can be performed without target class labels by minimizing the intra-class domain discrepancy (Kang et al., 2019). Pivot-based approaches should explore pivot extraction methods that account for higher-level semantics of abusive language across source and target corpora.

Acknowledgements

This work was supported partly by the french PIA project "Lorraine Université d'Excellence", reference ANR-15-IDEX-04-LUE.

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*, SIGIR'19, page 45–54, New York, NY, USA. Association for Computing Machinery.

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.

Nina Bauwelinck and Els Lefever. 2019. Measuring the impact of sentiment for hate speech detection on twitter. In *Proceedings of HUSO 2019, The fifth international conference on human and social analytics*, pages 17–22. IARIA, International Academy, Research, and Industry Association.

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*, 8:504–5221.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 120–128, Sydney, Australia. Association for Computational Linguistics.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. Hatebert: Retraining bert for abusive language detection in english. *arXiv preprint arXiv:2010.12472*.

Minmin Chen, Zhixiang Xu, Kilian Q. Weinberger, and Fei Sha. 2012. Marginalized denoising autoencoders for domain adaptation. In *Proceedings of the 29th International Conference on Machine Learning*, ICML'12, page 1627–1634, Madison, WI, USA. Omnipress.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17, pages 512–515.

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.

Chunning Du, Haifeng Sun, Jingyu Wang, Qi Qi, and Jianxin Liao. 2020. Adversarial and domain-aware BERT for cross-domain sentiment analysis. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4019–4028, Online. Association for Computational Linguistics.

Yaroslav Ganin and Victor Lempitsky. 2015. Unsupervised domain adaptation by backpropagation. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1180–1189, Lille, France. 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. *J. Mach. Learn. Res.*, 17(1):2096–2030.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *Proceedings of the 28th International Conference on Machine Learning*, ICML'11, page 513–520, Madison, WI, USA. Omnipress.

Geoffrey Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*.

Guoliang Kang, Lu Jiang, Yi Yang, and Alexander G Hauptmann. 2019. Contrastive adaptation network for unsupervised domain adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4893–4902.

Mladen Karan and Jan Šnajder. 2018. Cross-domain detection of abusive language online. In *Proceed-*

ings of the 2nd Workshop on Abusive Language Online (ALW2), pages 132–137, Brussels, Belgium. Association for Computational Linguistics.

Younghun Lee, Seunghyun Yoon, and Kyomin Jung. 2018. Comparative studies of detecting abusive language on twitter. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 101–106, Brussels, Belgium. Association for Computational Linguistics.

Zheng Li, Ying Wei, Yu Zhang, and Qiang Yang. 2018. Hierarchical attention transfer network for cross-domain sentiment classification. In *AAAI Conference on Artificial Intelligence*.

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 International Joint Conference on Artificial Intelligence (IJCAI 2017)*.

Edward Ombui, Moses Karani, and Lawrence Muchemi. 2019. Annotation framework for hate speech identification in tweets: Case study of tweets during kenyan elections. *2019 IST-Africa Week Conference (IST-Africa)*, pages 1–9.

Barbara Plank. 2011. *Domain adaptation for parsing*. Ph.D. thesis, University of Groningen.

Fabio Poletto, Valerio Basile, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2019. Annotating hate speech: Three schemes at comparison. In *Proceedings of the Sixth Italian Conference on Computational Linguistics, Bari, Italy, November 13-15, 2019*, volume 2481 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Santhosh Rajamanickam, Pushkar Mishra, Helen Yannakoudakis, and Ekaterina Shutova. 2020. Joint modelling of emotion and abusive language detection. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4270–4279, Online. Association for Computational Linguistics.

Alan Ramponi and Barbara Plank. 2020. Neural unsupervised domain adaptation in NLP—A survey. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6838–6855, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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.

Minho Ryu and Kichun Lee. 2020. Knowledge distillation for bert unsupervised domain adaptation. *arXiv preprint arXiv:2010.11478*.

Ankan Saha and Vikas Sindhwani. 2012. Learning evolving and emerging topics in social media: A dynamic nmf approach with temporal regularization. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, page 693–702, New York, NY, USA. Association for Computing Machinery.

Kuniaki Saito, Kohei Watanabe, Yoshitaka Ushiku, and Tatsuya Harada. 2018. Maximum classifier discrepancy for unsupervised domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3723–3732.

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.

E. Tzeng, J. Hoffman, K. Saenko, and T. Darrell. 2017. Adversarial discriminative domain adaptation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2962–2971.

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.

Zeerak Waseem, James Thorne, and Joachim Bingel. 2018. Bridging the gaps: Multi task learning for domain transfer of hate speech detection. In *Golbeck J. (eds) Online Harassment. Human–Computer Interaction Series*, pages 29–55, Cham. Springer International Publishing.

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.

Hu Xu, Bing Liu, Lei Shu, and Philip Yu. 2019. BERT post-training for review reading comprehension and aspect-based sentiment analysis. 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 2324–2335, Minneapolis, Minnesota. Association for Computational Linguistics.

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

Yftah Ziser and Roi Reichart. 2017. Neural structural correspondence learning for domain adaptation. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 400–410, Vancouver, Canada. Association for Computational Linguistics.

Yftah Ziser and Roi Reichart. 2018. Pivot based language modeling for improved neural domain adaptation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1241–1251, New Orleans, Louisiana. Association for Computational Linguistics.