Cross-Lingual Transfer Learning for Hate Speech Detection

Irina Bigoulaeva and **Viktor Hangya** and **Alexander Fraser**
Center for Information and Language Processing
LMU Munich
I.Bigoulaeva1@campus.lmu.de, {hangyav,fraser}@cis.lmu.de

Abstract

We address the task of automatic hate speech detection for low-resource languages. Rather than collecting and annotating new hate speech data, we show how to use cross-lingual transfer learning to leverage already existing data from higher-resource languages. Using bilingual word embeddings based classifiers we achieve good performance on the target language by training only on the source dataset. Using our transferred system we bootstrap on unlabeled target language data, improving the performance of standard cross-lingual transfer approaches. We use English as a high-resource language and German as the target language for which only a small amount of annotated corpora are available. Our results indicate that cross-lingual transfer learning together with our approach to leverage additional unlabeled data is an effective way of achieving good performance on low-resource target languages without the need for any target language annotations.

1 Introduction

Due to the increased digitization of society, the impact of online discourse on everyday life is becoming more pronounced. A single hateful message shared on social media now has the potential to incite violent offline movements, as well as exert a negative emotional impact on millions of readers. For this reason, platforms such as Twitter and Facebook have created community policies to ensure civil conduct on the part of their users. The goal is to filter hate speech, which unlike mere offensive or vulgar content, is exclusively designed to attack or denigrate entire groups of people and is generally agreed to add no constructive value to discussions. But with the sheer amount of posts being published, it is becoming difficult for humans to moderate them in a complete and timely manner. Different annotators are also not guaranteed to agree on every classification, even in the presence of well-defined annotation guidelines. Moreover, due to their repeated and prolonged exposure to negative content, many moderators experience a decline in mental health (Vidgen and Derczynski, 2020). For these reasons, automatic hate speech detection has become a field of high interest.

In general, the task of classifying hate speech has been acknowledged as difficult (de Gibert et al., 2018). One reason is data scarcity: there are currently few public hate speech datasets available, and the majority of them are for English. Thus, building systems for lower-resource languages is even more challenging (Vidgen and Derczynski, 2020). An additional difficulty of the task is the need to precisely define hate speech. While many people have an intuitive understanding of what hate speech is, this does not easily translate to a finite set of characteristics. Different hate speech datasets often deal with specific hate speech subtypes, leading to stark differences between the content of their hate speech classes and making the available resources for a given set of hate speech subtypes in a low-resource language even scarcer.

The goal of this paper is to reduce these difficulties by exploiting available resources from other languages. We address data scarcity in German, a generally high-resource language but a language for which there are not yet many hate speech datasets (only 4 labeled corpora compared to the 25 available for English (Vidgen and Derczynski, 2020)). Our method is a zero-shot setup that assumes we do not have any annotated training data in German. We develop a cross-lingual transfer learning approach based on bilingual word embeddings (BWEs) and neural classifiers to provide access to hate speech data in English. We use the English dataset of de Gibert et al. (2018) and the German dataset of the GermEval Shared Task on the Identification of Offensive Language (Ruppenhofer et al.,

Proceedings of the First Workshop on Language Technology for Equality, Diversity and Inclusion, pages 15–25
April 19, 2021.

2018) in our experiments. However, their annotation had to be modified using a few simple rules to ensure label compatibility.

In addition to training only on English, we bootstrap on two unlabeled German datasets, one of which we crawled from the web. Using a set of our English-only models we predict the labels of previously-unseen data and keep only those on which the used models agree. By using the newly-labeled data as well for training, we further improved the performance on the target language.

2 Previous Work

2.1 What is Hate Speech?

Hate speech detection is considered a subtype of offensive language detection. However, Davidson et al. (2017) note that the two concepts are often wrongly conflated due to a lack of a precise definition of hate speech. This is an undesirable strategy, given that social media is saturated with low-severity insults and comments that use rude language in a non-hostile way. Moreover, the absence of offensive words and slurs does not guarantee the absence of hate speech.

Multiple datasets have been created that feature hate speech as a distinct class from other forms of offensive language. However, these datasets nearly all differ in how they conceptualize hate speech. The dataset of Bretschneider and Peters (2017) focuses on hate speech in the sense of anti-foreigner comments. The dataset of Ross et al. (2016) contains hate speech exclusively against refugees. The three datasets of Majumder et al. (2019) do not focus on one particular target.

These discrepancies across hate speech datasets lead to an even smaller amount of available resources for building a system for hate speech detection.

Perhaps the most informative and broad definition of hate speech is given by de Gibert et al. (2018) who define it as "a deliberate attack directed towards a specific group of people motivated by aspects of the group's identity". With this definition in mind, de Gibert et al. (2018) present an English-language hate speech dataset featuring text scraped from the white-nationalist forum Stormfront. This dataset will be referred to as the Stormfront dataset. Due to the broad nature of its hate speech definition and its decent size (ca. 10,000 examples), it was chosen as the training set for this paper.

Table 1 illustrates some 'Hate' and 'noHate' sentences from the Stormfront dataset.[1] Sentences 2 and 4 are examples of racial hate speech, and Sentence 5 is an example of hate speech related to gender. Sentence 1 is not an example of hate speech, since it has a neutral sentiment and does not ascribe the qualities 'poor' and 'victimized' to an entire group of people. Finally, Sentence 3 expresses the writer's approval of a video and does not attack or mention any group.

In comparison to other languages, English has by far the largest amount of annotated hate speech data. A comprehensive online catalogue published by Vidgen and Derczynski (2020)[2] currently catalogues 25 datasets for English, with a grand total of 800,000 data samples. Other languages are left far behind: the next-best represented language, Arabic, has just over 48,000 total examples. German has four datasets totalling around 19,500 samples, found in: Bretschneider and Peters (2017), Ross et al. (2016), Wiegand et al. (2018b), and Majumder et al. (2019).

Of particular interest is the dataset of German-language tweets presented with the 2018 GermEval Shared Task on the Identification of Offensive Language (Ruppenhofer et al., 2018). The shared task focused on the detection of offensive language in general (the "coarse-grained task"), along with the detection of three of its subtypes (the "fine-grained task"): 'Insult', 'Profanity', and 'Abuse'. A tweet is assigned the 'Abuse' label if "... the target of judgment is seen as a representative of a group and it is ascribed negative qualities that are taken to be universal, omnipresent and unchangeable characteristics of the group" (Ruppenhofer et al., 2018). Importantly, this definition keeps the nature of the target group general and is therefore compatible with the hate speech definition of de Gibert et al. (2018). Thus, the GermEval dataset was used as our test set for evaluation.

2.2 Low-Resource Approaches

SVMs, CNNs and RNNs are the most commonly-used models for hate speech detection and offensive language detection in general (Waseem et al., 2017; Fišer et al., 2018; Roberts et al., 2019; Ruppenhofer et al., 2018). Two architectures from the 2018 GermEval task are particularly relevant to this paper. Xi et al. (2018) used a CNN following (Kim, 2014) to address the coarse-grained task. Wiedemann

[1] The data samples in this paper are shown for explanatory purposes and do not represent the views of the authors.

[2] https://hatespeechdata.com

	Sentence	Label
1.	This film tells the story of a poor victimised African boy (Joseph) who was allowed into Ireland .	noHate
2.	Whites in all their glory against impossibly large hordes of non-whites .	Hate
3.	The video about the “Day of honour”, is really cool .	noHate
4.	But unfortunately , Maine ’s become the dumping ground for Somalis and other African trash .	Hate
5.	Most White Western Women have the racial sense of a cabbage .	Hate

Table 1: Sample ‘Hate’ and ‘noHate’ comments according to de Gibert et al. (12).

et al. (2018) used a model combining CNN and BiLSTMs to achieve second-best performance in the coarse-grained task and top performance in the fine-grained task. The classifiers we use are based on these two approaches.

Data scarcity remains a relevant problem for hate speech detection, and recent work still aims to utilize resources from English to improve in languages with less hate speech data. Ranasinghe and Zampieri (2020) train transformer-based architectures on English data and use the learned weights to initialize models which are trained on target language data. Stappen et al. (2020) also utilize transformer architectures, training them on the source language and evaluating on the target language. Alternatively, they concatenate a small number of target language samples with the source language training data. Wiegand et al. (2018a) also augment their available German training set, using English-German BWEs to create a concatenated dataset with both the English and German training data. Mathur et al. (2018) utilize a cross-lingual transfer procedure for hate speech detection in Hinglish, a code-switched language that uses both Hindi and English words. By first training a CNN and an LSTM on an English dataset, then fine-tuning the models on Hinglish, better performance was achieved compared to a Hinglish-only model. In contrast to these works, our approach requires no target language annotations.

Kozareva (2006) present a bootstrapping-based approach that annotates new data for named entity recognition to improve the performance in low-resource scenarios. First a set of classifiers are trained, which are then applied to a unlabeled set with majority voting. The extended corpus is used to improve the performance by retraining the models from scratch. Rather than bootstrapping new target language data, Ghadery and Moens (2020) fine-tune an mBERT model (Devlin et al., 2019) on an augmented multilingual dataset constructed automatically by translating the English source dataset. Our approach does not require the additional resources of automatic translation, however.

We combine the bootstrapping procedure of Kozareva (2006) with the fine-tuning procedure of Mathur et al. (2018), first bootstrapping German-language data then using it to fine-tune our models. Variations of this bootstrapping procedure have been used for other tasks such as named entity recognition, using active learning instead of only source language trained models to annotate new data (Chaudhary et al., 2019). To our knowledge, we are the first to utilize such a procedure for detecting hate speech.

3 Experiment Setup

This section introduces the setup of our experiments. First the source and target datasets are discussed, along with the ways in which they were modified for maximum compatibility. This is followed by a presentation of our CNN and CNN/BiLSTM model architectures.

3.1 Datasets

The English and German datasets that we used in our experiments are listed in Table 2 and 3 respectively, along with their class distributions.

While the Stormfront and GermEval datasets have similar hate speech definitions, their annotation schemas differed. The Stormfront dataset features binary ‘Hate’ vs. ‘noHate’ labeling, along with a ‘Relation’ label for sentences that had to be considered in context with others to acquire a hateful meaning, and a ‘Skip’ label for when the sentence was either non-English or not meaningful enough to be given either of the binary labels. In contrast, the GermEval dataset features a two-tiered annotation schema: each tweet carries a coarse-grained label (‘Offense’ vs ‘Other’) as well as a

	noHate	Hate
Stormfront	9,580	1,364

Table 2: Ratio of hate to non-hate labels in the English Stormfront dataset.

	Other	Abuse	Insult	Prof.
Train	3,321	1,022	595	71
Test	2,330	773	381	48

Table 3: Distribution of the fine-grained class labels of the original German GermEval training and test datasets.

fine-grained label that specifies the subtype of offensiveness: either 'Insult', 'Profanity', or 'Abuse'. To ensure optimal cross-lingual transfer between these two datasets, we made certain modifications to their labeling schemas that were motivated by the datasets' specific class definitions.

First we simplified the annotation schema of the fine-grained GermEval data into a binary schema. As per the discussion in Section 2.1, we took GermEval's 'Abuse' label to be the counterpart of the Stormfront dataset's 'Hate', relabeling comments belonging to the 'Other', 'Insult', and 'Profanity' classes as 'noHate'.

Next, we relabeled all 'Skip' and 'Relation' samples from the Stormfront dataset to conform with the binary schema. The 92 comments that carried the label 'Skip', indicating that they were either non-English or not informative, were relabeled as 'noHate'. The 168 instances of the 'Relation' class were relabeled as 'Hate', since these sentences were always hateful when placed in context.

After relabeling was completed, we addressed the imbalanced class distributions of the English and German datasets. Examining Table 2, it is clear that the majority of samples in the Stormfront dataset are 'noHate'. This reflects the real-life observation that hate speech is less common than regular text. But this has been known to pose difficulties for machine learning models, which need plenty of data from both classes in order to be able to generalize (Madukwe et al., 2020; Vidgen and Derczynski, 2020). In our initial experiments, the 'noHate' class was so dominant over 'Hate' that the models were skewed towards assigning the 'noHate' label every time. Previous research suggests that oversampling the underrepresented class yields good model performance (De Smedt and Jaki, 2018). For this reason we oversampled the Stormfront dataset with multiple copies of the 'Hate' examples, yielding a nearly-balanced distribution that resulted in optimal performance for our models. This oversampled version of the Stormfront dataset will be referred to as EN-OS, and is the English dataset used for training.

	noHate	Hate
EN-OS	9,580	9,548
DE-REL	3,345	855
DE-DEV	642	167
DE-TEST	2,759	773

Table 4: Datasets used for our experiments.

We split the official GermEval training set into train and dev sets following the work of Wiedemann et al. (2018). First, we transferred the final 809 samples from the full set to a new development set named DE-DEV for hyperparameter tuning. The remaining samples formed the DE-REL dataset, which will be used in the fine-tuning experiments in Section 4.2. We did not use this data anywhere else. See Table 4.

3.2 Model Architectures

All our models described below are based on BWEs. BWEs are one of the most commonly-used tools for conducting cross-lingual transfer, the other being machine translation of source language data. However, ensuring accurate machine translation requires large amounts of resources, making BWEs more interesting in general for low-resource languages. BWEs ensure that source- and target language words are represented in a common vector space, enabling neural models based on BWEs to train on the source language and test on the target language without any intermediate steps.

To produce our bilingual English-German embeddings, monolingual embeddings were first trained using FastText SkipGram (Bojanowski et al., 2017) over English and German NewsCrawl corpora (Bojar et al., 2015) which contain text dating from 2007 to 2013 and were preprocessed with Moses tools (Koehn et al., 2007). The resulting embeddings were mapped with MUSE (Conneau et al., 2018). We used the default parameters of the above mentioned tools.

Our baseline classifier was a linear SVM over these BWEs. For each training example, we calculated the average of the vector representations of the words in the given text to obtain its global

Model	Accuracy	Hate			noHate			Macro-Average		
		P	R	F1	P	R	F1	P	R	F1
SVM	47.65	22.24	55.76	31.80	78.54	45.38	57.52	50.39	50.57	50.48
CNN1	59.17	22.15	34.41	26.95	78.25	66.11	71.67	50.20	50.26	50.23
CNN2	61.04	22.47	31.82	26.34	78.38	69.23	73.52	50.42	50.53	50.47
BiLSTM1	63.22	21.90	26.52	23.99	78.12	73.50	75.74	50.01	50.01	50.01
BiLSTM2	72.88	21.71	9.18	12.91	78.10	90.72	83.94	49.90	49.95	49.93

Table 5: Cross-lingual performance on DE-TEST after training on EN-OS.

representation. Wiegand et al. (2018a) showed that SVMs based on feature engineering, e.g., using word and character n-grams or lexicon based features, perform well on hate-speech detection. However, such systems are not easily applicable in cross-lingual settings, thus we use only the BWE based SVM.

Our second model was a CNN classifier following the widely-used architecture of Kim (2014). This model accepts an embedding layer as an input and feeds it into a convolution layer with a variable number of filters. Global max-pooling is performed on the convolution output, and the result is passed through a dense layer. The input word embeddings can either be randomly-initialized, preloaded from an outside source, or fine-tuned during training. Optionally, two of these variants may be used at the same time. We used only one form of embeddings, namely our BWEs, and did not update them during training. For the remaining model hyperparameters, we used the default values [3].

Our third model was based on the neural model of Wiedemann et al. (2018), with some modifications for compatibility with our cross-lingual setup. In our version, an input layer of our BWEs was fed into a BiLSTM layer of 100 units. The output of this BiLSTM layer was then fed into a convolution layer with three feature maps of 200 units each, with respective kernel sizes of 3, 4, and 5. Global max-pooling was applied after each convolution, and the output of this step was fed to a dense layer of 100 units.

4 Results

4.1 Cross-Lingual Experiments

In the first experiment phase, the SVM, CNN, and BiLSTM architectures were trained on EN-OS. Hyperparameters such as epoch count, learning rate, and class weights were optimized on DE-DEV. The two best-performing hyperparameter settings for each of the neural models were saved, resulting in a total of four English-trained models: CNN1, CNN2, BiLSTM1, and BiLSTM2. These will form the four-model ensemble used in Section 4.2.

The models' performance is summarized in Table 5. Despite the models being trained on forum posts they achieved good results on tweets as well.

With all three of our architectures, we observed that training with a proper class weight distribution was vital to ensuring good performance. Despite the fact that EN-OS was designed to be as balanced as possible, an improper weight ratio would result in a model always predicting either 'Hate' or 'noHate' on DE-DEV. This would achieve a high macro-average F1 score on DE-DEV but poor performance within one of the two classes. Therefore we chose the weight values that produced good F1 scores for both classes, even if it meant sacrificing average F1 performance.

Table 6 shows the optimal hyperparameter values during training on EN-OS. The CNN architecture benefited from high epoch counts and a class weight distribution that was skewed towards 'noHate'. The BiLSTM architecture in contrast required a reduced epoch count due to its overfitting behavior, and resulted in a class weight distribution that was skewed towards 'Hate'.

4.2 Bootstrapping

The next phase of cross-lingual experiments centered around data augmentation and fine-tuning. In the first augmentation experiment, we treated the samples in the DE-REL training set described in Table 4 as unlabeled and used an ensemble-based approach similar to Kozareva (2006) to label them. As mentioned in Section 3.1, we did not use DE-REL at any point during the training process.

Our ensemble consisted of the four neural models mentioned in Section 4.1. Of the relabeled sentences, we collected only those for which all four

[3] https://github.com/yoonkim/CNN_sentence/blob/master

	noHate	Hate	Dropout	Learn Rate	Epochs
SVM	0.17	0.83	n/a	n/a	n/a
CNN1	0.6	0.4	0.7	10^{-4}	30
CNN2	0.6	0.4	0.7	10^{-4}	40
BiLSTM1	0.2	0.8	0.7	10^{-4}	15
BiLSTM2	0.2	0.8	0.7	10^{-4}	15

Table 6: Optimal hyperparameters for training on EN-OS. The first two columns represent class weights.

	noHate	Hate
noHate	1,490	113
Hate	412	30
Total	1,902	143

Table 7: Confusion matrix of the ensemble-relabeled DE-REL* compared to the original annotations in DE-REL. Gold and predicted labels are shown in the rows and columns respectively.

models agreed into a new corpus called DE-REL*. We do not include the SVM in our ensemble and only fine-tune the neural models, since the neural architectures allow us to easily control how much of the original training data to forget by means of learning rate tuning.

Table 7 shows the confusion matrix for the labels of the dataset.

The Labels of DE-REL* Table 8 shows a selection of correct and incorrect classifications in DE-REL*, according to the original gold labels of the examples. Comment 1 shows that the ensemble was able to detect a covert kind of hate speech that involved contextual hints rather than overt vulgar language. One must understand what 'Mohrenkopf' is being used to mean in order to recognize the sentence as hate speech. Comment 2 was correctly classified as 'noHate' despite the presence of the controversial word 'Scheindemokratie' (en. *fake democracy*). This suggests that the model learned hate speech features that were more complex than lexical cues.

Comment 3 was falsely labeled as 'Hate' and Comment 4 was falsely labeled as 'noHate'. A possible reason for Comment 4 could be that its hate speech target (i.e. the group of German boys) differs from the groups typically targeted in a white supremacist forum, which was the source for the EN-OS training data. As a result, our models did not associate the insults contained in the comment with a hateful meaning.

The Fine-Tuning Phase We performed fine-tuning on the four English-trained models in Table 5, loading the dataset DE-REL* and resuming training for 3 epochs. All other hyperparameters, shown in Table 10, were selected based on both class-wise and macro-average F1 performance on DE-DEV. In particular, we had to modify the class weight ratio of each model from its original optimum.

Table 9 shows the models' scores on DE-TEST after fine-tuning on DE-REL*. Three of the four models improved their macro-average F1-score after fine-tuning. The largest improvement of 1.76 points was made by the BiLSTM1, while BiLSTM2 was close behind with an improvement of 1.45 points. CNN1 improved less, achieving only an additional 0.18 points.

CNN2 was the only model to perform worse after fine-tuning, decreasing in macro-average F1 by 0.27 points. A closer look at its class-wise F1 scores for 'Hate' and 'noHate' in Table 9 shows that the model's score in the latter class improved, while its score in the former decreased. This points to the fact that the model became skewed during fine-tuning towards predicting 'noHate'.

Crawled German Stormfront Data In the second experiment, we sought a source of completely new German data from the Internet to serve as a fine-tuning dataset. As mentioned previously, German hate speech data is scarce, and most of the existing datasets deviate significantly from (de Gibert et al., 2018) in hate speech definition. To keep to the stylistic tendencies of the Stormfront dataset, we took our data from a German-language thread within the Stormfront forum.

At the time of our experiments, the thread had around 5,500 posts. These consisted predominantly of comments written in German, although many were written in English. We programmed a simple webcrawler using the Python library Scrapy[4] to browse the thread and extract the content of the

[4] https://scrapy.org

		Sentence	Ensemble	Gold
DE-REL*	1.	@Nacktmagazin @DuHugonotte Und zum Nachtisch einen Mohrenkopf . *en. And for dessert a Mohrenkopf (head of a Moor / a kind of candy)*	Hate	Hate
	2.	Wir leben in einer Scheindemokratie aber in keiner #Demokratie! *en. We're living in a fake democracy, but not in a democracy.*	noHate	noHate
	3.	@Dora_Bromberger @lawyerberlin Hetze haben weder Fakten noch Argumente je ersetzt. *en. Hate has never replaced facts or arguments.*	Hate	noHate
	4.	@NegohotShlomo @Frank_Pasemann Pubertierende masturbierende Pickelträger...kein Wunder das deutsche Mädels keine deutsche Jungs wollen....! *en. Pubescent masturbating pimple-faces... no wonder German girls don't want German boys!*	noHate	Hate
DE-NEW	5.	Das Motto in Deutschland heutzutage ist nicht "Deutschland, Deutschland ueber alles", aber jedoch Jude, Jude ueber alles! *en. The motto in Germany these days isn't "Germany, Germany above all" but rather "Jew, Jew above all"!*	Hate	(Hate)
	6.	Ich persönlich komme mit allen (ausser juden) besten aus und halte es mit Dr.David Duke der sagt: "I do not hate any other people." *en. Personally I get along with all people fantastically (except Jews) and hold with Dr. David Duke, who says: "I do not hate any other people."*	Hate	(Hate)
	7.	Das ist wirklich wahr, für die meisten **NS** wäre die Erfindung der Zeitmaschine ein Segen; der Punkt ist das Leute die eine Ideologie vertreten und den **Führer** religiös verehren nur wenig Zugang zur Realität haben. da skann man ja hier im Forum auch beobachten... *en. That really is true, for the majority of the **Nazis**, the invention of the time machine would have been a blessing; the point is that people who support an ideology and religiously venerate the **Führer** have only a limited grasp on reality. yo ucan[sic] observe that right here in the forum...*	noHate	(noHate)
	8.	Der in Tutzing in Oberbayern lebende kanadische **Holocaust-Leugner** Alfred Schaefer ist wegen Volksverhetzung angeklagt. Der 63-Jährige selbst hat den Verhandlungstermin am 4. Mai vor dem Amtsgericht Dresden mit den Worten er sei "vor die Inquisition geladen" publik gemacht und angekündigt, den Prozess dazu zu nutzen, in langatmiger Form den **nationalsozialistischen Völkermord** an den Juden in Frage zu stellen. *en. The Canadian **Holocaust-denier** Alfred Schaefer, who lives in Tutzing in Upper Bavaria, has been charged with sedition. On his part, the 63-year-old made his trial appointment before the court of Dresden on the 4th of May public by saying he had been "invited by the Inquisition", and announced that he would use the process to verbosely call into question the **national-socialist genocide** of the Jewish people.*	noHate	(noHate)
	9.	Eine Partei, die immernoch gegen die "boesen amerikanischen Besatzer" wettert, kann und will ich nicht verstehen. *en. A party that still continues to fume over the "evil American occupants" I cannot understand and do not want to."*	Hate	(noHate)

Table 8: Correct and incorrect ensemble labels for either DE-REL* or DE-NEW. Gold labels for DE-NEW are given by the authors in brackets.

	Model	Accuracy	Hate			noHate			Macro-Average		
			P	R	F1	P	R	F1	P	R	F1
DE-REL*	CNN1	**61.95**	**22.42**	30.01	25.66	**78.33**	**70.90**	**74.43**	**50.37**	**50.45**	**50.41**
	CNN2	**68.15**	22.24	18.24	20.04	78.19	**82.13**	**80.11**	50.22	50.19	50.20
	BiLSTM1	51.22	**23.31**	**53.69**	**32.51**	**79.57**	50.53	61.80	**51.44**	**52.11**	**51.77**
	BiLSTM2	66.53	**24.02**	**24.45**	**24.23**	**78.72**	78.33	78.52	**51.37**	**51.39**	**51.38**
DE-NEW	CNN1	**72.71**	21.99	9.70	13.46	78.13	**90.36**	**83.80**	50.06	50.03	50.05
	CNN2	**73.53**	**23.18**	9.06	13.02	78.24	**91.59**	**84.39**	**50.71**	50.32	**50.51**
	BiLSTM1	50.03	**22.63**	**53.04**	**31.72**	**78.90**	49.18	60.59	**50.76**	**51.11**	**50.94**
	BiLSTM2	22.08	**21.93**	**100.00**	**35.97**	**100.00**	0.25	0.51	**60.96**	**50.13**	**55.02**

Table 9: Model performance on DE-TEST with bootstrapping, i.e., after training on EN-OS and fine-tuning on either DE-REL* or DE-NEW. Improvements compared to the models without fine-tuning shown in bold.

		noHate	Hate	Dropout	Learn Rate	Epochs
DE-REL*	CNN1	0.1	0.9	0.4	10^{-7}	3
	CNN2	0.2	0.8	0.7	10^{-6}	3
	BiLSTM1	0.08	0.92	0.8	10^{-7}	3
	BiLSTM2	0.1	0.9	0.4	10^{-6}	3
DE-NEW	CNN1	0.1	0.9	0.2	10^{-7}	4
	CNN2	0.2	0.8	0.2	10^{-7}	4
	BiLSTM1	0.01	0.99	0.9	10^{-12}	2
	BiLSTM2	0.01	0.99	0.9	10^{-12}	3

Table 10: Optimal hyperparameters for fine-tuning on DE-REL* and DE-NEW. The first two columns represent class weights.

posts. To account for the typical prevalence of lengthy posts in a forum setting, the crawler considered each paragraph distinguished by a newline to be a separate text sample.

Before the data could be used for training, we performed some manual preprocessing to ensure compatibility with the format of a tweet. We removed the following things from the data, as much as was feasible to detect manually:

- Non-German text
- Bullet-point lists
- Quotes from books, articles, etc. over 1000 characters
- Links, i.e. to Youtube videos
- Extremely short lines: names, one-word responses, timestamps, letter salutations
- Lines or sentences that were cut off without any clear continuation

The following things were kept:

- Quotes or news article snippets under 1000 characters.
- Multi-line interview dialogue, with each line considered as a distinct text sample.
- Mixed English/German sentences, since Anglicisms were common in the GermEval tweets as well.

The following things were manually corrected and kept:

- Sentence parts that the crawler considered separate but clearly belonged together:
  - 'tut mir' and 'leid' → 'tut mir leid'
- Words broken apart by spaces:
  - 'd aß' → 'daß'

The result of this preprocessing was a corpus of 6,586 text samples, all or nearly all written in German. We refer to this corpus as DE-NEW and use it as the training set during fine-tuning.

The Labels of DE-NEW Since we had no gold labels of DE-NEW to evaluate our ensemble's classifications, we manually examined several examples and judged them strictly according to the points of the hate speech definition in de Gibert et al. (2018). Table 8 shows four classifications made by the ensemble. Comments 5 and 6 are clearly 'Hate' and were intuitively classified as such. Comments 7 and 8 were intuitively classified as 'noHate', although the comments contain several words which we bolded that might be associated with a hateful context: "NS" (*en. National-Socialist/Nazis*), "Führer", "Holocaust-Leugner", (*en. Holocaust-denier*). Despite the occurrence of these words, all four models of our ensemble were able to classify the whole text as 'noHate', indicating that the models learned deeper hate speech features than lexical cues. Finally, Comment 9 was mistakenly classified as 'Hate'. Although the statement is critical of a political party, it is not severe enough to be an attack. Table 9 illustrates the models' performance after fine-tuning on D-NEW.

As with DE-REL*, fine-tuning on DE-NEW improved the macro-average F1 scores for three of the four models. Once again a CNN performed worse, namely CNN1, whose average F1 score dropped by 0.18 points. The LSTMs once again showed the greatest improvement. Notably, BiLSTM2 managed to achieve the highest macro-average F1 score among both fine-tuned groups, due to its high precision in the 'Hate' class and high recall in the 'noHate' class.

These results were also dependent upon proper tuning of hyperparameters. The optimal values are shown in Table 10. Importantly, we decreased the dropout rates of the CNNs compared to the first fine-tuning setup and trained for four epochs instead of three. This procedure did not work for the

BiLSTMs, however, as it caused them to ovefit to DE-NEW and predict only 'Hate'. This resulted in a rapid decrease in macro-average F1 performance on DE-DEV. Therefore, to optimize performance for the BiLSTMs, we significantly lowered the learning rate, increased the dropout, and skewed the class weights strongly towards 'Hate', leaving the weight for 'noHate' nearly zero.

Discussion Our results align with the findings in Mathur et al. (2018), where fine-tuning on cross-lingual data improved classifier performance. In the same vein, both of our fine-tuning experiments improved the macro-average F1 performance of English-trained neural hate speech classifiers. However, ensuring good performance in our fine-tuning setups required proper hyperparameter tuning. Sub-optimal hyperparameter settings caused the models to overfit to the fine-tuning data and decrease in macro-average F1 performance. We noted that in each of the fine-tuning setups, one of our two CNN models worsened in macro-average F1 performance despite hyperparameter tuning. Moreover, borrowing the hyperparameter settings from the CNN that did improve its score did not help. A reason for this could be that the CNNs were originally trained for a different number of epochs. In general, however, we found that the CNN architecture was robust at handling the fine-tuning data without overfitting. In both fine-tuning rounds, the optimal class weight distributions for CNN1 and CNN2 remained the same as during their initial training, and when fine-tuning on DE-NEW they could afford to be trained for four epochs instead of three. The BiLSTMs in contrast required low learning rates, low epoch counts, and high dropout rates, particularly when being fine-tuned on DE-NEW. Nevertheless, when we tuned the hyperparameters for these models correctly, we achieved macro-average F1 improvements by far surpassing those of the CNNs.

We observe that the average F1 scores of the SVM, CNNs and BiLSTMs were relatively similar before fine-tuning, and also to a lesser degree after fine-tuning the neural models. This is in line with previous results in (Bachfischer et al., 2018), where the macro-average F1 scores of an SVM, CNN, and LSTM differed in around three points from each other. This is most likely the result of too little data for the neural networks, and may indicate the merits of procuring even more training data than we presently did to benefit from these models' full potential.

A caveat to training on a specialized dataset such as the Stormfront dataset is that the hate speech examples will typically have a characteristic target group, in this case non-white individuals. The effects of this were observed in our ensemble's mislabeling of Comment 4 in Table 8, where the hate speech target was the group of German boys. This underscores the importance of training on multiple subtypes of hate speech.

5 Conclusion

Building automatic hate speech detection systems for low-resource languages is problematic due to the small amount of available datasets. Our goal was to investigate whether cross-lingual transfer learning could be used to mitigate the problem of data scarcity. We chose an English dataset with a broad hate speech definition for training and a similar German corpus for testing. Although the datasets were similar, we had to simplify the complex annotation schema of the target language dataset into the binary schema of the source dataset to make them compatible for the cross-lingual experiments. Our results showed that cross-lingual transfer learning is indeed an effective tool for hate speech detection in low-resource languages. Additionally, we assembled two corpora of previously-unseen, unlabeled target language data and applied an ensemble of trained classifiers to them. We showed that fine-tuning on these automatically labeled examples improves classification performance. Our goal for the future is to apply cross-lingual transfer learning to other language pairs with greater syntactic differences than German and English, as well as employing current state-of-the-art transformer models such as BERT. In addition, since the differences of labeling schemas across various datasets could prevent the application of transfer learning methods, we aim at bridging the gap between incompatibly-labeled train and test datasets.

Acknowledgments

This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No. 640550) and by German Research Foundation (DFG; grant FR 2829/4-1).

References

Matthias Bachfischer, Uchenna Akujuobi, and Xiangliang Zhang. 2018. KAUSTmine - Offensive Comment Classification on German Language Microposts. In *Proceedings of the GermEval 2018 Workshop*, pages 33–37.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, pages 135–146.

Ondřej Bojar, Rajan Chatterjee, Christian Federmann, Barry Haddow, Chris Hokamp, Matthias Huck, Varvara Logacheva, and Pavel Pecina, editors. 2015. *Proceedings of the Tenth Workshop on Statistical Machine Translation*.

Uwe Bretschneider and Ralf Peters. 2017. Detecting offensive statements towards foreigners in social media. In *50th Hawaii International Conference on System Sciences, HICSS 2017, Hilton Waikoloa Village, Hawaii, USA, January 4-7, 2017*, pages 1–10. ScholarSpace / AIS Electronic Library (AISeL).

Aditi Chaudhary, Jiateng Xie, Zaid Sheikh, Graham Neubig, and Jaime G. Carbonell. 2019. A little annotation does a lot of good: A study in bootstrapping low-resource named entity recognizers. 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 5164–5174.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2018. Word Translation Without Parallel Data. In *Proceedings of the International Conference on Learning Representations*.

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.

Tom De Smedt and Sylvia Jaki. 2018. Challenges of Automatically Detecting Offensive Language Online: Participation Paper for the Germeval Shared Task 2018 (HaUA). In *Proceedings of the GermEval 2018 Workshop*, pages 27–32.

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*, pages 4171–4186.

Darja Fišer, Ruihong Huang, Vinodkumar Prabhakaran, Rob Voigt, Zeerak Waseem, and Jacqueline Wernimont, editors. 2018. *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*. Association for Computational Linguistics, Brussels, Belgium.

Erfan Ghadery and Marie-Francine Moens. 2020. Liir at semeval-2020 task 12: A cross-lingual augmentation approach for multilingual offensive language identification. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 2073–2079.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate speech dataset from a white supremacy forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180.

Zornitsa Kozareva. 2006. Bootstrapping named entity recognition with automatically generated gazetteer lists. In *Student Research Workshop*.

Kosisochukwu Madukwe, Xiaoying Gao, and Bing Xue. 2020. In data we trust: A critical analysis of hate speech detection datasets. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 150–161, Online. Association for Computational Linguistics.

Prasenjit Majumder, Daksh Patel, Sandip Modha, and Thomas Mandl. 2019. Overview of the HASOC track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages. In *Proceedings of the 11th forum for information retrieval evaluation*, pages 14–17.

Puneet Mathur, Ramit Sawhney, Meghna Ayyar, and Rajiv Shah. 2018. Did you offend me? classification of offensive tweets in Hinglish language. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 138–148, Brussels, Belgium. Association for Computational Linguistics.

Tharindu Ranasinghe and Marcos Zampieri. 2020. Multilingual offensive language identification with cross-lingual embeddings. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5838–5844.

Sarah T. Roberts, Joel Tetreault, Vinodkumar Prabhakaran, and Zeerak Waseem, editors. 2019. *Proceedings of the Third Workshop on Abusive Language Online*. Association for Computational Linguistics, Florence, Italy.

Björn Ross, Michael Rist, Guillermo Carbonell, Benjamin Cabrera, Nils Kurowsky, and Michael Wojatzki. 2016. Measuring the Reliability of Hate Speech Annotations: The Case of the European Refugee Crisis. In *Proceedings of NLP4CMC III: 3rd Workshop on Natural Language Processing for Computer-Mediated Communication*, volume 17 of *Bochumer Linguistische Arbeitsberichte*, pages 6–9.

Josef Ruppenhofer, Melanie Siegel, and Michael Wiegand, editors. 2018. *Proceedings of the GermEval 2018 Workshop*. Austrian Academy of Sciences.

Lukas Stappen, Fabian Brunn, and Björn Schuller. 2020. Cross-lingual Zero- and Few-shot Hate Speech Detection Utilising Frozen Transformer Language Models and AXEL.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):e0243300.

Zeerak Waseem, Wendy Hui Kyong Chung, Dirk Hovy, and Joel Tetreault, editors. 2017. *Proceedings of the First Workshop on Abusive Language Online*. Association for Computational Linguistics, Vancouver, BC, Canada.

Gregor Wiedemann, Eugen Ruppert, Raghav Jindal, and Chris Biemann. 2018. Transfer Learning from LDA to BiLSTM-CNN for Offensive Language Detection in Twitter. In *Proceedings of the GermEval 2018 Workshop*.

Michael Wiegand, Anastasija Amann, Tatiana Anikina, Aikaterini Azoidou, Anastasia Borisenkov, Kirstin Kolmorgen, Insa Kröger, and Christine Schäfer. 2018a. Saarland University's Participation in the GermEval Task 2018 (UdSW) - Examining Different Types of Classifiers and Features. In *Proceedings of the GermEval 2018 Workshop*, pages 21–26.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018b. Overview of the germeval 2018 shared task on the identification of offensive language. In *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018)*, pages 1–10.

Jian Xi, Michael Spranger, and Dirk Labudde. 2018. CNN-Based Offensive Language Detection. In *Proceedings of the GermEval 2018 Workshop*.