XHATE-999: Analyzing and Detecting Abusive Language Across Domains and Languages

Goran Glavaš[1]*, **Mladen Karan**[2]*, **Ivan Vulić**[3]*
[1]Data and Web Science Group, University of Mannheim, Germany
[2]Text Analysis and Knowledge Engineering Lab, University of Zagreb, Croatia
[3]Language Technology Lab, TAL, University of Cambridge, UK
goran@informatik.uni-mannheim.de
mladen.karan@fer.hr
iv250@cam.ac.uk

Abstract

We present XHATE-999, a multi-domain and multilingual evaluation data set for abusive language detection. By aligning test instances across six typologically diverse languages, XHATE-999 for the first time allows for disentanglement of the domain transfer and language transfer effects in abusive language detection. We conduct a series of domain- and language-transfer experiments with state-of-the-art monolingual and multilingual transformer models, setting strong baseline results and profiling XHATE-999 as a comprehensive evaluation resource for abusive language detection. Finally, we show that domain- and language-adaptation, via intermediate masked language modeling on abusive corpora in the target language, can lead to substantially improved abusive language detection in the target language in the zero-shot transfer setups.

1 Introduction

In the era of ever-growing amounts of user-generated online content it is becoming increasingly difficult to scale up moderation efforts (Nobata et al., 2016). However, the need for moderation is rapidly increasing due to escalated toxic behavior online, enabled by "hiding" behind anonymous profiles, lack of physical contact between participants (i.e., the communication is then typically perceived as less personal), and lack of direct negative societal consequences (Perse and Lambe, 2016). Consequently, research on automated methods for detecting abusive language in user-generated content is becoming increasingly important. While such methods cannot completely replace human moderators, they are very helpful as assistance tools, offering moderation suggestions, thus partially automating and expediting human moderation work.

The focus of abusive language detection is still predominantly on a single language – English, and single-domain setups (e.g., Twitter). However, some recent initiatives have aspired to broaden the scope of abusive detection methodology to other languages, showcasing the usefulness of cross-lingual transfer for the task (Sohn and Lee, 2019; Stappen et al., 2020; Pamungkas and Patti, 2019; Wiedemann et al., 2020, *inter alia*). Another line of research (Wiegand et al., 2018a; Karan and Šnajder, 2018; Waseem et al., 2018, *inter alia*) focuses on benefits of cross-domain transfer in monolingual settings. An interesting aspect, currently lacking in prior work, is the interaction of cross-lingual and cross-domain settings. Furthermore, except for some notable exceptions discussed in §2, previous work in cross-lingual setups is still tied to resource-rich and typologically similar languages (e.g., English, German, Spanish, Italian) (Stappen et al., 2020). We aim to fill both these gaps by introducing XHATE-999, a multilingual data set annotated for abusive language in three domains, and carefully manually translated from English to 5 typologically diverse languages, with 999 semantically aligned test instances across all languages.

Unlike other abusive language detection data (surveyed in §2), XHATE-999 allows us to separate the effects that occur due to *domain shift* from the effects related to *language shift*. Current data sets typically confound the two: i.e., a switch to a test set from a different language also implies a change of the topic/domain. Having an identical domain in the source language and the target language enables research questions such as: Is domain shift or language shift more instrumental to performance decrease

*Equal contribution.

This work is licensed under a Creative Commons Attribution 4.0 International License. License details: http://creativecommons.org/licenses/by/4.0/.

Proceedings of the 28th International Conference on Computational Linguistics, pages 6350–6365
Barcelona, Spain (Online), December 8-13, 2020

in transfer settings? Are these patterns consistent across different domains and languages? Moreover, covering semantically aligned test instances, XHATE-999 enables direct comparisons on different target languages, giving rise to the following research questions: How consistent is model behavior across languages? *Ceteris paribus*, is abusive language detection in some languages inherently more difficult than in the others? Is this behavior consistent across different domains?

Besides offering XHATE-999 as the new evaluation resource to the community, in this work we also aim to provide answers to the questions posed above. We evaluate state-of-the-art transfer learning methodology based on pretrained monolingual and multilingual Transformer models – RoBERTa (Liu et al., 2019), monolingual English BERT and multilingual BERT (mBERT) (Devlin et al., 2019), and XLM-R (Conneau et al., 2020) – in a range of in-domain and cross-domain monolingual and cross-lingual experimental setups. These evaluations set strong and challenging baseline results on XHATE-999, and show that cross-lingual performance drops depend, among other aspects, 1) on the actual domain and corresponding abusive language properties (e.g., abusive unigrams versus abusive phrases), and 2) on typological properties of the target language and linguistic distance to English as the source language (e.g., smaller drops are observed on German as the target language than for Turkish or Albanian). We also empirically verify that training data augmentation by merging training examples from different domains can be very detrimental to both monolingual and cross-lingual performance in cases where the abusive language domains are too distant.

Finally, we introduce a simple transfer model adaptation that yields improved performance in cross-lingual transfer for abusive language detection. Inspired by recent work on additional domain-adaptive pretraining (Gururangan et al., 2020) as well as additional target language pretraining (Ponti et al., 2020; Glavaš and Vulić, 2020a), we propose to continue training mBERT and XLM-R via masked language modeling (MLM) (i.e., the so-called *intermediate MLM-ing*) on automatically extracted "hateful" raw text in the target languages.[1] We show that this additional language and domain adaptation of the base massively multilingual model can yield further performance gains: we obtain higher scores than MLM-ing on randomly sampled raw text of the same size, confirming that both language adaptation and adaptation to abusive language are required to boost transfer performance.

2 Related Work and Motivation

Variants of Abusive Language. Abusive language appears in many flavors, including sexism, racism (Waseem and Hovy, 2016; Waseem, 2016), toxicity (Kolhatkar et al., 2019), hatefulness (Gao and Huang, 2017), aggression (Kumar et al., 2018), attack (Wulczyn et al., 2017), cyberbullying (Van Hee et al., 2015; Sprugnoli et al., 2018), misogyny (Fersini et al., 2018), obscenity, threats, and insults. Waseem et al. (2017) proposed a systematic typology of toxic language. Another typology focusing more on the nature of targets of abusive texts was proposed by Zampieri et al. (2019). A similar scheme, expanded to include the personal sentiments of annotators, was introduced by Ousidhoum et al. (2019). A very fine-grained hierarchical annotation scheme including 81 different types of annotations was used to label the data set of Fortuna et al. (2019). Furthermore, Founta et al. (2018) propose an iterative crowdsourcing-based approach to derive a set of high-quality abusive language labels. Recently, it has been pointed out that existing abusive language data sets are biased towards certain types of abuse (Jurgens et al., 2019; Vidgen and Derczynski, 2020) and domains/topics (Wiegand et al., 2019). In this work, we combine three different abusive language variants – hatefulness (Gao and Huang, 2017), aggression (Kumar et al., 2018), and attack (Wulczyn et al., 2017) – spanning three distinct data sources (comments under Fox News stories, Twitter/Facebook posts, and Wikipedia edit messages, respectively) into an integrated and cross-language aligned multilingual evaluation resource.

Multilingual and Cross-Lingual Abusive Language Detection. There is a growing body of work on abusive language detection for other languages, realized mostly through shared tasks. The recent OffenseEval task (Zampieri et al., 2020) introduced a multilingual data set for 5 languages (English, Arabic, Danish, Hebrew, Turkish), which was expanded to German and Italian by Casula (2020). The

[1] In the actual experiments, we do not assess if the raw text is considered abusive - the criterion for sentence inclusion is that it simply contains at least one cue word that is considered abusive - e.g., *stupid, racist, hate, fool, kill, ridiculous*.

HatEval shared task (Basile et al., 2019) spans only English and German, and other works (Steinberger et al., 2017; Sohn and Lee, 2019; Ousidhoum et al., 2019; Steimel et al., 2019; Stappen et al., 2020; Corazza et al., 2020, *inter alia*) similarly target only major European languages such as French, German, Italian, Czech, and Spanish.[2] As indicated by Stappen et al. (2020), annotated evaluation data for more diverse and resource-poor languages is a prerequisite to develop portable and widely reachable abusive language detection methodology. With XHATE-999, we make a step towards reaching out also to such languages.

In the cross-lingual settings, Steinberger et al. (2017) train separate detection models for several languages, but link results via named entities and dictionaries for use in a search engine. A more sophisticated transfer between languages is explored by Corazza et al. (2020) based on shared cross-lingual word embeddings (Ruder et al., 2019). The most recent work (Sohn and Lee, 2019; Stappen et al., 2020; Wiedemann et al., 2020) has naturally shifted towards the current state-of-the-art cross-lingual transfer paradigm (Hu et al., 2020): large multilingual Transformer-based (Vaswani et al., 2017) models such as multilingual BERT (Devlin et al., 2019) and XLM(-R) (Conneau and Lample, 2019; Conneau et al., 2020), pretrained via masked language modeling (MLM). The usefulness of machine translation systems (Sohn and Lee, 2019), cross-attention (Stappen et al., 2020), and readily available dictionaries such as HurtLex (Bassignana et al., 2018) has also been explored (Pamungkas and Patti, 2019).

Cross-Domain Abusive Language Detection. A comprehensive analysis of cross-domain models is provided by Waseem et al. (2018), who experiment with multi-task learning for domain transfer on three data sets. In a similar vein, Karan and Šnajder (2018) employ frustratingly easy domain adaptation (Daumé III, 2007) to experiment with domain transfer on a wide range of abusive language data sets. Some cross-domain approaches rely on term analysis, e.g., Wiegand et al. (2018a) start from a manually constructed sample of abusive terms and augment it automatically to aid domain adaptation, while Rizoiu et al. (2019) aim to construct task-agnostic representations of abusive language. This stands in contrast with insights from Swamy et al. (2019), which suggest that the high variation in abusive language typically precludes wide generalisations and domain adaptation. The work of Pamungkas and Patti (2019) is closest to ours, as they provide some preliminary experiments on domain transfer across languages, mostly indicating its complexity, key challenges, and usefulness of available abusive language lexicons. However, they focus on readily available and unaligned data sets in major European languages (English, German, Italian, Spanish), do not provide direct comparisons across languages, now enabled by XHATE-999, and do not investigate isolating the effects of language versus domain transfer.

3 New Multilingual Data Set: XHATE-999

Initial Data Preparation. In order to build a data set that comprises multiple variants (i.e., tasks) of abusive language detection, we sampled annotated examples from three well-known and diverse English data sets: (Gao and Huang, 2017) (termed GAO henceforth, capturing hatefulness), (Kumar et al., 2018) (TRAC, agression), and (Wulczyn et al., 2017) (WUL, attack). The motivation for these particular data sets is twofold. First, they span three distinct data domains: Fox News (GAO), Twitter/Facebook (TRAC), and Wikipedia (WUL). Second, they focus on three domains with varying amounts of annotated data available for training in English: WUL comprises 71,754 training examples (and 24,130 validation examples), while the respective numbers are 10,341 (2,593) for TRAC, and only 919 (218) for GAO. Training and validation splits from the original work were retained for all three data sets. We next map the labels of each data set into the binary labels: *abusive* vs. *non-abusive*. GAO and WUL already come with binary labels, while the original TRAC uses three labels: *non-aggressive*, *covertly-aggressive*, and *openly-aggresive*. We relabel the first as *non-abusive*, and the other two as *abusive*.

We then sample test instances from the respective test portions of all three data sources.[3] For quality

[2]There are also approaches which target monolingual settings with a language that is not English. A non-exhaustive list includes Arabic (Mubarak et al., 2017; Chowdhury et al., 2020), Danish (Sigurbergsson and Derczynski, 2020), German (Jaki and De Smedt, 2019; Wiegand et al., 2018b), Hindi (Saroj and Pal, 2020), Italian (Bosco et al., 2018; Fersini et al., 2018), Polish (Ptaszynski et al., 2019), Portuguese (Fortuna et al., 2019), Dutch (Tulkens et al., 2016), and Slovene (Fišer et al., 2017). There is also some work on specific language variants, like Hindi–English code-switched language (Mathur et al., 2018a; Mathur et al., 2018b) or South African English (Oriola and Kotzé, 2020).

[3]Downsampling was conducted to enable translation into a sufficiently large sample of target language under budget

Text	Data	Label
Now Indian government must have to take powerful Action against Bluddy Pakistan....	TRAC	*Y*
Leonidas was sent into the wild... to learn how to survive	TRAC	*N*
get off me you disease, i want my name of this crappy site. Avoid what? everyone knows what Ive done	WUL	*Y*
Of course, I am willing to negotiate this in a way that we can agree on. I just want to say I'm sorry.	WUL	*N*
Cicero Black lives of murderers, criminals and rioters do not matter.	GAO	*Y*
MarineAssassin That stuff is funny dont care who ya are	GAO	*N*

Table 1: Examples from the test partition of English XHATE-999. *Y=abusive*; *N=non-abusive*.

assurance, each test example candidate has been manually checked by the authors, and replaced by another sample if it (i) comprises only a single non-indicative word, (ii) it is not written in English, or (iii) it relies on world knowledge which is too specific or geographically localized or on contextual information which hinders proper translation. The final English XHATE-999 test set comprises 600, 300 and 99 instances from WUL, TRAC, and GAO, respectively. Some English test examples are provided in Table 1.

Manual Translation. The pivotal objectives in XHATE-999 creation were: **1)** to create a multilingual data set that is aligned across diverse target languages, in order to enable direct performance comparisons across languages, and **2)** to ensure high quality, fluency, naturalness, and idiomacity of monolingual data sets in each target language. To achieve this, we followed a carefully monitored translation-based approach, recently used to collect a multilingual commonsense reasoning evaluation resource (Ponti et al., 2020). The main idea is, instead of using fast-turnaround, but low-quality crowdsourcing solutions (Lavee et al., 2019), **1)** to run the translation task with a small number of carefully selected translators per target language, and **2)** to provide opportunity for necessary target-language adjustments (e.g., using multi-word paraphrases, culturally more adequate substitutes or near-synonyms) without hurting "the abusiveness level" of the English instance. To this effect, the translators were allowed to introduce slight modifications into their translation in order to reflect and maintain the level of abuse present in the original instance. Such modifications were necessary in cases where a literal translation would lose its abusive nature in the target language. For instance, this happens with English-specific phrases that do not exist in the target language. Another example are English word plays (e.g., merging personal names with terms for animal species and using the portmanteau as an insult, see an example in the Appendix); in these cases the translators were instructed to make up a roughly equivalent insult of the same type in the target language. The chosen translators were human experts who were fluent in English while the target language was their native language. While detailed translation guidelines are available in the Appendix, the crucial guidelines can be summarized as follows:

Given a piece of text, translate it from English (the source language) into your mother tongue (the target language). The translation should be as accurate as possible, but under the constraint that the level of abuse present in the original text is well preserved in the translation.

We have translated the 999 test instances from the English (EN) XHATE-999 to five target languages: Albanian (SQ), Croatian (HR), German (DE), Russian (RU), and Turkish (TR). The choice of the target languages has been guided by the following (sometimes competing) criteria: **a)** availability of trusted translators per target language; **b)** translation budget; and **c)** relative typological and etymological diversity of the language sample, along with the general availability of linguistic resources for the language (e.g., English and German as resource-rich languages versus Albanian as a resource-lean language). The translation effort was approximately 45 person-hours per target language.

The advantage of the translation-based approach adapted from Ponti et al. (2020) is twofold. First, it allows for disentangling the impact of language versus domain shift: the alignment between the source and the target language test data ensures that any performance loss of a cross-lingual transfer approach is solely due to language shift. Second, the alignment of test data across languages allows for a cleaner and more meaningful cross-language comparison of (transfer) results. This opens up new research opportunities related to studying abusive language detection across a larger number of typologically diverse languages.

restrictions. The number of final test instances also partially reflects the size differences of the original test sets.

Train & Dev / Test	BERT (Base)			RoBERTa (Base)		
	WUL	TRAC	GAO	WUL	TRAC	GAO
WUL	89.0	53.8	56.0	87.8	50.4	44.1
TRAC	77.1	76.0	60.5	82.6	<u>77.2</u>	<u>62.6</u>
GAO	43.9	71.1	48.9	54.8	57.1	**70.3**
ALL	<u>90.6</u>	75.9	54.3	**90.7**	**77.3**	59.8

Table 2: Monolingual evaluation on English subset of XHATE-999; we used the corresponding readily available English training sets. ALL: training on the concatenation of all three training sets. All scores are $F_1 \times 100\%$. Highest scores for each test subset are in boldface, the second best scores are underlined.

4 Monolingual Evaluation: Analyses across Domains

We rely on English training data for WUL, TRAC, and GAO (see §3) in all monolingual and cross-lingual experiments. We first focus on cross-domain experiments in monolingual English settings. In short, we analyze the difference in performance when training **1)** on all available training data from all three data sets (WUL+TRAC+GAO; this setup is labeled ALL); **2)** only on the training set which corresponds to the particular test subset (e.g., when testing on WUL we train only on WUL training data; this setup is labeled SAME); and **3)** on a non-corresponding training set (e.g., when testing on WUL, we train on TRAC or GAO training data), probing the impact of domain shift.

Experimental Setup. We experiment with two pretrained monolingual English transformer models: BERT (Devlin et al., 2019) *Base Cased*, and RoBERTa (Liu et al., 2019) *Base*,[4] both with $L = 12$ transformer layers, hidden state size of $H = 768$, and $A = 12$ self-attention heads. We adopt the standard fine-tuning architecture for sequence classification tasks: we add a simple feed-forward classification head taking as input the transformed representation of the sequence start token (`[CLS]` for BERT, `<s>` for RoBERTa) $\mathbf{x}_{ss} \in \mathbb{R}^H$, i.e., $\hat{\mathbf{y}} = softmax(\mathbf{x}_{ss}\mathbf{W}_{cl} + \mathbf{b}_{cl})$, with $\mathbf{W}_{cl} \in \mathbb{R}^{H \times 2}$ and $\mathbf{b}_{cl} \in \mathbb{R}^2$ as classifier's parameters. We tune the parameters by minimizing the standard cross-entropy loss.

For both BERT and RoBERTa we search the following hyperparameter grid: learning rate $\in \{5 \cdot 10^{-6}, 10^{-5}, 3 \cdot 10^{-5}\}$, dropout rate (applied to the output layer of the transformer) $\in \{0, 0.1\}$, and batch size $\in \{16, 32\}$. We found the following hyperparameter configuration to be optimal in all experiments: learning rate = 10^{-5}, batch size = 32, and dropout rate = 0.1. We opt for early stopping based on the development set performance (F_1 score). We measure the development set performance after every 500 updates for the WUL training set (as well as the ALL setup in which we train on the concatenation of all three training sets), every 100 updates for TRAC, and every 20 updates for GAO. We stop training if there is no development set performance improvement over 10 consecutive evaluations. We optimize the parameters with the Adam algorithm (Kingma and Ba, 2015) ($\epsilon = 10^{-8}$, no weight decay nor warmup) and clip the norms of gradients for individual updates to 1.0. We report the results in terms of F_1 scores.

Results and Discussion. The main results of all in-domain and cross-domain experiments are summarized in Table 2. We observe several interesting phenomena. First, as expected, RoBERTa provides peak scores across all three test subsets, but there is some variation in performance; BERT outperforms RoBERTa for a few training–test combinations. For instance, BERT has a slight edge over RoBERTa in the WUL-WUL SAME setup, and it is also on-par in the ALL-WUL setup. However, RoBERTa seems as a more robust choice overall, especially in the ALL and SAME (WUL-WUL, TRAC-TRAC, and GAO-GAO) setups. We observe a particularly substantial gain in the low-data GAO-GAO setup (with only 919 training instances available): this confirms recent findings in other language understanding tasks (Lauscher et al., 2020; Brown et al., 2020) that few-shot fine-tuning works much better with pretrained language models which were exposed to more text during pretraining.

Cross-domain experiments also lead to several insights. Augmenting heterogeneous training data (as done in the ALL setup) is not necessarily useful: for instance, we do not see any gains moving from

[4]Our code is built on top of the HuggingFace Transformers framework: `https://github.com/huggingface/transformers`. We used these monolingual English models: `bert-base-cased` and `roberta-base`.

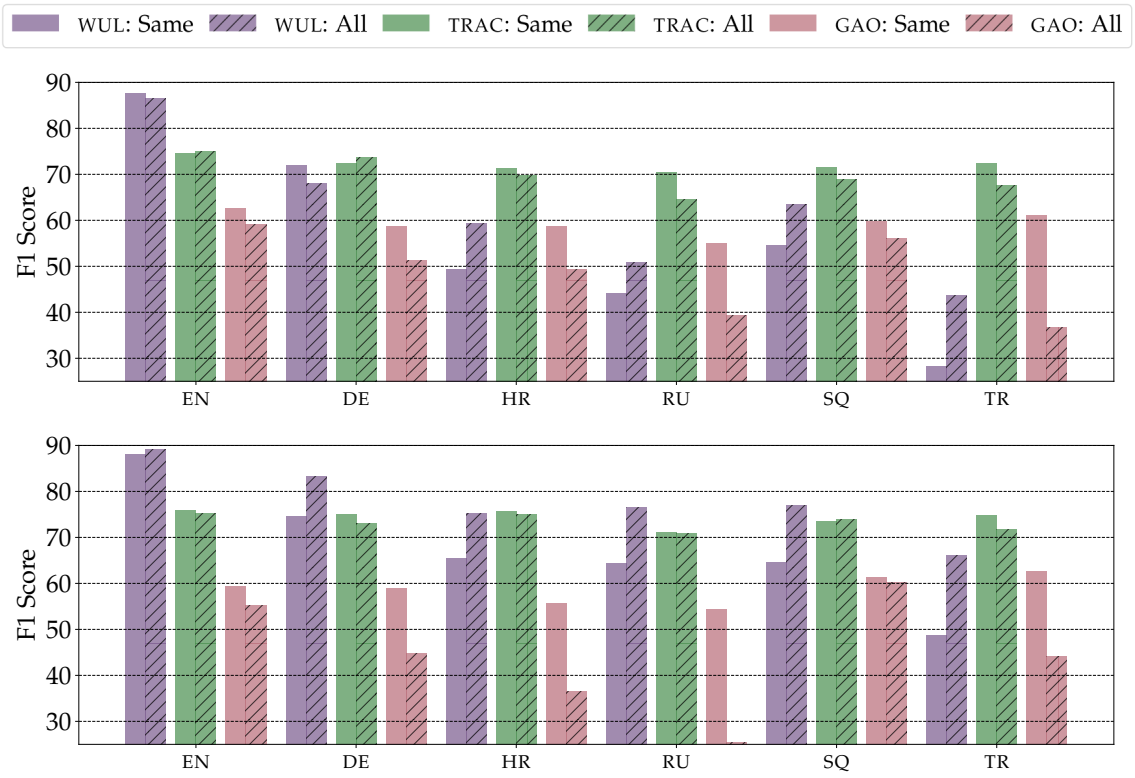

Figure 1: Zero-shot cross-lingual transfer of abusive language detection. EN is the source language in all experiments. **Upper Figure**: transfer via mBERT; **Lower Figure**: transfer via XLM-R. Stripped bars denote the training setup in English where we merge ALL three training sets; the other bars denote the training setup where the training and test data are from the SAME domain: i.e., WUL, GAO, or TRAC.

the TRAC-TRAC setup to the ALL-TRAC setup, but we see small benefits moving from WUL-WUL to ALL-WUL. ALL and SAME setups score much higher than cross-domain training for WUL and TRAC. A large drop in performance on TRAC when training on the large WUL training data set is particularly indicative: it suggests that having more training data (from another abusive language detection task) does not imply better detection scores if there is a domain/task mismatch such as the one between WUL (detecting attacks in Wikipedia comments) and TRAC (detecting aggression in social media). The same trend is visible also when training on TRAC and testing on WUL, but this could also be partially attributed to a smaller TRAC training set (see §3). The scores also indicate that, due to a similar domain mismatch, WUL is a less appropriate training set for GAO test data: we see higher scores when training on smaller TRAC data than on larger WUL training data, both for BERT and RoBERTa. Interestingly, we also see smaller drops on the TRAC test data when training on the extremely small GAO training data versus large WUL data. This again hints that having more similar data domains for cross-domain transfer is more important than having very large data sets in more distant abusive language domains.

5 Cross-Lingual Transfer and Evaluation

In our zero-shot language transfer experiments, the focus is on the ALL and SAME setups, which provided the peak scores in the English monolingual experiments in §4.[5] In the cross-lingual ALL setup, we again train on the merged WUL+TRAC+GAO English training data, while in the SAME setup, we train on one of the three English training portions, and test on the corresponding test subset in the target language (i.e., we evaluate WUL-WUL, TRAC-TRAC, and GAO-GAO training–test combinations).

[5]We provide the detailed results of simultaneous domain- and language-transfer results in the Appendix.

Transfer Setup. We adopt the massively multilingual pretrained Transformers as a state-of-the-art mechanism for zero-shot language transfer. Concretely, we conduct our experiments using multilingual BERT (mBERT) (Devlin et al., 2019) and XLM-on-RoBERTa (XLM-R) (Conneau et al., 2020).[6,7] Training, validation, and optimization procedures as well as the hyperparameters are exactly the same as those reported for monolingual English experiments (see Experimental Setup in §4).

Results and Discussion. A summary of the cross-lingual transfer results on all test XHATE-999 subsets in all target languages is provided in Figure 1. First, comparing the results of massively multilingual models versus English-specific pretrained LMs on the English test sets (still no transfer involved), we now report slight performance drops: peak scores with EN RoBERTa versus XLM-R decrease from 90.7 to 89.2 on WUL, from 77.3 to 76.0 on TRAC, and from 70.3 to 59.3 on GAO. Similar trends are observed in the BERT versus mBERT comparison. This is expected and can be explained through the well-known "curse of multilinguality" (Conneau et al., 2020; Lauscher et al., 2020; Pfeiffer et al., 2020), where multilingual models with limited capacity trade off their performance in a particular language for much higher portability and transfer ability.[8] XLM-R generally offers stronger transfer performance than mBERT for abusive language detection, which is in line with findings from other tasks (Hu et al., 2020).

We observe large performance drops in the WUL SAME and ALL setups with mBERT. The drops, although smaller, are also pronounced when using XLM-R. However, transfer performance is much higher and more stable in the SAME setup for the other two domains – TRAC and GAO – while there are still conspicuous performance drops in the ALL setup. We hypothesise that, due to a large English WUL training set, the models overfit to the English data much more than when trained on significantly smaller TRAC and GAO training sets. We also speculate that the drops are due to the nature of the WUL data, which does not come from social media, so it typically contains longer and more complex utterances (WUL utterances are on average 25% and 30% longer than TRAC and GAO utterances, respectively). This means that the abusive language detection model is more likely to overfit to abusive idiomatic expressions in English, which are more difficult to semantically align to similar expressions in target languages via the shared multilingual semantic representation space.

When testing on TRAC and GAO, the results suggest that it is more effective to transfer the model trained on their respective in-domain training portions than the model trained on the merged ALL data. Effectively, this implies that the detection model cannot handle both domain and language transfer simultaneously for these two domains. Language transfer without any domain shift outperforms transfer with more (but also more out-of-domain) training data: the useful signal for the two test sets gets overwritten by the much larger WUL training data, and the model then tends to overfit to WUL. However, the opposite is true with WUL test data: we see improvements in the ALL setup over the SAME setup for all target languages. We hypothesize that, again due to a large WUL EN training corpus, adding training data from other two domains in fact acts as a regularization mechanism: it can impede the idiomatic overfitting to English which is difficult to transfer to other languages.

Besides the actual domain (and its shift), the properties of the target language also impact transfer performance. The pattern is especially visible in WUL evaluations for both training setups: performance drops are much lower for German, the target language most similar to EN (i.e., both are Germanic languages), while we note the highest drop on TR as the only non-Indo-European language and agglutinative language in our target language sample. However, this pattern does not hold on TRAC and GAO evaluations: e.g., absolute results on TR GAO are higher than in any other target language, and we observe a similarly strong result on TR TRAC. While the variation might be partially due to small training and test GAO data,

[6]We have also benchmarked another strand of cross-lingual models that conduct the transfer via static projection-based cross-lingual word embeddings (Artetxe et al., 2018; Glavaš et al., 2019; Ruder et al., 2019; Vulić et al., 2019; Glavaš and Vulić, 2020b). Similar to what was observed in other language transfer tasks recently (Hu et al., 2020), these methods have in our experiments been consistently outperformed by transfer methods based on massively multilingual transformers (mBERT and XLM-R). Therefore, we do not report these results for brevity and to avoid clutter.

[7]Models from HuggingFace Transformers: `bert-base-multilingual-cased` and `xlm-roberta-cased`.

[8]The only exception is mBERT outperforming English BERT on GAO: while it is difficult to draw general conclusions due to the small respective training and test set, this could mean that multilingual pretraining with lower capacity for English-specific representations avoids overfitting to small training sets during fine-tuning.

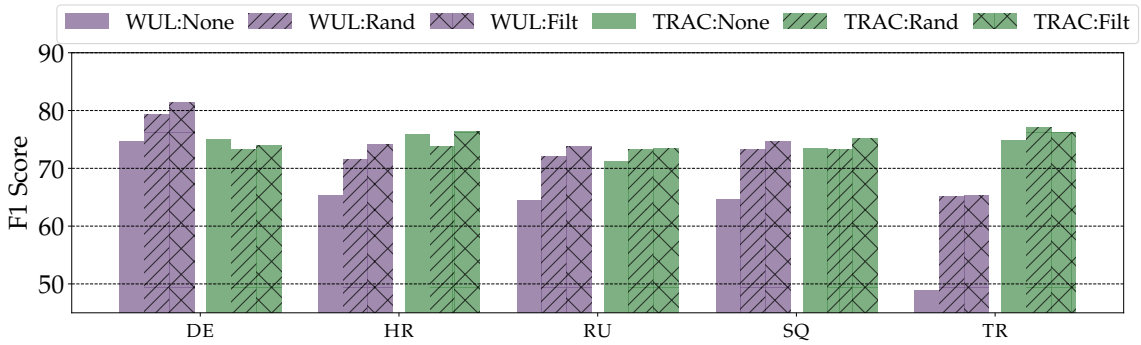

Figure 2: Cross-lingual transfer of abusive language detection with different variants of intermediate masked language modeling (MLM), see §6: no additional intermediate MLM-ing (*None*); intermediate MLM-ing in the target language on N randomly sampled sentences (*Rand*); intermediate MLM-ing on N non-randomly *filtered* sentences that contain abusive cue words (*Filt*). We report the results in the SAME setup (e.g., training on EN WUL, testing on target language WUL portions) with XLM-R.

stable results in transfer experiments on TRAC across all target languages suggest that the complexity (or rather simplicity) of the abusive language domain and data also plays a role in transfer capability.

Overall, the results indicate that the success of cross-lingual transfer depends on a multitude of factors such as the actual task formulation (i.e., abusive language domain) and the nature of abusive language data, (dis)similarity between the source and the target language, the actual transfer methodology, and in-domain versus cross-domain training. Our work has aimed to disentangle all these factors in order to measure their contribution to the final transfer performance, and future work should pay more attention to their complex interactions in transfer experiments for abusive language detection.

Brief Error Analysis. The instance-level alignment of XHATE-999 portions in different languages now enables a comparative cross-language analysis of classification errors. Target language misclassifications, with correctly classified English counterparts, in 90% of the cases represent false negatives, i.e., cases with undetected abusive language. We identified these to predominantly be the instances: 1) requiring extensive world knowledge (e.g., *give the old yeller treatment*), 2) containing idiomatic EN abusive phrases (*bird brain*), 3) containing (deliberately) mistyped insults and profane words (*id1ot*), 4) with polysemous EN words used in abusive sense, but with a non-abusive dominant sense (*balls*), and 5) with abusive content packed into compounds (*feminazi*). In all these cases, the multilingual transformers (mBERT and XLM-R) fail to align the meaning of the abusive clue from the original English utterance with the meaning of the corresponding (in most cases non-literal) abusive translation in the target language. We leave more extensive qualitative analyses for future work.

6 Intermediate Masked Language Modeling on Filtered Text

Motivation and Approach. Language models such as mBERT and XLM-R are pretrained on large general-purpose and massively multilingual corpora (100+ languages). While this makes them versatile and widely applicable, it does not make them acquire "abusive language" and also leads to the "curse of multilinguality", i.e., suboptimal representations for individual languages, due to constrained model capacity (Conneau et al., 2020). We thus hypothesize that **1)** adapting them to particular target languages, and **2)** exposing them to additional abusive (instead of general-purpose) language might lead to performance gains, especially in cross-lingual transfer. We opt to achieve these adaptations through additional *intermediate* masked language modeling in the target languages as follows.

We explore three scenarios: **1)** no intermediate MLM-ing (*None*; results from §5), **2)** intermediate MLM-ing on N randomly sampled sentences in the target language (*Rand*), and **3)** intermediate MLM-ing on the same number of target language sentences N, but now filtered from large corpora to contain salient abusive terms (*Filt*), which should consequently better adapt the models to abusive language. The *Rand* MLM-ing provides target language adaptation of a massively multilingual model (Pfeiffer et al.,

English	German	Russian	Croatian	Albanian	Turkish
		GAO			
slave	Sklave	рабыня	rob	rob	köle
ugly	hässlich	уродливый	ružan	i shëmtuar	çirkin
useless	nutzlos	бесполезный	beskoristan	i kotë	Faydasız
racist	Rassist	расист	rasistički	racist	ırkçı
immigrants	Einwanderer	иммигранты	imigranti	emigrantët	göçmenler
harassment	Belästigung	домогательство	uznemiravanje	ngacmim	rahatsızlık
		TRAC			
stupid	blöd	глупый	glup	budalla	Aptal
fool	täuschen	дурачить	budala	budalla	aptal
terrorist	Terrorist	террорист	terorista	terrorist	terörist
communists	Kommunisten	коммунисты	komunisti	komunistët	komünistler
hell	Hölle	ад	pakao	ferr	cehennem
gay	Fröhlich	гей	homoseksualac	homoseksual	eşcinsel
dick	Schwanz	Дик	kurac	kar	çük
		WUL			
hell	Hölle	ад	pakao	ferr	cehennem
gay	Fröhlich	гей	homoseksualac	homoseksual	eşcinsel
dick	Schwanz	Дик	kurac	kar	çük
nazis	Nazis	гитлеровцы	nacisti	nazistët	Naziler
screw	Schraube	винт	vijak	vidhos	vida
fascist	faschistisch	фашист	fašistički	fashist	faşist

Table 3: Some of the abusive English terms and corresponding target language translations obtained automatically using Google Translate, used for filtering corpora for intermediate MLM-ing.

2020), which should partially alleviate the issues arising due to limited model capacity and the "curse of multilinguality" (Conneau et al., 2020; Lauscher et al., 2020), but without any domain adaptation. The *Filt* variant ideally offers both language adaptation and (at least crude) adaptation to abusive language.

Corpora, Filtering, and MLM Training. We first semi-automatically obtain lists of abusive terms related to our abusive language domains, based on the English WUL, GAO, and TRAC training sets. In short, we train a logistic regression classifier on each training set separately, rank the words according to the weights associated with the *abusive* class, and retain only the ones which occur in the top 10k most frequent English words in the ukWaC corpus (Ferraresi et al., 2008). A manual inspection reveals that many words in the lists are only topically related without being abusive terms (e.g., *mother*, *cake*, *vision*): therefore, in the next step the list is manually filtered to retain only salient abusive terms. This yields the final lists of 10 (GAO), 8 (TRAC), and 27 (WUL) abusive terms in English, which are automatically translated to the target languages via Google Translate without any subsequent manual correction. Some examples of abusive clues obtained through this semi-automatic procedure are shown in Table 3.

For the *Filt* intermediate MLM-ing, we then extract at most 200K sentences that contain at least one term from at least one list of abusive terms from a large corpus. For all languages we rely on readily available web-crawled corpora: ukWaC and deWaC (Ferraresi et al., 2008) for EN and DE, hrWaC (Ljubešić and Erjavec, 2011; Šnajder et al., 2013) for HR, the OSCAR data (Suárez et al., 2019) for TR and SQ, and the Araneum corpora (Benko, 2014) for RU. The total number of extracted sentences per language is 193K (DE), 200K (EN), 97K (HR), 65K (RU), 27K (SQ), and 200K (TR). For *Rand* MLM-ing, we simply randomly sample the same number of sentences as for *Filt* from the same web-crawled corpora.

We execute the intermediate MLM training in *Rand* and *Filt* scenarios by dynamically masking 15% of the subword tokens in order to predict them from the context. We train for 30 epochs, in batches of 32 sentences, by minimizing the cross-entropy loss with the Adam algorithm (Kingma and Ba, 2015).

Results and Discussion. The cross-lingual experimental setup is identical to the one in §5, with the exception of experimenting with mBERT and XLM-R under different intermediate MLM-ing scenarios (*None*, *Rand*, *Filt*). We show results only with XLM-R as the better-performing multilingual transformer. The results for WUL and TRAC in the SAME setup (see §5) are summarized in Figure 2 (full results available in the Appendix). The scores clearly suggest the usefulness of the language and domain adaptation, especially on WUL, while the positive trends, although present, are less pronounced on TRAC. On WUL, we observe improvements over the baseline (*None*; no intermediate MLM-ing) for all 5/5 target languages for both *Rand* and *Filt*. On top of this, *Filt* offers some gains over *Rand* in 5/5 transfer experiments, with the average of 73.9 F_1 for *Filt* and 72.2 for *Rand*. On TRAC, *Filt* outperforms *None* on

4/5 target languages (i.e., the only exception is German).

Furthermore, intermediate MLM-ing (*Rand* and *Filt*) yields the highest gains for two target languages that are most distant from EN: SQ and TR. The gains for Turkish are particularly large: e.g., *Filt* gains on WUL amount to +8.1 F_1 points in the ALL setup and +16.5 points in the SAME training setup. This is mostly due to language adaptation to TR. Turkish is an extremely morphologically rich language written in Latin script, which means that it is not sufficiently represented in the joint multilingual subword vocabulary of XLM-R: both *Rand* and *Filt* intermediate MLM-ing therefore adapt/specialize the shared subwords towards Turkish; additional slight gains with *Filt* are due to more in-domain MLM-ing. In sum, the improvements with *Rand* indicate that adaptation to the particular target language is important for enhanced abusive language detection, but further improvements can be achieved by customizing XLM-R with filtered sentences containing lexical clues of abusive language.

From another perspective, our experiments have verified that both language-adaptive additional pretraining (Pfeiffer et al., 2020; Ponti et al., 2020; Glavaš and Vulić, 2020a) as well as domain-adaptive additional pretraining (Gururangan et al., 2020) of general-purpose language models have a synergistic positive impact on cross-lingual transfer for abusive language detection. However, the scores from Figure 2 also indicate that there is still ample room for improvement, particularly in resource-lean and distant languages (SQ and TR). Additional advances might be met through techniques such as selective sharing and more sophisticated typologically driven adaptation in transfer (Ponti et al., 2018; Nikolaev et al., 2020), using larger and manually compiled lexicons of abusive language (Bassignana et al., 2018) instead of small, noisy and inexpensively built lexicons as in this work, or choosing more suitable source languages (Lin et al., 2019) and source domains (Gururangan et al., 2020). Few-shot transfer, requiring a small number of labeled target-language instances, is another strategy that has, in the context of language transfer via multilingual transformers, been shown to lead to large gains over zero-shot transfer (Lauscher et al., 2020).

Brief Error Analysis. Finally, we perform a closer inspection of target language test instances that have been misclassified after the intermediate MLM training on the random corpus (*Rand*), yet correctly classified after MLM-training on the corpus filtered with the list of abusive words (*Filt*). We discovered that many of such instances – 47% for SQ, 57% for HR, 40% for DE, 48% for RU, and 53% for TR– contain at least one of the abusive cue words that we used to create the corpus for MLM-ing in *Filt*.

7 Conclusion and Future Work

We have presented XHATE-999, a data set enabling evaluation of both cross-domain and cross-lingual abusive language detection, and in-depth explorations of the interplay between language shift and domain shift. XHATE-999 spans three diverse abusive language domains and six diverse languages. The semantic alignment between test instances in all languages for the first time enables comparative analyses of model behavior across domains and languages. We have also profiled the potential of XHATE-999 as a comprehensive resource for evaluating abusive language detection through a series of in-domain and cross-domain experiments in monolingual and cross-lingual setups with state-of-the-art transfer learning models. We have then demonstrated that domain-adaptive and language-adaptive additional pretraining of general-purpose multilingual models (multilingual BERT and XLM-R) can yield further performance gains in transfer experiments, especially for resource-lean languages.

We hope that XHATE-999 will inspire and instigate deeper understanding of the underlying phenomena and further research on cross-lingual and cross-domain abusive language detection, with a stronger focus towards diverse and resource-lean languages and domains. We make the XHATE-999 data set publicly available at `https://github.com/codogogo/xhate`.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of ACL*, pages 789–798.

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 *Proceedings of SEMEVAL*, pages 54–63.

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. HurtLex: A multilingual lexicon of words to hurt. In *Proceedings of the 5th Italian Conference on Computational Linguistics*, volume 2253, pages 1–6.

Vladimír Benko. 2014. Aranea: Yet another family of (comparable) Web corpora. In *Proceedings of the International Conference on Text, Speech, and Dialogue*, pages 247–256.

Cristina Bosco, Dell'Orletta Felice, Fabio Poletto, Manuela Sanguinetti, and Tesconi Maurizio. 2018. Overview of the evalita 2018 hate speech detection task. In *Proceedings of EVALITA*, volume 2263, pages 1–9.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *CoRR*, abs/2005.14165.

Camilla Casula. 2020. Transfer learning for multilingual offensive language detection with BERT.

Shammur Absar Chowdhury, Hamdy Mubarak, Ahmed Abdelali, Soon-gyo Jung, Bernard J Jansen, and Joni Salminen. 2020. A multi-platform arabic news comment dataset for offensive language detection. In *Proceedings of LREC*, pages 6203–6212.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Proceedings of NeurIPS*, pages 7057–7067.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of ACL*.

Michele Corazza, Stefano Menini, Elena Cabrio, Sara Tonelli, and Serena Villata. 2020. A multilingual evaluation for online hate speech detection. *ACM Transactions on Internet Technology (TOIT)*, 20(2):1–22.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of ACL*, pages 256–263.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Adriano Ferraresi, Eros Zanchetta, Marco Baroni, and Silvia Bernardini. 2008. Introducing and evaluating ukWaC, a very large web-derived corpus of English. In *Proceedings of the 4th Web as Corpus Workshop (WAC-4)*, pages 47–54.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the EVALITA 2018 task on automatic misogyny identification (AMI). *Proceedings of EVALITA*, 12:59.

Darja Fišer, Tomaž Erjavec, and Nikola Ljubešić. 2017. Legal framework, dataset and annotation schema for socially unacceptable on-line discourse practices in Slovene. In *Proceedings of the First Workshop on Abusive Language Online*, pages 46–51.

Paula Fortuna, Joao Rocha da Silva, Leo Wanner, Sérgio Nunes, et al. 2019. A hierarchically-labeled Portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. *arXiv preprint arXiv:1802.00393*.

Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. In *Proceedings of RANLP*, pages 260–266.

Goran Glavaš and Ivan Vulić. 2020a. Is supervised syntactic parsing beneficial for language understanding? an empirical investigation. *arXiv preprint arXiv:2008.06788*.

Goran Glavaš and Ivan Vulić. 2020b. Non-linear instance-based cross-lingual mapping for non-isomorphic embedding spaces. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7548–7555.

Goran Glavaš, Robert Litschko, Sebastian Ruder, and Ivan Vulić. 2019. How to (properly) evaluate cross-lingual word embeddings: On strong baselines, comparative analyses, and some misconceptions. In *Proceedings of ACL*, pages 710–721.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of ACL*, pages 8342–8360.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. In *Proceedings of ICML*.

Sylvia Jaki and Tom De Smedt. 2019. Right-wing German hate speech on Twitter: Analysis and automatic detection. *arXiv preprint arXiv:1910.07518*.

David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 2019. A just and comprehensive strategy for using NLP to address online abuse. In *Proceedings of ACL*, pages 3658–3666.

Mladen Karan and Jan Šnajder. 2018. Cross-domain detection of abusive language online. In *Proceedings of the Second Workshop on Abusive Language Online*, pages 132–137.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR*.

Varada Kolhatkar, Hanhan Wu, Luca Cavasso, Emilie Francis, Kavan Shukla, and Maite Taboada. 2019. The SFU opinion and comments corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics*, pages 1–36.

Ritesh Kumar, Aishwarya N. Reganti, Akshit Bhatia, and Tushar Maheshwari. 2018. Aggression-annotated corpus of Hindi-English code-mixed data. In *Proceedings of LREC*, pages 1425–1431.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot cross-lingual transfer with multilingual transformers. *CoRR*, abs/2005.00633.

Tamar Lavee, Lili Kotlerman, Matan Orbach, Yonatan Bilu, Michal Jacovi, Ranit Aharonov, and Noam Slonim. 2019. Crowd-sourcing annotation of complex NLU tasks: A case study of argumentative content annotation. In *Proceedings of the First Workshop on Aggregating and Analysing Crowdsourced Annotations for NLP*, pages 29–38.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of ACL*, pages 3125–3135.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Nikola Ljubešić and Tomaž Erjavec. 2011. hrWaC and slWaC: Compiling web corpora for Croatian and Slovene. In *Proceedings of the International Conference on Text, Speech and Dialogue*, pages 395–402.

Puneet Mathur, Ramit Sawhney, Meghna Ayyar, and Rajiv Shah. 2018a. Did you offend me? Classification of offensive tweets in Hinglish language. In *Proceedings of the Second Workshop on Abusive Language Online*, pages 138–148.

Puneet Mathur, Rajiv Shah, Ramit Sawhney, and Debanjan Mahata. 2018b. Detecting offensive tweets in hindi-english code-switched language. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 18–26.

Hamdy Mubarak, Kareem Darwish, and Walid Magdy. 2017. Abusive language detection on Arabic social media. In *Proceedings of the First Workshop on Abusive Language Online*, pages 52–56.

Dmitry Nikolaev, Ofir Arviv, Taelin Karidi, Neta Kenneth, Veronika Mitnik, Lilja Maria Saeboe, and Omri Abend. 2020. Fine-grained analysis of cross-linguistic syntactic divergences. In *Proceedings of ACL*, pages 1159–1176.

Chikashi Nobata, Joel Tetreault, Achint Thomas, Yashar Mehdad, and Yi Chang. 2016. Abusive Language Detection in Online User Content. In *Proceedings of the 25th International Conference on World Wide Web*, pages 145–153.

Oluwafemi Oriola and Eduan Kotzé. 2020. Evaluating machine learning techniques for detecting offensive and hate speech in South African tweets. *IEEE Access*, 8:21496–21509.

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

Endang Wahyu Pamungkas and Viviana Patti. 2019. Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon. In *Proceedings of ACL: Student Research Workshop*, pages 363–370.

Elizabeth M. Perse and Jennifer Lambe. 2016. *Media Effects and Society*. Routledge.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An adapter-based framework for multi-task cross-lingual transfer. *CoRR*, abs/2005.00052.

Edoardo Maria Ponti, Roi Reichart, Anna Korhonen, and Ivan Vulić. 2018. Isomorphic transfer of syntactic structures in cross-lingual NLP. In *Proceedings of ACL*, pages 1531–1542.

Edoardo Maria Ponti, Goran Glavas, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. *CoRR*, abs/2005.00333.

Michal Ptaszynski, Agata Pieciukiewicz, and Paweł Dybała. 2019. Results of the PolEval 2019 shared task 6: First dataset and open shared task for automatic cyberbullying detection in Polish Twitter.

Marian-Andrei Rizoiu, Tianyu Wang, Gabriela Ferraro, and Hanna Suominen. 2019. Transfer learning for hate speech detection in social media. *arXiv preprint arXiv:1906.03829*.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual embedding models. *Journal of Artificial Intelligence Research*, 65:569–631.

Anita Saroj and Sukomal Pal. 2020. An Indian language social media collection for hate and offensive speech. In *Proceedings of the Workshop on Resources and Techniques for User and Author Profiling in Abusive Language*, pages 2–8.

Gudbjartur Ingi Sigurbergsson and Leon Derczynski. 2020. Offensive language and hate speech detection for Danish. In *Proceedings of LREC*, pages 3498–3508.

Jan Šnajder, Sebastian Padó, and Željko Agić. 2013. Building and evaluating a distributional memory for Croatian. In *Proceedings of ACL*, pages 784–789.

Hajung Sohn and Hyunju Lee. 2019. MC-BERT4HATE: Hate speech detection using multi-channel BERT for different languages and translations. In *Proceedings of the International Conference on Data Mining Workshops*, pages 551–559.

Rachele Sprugnoli, Stefano Menini, Sara Tonelli, Filippo Oncini, and Enrico Piras. 2018. Creating a WhatsApp dataset to study pre-teen cyberbullying. In *Proceedings of the Second Workshop on Abusive Language Online*, pages 51–59.

Lukas Stappen, Fabian Brunn, and Björn Schuller. 2020. Cross-lingual zero-and few-shot hate speech detection utilising frozen transformer language models and AXEL. *arXiv preprint arXiv:2004.13850*.

Kenneth Steimel, Daniel Dakota, Yue Chen, and Sandra Kübler. 2019. Investigating multilingual abusive language detection: A cautionary tale. In *Proceedings of RANLP*, pages 1151–1160.

Josef Steinberger, Tomás Brychcín, Tomás Hercig, and Peter Krejzl. 2017. Cross-lingual flames detection in news discussions. In *Proceedings of RANLP*, pages 694–700.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous pipeline for processing huge corpora on medium to low resource infrastructures. In *Proceedings of the Seventh Workshop on the Challenges in the Management of Large Corpora (CMLC-7)*.

Steve Durairaj Swamy, Anupam Jamatia, and Björn Gambäck. 2019. Studying generalisability across abusive language detection datasets. In *Proceedings of CoNLL*, pages 940–950.

Stéphan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. 2016. A dictionary-based approach to racism detection in Dutch social media. In *Proceedings of the Workshop on Text Analytics for Cybersecurity and Online Safety*.

Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Véronique Hoste. 2015. Detection and fine-grained classification of cyberbullying events. In *Proceedings of RANLP*, pages 672–680.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of NeurIPS*, pages 5998–6008.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data: Garbage in, garbage out. *arXiv preprint arXiv:2004.01670*.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Do we really need fully unsupervised cross-lingual embeddings? 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 4398–4409.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? Predictive features for hate speech detection on Twitter. In *Proceedings of NAACL-HLT: Student Research Workshop*, pages 88–93.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84.

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.

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.

Gregor Wiedemann, Seid Muhie Yimam, and Chris Biemann. 2020. UHH-LT & LT2 at SemEval-2020 Task 12: Fine-tuning of pre-trained transformer networks for offensive language detection. *arXiv preprint arXiv:2004.11493*.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018a. Inducing a lexicon of abusive words–a feature-based approach. In *Proceedings of NAACL-HLT*, pages 1046–1056.

Michael Wiegand, Melanie Siegel, and Josef Ruppenhofer. 2018b. Overview of the GermEval 2018 shared task on the identification of offensive language. In *Proceedings of KONVENS*.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: the problem of biased datasets. In *Proceedings of NAACL-HLT*, pages 602–608.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th International Conference on World Wide Web*, pages 1391–1399.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. pages 1415–1420.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. Semeval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). *arXiv preprint arXiv:2006.07235*.

Appendix

A Translation Instructions

We give below the full unedited version of the translation instructions given to the translators and discussed with them in the initial meeting.

These are the guidelines for the task of translating datasets annotated for the presence of abusive language from English (the source language) to other (target) languages. The purpose of this annotation is to obtain annotated data in the target languages which can be used to develop and evaluate cross-lingual machine learning models to automate the detection task.

Disclaimer: The rest of these instructions and (somewhat more so) the data in this task contain some very explicit and strong language. This is unavoidable due to the nature of the task. If you feel this would be too unsettling for you, please stop reading after this paragraph and simply let us know you are no longer interested.

Your task in this annotation is as follows:

Given a piece of text, translate it from English (the source language) into your mother tongue (the target language). The translation should be as accurate as possible, but under the constraint that the level of abuse present in the original text is well preserved in the translation.

You will not have to do the entire translation manually from scratch, but you will rather edit and correct the outputs of Google Translate. This should save you some time. We recommend that you go about translating each text in two steps. First, correct any mistakes in the outputs of Google Translate and make sure that the translation is accurate. Second, asses whether the level of abuse you perceive in the original English text is roughly equivalent to the level of abuse in your translation. If there is a notable difference then tweak your translation a bit to reduce the difference. The tweaked translation will be slightly less accurate (as in less literal) but is preferable since it better preserves the level of abuse.

The first step above is simply translation. The second step though warrants some examples (all examples are from English to Croatian):

1. Sometimes you will need to make no modifications e.g. “shit” can be literally translated as “sranje”, wich carries roughly the same level of abuse in Croatian.

2. Sometimes you will need to make minor semantic modifications, e.g. “piece of shit” cannot be translated directly (there is no croatian phrase for that), but you can translate it as “govno” which literally means “shit” (without the “piece of” semantics) and is used in roughly the same context as the english phrase. This slightly changes the meaning but retains the level of abuse.

3. Sometimes major semantic changes will be needed e.g., “scumbag” also has no Croatian translation, and has no croatian insult that is similar. You could simply translate it as “gad” or “đubre” (“bastard” or “shit”). These have a different meaning but carry roughly the same level of abuse and are appropriate translations.

4. You will often encounter abuse in much more subtle ways. A text can be aggressive without explicitly using swearing. E.g., someone might refer to a person as a “failed abortion”. Even more subtly someone might consistently use the “It” pronoun when referring to a particular person. Be sure to preserve these subtle nuances when translating. This is particularly important, as we expect that Google Translate will often fail to preserve these details which can mean the difference between abuse and non-abuse.

5. You will sometimes need to dig deeper to fully understand some of the text. For example the word “Hildebeest” is not clear at first glance. A short search reveals it is mocking Hillary Clinton via fusing “Hillary” with “wildebeest”. This information is important to properly retain the level of abuse in the translation. In this particular case you would ideally have to come up with a similar mocking play on words in your own language. If that is difficult or unnatural in your language, you could alternatively translate this using any “standard” insult in your language, which carries roughly the same level of abuse as “Hildebeest”.

Finally, keep in mind that not all texts you need to translate are abusive (around 50% of them are). So try to not have any bias that would push your translations to become generally more or less abusive than they need to be. Simply read the texts carefully and recreate the level of abuse in your translations. In an ideal scenario the translations are exactly as abusive (or non abusive) as the original texts.

Train & Dev / Test	*None*			*Rand*			*Filt*		
	WUL	TRAC	GAO	WUL	TRAC	GAO	WUL	TRAC	GAO
WUL	87.8	50.4	44.1	89.8	58.6	56.5	88.4	46.4	31.3
TRAC	82.6	77.2	62.6	79.5	76.2	64.0	80.8	75.6	61.1
GAO	54.8	57.1	70.3	58.8	58.1	60.8	46.6	59.4	61.7
ALL	90.7	77.3	59.8	90.1	76.5	63.1	90.3	76.9	63.2

Table 4: Domain transfer performance of monolingual English RoBERTa-based models; we used the corresponding readily available English training sets. ALL: training on the concatenation of all three training sets. All scores are $F_1 \times 100\%$.

Train setup / Test set	*None*			*Rand*			*Filt*		
	WUL	TRAC	GAO	WUL	TRAC	GAO	WUL	TRAC	GAO
DE-ALL	83.2	73.1	44.7	79.7	72.8	50.0	82.7	73.9	60.0
DE-SAME	74.6	75.1	58.9	79.3	73.2	63.8	81.4	73.9	54.1
HR-ALL	75.1	74.9	36.4	76.6	72.9	58.1	74.1	73.9	54.0
HR-SAME	65.4	75.8	55.6	71.5	73.8	59.8	74.2	76.3	60.9
RU-ALL	76.6	70.9	25.4	75.5	70.8	37.1	72.3	66.0	24.6
RU-SAME	64.4	71.2	54.3	72.0	73.3	51.9	73.8	73.4	44.2
SQ-ALL	77.0	74.0	60.2	80.4	73.9	59.1	79.5	74.9	56.1
SQ-SAME	64.7	73.5	61.4	73.2	73.3	54.3	74.6	75.1	66.7
TR-ALL	66.1	71.8	44.2	72.0	74.0	43.2	74.2	74.9	49.4
TR-SAME	48.8	74.9	62.6	65.2	77.1	62.5	65.3	76.2	63.6

Table 5: Full domain- and language-transfer results for the XLM-R-based models. The ALL setup denotes the model was trained on all train data, while the SAME setup denotes the model was trained on the train set corresponding to the test set. All scores are $F_1 \times 100\%$.

B Full Domain- and Language-Transfer Results

Full monolingual (English) domain-transfer results and full cross-lingual (domain- **and** language-transfer) results are displayed in Tables 4 and 5, respectively.