When More is not Necessary Better: Multilingual Auxiliary Tasks for Zero-Shot Cross-Lingual Transfer of Hate Speech Detection Models

Anonymous ACL submission

Abstract

Zero-shot cross-lingual transfer learning has been shown to be highly challenging for tasks involving a lot of linguistic specificities or when a cultural gap is present between languages, such as in hate speech detection. In this paper, we highlight this limitation on several datasets and investigate how training on multilingual auxiliary tasks – sentiment analysis, named entity recognition, and tasks relying on syntactic information – impacts the zero-shot transfer of hate speech detection models across languages. We show the positive impact of these tasks, particularly named entity recognition, for bridging the gap between languages. Then, we present cases where the language model training data prevents hate speech detection models from benefiting from a *knowledge proxy* brought by auxiliary tasks fine-tuning. Our results warrant further investigation on how to best address cultural gap issues in resource-scarce scenario.

1 Introduction

Given the impact social media hate speech can have on our society as a whole – leading to many small-scale “Overton window” effects – the NLP community has devoted considerable efforts to automatic hate speech detection using machine learning-based approaches, and proposed different benchmarks and datasets to evaluate their techniques (Dinakar et al., 2011; Sood et al., 2012; Waseem and Hovy, 2016; Davidson et al., 2017; Fortuna and Nunes, 2018; Kennedy et al., 2020).

However, these systems are designed to be efficient at a given point in time for a specific type of online content they were trained on. As hate speech evolves significantly across time, at different granularities (Florio et al., 2020), hate-speech models need to cope with its diachronic and synchronic variations. For example, more efficient online platforms censorship has forced social media users to adapt to these platforms’ filtering systems to avoid detection, through the evolution of their lexicon and spelling variations, in an endless cat and mouse game (Berger and Perez, 2016; Vidgen et al., 2019). Furthermore, as noted by Markov et al. (2021), the occurrence of *new* hate speech domains, and their associated lexicons, hashtags, etc. can be triggered by events, from local scope incidents to world-wide crisis.[1] The cultural aspect also impacts hate speech in a synchronic fashion. In particular, hate speech perception is highly variable across languages, meaning that some slur expressions can be considered not offensive in one language, denoting an informal register nonetheless, but will be considered offensive, if note hateful, in another (Nozza, 2021). In this paper, we denote this divergence as *cultural gap*. This notion is also introduced by Cabrio et al. (2014), who study the gap between languages in DBpedia entries.

In this work, we focus on cross-lingual variations of hate speech. In resource-scarce scenarios, several works propose methods to transfer hate speech detection models from one language to another (Basile and Rubagotti, 2018; van der Goot et al., 2018; Pamungkas and Patti, 2019; Ranasinghe and Zampieri, 2020). The main options are either to use a multilingual language model-based transfer learning architecture, or to train models on a translation of the initial training data to the target language (Rosa et al., 2021). However, hate speech cultural and linguistic variations can lower the transferability of hate speech detection models across languages. To overcome this limitation, those methods need a certain amount of target language training data, or efficient translation models, that are not available in low-resource scenarios. The cultural and linguistic information specific to the monolingual hate speech target data needs to be found elsewhere in our zero-shot transfer setting. We hypothesize that this information

[1] e.g. Hate speech towards the Chinese communities in 2020 with the emergence of the COVID-19 Pandemic.

may be captured through fine-tuning the languages model on resource-rich tasks in both the source and the target language of the transfer (van der Goot et al., 2021a). For example, sentiment features are sometimes used to support hate speech detection. We also hypothesize that syntactic features learned through tasks such as dependency parsing can provide support for adapting hate speech detection models to new languages. Thus, our work focuses on zero-shots cross-language multitask architectures where no annotated hate speech data is available for our target languages but some annotated data for other tasks can be accessed. Using a multitask architecture (van der Goot et al., 2021b) on top of a multilingual model (XLM-R, Conneau et al., 2020), we investigate the impact of auxiliary tasks operating at different sentence linguistics levels (POS Tagging, Named Entity Recognition (NER), Dependency Parsing and Sentiment analysis) on the transfer effectiveness. Following Nozza (2021)'s original set of languages and datasets (hate speech against women and immigrants, from Twitter datasets in English, Italian and Spanish), our main contributions are as follow:

- Using a strictly comparable setting, we confirm and highlight *cultural gap* issues for zero-shot cross-lingual transfer of hate speech detection.
- Through our experiments, we identify auxiliary tasks with a positive impact on cross-lingual transfer when trained jointly with hate speech detection: sentiment analysis and NER. The impact of syntactic tasks is more mitigated, as the information they bring is highly language and domain-specific.
- We confirm that as expected using a domain-adapted language model such as XLM-T (Barbieri et al., 2021) brings an orthogonal improvement to all configurations. Additionally, running random initialization of XLM-R results allowed us to precisely identify the cases where the XLM-R pre-training data actually prevented the auxiliary tasks to favor an efficient cross-lingual transfer.

2 Related work

2.1 Enhancing language models with other tasks

In order to improve the efficiency of a pre-trained language model for a given task, this model can undergo a preliminary fine-tuning on an intermediate task, before fine-tuning again on the downstream task (Pruksachatkun et al., 2020). This system, also called STILT, was formalized by Phang et al. (2018), who perform sequential task-to-task pretraining. More recently, Phang et al. (2020) deepen the investigation by turning towards cross-lingual STILT. They fine-tune a language model on nine intermediate language-understanding tasks in English, and apply it to a set of non-English target tasks. The efficiency of the systems differs a lot depending on the tasks. One might expect that English STILT would be less efficient than using tasks in the target language. The authors show that machine-translating intermediate task data for training, or using a multilingual language model, does not improve the transfer compared to English training data. They also showed that multi-task training on all intermediate tasks slightly outperforms separately training on the tasks.

Finally, Pruksachatkun et al. (2020) perform a survey of intermediate and target task pairs to analyze the usefulness of this intermediary fine-tuning. To isolate the training effect on each task, they train the language models on each task separately before fine-tuning it on the target (downstream) task. They find a low correlation between the acquisition of low-level skills and downstream task performance, while tasks that require complex reasoning and high-level semantic abilities such as common-sense-oriented tasks have a higher benefit. But overall, their results do not allow to draw precise conclusions.

In the context of hate speech detection, when auxiliary task training is applied, it is done almost exclusively with the sentiment analysis task (Bauwelinck, Nina and Lefever, Els, 2019; Aroyehun and Gelbukh, 2021), and only in monolingual scenarios. But additional information is sometimes added to the hate speech classifier differently. Gambino and Pirrone (2020), among the best systems on the HaSpeeDe task of the EVALITA 2020, use the POS-tagged text as input of the classification systems, which is highly beneficial for Spanish and a bit less for German and English. Furthermore, the effect of syntactic information also is investigated by Narang and Brew (2020), using classifiers based on the syntactic structure of the text for abusive language detection. Markov et al. (2021) evaluate the impact of manually extracted POS, stylometric and emotion-based features on hate speech detection, showing that the latter two are robust features for hate speech detection across languages.

2.2 Zero-shot cross-lingual transfer learning for hate speech detection

Due to the lack of annotated data on many languages and domains for hate speech detection, zero-shot cross-lingual transfer has been tackled a lot in the literature. Among the most recent work, Pelicon et al. (2021) investigate the impact of a preliminary training of a classification model on hate speech data languages different from the target language; they show that language models pre-trained on a small number of languages benefit more of this intermediate training, and often outperforms massively multilingual language models. Nozza (2021), on which this paper builds upon, demonstrates the limitation of cross-lingual transfer for domain-specific hate speech – in particular, hate speech towards women – and explains it by showing examples of cultural variation between languages. Some notable hate speech vocabulary in one language may be used as an intensifier in another language[2].

3 The bottleneck of zero-shot cross-lingual transfer

3.1 Hate speech corpora

We use the same hate speech datasets as Nozza (2021), who relied on it to point out the limitations of zero-shot cross-lingual transfer. The corpora are in three languages: English (en), Spanish (es) and Italian (it). Each language is divided into corpora from two domains: hate speech towards immigrants and hate speech towards women.[3]

Comparable settings. The corpora do not have the same size in the different languages and domains. Therefore, we build comparable corpora in each language and domain to ensure the comparability of the transfer settings. We reduce all datasets to a total size of 2 591 tweets, the size of the smallest one, sampling from each original split separately; each train set has 1 618 tweets, each development set 173, and each test set 800. When sub-sampling the corpora, we make sure they stay comparable: we use the Kolmogorov–Smirnov test to compare the sentence length distribution (number of tokens) between the sampled and the original datasets. We do the same for the percentage of hate speech. The sampling is done randomly until the similarity conditions with the original dataset are met. Finally, we use the remaining observations from the original datasets, that were not sampled to be used in the train-dev-test sets, to create a blind set. We down-sample each blind set to 1000 tweets, except for the one in Spanish on the immigrants domain, which is the smallest of all, thus having no blind test. The original size for each dataset as well as the sampling size for building the comparable datasets can be found in Table 4 in Appendix A.

Pre-processing. We process the datasets by replacing all mentions and urls with specific tokens, and segmenting the hashtags into words.[4] Given the compositional nature of hashtags (a set of concatenated words), hashtag segmentation is frequently done as a processing step in the literature when processing tweets (e.g. Röttger et al. (2021)); it can improve tasks such as tweet clustering (Gromann and Declerck, 2017).

3.2 Experimental settings

For all our experiments, we use the MaChAmp v0.2 framework[5] (van der Goot et al., 2021b), a multi-task toolkit based on AllenNLP (Gardner et al., 2018). We keep the default hyperparameters of MaChAmp for all experiments, which the authors optimized on a wide variety of tasks.

We fine-tune a multilingual language model on the hate speech detection task for each of the six training corpora described in the previous section. Then, we apply the model on the test set of each target language, investigating two settings: 1) monolingual, i.e, training and testing on the same language and domain for hate speech; 2) zero-shot, cross-lingual, i.e. training on one and testing on another.

The test sets sampled from the original corpora are relatively small (800 observations). To increase the robustness of the results, we use five different seeds when fine-tuning a language model on the hate speech detection task and report the average macro-F1 over the five runs. We keep the best out of 20 epochs for each run according to the macro-F1 score on the development set.

[2] Nozza (2021) gives the example of the Spanish word *puta* often used as an intensifier without any misogynistic connotation, while it translates to a slang version of "prostitute" in English.

[3] The corpora come from various shared tasks that are listed in Table 7 in Appendix A.

[4] Using the Python package `wordsegment`.

[5] `https://github.com/machamp-nlp/machamp`, under the MIT license.

3.3 Hate speech detection baseline

In this section, we compare two language models: m-BERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020). The latter shows slightly higher scores on average. We also ensure that switching to comparable corpora using the same model and hyperparameters leads to similar tendencies. The detailed results of these experiments can be found in Table 13 in Appendix C. Given the results that show similar tendencies, in all following experiments, we use the comparable datasets and the language model XLM-R.

We report the results with the XLM-R model, following the settings described in the previous section, in Table 1a. To summarize the results, we aggregate them according to the two settings exposed in Section 3.2: monolingual (*mono*), and zero-shot cross-lingual (*cross*). Table 1b is the aggregated equivalent of Table 1a. For each domain (immigrants and women), we average the scores by setting: the "monolingual" columns show the average of all scores in italic in Table 1a, while the "cross-lingual" column is the average of the results for all scores in normal case. We also report the results from Nozza (2021), which show similar tendencies. In particular, we observe the phenomenon that raised the issue of zero-shot cross-lingual transfer: for the *women* domain, the models tested on Spanish and Italian test sets in a zero-shot setting have much lower scores compared to the *immigrants* domain (the cells are highlighted in red in the table). This is caused, as demonstrated by Nozza (2021), by the language-specific taboo interjections that lead the model to wrongly classify text as hateful towards women.

On a side note, models trained and tested on the English corpus on the immigrants domain have particularly low scores. This phenomenon was also observed by Nozza (2021), and is explained by the authors by the presence of specific words and hashtags that were used for scraping the tweets and that lead the model to over-fit. This issue is further explored in the Discussion section.

We build on these results to set up an experiment pipeline to study the impact of auxiliary task training on this problem. We describe the multi-task training pipeline in the following section. We experiment with tasks that rely on syntactic information, sentiment analysis which is related to hate speech detection, and named entity recognition. By using data for auxiliary tasks in both the source and the target language, we expect the auxiliary task training to work as a bridge between the source and target language, helping the cross-lingual transfer by providing more information on the target language and the difference between the two languages.

4 Auxiliary tasks training

We define several training tasks whose effect on cross-lingual transfer of hate speech detection models is to be evaluated: a sequence-level task, sentiment analysis, and several token-level tasks: Named Entity Recognition (NER) and a set syntactic tasks that we gather – by misnomer – under the term "Universal Dependency" (UD). The sentiment analysis and NER tasks allow the model to learn high-level, semantic information, while the UD tasks convey syntactic skills to the model.

4.1 Auxiliary tasks

Syntactic tasks. We investigate the effect of adding syntactic information by using all Universal Dependency (UD, Nivre et al., 2020) tasks (Dependency Parsing, Part-Of-Speech (POS) tagging, lemmatization and morphological tagging). We use the dataset EWT (Silveira et al., 2014), GSD and ISDT (Bosco et al., 2014), for English, Spanish and Italian respectively. The datasets being of different sizes, we sample them to obtain the same training size in all languages. We use a train set size of 12 543 lines, the size of the smallest dataset. Detailed statistics about the datasets can be found in Table 10 in Appendix A.

Sentiment analysis. We use Twitter sentiment analysis datasets on each of our three target languages.[6] They have been gathered and unified by Barbieri et al. (2021), with a unique split size (training 1 839, development 324, test 870) and a balanced distribution across the three sentiment labels (positive, negative and neutral). Precise statistics and additional information on each dataset can be found in Table 8 in Appendix A.

Named Entity Recognition (NER). An advantage of this task, which consists in identifying entities in a sequence, is that it is more language-agnostic than the others. Indeed, named entities are often transparent between languages, making it a good choice for cross-lingual transfer. We use the NER WikiANN dataset from (Pan et al., 2017;

[6] https://github.com/cardiffnlp/xlm-t

Model	Source lang	immigrants			women		
		en	es	it	en	es	it
m-BERT Nozza (2021)	en	*36.8*	63.3	59.0	*55.9*	54.6	44.9
	es	59.6	*63.0*	68.3	55.8	*83.9*	33.7
	it	63.5	66.6	*77.7*	54.5	46.3	*80.8*
XLM-R (comparable data)	en	*52.8*	44.2	64.6	*49.4*	46.6	44.7
	es	68.0	*75.1*	64.9	47.1	*60.0*	43.8
	it	64.4	44.2	*74.5*	53.3	52.6	*83.7*

(a) Detailed baseline results. Monolingual transfer settings are highlighted in *italic*.

Model	immigrants		women	
	mono	cross	mono	cross
m-BERT	59.2 (16.92)	63.3 (3.37)	73.5 (12.53)	48.3 (7.78)
XLM-R	67.5 (10.38)	58.4 (10.09)	64.4 (14.34)	48.0 (3.67)

(b) Aggregated baseline results, with standard deviation of the scores under each average.

Table 1: Baseline results, both in a detailed and an aggregated fashion. All results except for the one from Nozza (2021) are macro-F1 (%) averaged over 5 runs. All use 20 epochs. Numbers in brown highlight cases when the zero-shot cross-lingual transfer fails.

Rahimi et al., 2019), which covers our three languages. The sets of our languages have a unique split size (training 20 000 examples, development 10 000, test 10 000).

4.2 Multi-task learning pipeline

We perform multi-task learning using the MaChAmp framework; it offers code to fine-tune contextual embeddings for several tasks and several datasets using a shared encoder and different decoders depending on the target task. As our datasets have different sizes, we use a “smooth sampling” method proposed in the framework to avoid having under-represented datasets during training. It consists of re-sampling the datasets according to a multinomial distribution for each batch.

We fine-tune the selected multilingual model, XLM-R, on the different auxiliary tasks. For each auxiliary task, the training is done jointly on the datasets in the three languages, which have been sampled to be of the same size. Only the language of the hate speech detection training corpus varies across the experiments.

All combinations of the three auxiliary tasks are tested. In practice, the language model can be trained on the auxiliary tasks either in an intermediary fashion before being fine-tuned on the downstream task (“sequential” setting, similarly to Pruksachatkun et al. (2020)), or jointly with the hate speech detection task (“joint” setting). According to our experiments, the latter shows the best performance; we report only results with joint training in the paper. All results involving hate speech are obtained using the pipeline described in Section 3.2, averaging the macro-F1 over five different runs. The last encoder layer of the language model is used to compute the scores.

Auxiliary Task	immigrants		women	
	mono	cross	mono	cross
XLM-R				
None	67.5	58.4	64.4	48.0
Sent	3.0	3.7	1.7	-2.0
UD	2.9	-2.5	1.8	-3.2
NER	0.9	3.1	2.8	1.7
Sent+UD	4.0	5.1	1.1	-2.0
Sent+NER	0.2	3.7	3.3	-2.0
UD+NER	-1.0	-1.6	2.0	-6.5
Sent+UD+NER	-0.1	5.2	2.8	-2.6
XLM-T				
None	67.8	60.3	68.1	53.7

Table 2: Hate speech detection macro-F1 scores (%) of the XLM-R and XLM-T baselines, and deltas between the score of each XLM-R model fine-tuned on the auxiliary tasks and the XLM-R baseline. Green scores indicates when auxiliary task training has a positive impact on hate speech detection. *Sent* stands for Sentiment.

5 Results

We analyze the training effect of different combinations of auxiliary tasks on top of XLM-R, jointly with monolingual hate speech detection. Aggregated results for all tasks combinations can be found in Table 2. Instead of the raw scores, we compute the deltas between the baseline system (no auxiliary task) and the augmented system with training on various auxiliary tasks. For each topic (immigrants and women), we average the deltas by settings (monolingual and zero-shot cross-lingual), as explained for Table 1b. For all of our experiments, the non-aggregated results with the full scores instead of the delta can be found in Appendix C, Table 14, Table 19 and Table 15.

Training on auxiliary tasks globally improves monolingual hate speech detection in both domains, immigrants and women (Table 2, columns *mono*).

In the zero-shot cross-lingual transfer scenario (*cross*), we hypothesized that the additional information on the source and target languages could bridge the gap between the languages and improve the transfer for hate speech detection. Looking at the aggregated scores in Table 2, for the *immigrants* domain, all the auxiliary tasks lead to an improvement, except for UD. Jointly training on our three tasks gives the best improvements. However, for the *women* domain, auxiliary task training often has a detrimental effect on zero-shot cross-lingual transfer; for example, Table 3 shows the results on the problematic cases for the sentiment analysis auxiliary task. We only have an improvement when using the NER task. The cultural gap between the languages for this domain appears to be the cause of this discrepancy.

As cross-lingual transfer using UD as auxiliary task often leads to a degradation of the performances of hate speech detection, we performed an ablation study to investigate the effect of lemmatization, dependency parsing, and POS-tagging separately (see Table 18 in Appendix C). Most UD tasks lead to similarly low performances; only dependency parsing has a slightly better impact on hate speech detection.

6 Impact of Language Model Pre-training on Auxiliary Tasks

6.1 Randomly initialized language model

Following the behavioral approach of Muller et al. (2021b) in order to identify the effect of the pre-trained contextual embeddings on our auxiliary tasks, we randomly initialize the parameters of all layers of XLM-R. Then, we fine-tune this model on our auxiliary tasks and conduct the same experiment as in the previous section.

The macro-F1 scores are obviously lower than XLM-R scores due to the random initialisation. Moreover, the impact of the different auxiliary tasks (the deltas with the baseline score, with no auxiliary task) is most of the time lower when randomly initializing the XLM-R model. However, for the zero-shot cross-lingual transfer problematic cases, the effect of auxiliary tasks training is almost always higher. Thus, the XLM-R pre-training with a large amount of out-of-domain textual data seems to be detrimental to the knowledge transfer of the auxiliary tasks when a cultural gap between languages hinders the zero-shot cross-lingual transfer. Table 20 in Appendix shows this behaviour for the aggregated results. A detailed example with sentiment analysis as auxiliary task for these problematic cases can be found in Table 3.

6.2 Using Twitter-specific language models and training datasets

We experiment with XLM-T (Barbieri et al., 2021), an off-the-shelf XLM-R model fine-tuned on 200 million tweets (1,724 million tokens) scraped between 05/2018 and 03/2020, in more than 30 languages, including our three target languages. As expected, without auxiliary task training, the in-domain model leads to significantly better scores overall compared to XLM-R (see the last line of Table 2). However, the impact of auxiliary tasks on the performance of hate speech detection is comparable to the one observed with XLM-R: positive in all settings except for the zero-shot cross-lingual transfer in the women domain. One of the only auxiliary tasks for which the auxiliary training benefits from the Twitter-fine-tuned language model is sentiment analysis (Table 3, aggregated results for all tasks can be found in Table 20 in the Appendix).

Contrarily to Wikipedia, where data are highly similar from one high-resource language to another, Twitter data can significantly differ between languages due to cultural differences and events in the respective countries. This issue is further discussed in the next section.

Given the positive impact of using a language model fine-tuned on Twitter data, we also use Twitter data for our auxiliary tasks. The sentiment analysis corpora are already built from tweets. However, very few Twitter datasets are annotated in Universal Dependencies or NER.

For our UD experiments, we use two corpora of small size, in English and Italian only. We add to our existing out-of-domain UD corpora for auxiliary training given their reduced size. However, training on these augmented UD datasets does not lead to any significant improvement. Details on the Twitter UD datasets can be found in Table 9, and hate speech detection results in Table 16 in Appendix.

For our NER experiments, we use the WNUT16 NER shared task[7] Twitter dataset (Strauss et al., 2016), only available in English (cf. Table 11 for statistics of this dataset). Auxiliary training on this dataset has a significant positive impact on hate

[7] The task focuses on finding 10 types of target entities, including company, facility, geo-location, movie, music-artist, other, person, product, sport team and TV show

Model	Source lang	es	it
XLM-R	en	44.9	44.1
	es	61.1	34.2†
	it	47.3‡	83.4
XLM-T	en	50.0	64.3
	es	62.9	45.9†
	it	48.0‡	84.8†
XLM-R-random-init	en	44.2	49.9
	es	51.7	56.0‡
	it	46.8	83.4

Table 3: Hate speech detection macro-F1 scores (%) of all 3 models trained jointly with **sentiment analysis**, for the problematic cases. Each model is compared with its associated baseline, green scores indicates when auxiliary task training has a positive impact on hate speech detection, red otherwise. The subscript indicates whether the score is significantly higher or lower compared to the baseline (see Appendix B).

speech detection; the impact is higher than auxiliary training on the out-of-domain NER corpus (NER WikiANN), even for non-English test sets. Results can be found in Table 17 in Appendix.

7 Discussion

On syntactic information, and the impact of UD auxiliary task training. We demonstrate the effect variation when jointly training hate speech detection with auxiliary tasks: sentiment analysis, NER and UD tasks. Compared with the first two tasks, adding syntactic information has the lowest positive impact on hate speech detection, often decreasing the performance for zero-shot cross-lingual settings. This is in line with results from the literature that agree on the positive effect on sentiment analysis (del Arco et al., 2021; Aroyehun and Gelbukh, 2021), but face varying conclusions when it comes to UD tasks. Narang and Brew (2020) show the positive impact of syntactic features on top of non-contextualized embeddings for hate speech detection; Gambino and Pirrone (2020), among the best systems on the EVALITA2020 hate speech detection task, use POS-tagged text as input for classification. On the contrary, in a monolingual setting, Klemen et al. (2020) show that morphological features added to LSTM and BERT-based hate speech detection models do not help for comment filtering. Similarly, using sequential auxiliary training of tasks such as POS tagging, Pruksachatkun et al. (2020) show that the resulting additional low-level information often leads to negative transfer for many downstream tasks.

In our cross-lingual setting, our goal was to use these tasks as a proxy to fill the mismatch between languages and facilitate the transfer. We hypothesize that when working on tweets, their constrained styles – short sentences, generally with low syntactic complexity – make additional syntactic knowledge unhelpful for a downstream task such as hate speech detection, which benefits more from semantic information.

On the quality of training datasets. We started this work to understand the cultural gap in multilingual hate speech datasets highlighted by Nozza (2021). However, our results on the impact of auxiliary task training to bridge this cultural gap have to be mitigated with the limitations of the datasets.

First, the baseline macro-F1 score for the hate speech detection task (Table 1a) are often quite low, even when the models are tested and evaluated on the same language (e.g. a score of 59.2 on average on the *immigrants* domain using m-BERT). The scores of hate speech detection models trained on the immigrants-English dataset are particularly low (Table 1a). This might be due to the specificity of this dataset: Two measures, namely the average number of hashtags per tweet and the total number of unique hashtags, show the large discrepancy in hashtag distributions compared to the other datasets (e.g. 2193 unique hashtags in the immigrants English dataset, compared to only 301 in Spanish). Both statistics are around five times stronger in English compared to the other languages (detailed statistics can be found in Table 5 in the Appendix). To confirm the influence of the source language hashtag distribution on the model cross-lingual performance, we removed all hashtags in all datasets, and compared the hate speech detection score with the ones on the datasets with hashtags[8] (see Table 6 in Appendix). Removing hashtags led to an overall improvement of macro-F1 scores when training models on the immigrants English dataset, as noted by Nozza (2021). The results are similar when training on the immigrants Italian data, which has the second-highest number of unique hashtags (693). In a shared domain (e.g. international football events, worlwide movie release, etc.) , hashtags are often the same between languages; thus, removing them should intuitively decrease the cross-lingual transferability of the models. Here, our results indicate that too many hashtags, even

[8] Recall that the default preprocessing in this work is the segmentation of hashtags into words.

segmented as words, possibly lead the language model to overfit and hinder a tweet-level task's accuracy, such as hate speech detection.

Cross-lingual zero-shot transfer on a domain with cultural gap between languages. We highlight that despite all the performance improvements from auxiliary task training for monolingual hate speech detection and cross-lingual zero-shot transfer, some "problematic cases" remain: situations when cultural variations negatively impact the model's ability to transfer knowledge from one language to another. It is especially the case in the *women* domain.

Using XLM-R, the NER auxiliary tasks (with both WikiANN and WNUT16 datasets) lead to the best improvement of hate speech detection for cultural gap cases. Indeed, the cross-lingual transferability of NER is facilitated by the fact that many named entities are the same across languages (e.g. person and organisation names); besides, many successful unsupervised cross-lingual transfer systems for this task can be found in the literature (Rahimi et al., 2019; Bari et al., 2020). Thus, this task is well-suited to bridge the gap between languages suffering from a variation between their hate speech. The other efficient method to improve cross-lingual zero-shot transfer is to use a language model fine-tuned on Twitter data. With and without auxiliary tasks fine-tuning, we showed this adaptation's significant and consistent positive impact. This is in line with the findings of Muller et al. (2021a); similarly, Bose et al. (2021) show the superiority of MLM over other tasks in a cross-corpora transfer setting. In a cross-lingual setting, van der Goot et al. (2021a) jointly train auxiliary tasks with a downstream task (in their case, spoken language understanding) to find that MLM fine-tuning consistently improves the downstream task.

Beyond the obvious improvement due to the MLM training on more adapted data, we would have expected XLM-T to increase the impact of auxiliary tasks fine-tuning for the problematic cases, cross-lingual transfer on the women domain; a more adapted language model helping to bridge the gap between hate speech in the source and target languages. Here, the Twitter data used for the XLM-T training may not be optimal for the observed cultural gap. It was trained on tweets published between 05/2018 and 03/2020, while the hate speech corpora range from 2017 to 2018, depending on the language; moreover, some events were specifically targeted when scraping Twitter for hate speech detection (e.g., Gamergate victims for the Italian datasets on hate speech towards women (Fersini et al., 2018)). Overall, when we use XLM-T, we only adapt the model to the form and style of Twitter data (small sentences, with mentions and urls. . .). The tweets' content, topic, and vocabulary might differ a lot between the hate speech corpora, the XLM-T training data, and the sentiment analysis corpora. We can only hypothesize on these variations. However, they should be quantified to understand better the impact of fine-tuning on these data and to distinguish this phenomenon from the actual cultural gap between languages.

8 Conclusion

We showed the positive impact of jointly training hate speech detection with NER, sentiment analysis, and UD tasks. We highlighted problematic cases where zero-shot cross-lingual transfer of hate speech models towards women fails because of the cultural gap between languages. In these cases, we identify the two most efficient solutions: auxiliary training on the NER task and using a language model fine-tuned on more adapted data. Finally, we discussed limitations related to training data for language model pre-training, auxiliary tasks, and hate speech detection. In particular, we highlighted how Twitter-specific fine-tuning acts as a "knowledge proxy" to bridge the gap between languages. In cases when not enough training data is available for domain-adapted MLM fine-tuning, tasks such as NER allow the model to discern common patterns between the source and target languages and transfer knowledge from one to another.

In a low-resource situation where in-domain hate speech data is not available for training, generating synthetic hate speech data has been investigated to improve training (Wullach et al., 2021) and for domain adaptation (Sarwar and Murdock, 2021), but only in a monolingual setting, for English. We plan to investigate this research avenue in future work. Meanwhile, we will release our code and models upon publication in the hope that the community will investigate the best way to tackle this cultural gap between languages for hate speech detection.

9 Ethical considerations

This paper is part of a line of work aiming to tackle hate speech detection when we have no training data in the target language, fight the spread of offensive and hateful speech online, and have a positive global impact on the world. Its goal is to understand if hate speech is transferable from a language to another; as such, it has been approved by our institutional review board (IRB), and follows the national and European General Data Protection Regulation (GDPR).

We did not collect any data from online social media for this work. We only used publicly available datasets – exclusively diffused for shared tasks that were tackled by a large number of participants (see Table 7 in Appendix A). These datasets do not include any metadata, only the tweet's text associated with the hate speech label. Thus, linking the annotated data to individual social media users is not straightforward.

All our experiments were executed on clusters whose energy mix is made of nuclear (65–75%), 20% renewable, and the remaining with gas (or more rarely coal when imported from abroad). More details on computational costs can be found in Table 12 in Appendix B.

The presence of bias in the pre-trained languages models we use, due to the bias in the data they were trained on, may have an impact on the hate speech detection; particularly on the topic of hate speech towards women. This area of research is currently under heavy scrutiny by the community.

References

Segun Taofeek Aroyehun and Alexander Gelbukh. 2021. Evaluation of intermediate pre-training for the detection of offensive language. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021), CEUR Workshop Proceedings. CEUR-WS. org*.

Francesco Barbieri, Valerio Basile, Danilo Croce, Malvina Nissim, Nicole Novielli, and Viviana Patti. 2016. Overview of the evalita 2016 sentiment polarity classification task. In *Proceedings of third Italian conference on computational linguistics (CLiC-it 2016) & fifth evaluation campaign of natural language processing and speech tools for Italian. Final Workshop (EVALITA 2016)*.

Francesco Barbieri, Luis Espinosa-Anke, and Jose Camacho-Collados. 2021. A Multilingual Language Model Toolkit for Twitter. In *arXiv preprint arXiv:2104.12250*.

M Saiful Bari, Shafiq Joty, and Prathyusha Jwalapuram. 2020. Zero-resource cross-lingual named entity recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7415–7423.

Angelo Basile and Chiara Rubagotti. 2018. Crotonemilano for ami at evalita2018. a performant, cross-lingual misogyny detection system. *EVALITA Evaluation of NLP and Speech Tools for Italian*, 12:206.

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

JM Berger and Heather Perez. 2016. The Islamic State's diminishing returns on Twitter: How suspensions are limiting the social networks of English-speaking ISIS supporters. Technical report, George Washington University.

Cristina Bosco, Felice Dell'Orletta, Simonetta Montemagni, Manuela Sanguinetti, and Maria Simi. 2014. The evalita 2014 dependency parsing task. *The Evalita 2014 Dependency Parsing task*, pages 1–8.

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

Elena Cabrio, Julien Cojan, and Fabien Gandon. 2014. Mind the cultural gap: Bridging language-specific dbpedia chapters for question answering. In *Towards the Multilingual Semantic Web*, pages 137–154. Springer.

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 the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 11.

Flor Miriam Plaza del Arco, Sercan Halat, Sebastian Padó, and Roman Klinger. 2021. Multi-task learning with sentiment, emotion, and target detection to recognize hate speech and offensive language. In *Forum for Information Retrieval Evaluation, Virtual Event*.

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

Manuel Carlos Díaz Galiano, Eugenio Martínez Cámara, Miguel Ángel García Cumbreras, Manuel García Vega, and Julio Villena Román. 2018. The democratization of deep learning in tass 2017. -.

Karthik Dinakar, Roi Reichart, and Henry Lieberman. 2011. Modeling the detection of textual cyberbullying. In *fifth international AAAI conference on weblogs and social media*.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the evalita 2018 task on automatic misogyny identification (ami). *EVALITA Evaluation of NLP and Speech Tools for Italian*, 12:59.

Komal Florio, Valerio Basile, Marco Polignano, Pierpaolo Basile, and Viviana Patti. 2020. Time of your hate: The challenge of time in hate speech detection on social media. *Applied Sciences*, 10:4180.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Giuseppe Gambino and Roberto Pirrone. 2020. Chilab@ haspeede 2: Enhancing hate speech detection with part-of-speech tagging. -.

Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*, pages 1–6, Melbourne, Australia. Association for Computational Linguistics.

Dagmar Gromann and Thierry Declerck. 2017. Hashtag processing for enhanced clustering of tweets. In *RANLP*, pages 277–283.

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

Matej Klemen, Luka Krsnik, and Marko Robnik-Šikonja. 2020. Enhancing deep neural networks with morphological information. *arXiv preprint arXiv:2011.12432*.

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.

Ilia Markov, Nikola Ljubešić, Darja Fišer, and Walter Daelemans. 2021. Exploring stylometric and emotion-based features for multilingual cross-domain hate speech detection. In *Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*, pages 149–159, Online. Association for Computational Linguistics.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021a. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online. Association for Computational Linguistics.

Benjamin Muller, Yanai Elazar, Benoît Sagot, and Djamé Seddah. 2021b. First align, then predict: Understanding the cross-lingual ability of multilingual BERT. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2214–2231, Online. Association for Computational Linguistics.

Kanika Narang and Chris Brew. 2020. Abusive language detection using syntactic dependency graphs. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 44–53, Online. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Debora Nozza. 2021. Exposing the limits of zero-shot cross-lingual hate speech detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 907–914, Online. Association for Computational Linguistics.

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 the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 363–370, Florence, Italy. Association for Computational Linguistics.

Xiaoman Pan, Boliang Zhang, Jonathan May, Joel Nothman, Kevin Knight, and Heng Ji. 2017. Cross-lingual name tagging and linking for 282 languages. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1946–1958, Vancouver, Canada. Association for Computational Linguistics.

Andraž Pelicon, Ravi Shekhar, Blaž Škrlj, Matthew Purver, and Senja Pollak. 2021. Investigating cross-lingual training for offensive language detection. *PeerJ Computer Science*, 7:e559.

Jason Phang, Iacer Calixto, Phu Mon Htut, Yada Pruksachatkun, Haokun Liu, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. English intermediate-task training improves zero-shot cross-lingual transfer too. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 557–575, Suzhou, China. Association for Computational Linguistics.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Yada Pruksachatkun, Jason Phang, Haokun Liu, Phu Mon Htut, Xiaoyi Zhang, Richard Yuanzhe Pang, Clara Vania, Katharina Kann, and Samuel R. Bowman. 2020. Intermediate-task transfer learning with pretrained language models: When and why does it work? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5231–5247, Online. Association for Computational Linguistics.

Afshin Rahimi, Yuan Li, and Trevor Cohn. 2019. Massively multilingual transfer for NER. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 151–164, Florence, Italy. 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, Online. Association for Computational Linguistics.

Guilherme Moraes Rosa, Luiz Henrique Bonifacio, Leandro Rodrigues de Souza, Roberto Lotufo, and Rodrigo Nogueira. 2021. A cost-benefit analysis of cross-lingual transfer methods. *arXiv preprint arXiv:2105.06813*.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. Semeval-2017 task 4: Sentiment analysis in twitter. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 502–518.

Paul Röttger, Bertie Vidgen, Dong Nguyen, Zeerak Waseem, Helen Margetts, and Janet Pierrehumbert. 2021. HateCheck: Functional tests for hate speech detection models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 41–58, Online. Association for Computational Linguistics.

Sheikh Muhammad Sarwar and Vanessa Murdock. 2021. Unsupervised domain adaptation for hate speech detection using a data augmentation approach. -.

Natalia Silveira, Timothy Dozat, Marie-Catherine de Marneffe, Samuel Bowman, Miriam Connor, John Bauer, and Chris Manning. 2014. A gold standard dependency corpus for English. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2897–2904, Reykjavik, Iceland. European Language Resources Association (ELRA).

Sara Sood, Judd Antin, and Elizabeth Churchill. 2012. Profanity use in online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 1481–1490, New York, NY, USA. Association for Computing Machinery.

Benjamin Strauss, Bethany Toma, Alan Ritter, Marie-Catherine De Marneffe, and Wei Xu. 2016. Results of the wnut16 named entity recognition shared task. In *Proceedings of the 2nd Workshop on Noisy User-generated Text (WNUT)*, pages 138–144.

Rob van der Goot, Nikola Ljubešić, Ian Matroos, Malvina Nissim, and Barbara Plank. 2018. Bleaching text: Abstract features for cross-lingual gender prediction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 383–389, Melbourne, Australia. Association for Computational Linguistics.

Rob van der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Üstün, Marija Stepanović, Alan Ramponi, Siti Oryza Khairunnisa, Mamoru Komachi, and Barbara Plank. 2021a. From masked language modeling to translation: Non-English auxiliary tasks improve zero-shot spoken language understanding. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2479–2497, Online. Association for Computational Linguistics.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021b. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Bertie Vidgen, Alex Harris, Dong Nguyen, Rebekah Tromble, Scott Hale, and Helen Margetts. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 80–93, Florence, Italy. Association for Computational Linguistics.

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

Tomer Wullach, Amir Adler, and Einat Minkov. 2021. Fight fire with fire: Fine-tuning hate detectors using large samples of generated hate speech. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4699–4705, Punta Cana, Dominican Republic. Association for Computational Linguistics.

A Datasets overview

A.1 Hate speech datasets overview

Domain-language	train	dev	test	blind
immigrants-it	2000	500	1000	.
immigrants-en	4500	500	1499	.
immigrants-es	1618	173	800	.
women-it	2500	500	1000	.
women-en	4500	500	1472	.
women-es	2882	327	799	.
Comparable size	1618	173	800	1000

Table 4: Hate speech detection datasets: Size of full datasets (number of sentences) and new split with comparable data size. Only the immigrants-es dataset has no blind set.

	immigrants			women		
	en	es	it	en	es	it
Nb of tokens per tweet						
avg	26.1	21.5	17.4	17.9	20.9	18.6
median	25	19	18	18	19	15
max	55	59	30	63	59	55
min	1	1	2	2	1	2
Nb of hashtags (avg per tweet, total unique nb)						
avg	1.33	0.2	0.6	0.18	0.18	0.2
unique	2193	301	693	510	451	369
Train/test OOV Ratio						
	0.34	0.43	0.41	0.41	0.51	0.48

Table 5: Descriptive statistics on hate speech detection training datasets.

A.2 Auxiliary tasks datasets overview

Treebanks additional pre-processing As the MACHAMP framework does not support the Connl UD format, treebanks must be converted back to the connl06 format, which most notably involved the removal of all contracted tokens, potentially leading to tokenization mismatches between our data sources. However, a rapid analysis showed that it has a very limited impact because of their low frequency and the generalization of sub-word tokenization.

B Experimental Details.

Language Models. In this paper, we use two language models: m-BERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020). The former is the multilingual version of BERT, trained on Wikipedia content in 104 languages, with 100 parameter. The latter has the same architecture as RoBERTa (Liu et al., 2019) with 550M parameters, and is trained on the publicly available 2.5 TB CommonCrawl Corpus, covering 100 languages. The checkpoints are loaded from the Transformers library. We conducted experiments with scalar mixing the 12 encoder layers of the language model XLM-R instead of using only the last one, without obtaining better performances in the downstream task on average.

Statistical testing. To increase the robustness of the results, we use five different seeds when fine-tuning a language model on the hate speech detection task and report the average macro-F1 over the five runs. In all results tables, for each score of experiments using auxiliary tasks, the subscript indicates whether the score is significantly higher or lower compared to the baseline. The comparison is made using a one-sided t-test over the list of scores of the five runs of each model.[12] A dagger (†) as exponent indicates that the p-value is smaller than 0.05, while a double-dagger (‡) indicates a p-value smaller than 0.01.

Computational Costs. We conduct our experiments on RTX8000 GPUs. Our experiments were performed on a cluster hosted in a country with a nuclear mix of 80%. We have three models (XLM-R, XLM-T and XLM-R with random initialization) that we test in 7 different auxiliary tasks com-

[12] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.ttest_ind.html.

Preprocessing	Source lang	immigrants			women		
		en	es	it	en	es	it
segmented-hashtags	en	52.8	44.2	64.6	49.4	46.6	44.7
	es	68.0	75.1	64.9	47.1	60.0	43.8
	it	64.4	44.2	74.5	53.3	52.6	83.7
removed-hashtags	en	56.6 (3.8)	50.7 (6.5)	68.2 (3.6)	49.2 (-0.3)	44.7 (-1.8)	48.1 (3.4)
	es	54.0 (-14.0)	64.4 (-10.7)	55.9 (-8.9)	50.8 (3.6)	52.6 (-7.4)	43.3 (-0.5)
	it	65.1 (0.7)	48.5 (4.2)	74.6 (0.1)	54.0 (0.7)	51.0 (-1.6)	83.9 (0.2)

Table 6: Comparison of hate speech detection depending on the hashtags processing method: segmented-hashtags (removing the "#" and tokenizing the hashtag to divide it into words) and removed-hashtags (removing all hashtags). Number in parenthesis are the deltas with the scores on segmented-hashtags data. The colour indicated whether the deltas are positive or negatives.

Shared task	Link
Hateval	https://github.com/msang/hateval
EVALITA AMi 2018	https://github.com/MIND-Lab/ami2018
HaSpeeDe 2018	https://github.com/msang/haspeede/tree/master/2018

Table 7: Shared tasks used for the Hate speech corpora.

Language	Shared task	Reference	Scraping period
English	SemEval 2017	Rosenthal et al. (2017)	01/2012–12/2015
Italian	Intertass 2017	Díaz Galiano et al. (2018)	07/2016–01/2017
Spanish	Sentipolc 2016	Barbieri et al. (2016)	2013–2016

Table 8: Data overview for the sentiment analysis task. All datasets contain text scraped from Twitter. They have been unified to a common train / dev / test split size: 1 839 / 324 / 870.

Dataset	Language	train/dev/test size	Period
Tweebank	English	1,639 / 710 / 1,201	02/2016 – 07/2016
PoSTWITA	Italian	5,368 / 671 / 674	07/2009 – 02/2013

Table 9: Twitter UD data overview.

Dataset	Language	train	dev	test
EWT[9]	English	12 543	2 001	2 077
GSD[10]	Spanish	14 187	1 400	426
ISDT[11]	Italian	13 121	564	482
Comparable size		12543	564	426

Table 10: Universal Dependencies (UD) datasets and size of their respective splits.

	Train	Dev
# tweets	2,349	1,000
# tokens	46,469	16,261
# entity tokens	2,462	1,128

Table 11: Statistics of the WNUT 2016 NER shared task dataset.

Task	Duration
Hate only	0:14
Sentiment+Hate	0:21
UD+Hate	1:57
NER+Hate	2:18

Table 12: Duration of training one seed per model

binations, with 5 seeds each. Details on the average GPU time for the basic task combinations (jointly training hate speech with one task) are in Table 12.

C Complementary results

C.1 Ablation study on UD tasks.

As cross-lingual transfer using UD as an auxiliary task often leads to a degradation of the performances of hate speech detection, we investigate the effect of lemmatization, dependency parsing, and POS-tagging separately. Table 18 shows the full results for all tasks.

The results with each task separately are sensibly the same as with the 4 tasks trained jointly, often with no significant change, sometimes with

	Model	Source lang	immigrants			women		
			en	es	it	en	es	it
Full raw datasets	m-BERT Nozza (2021)	en	36.8	63.3	59.0	55.9	54.6	44.9
		es	59.6	63.0	68.3	55.8	83.9	33.7
		it	63.5	66.6	77.7	54.5	46.3	80.8
	m-BERT	en	32.7	47.0	67.1	52.4	43.7	45.4
		es	68.0	81.2	64.6	40.3	61.9	38.9
		it	62.4	48.4	77.3	52.8	51.6	84.8
	XLM-R	en	36.0	50.1	68.5	54.9	45.9	50.7
		es	54.6	74.2	60.1	46.3	62.9	40.9
		it	73.2	51.4	77.9	56.4	53.5	85.2
Full processed datasets	m-BERT	en	38.4	41.0	68.6	54.0	40.5	36.2
		es	68.2	82.3	63.6	47.4	64.7	40.5
		it	71.3	48.3	78.3	46.9	49.2	84.4
	XLM-R	en	38.8	42.4	66.8	55.1	46.4	51.8
		es	67.3	81.4	64.8	48.4	65.3	38.2
		it	71.2	51.1	77.8	57.3	55.6	85.2
Comparable processed datasets	m-BERT	en	43.3	46.5	64.5	51.2	41.2	42.7
		es	70.0	82.9	64.7	45.7	64.6	37.7
		it	71.6	46.8	77.4	48.9	48.9	84.5
	XLM-R	en	52.8	44.2	64.6	49.4	46.6	44.7
		es	68.0	75.1	64.9	47.1	60.0	43.8
		it	64.4	44.2	74.5	53.3	52.6	83.7

Table 13: Baseline results on full un-processed (raw) datasets, full pre-processed datasets (special tokens for mentions and urls, hashtags segmentation), and comparable pre-processed datasets. All results except for the one from Nozza (2021) are macro-F1 (%) averaged over 5 runs. All use 20 epochs. Numbers in brown highlight cases when the zero-shot cross-lingual transfer fails.

performance degradation compared to the absence of auxiliary task. In the case of monolingual hate speech detection, UD auxiliary training sometimes degrades the hate speech detection performance, but mostly for English. Overall, UD-dependency auxiliary task training leads to the highest and most consistent improvement compared to the baseline.

C.2 In-domain UD datasets

We use two Twitter datasets annotated in Universal Dependencies: the Tweebank corpus[13] in English, and PoSTWITA [14] in Italian. Detailed information can be found in Table 9. As stated before, we process the treebanks to remove the cases where words are splitted or inserted. This situation happens for 17% of the lines in the raw English Treebank and 14% on the Italian one; we conclude that the impact of this processing should be small.

We add these datasets, sampled to around 1600 sentences – the size of the smallest one – to our UD corpora. Training on this augmented dataset does not improve the hate speech classification transfer compared to the original UD datasets. However, the small size of this additional data, besides the fact that one language is missing (Spanish), does not allow us to draw a meaningful conclusion on the impact of domain-specific UD data on cross-lingual transfer.

[13] https://github.com/Oneplus/Tweebank

[14] https://github.com/UniversalDependencies/UD_Italian-PoSTWITA

Auxiliary task	Source lang	immigrants			women		
		en	es	it	en	es	it
None	en	52.8	44.2	64.6	49.4	46.6	44.7
	es	68.0	75.1	64.9	47.1	60.0	43.8
	it	64.4	44.2	74.5	53.3	52.6	83.7
sentiment	en	50.1	55.8‡	67.7†	53.6	44.9	44.1
	es	70.6†	85.1	64.1	57.3†	61.1	34.2†
	it	65.9	48.5	76.2	48.1‡	47.3‡	83.4
UD	en	50.3	41.9	60.2†	54.9	43.0	50.4
	es	61.0†	84.4	60.7	54.3†	58.8	34.9†
	it	62.3	49.4	76.2	36.9‡	49.3	84.6
NER	en	43.7‡	46.9	69.3‡	52.0	46.8	58.0†
	es	70.9†	85.0	63.5	43.1	65.4†	41.5
	it	63.3	54.9†	76.4	55.1	53.7	84.1
sentiment+UD	en	53.4	56.3‡	71.8‡	52.7	51.7†	47.7
	es	66.7	85.1	66.8†	52.8	59.6	30.9†
	it	65.7	53.9†	76.0	46.3‡	47.0†	84.0
sentiment+NER	en	40.4‡	48.0	70.1‡	56.2	41.3†	48.4
	es	71.2‡	85.5	66.0	45.8	63.5	39.5
	it	66.9	50.3	77.2†	49.4†	51.6	83.3
UD+NER	en	37.9‡	48.1	68.4‡	53.5	44.0	53.4
	es	68.1	84.7	52.6	44.7	62.1	34.5
	it	63.9	54.7†	77.5	31.5‡	46.9	85.7†
sentiment+UD+NER	en	38.2‡	54.5‡	71.5‡	53.9	44.0	53.0
	es	67.1	85.8	67.4†	49.3	63.1†	32.9†
	it	67.4	53.9†	78.0†	45.3‡	48.0†	84.4

Table 14: Hate speech detection macro-F1 scores (%) of XLM-R fine-tuned on various combinations of auxiliary tasks jointly with the hate speech detection task. We compare each macro-F1 score with the baseline score (without auxiliary task). Green values indicate an increase in score, red values a decrease. The subscript indicates whether the macro-F1 of the model trained with the auxiliary tasks is significantly higher or lower compared to the model without auxiliary task. The comparison is made using a one-sided t-test over the list of scores of the five runs of each model. A dagger (†) as exponent indicates that the p-value is smaller than 0.05, while a double-dagger (‡) indicates a p-value smaller than 0.01.

Auxiliary task	Source lang	immigrants			women		
		en	es	it	en	es	it
XLMR-none	en	52.8	44.2	64.6	49.4	46.6	44.7
	es	68.0	75.1	64.9	47.1	60.0	43.8
	it	64.4	44.2	74.5	53.3	52.6	83.7
None	en	46.4	56.7	59.8	60.7	49.6	61.1
	es	65.7	84.2	66.2	60.6	59.9	38.8
	it	58.9	54.6	72.8	57.0	54.8	83.7
sentiment	en	43.8	58.6	70.2†	57.0†	50.0	64.3
	es	65.8	86.9‡	67.2	54.0†	62.9	45.9†
	it	58.6	55.7	76.2‡	56.9	48.0‡	84.8†
UD	en	52.2	57.6	65.4	57.4†	43.7‡	59.2
	es	65.5	85.6†	62.0†	55.7	58.3	33.0†
	it	55.6	54.3	73.9	36.1‡	45.3‡	85.3‡
NER	en	42.3	51.2	69.4†	59.3	42.3‡	59.4
	es	67.9	86.1‡	66.0	63.4	64.4†	43.4
	it	66.2†	61.8†	78.5‡	52.7	57.6†	85.5‡
sentiment+UD	en	45.7	53.2	68.6†	59.5	46.5	64.4
	es	65.9	87.1‡	67.0	57.6	61.8	37.9
	it	62.0	55.0	76.7‡	54.3	48.4‡	84.8†
sentiment+NER	en	40.7†	53.2	70.3†	60.5	43.8†	56.7
	es	69.9‡	86.9‡	66.9	61.4	66.5‡	43.6
	it	65.5	58.9	77.9‡	56.7	51.0‡	84.6
sentiment+UD+NER	en	38.7‡	56.4	71.2†	59.0	46.0†	57.6
	es	67.0	86.3†	67.8	58.7	64.8†	39.7
	it	61.4	55.6	79.1‡	54.9	47.7‡	84.6

Table 15: Hate speech detection macro-F1 scores (%) of XLM-T fine-tuned on various combinations of auxiliary tasks jointly with the hate speech detection task. We compare each macro-F1 score with the baseline score (without auxiliary task). Green values indicate an increase in score, red values a decrease. The subscript indicates whether the macro-F1 of the model trained with the auxiliary tasks is significantly higher or lower compared to the model without auxiliary task. The comparison is made using a one-sided t-test over the list of scores of the five runs of each model. A dagger (†) as exponent indicates that the p-value is smaller than 0.05, while a double-dagger (‡) indicates a p-value smaller than 0.01.

Source lang	immigrants			women		
	en	es	it	en	es	it
en	50.1	44.8	64.5	56.5	41.1‡	61.7†
es	62.5‡	83.7	54.5‡	51.3	59.5	36.9
it	56.1	46.0	76.7	41.7‡	48.1	84.7

Table 16: Hate speech detection macro-F1 scores (%) of XLM-T fine-tuned on **UD Twitter** jointly with the hate speech detection task. We compare each macro-F1 score with the baseline score (without auxiliary task). Green values indicate an increase in score, red values a decrease. The subscript indicates whether the macro-F1 of the model trained with the auxiliary tasks is significantly higher or lower compared to the model without auxiliary task. The comparison is made using a one-sided t-test over the list of scores of the five runs of each model. A dagger (†) as exponent indicates that the p-value is smaller than 0.05, while a double-dagger (‡) indicates a p-value smaller than 0.01.

Source lang	immigrants			women		
	en	es	it	en	es	it
en	47.5	49.7	67.8‡	58.4†	45.7	61.8†
es	66.2	83.1	65.2	47.3	61.9	40.9
it	63.5	50.3	76.2	54.4	54.0	83.7

Table 17: Hate speech detection macro-F1 scores (%) of XLM-T fine-tuned on **NER Twitter** jointly with the hate speech detection task. We compare each macro-F1 score with the baseline score (without auxiliary task). Green values indicate an increase in score, red values a decrease. The subscript indicates whether the macro-F1 of the model trained with the auxiliary tasks is significantly higher or lower compared to the model without auxiliary task. The comparison is made using a one-sided t-test over the list of scores of the five runs of each model. A dagger (†) as exponent indicates that the p-value is smaller than 0.05, while a double-dagger (‡) indicates a p-value smaller than 0.01.

Auxiliary task	Source lang	immigrants			women		
		en	es	it	en	es	it
UD-dependency	en	50.6	45.8	67.4‡	54.5	47.3	52.0
	es	62.3‡	84.9	61.7†	51.2	60.1	34.3†
	it	64.0	48.0	74.8	53.7	51.7	84.2
UD-lemma	en	51.0	37.7†	54.7†	54.4	40.2‡	51.5
	es	61.8‡	84.0	61.8‡	42.7	56.1	33.3†
	it	57.1†	43.0	76.0	51.4	50.6	84.3
UD-UPOS	en	52.9	43.8	64.7	53.6	41.7	50.0
	es	65.6†	84.4	63.6	54.4	56.3	36.1
	it	67.3	44.1	75.3	51.6	57.0‡	84.6
UD-UPOS+dependency	en	53.7	44.4	65.6	50.2	46.7	48.5
	es	61.1‡	83.0	64.4	53.5	59.3	36.3
	it	64.0	47.3	75.7	47.5†	46.7†	83.7

Table 18: **Ablation study**: Hate speech detection macro-F1 scores (%) of XLM-R fine-tuned on the different UD tasks jointly with the hate speech detection task. We compare each macro-F1 score with the baseline score (without auxiliary task). Green values indicate an increase in score, red values a decrease. The subscript indicates whether the macro-F1 of the model trained with the auxiliary tasks is significantly higher or lower compared to the model without auxiliary task. The comparison is made using a one-sided t-test over the list of scores of the five runs of each model. A dagger (†) as exponent indicates that the p-value is smaller than 0.05, while a double-dagger (‡) indicates a p-value smaller than 0.01.

Auxiliary task	Source lang	immigrants			women		
		en	es	it	en	es	it
None	en	44.4	46.7	58.2	53.6	41.2	46.0
	es	52.5	74.4	52.2	47.1	53.2	48.0
	it	52.7	50.7	73.4	47.4	48.3	83.9
sentiment	en	44.5	$53.4^{\dagger}$	$62.2^{\dagger}$	54.7	44.2	49.9
	es	$48.7^{\dagger}$	74.7	$57.5^{\ddagger}$	$41.2^{\ddagger}$	51.7	$56.0^{\ddagger}$
	it	52.6	$54.8^{\ddagger}$	$71.0^{\ddagger}$	$40.0^{\ddagger}$	46.8	83.4
UD	en	$55.2^{\ddagger}$	45.6	52.7	$50.6^{\dagger}$	42.5	42.5
	es	54.7	74.2	49.9	46.7	52.4	$52.5^{\ddagger}$
	it	50.8	48.0	73.4	44.9	45.7	$81.2^{\ddagger}$
NER	en	41.5	46.3	57.2	54.4	$49.4^{\ddagger}$	45.6
	es	$55.2^{\dagger}$	74.1	53.6	48.9	$55.1^{\dagger}$	49.9
	it	50.9	$45.3^{\ddagger}$	$72.5^{\dagger}$	$41.1^{\ddagger}$	50.0	83.9
sentiment+UD	en	$52.5^{\dagger}$	$52.4^{\dagger}$	$61.9^{\dagger}$	51.8	$51.7^{\ddagger}$	48.1
	es	49.6	74.6	54.8	46.1	52.2	$56.6^{\ddagger}$
	it	53.0	$57.8^{\ddagger}$	$72.5^{\dagger}$	$42.5^{\ddagger}$	46.0	$81.6^{\ddagger}$
sentiment+NER	en	38.4	$57.1^{\dagger}$	60.2	49.6	45.6	51.6
	es	56.0	73.3	53.2	42.1	54.9	54.3
	it	$49.3^{\dagger}$	51.8	72.7	$37.5^{\ddagger}$	46.4	$81.7^{\dagger}$
sentiment+UD+NER	en	$37.9^{\dagger}$	$51.2^{\dagger}$	$60.6^{\dagger}$	51.9	$52.7^{\ddagger}$	48.1
	es	51.4	74.2	55.1	47.5	54.8	$56.6^{\ddagger}$
	it	51.8	$54.9^{\dagger}$	$72.0^{\dagger}$	$39.7^{\ddagger}$	45.6	$82.0^{\ddagger}$

Table 19: Hate speech detection macro-F1 scores (%) of a **randomly initialized** XLM-R model, fine-tuned on various combinations of auxiliary tasks jointly with the hate speech detection task. We compare each macro-F1 score with the baseline score (without auxiliary task). Green values indicate an increase in score, red values a decrease. The subscript indicates whether the macro-F1 of the model trained with the auxiliary tasks is significantly higher or lower compared to the model without auxiliary task. The comparison is made using a one-sided t-test over the list of scores of the five runs of each model. A dagger (†) as exponent indicates that the p-value is smaller than 0.05, while a double-dagger (‡) indicates a p-value smaller than 0.01.

Auxiliary Task	immigrants		women	
	mono	cross	mono	cross
XLM-R				
None	67.5$_{(10.38)}$	58.4$_{(10.09)}$	64.4$_{(14.34)}$	48.0$_{(3.67)}$
UD	2.9$_{(4.87)}$	-2.5$_{(3.79)}$	1.8$_{(2.80)}$	-3.2$_{(8.10)}$
NER	0.9$_{(7.79)}$	3.1$_{(4.05)}$	2.8$_{(2.04)}$	1.7$_{(5.55)}$
sentiment	3.0$_{(5.27)}$	3.7$_{(3.85)}$	1.7$_{(1.87)}$	-2.0$_{(6.18)}$
sentiment+UD	4.0$_{(4.23)}$	5.1$_{(4.84)}$	1.1$_{(1.58)}$	-2.0$_{(6.96)}$
sentiment+NER	0.2$_{(9.44)}$	3.7$_{(1.68)}$	3.3$_{(2.92)}$	-2.0$_{(2.98)}$
sentiment+UD+NER	-0.1$_{(10.62)}$	5.2$_{(4.05)}$	2.8$_{(1.56)}$	-2.6$_{(6.36)}$
UD+NER	-1.0$_{(10.25)}$	-1.6$_{(11.48)}$	2.0$_{(0.94)}$	-6.5$_{(8.05)}$
XLM-T				
None	67.8$_{(15.80)}$	60.3$_{(4.30)}$	68.1$_{(11.01)}$	53.7$_{(7.67)}$
UD	2.8$_{(2.13)}$	-0.2$_{(3.19)}$	-1.1$_{(2.07)}$	-8.2$_{(6.12)}$
NER	1.2$_{(4.05)}$	3.4$_{(5.17)}$	1.7$_{(2.44)}$	-0.6$_{(4.30)}$
sentiment	1.1$_{(2.70)}$	2.4$_{(3.66)}$	0.1$_{(2.85)}$	-0.5$_{(5.01)}$
sentiment+NER	0.7$_{(4.66)}$	3.8$_{(4.38)}$	2.4$_{(2.99)}$	-1.5$_{(3.60)}$
sentiment+UD	2.1$_{(1.99)}$	1.6$_{(3.75)}$	0.6$_{(1.35)}$	-2.2$_{(2.92)}$
sentiment+UD+NER	0.2$_{(5.85)}$	2.9$_{(3.90)}$	1.4$_{(2.76)}$	-2.9$_{(2.39)}$
UD+NER	-0.7$_{(6.59)}$	1.7$_{(3.92)}$	-0.3$_{(2.79)}$	-9.1$_{(8.71)}$
XLM-R-random-init				
None	64.1$_{(13.91)}$	52.2$_{(3.41)}$	63.5$_{(14.38)}$	46.3$_{(2.42)}$
UD	3.5$_{(5.13)}$	-1.9$_{(2.27)}$	-2.2$_{(0.98)}$	-0.5$_{(2.76)}$
NER	-1.4$_{(1.10)}$	-0.8$_{(2.55)}$	0.9$_{(0.80)}$	1.1$_{(4.25)}$
sentiment+NER	-2.6$_{(2.40)}$	2.4$_{(4.14)}$	-1.5$_{(2.36)}$	-0.1$_{(6.04)}$
sentiment	-0.7$_{(1.19)}$	2.7$_{(3.59)}$	-0.3$_{(1.07)}$	-0.0$_{(5.51)}$
sentiment+UD	2.5$_{(4.01)}$	2.8$_{(3.35)}$	-1.7$_{(0.54)}$	2.2$_{(5.64)}$
sentiment+UD+NER	-2.7$_{(2.71)}$	2.0$_{(2.25)}$	-0.7$_{(1.68)}$	1.8$_{(6.36)}$

Table 20: Aggregated results for all models. For each model, we compute the deltas between the auxiliary task training and the baseline hate speech detection. For each domain (immigrants and women), we average the scores by setting: monolingual (*mono*), and zero-shot cross-lingual (*cross*). Variances between averaged results in each settings (e.g., the three cases where the training and test data are in the same language for each domain, for the *mono* setting) are in gray next to the deltas.)