Multilingual Auxiliary Tasks Training: Bridging the Gap between Languages for Zero-Shot Transfer of Hate Speech Detection Models

Syrielle Montariol* **Arij Riabi*** **Djamé Seddah**
INRIA Paris, France
firstname.lastname@inria.fr

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 for hate speech detection in several domains and languages using strict experimental settings. Then, we propose to train on multilingual auxiliary tasks – sentiment analysis, named entity recognition, and tasks relying on syntactic information – to improve zero-shot transfer of hate speech detection models across languages. We show how hate speech detection models benefit from a cross-lingual *knowledge proxy* brought by auxiliary tasks fine-tuning and highlight these tasks' positive impact on bridging the hate speech linguistic and cultural gap between languages.

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 varies significantly diachronically (Florio et al., 2020) and synchronically (Yin and Zubiaga, 2021), hate speech detection models need to be constantly adapted to new contexts. For example, as noted by Markov et al. (2021), the occurrence of new hate speech domains and their associated lexicons and expressions can be triggered by real-world events, from local scope incidents to worldwide crisis.[1] New annotated datasets are needed to optimally capture all these domain-specific, target-specific hate speech types. The possibility of creating and constantly updating exhaustively annotated datasets, adapted to every possible language and domain, is chimerical. Thus, the task of hate speech detection is often faced with low-resource issues.

In this low-resource scenario for a given target language and domain, if annotated data is available in another language, the main option for most NLP tasks is to perform *zero-shot* transfer using a multilingual language model (Conneau et al., 2020). However, in our case, hate speech perception is highly variable across languages and cultures; for example, some slur expressions can be considered not offensive in one language, denoting an informal register nonetheless, but will be considered offensive, if not hateful, in another (Nozza, 2021). Despite the cross-lingual transfer paradigm being extensively used in hate speech detection to cope with the data scarcity issue (Basile and Rubagotti, 2018; van der Goot et al., 2018; Pamungkas and Patti, 2019; Ranasinghe and Zampieri, 2020) or even the use of models trained on a translation of the initial training data (Rosa et al., 2021) , this strong hate speech cultural and linguistic variation can lower the transferability of hate speech detection models across languages in a zero-shot setting.

To overcome this limitation, in the absence of training data or efficient translation models for a target language, the cultural and linguistic information specific to this language needs to be found elsewhere. In this paper, we propose to capture this information by fine-tuning the language model on resource-rich tasks in both the transfer's source and target language. Indeed, even though hate-annotated datasets are not available in both lan-

*These authors contributed equally.

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

Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022, pages 347–363
November 20–23, 2022.

guages, it is likely that similarly annotated data in the source and target language exist for other tasks. A language model jointly fine-tuned for this other task in the two languages can learn some patterns and knowledge, bridging the gap between the languages, and helping the hate speech detection model to be transferred between them.

In summary, our work focuses on zero-shots cross-language multitask architectures where annotated hate speech data is available only for one source language, but some annotated data for other tasks can be accessed in both the source and target languages. Using a multitask architecture (van der Goot et al., 2021b) on top of a multilingual model, 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. Using 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 follows.

- Building strictly comparable corpora across languages,[2] leading to a thorough evaluation framework, we highlight cases where zero-shot cross-lingual transfer of hate speech detection models fails and diagnose the effect of the choice of the multilingual language model.
- 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.
- Using the HateCheck test suite (Röttger et al., 2021, 2022), we identify which hate speech *classes of functionalities* suffer the most from cross-lingual transfer, highlighting the impact of *slurs*; and which ones benefit from joint training with multilingual auxiliary tasks.

2 Related Work

Intermediate task training. In order to improve the efficiency of a pre-trained language model for a given task, this model can undergo preliminary fine-tuning on an intermediate task before fine-tuning again on the downstream task. This idea was formalized as Supplementary Training on Intermediate Labeled-data Tasks (STILT) by Phang et al. (2018), who perform sequential task-to-task pre-training. More recently, Pruksachatkun et al. (2020) perform a survey of intermediate and target task pairs to analyze the usefulness of this intermediary fine-tuning, but only in a monolingual setting. Phang et al. (2020) turn 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. They 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. However, to the best of our knowledge, using intermediate task training data on both the source and the target language for transfer has not been tested in the literature.

Auxiliary tasks for hate speech detection. Auxiliary task training for hate speech detection has been 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 EVALITA 2020, use 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 is also 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.

Zero-shot cross-lingual transfer 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) investigates 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 out-

[2] Our comparable datasets are available at https://github.com/ArijRB/Multilingual-Auxiliary-Tasks-Training-Bridging-the-Gap-between-Languages-for-Zero-Shot-Transfer-of-/.

performs massively multilingual language models. To perform cross-lingual experiment, Glavaš et al. (2020) create a dataset with aligned examples in six different languages, avoiding the issue of hate speech variation across languages that we tackle in this paper. On their aligned test set, they show the positive impact of intermediate masked language model fine-tuning on abusive corpora in the target language. Using aligned corpora allows the authors to focus on the effect of the intermediate finetuning without the noise of inter-language variability. On the contrary, in our case, we investigate the issue of limited transferability of hate speech detection models across languages. 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.[3] Stappen et al. (2020) perform zero- and few-shots cross-lingual transfer on some of the datasets we use in this paper, with an attention-based classification model; but contrarily to us, they do not distinguish between the hate speech targets.

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 them to point out the limitations of zero-shot cross-lingual transfer. The corpora are in three languages: English (en), Spanish (es) and Italian (it); and two domains: hate speech towards immigrants and hate speech towards women. The corpora come from various shared tasks; For English and Spanish, we use the dataset from a shared task on hate speech against immigrants and women on Twitter (HatEval). For the Italian corpora, we use the automatic misogyny identification challenge (AMI) (Fersini et al., 2018) for the women domain and the hate speech detection shared task on Facebook and Twitter (HaSpeeDe) (Bosco et al., 2018) for the immigrants domain. Links to the resources are listed in Table 6 in Appendix A.

[3] 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.

The hate speech detection task is a binary classification task where each dataset is annotated with two labels: *hateful* and *non hateful*. We train binary classification models on the train sets in each language and predict on the test set of each 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. We evaluate the models using macro-F1 as metric.

3.2 Original baseline results

The original results reported by Nozza (2021) can be found in the first rows of Table 1. In the table, we highlight in brown zero-shot cross-lingual cases where the macro-F1 score drops by more than 25% compared to the monolingual setting: these are cases for which we consider that the cross-lingual transfer failed. We observe the phenomenon that raised the issue of zero-shot cross-lingual transfer: in the *women* domain, the models trained on Spanish and Italian in a zero-shot setting have much lower scores compared to the monolingual results; 4 out of the 6 cross-lingual cells are highlighted in brown. One possible cause, as explained by Nozza (2021), is the presence of language-specific offensive 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 (macro-F1 of 36.8 in the monolingual setting). This phenomenon was also observed by Nozza (2021) and Stappen et al. (2020), 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 overfit, linked with a large discrepancy between the train and test set.

3.3 Experimental settings

Building comparable corpora. We started this work to investigate the failure of cross-lingual hate speech datasets for the women domain highlighted by Nozza (2021). However, these experiments were not realized in comparable settings; the corpora do not have the same size in the different languages and domains. Our goal is to confirm these results under a strictly comparable setting, and a multi-seed robust experimental framework. Therefore, we build comparable corpora in each language and domain to ensure the comparability of the transfer settings. We reduce all datasets to

Model	Src 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
Comparable corpus size and new random split							
m-BERT	en	72.5	48.5	63.8	75.2	41.7	43.4
	es	59.4	80.9	58.5	54.5	76.9	40.5
	it	62.8	54.8	76.3	46.3	53.6	88.3
XLM-R	en	75.3	51.9	70.1	76.6	51.6	49.9
	es	62.0	83.4	65.4	63.4	77.8	46.9
	it	69.2	51.3	78.6	60.3	57.3	89.0
XLM-T	en	76.8	48.5	73.5	78.6	61.5	60.6
	es	65.9	84.2	60.7	72.5	80.3	51.9
	it	71.5	56.8	78.4	63.4	58.2	90.3

Table 1: Monolingual and cross-lingual hate speech detection macro-F1 scores on all corpora. 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 loss in performance in the zero-shot cross-lingual case compared to the monolingual case is higher than 25%.

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. We use the Kolmogorov–Smirnov test to compare the sentence length distribution (number of tokens) and the percentage of hate speech between the sampled and the original datasets, to make sure they stay comparable. The sampling is done randomly until the similarity conditions with the original dataset are met. The original size for each dataset as well as the sampling size for building the comparable datasets and the percentage of hateful examples can be found in Table 7 and Table 8 in Appendix A.

On top of this, before the sub-sampling of the corpora, we merge the development, test and train dataset for each language and domain before performing a new random split. This allows us to overcome the train-test discrepancy observed in the English-immigrants dataset we mentioned above.

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 pre-processing step in the literature when handling tweets (e.g. (Röttger et al., 2021)); it can improve tasks such as tweet clustering (Gromann and Declerck, 2017).

Models training. 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 most of 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. We keep the best out of 20 epochs for each run according to the macro-F1 score on the development set.

Note that the new comparable 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.

Language Models. We use two general-domain large-scale multilingual language models: m-BERT (Devlin et al., 2019) following Nozza (2021) and XLM-R (Conneau et al., 2020). The former is the multilingual version of BERT, trained on Wikipedia content in 104 languages, with 100M parameters. 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.

Then, 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.

3.4 Setting a new baseline

We compare the scores for m-BERT from Nozza (2021) to the scores obtained using our comparable corpora, reported in Table 1. First, our experiment with m-BERT on comparable corpora allows us to highlight additional cases where zero-shot cross-lingual transfer "fails" (macro-F1 dropping by more than 25% compared to monolingual score) in the *immigrants* domain, that were not visible in the previous study due to variations in training corpus size. On top of this, with the new splits,

[4] Using the Python package `wordsegment`.

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

we do not observe the extremely low scores on English for the immigrant domain anymore, allowing us to draw more reliable conclusions on the monolingual/cross-lingual performance gap.

Comparing m-BERT and XLM-R, the latter shows higher scores for almost all languages and domains. It also shows, in general, slightly lower macro-F1 loss between monolingual and cross-lingual settings; which is related to its much larger number of parameters and training corpus size compared to m-BERT.

Fine-tuning XLM-T leads to higher macro-F1 scores for almost all languages and domains compared to XLM-R; which is expected, as it was fine-tuned using the Masked Language Modeling (MLM) task on tweets, which is much more similar to the hate speech datasets, at least stylistically due to the Twitter platform constraints (e.g. number of characters). In terms of monolingual/cross-lingual discrepancy, we also observe in general a much lower macro-F1 drop. Having seen a large amount of similar data in all languages, the model can much more easily bridge the gap between languages when performing zero-shot cross-lingual transfer for this highly domain-specific task.

However, such a large amount of training data from a similar source in different languages is not so easy to come by. To bridge the language gap in very context-specific tasks such as hate speech detection, in the case of absence of an adequately trained multilingual language model, we turn towards other sources of multilingual information for the model: using annotated corpora for other *auxiliary* tasks in the source and target languages.

In all following experiments, we use the comparable datasets and the general-domain multilingual language model XLM-R to study the impact of auxiliary task training on this problem[6]. 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 Experiments

We define several training tasks whose effects on cross-lingual transfer of hate speech detection models are 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 group – by misnomer – under the term "Universal Dependency" (UD). We hypothesize that 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 sentences, the size of the smallest dataset. Detailed statistics about the datasets can be found in Table 12 in Appendix A.

Sentiment analysis. We use Twitter sentiment analysis datasets on each of our three target languages. 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)[7]. Detailed statistics and additional information on each dataset can be found in Table 10 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; Rahimi et al., 2019), which covers our three languages. The sets have a unique split size (training 20k examples, development 10k, test 10k).

4.2 Multi-task learning pipeline

We perform multi-task learning using the MaChAmp framework (van der Goot et al., 2021b); it fine-tunes contextual embeddings for

[6]The results for XLM-T display similar tendencies with higher scores compared to XLM-R, Detailed and summarized tables can be found in 499 Appendix B, Table 14

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

several tasks and several datasets using a shared encoder and different decoders depending on the target task. As the datasets associated with the different tasks have varying sizes, we use a "smooth sampling" method 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 multilingual model XLM-R on the different auxiliary tasks. The training is done jointly on the auxiliary task datasets in the three languages, in order to allow the model to learn patterns between languages, and on the hate speech dataset in the *source* language, before being tested on the *target* language. 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 (similarly to Pruksachatkun et al. (2020)), or jointly with the hate speech detection task. According to our experiments, the latter exhibits 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 3.3, averaging the macro-F1 over five different runs.

5 Results on Auxiliary Tasks Training

We analyze the training effect of adding different auxiliary tasks on top of XLM-R, jointly with monolingual hate speech detection. Results can be found in Table 2. Instead of raw scores, we compute the deltas between the baseline system (no auxiliary task, same as Table 1) and the augmented system with training jointly with auxiliary tasks: NER, sentiment analysis (*Sent*) and syntactic tasks (UD), for each language pair (Table 2a).

To help with the interpretation, we aggregate the results according to the monolingual (*mono*), and zero-shot cross-lingual (*cross*) settings. Table 2b is the aggregated equivalent of Table 2a. For each domain (immigrants and women), we average the scores by setting: the *mono* columns show the average of all scores in the diagonal in Table 2a, while the *cross* column is the average of all the rest.

In the zero-shot cross-lingual transfer scenario, we hypothesized that the additional information on the source and target languages could bridge the gap between the languages and improve the transfer

Aux. task	Src lang	immigrants			women		
		en	es	it	en	es	it
None	en	*75.3*	51.9	70.1	*76.6*	51.6	49.9
	es	62.0	*83.4*	65.4	63.4	*77.8*	46.9
	it	69.2	51.3	*78.6*	60.3	57.3	*89.0*
Sentiment	en	-1.0	-1.2	0.0	$2.0^{\dagger}$	0.9	$-6.2^{\dagger}$
	es	$5.1^{\dagger}$	0.6	1.5	0.7	$2.1^{\ddagger}$	$-9.6^{\ddagger}$
	it	$1.4^{\dagger}$	1.7	-0.9	$-8.3^{\ddagger}$	-0.7	0.1
NER	en	$1.4^{\dagger}$	1.0	-1.9	0.4	0.2	1.9
	es	3.1	0.4	-1.1	$-8.7^{\dagger}$	$2.2^{\ddagger}$	-4.9
	it	$3.3^{\ddagger}$	$4.5^{\ddagger}$	$-1.4^{\dagger}$	$-2.8^{\dagger}$	-0.5	$1.1^{\dagger}$
UD	en	$1.7^{\dagger}$	-2.4	-1.2	0.7	-0.4	$-10.6^{\dagger}$
	es	-3.6	-1.1	$-6.5^{\dagger}$	-4.9	-0.4	$-10.9^{\ddagger}$
	it	$-14.4^{\ddagger}$	$5.0^{\ddagger}$	$-1.6^{\dagger}$	$-14.7^{\ddagger}$	-5.6	-0.3

(a) Detailed view.

Auxiliary Task	immigrants		women	
	mono	cross	mono	cross
None	79.1	61.6	81.1	54.9
Sentiment	-0.4	1.4	1.4	-3.9
NER	0.1	1.5	1.3	-2.5
UD	-0.3	-3.8	0.0	-7.8
Sentiment + NER	0.4	2.5	1.3	-4.7

(b) Aggregated view.

Table 2: Effect (delta with hate speech detection baseline, averaged over 5 runs) of fine-tuning XLM-R on the three auxiliary tasks, on hate speech detection macro-F1 scores (%). Green values indicate an increase in score, red values a decrease. 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.[8] 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.

for hate speech detection. Looking at the scores for cross-lingual transfer, sentiment analysis and NER lead to an average improvement of respectively of 1.42 and 1.48 points for the immigrants domains; combined (last row of Table 2b), they lead to an even greater improvement of 2.5 percentage points. On the contrary, for the women domain, these two tasks lead to significant improvements almost only in the monolingual setting. As underlined before, zero-shot cross-lingual transfer is especially hard in this domain due to cultural and linguistic variations (Nozza, 2021) that auxiliary task training fails to capture. Finally, UD tasks auxiliary training leads to a large drop of performance in most cases. The impact of auxiliary tasks on the performance of hate speech detection using the XLM-T model

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

is comparable to the one observed with XLM-R. Detailed and summarized tables can be found in Appendix B, Table 14.

6 Diagnosis: Effect of Auxiliary Task Training

There is an extensive literature on how performance metrics aggregated over the full test set are far from conveying enough information to fully evaluate and compare the strengths and weaknesses of models (Ribeiro et al., 2020), including for the task of hate speech detection (Röttger et al., 2021). Here, we use the HateCheck test suite in English (Röttger et al., 2021) and its recent multilingual version MHC (Röttger et al., 2022), which includes our two other target languages, Spanish and Italian. These are test sets covering a wide range of hate speech detection aspects that the authors call *functionalities*, testing detection models with hateful and non-hateful sentences of various styles, vocabulary, syntax and hate speech targets. All 29 *functionalities* are grouped into 11 classes and 7 protected groups as targets[9], and the various test cases of each functionality lead to a total of 3,901 sentences classified as hateful or not hateful. The protected groups vary across languages in the MHC test set; the authors selected them to better adapt to the cultural context of each language. The target group "women" is covered for our three languages, but the target group "immigrants" is not covered in Spanish; instead, we match it with the group "indigenous people".[10] Moreover, to ease the interpretation, we perform the analysis on the aggregated 11 classes of functionalities.

We do not evaluate the performance of our various models on the test suite intrinsically: what we want to measure is the *effect* of zero-shot cross-lingual transfer and auxiliary tasks training on the hate speech functionalities. First, we measure the difference between monolingual and zero-shot cross-lingual training on the various functionalities: what the model "loses" by not being trained on the same language as the test set. We rank the functionalities by average difference across the two domains (Table 3). The largest loss in performance when performing zero-shot transfer is found for functionalities involving slurs: -14.72 of macro-F1 for the immigrants domain and -17.22 for the women domain. Indeed, slurs are extremely cultural and language-specific. Second, we measure

functionality	immigrants	women
slur	-14.72	-17.22
negate	-10.34	0.82
spell	-7.56	5.78
derog	-9.37	7.92
threat	-2.61	1.63
ident	5.57	-3.22
counter	-2.43	10.03
ref	6.62	7.11
profanity	-3.75	18.33
phrase	18.57	5.63

Table 3: Difference between monolingual and zero-shot cross-lingual performance by functionality when fine-tuning XLM-R on hate speech detection (no auxiliary task), averaged over all language pairs, by domain.

the impact of multilingual auxiliary task training compared to training on hate speech detection only (baseline model), on the various functionalities. For the two domains and for each source-target language pair, we measure the HateCheck functionality score of the baseline model, and jointly on every auxiliary task. For each auxiliary task, we compute the *relative* difference in score with the baseline model; this difference represents the effect of the joint training. However, we focus here on the joint training impact for zero-shot cross-lingual transfer; thus, we separate the impact of auxiliary task training in a monolingual setting and in a cross-lingual setting. In Table 4, we display the effect of auxiliary task training on zero-shot transfer *on top of* the effect of these tasks on monolingual transfer. To designate the functionalities, we use the same denomination as in the HateCheck test suite. Detection of hate speech involving *slurs*, which suffers the most from zero-shot cross-lingual transfer, is improved by training with NER or UD. Training on UD tasks is especially helpful on cases involving spelling variations (*spell*), contrarily to the two other tasks, and phrasing variations (*phrasing*). Counter-speech detection, an extremely hard task involving *not* classifying counter-speech (e.g. denouncement of hate by quoting it) as hateful, is

[9] We refer the reader to (Röttger et al., 2022), pp.45, for an extensive definition of these classes and groups.

[10] This choice stems from measuring the similarity between Spanish immigrants train set and the test cases of each target group in Spanish Hatecheck using tf-idf representation. Indigenous people ("indígenas" in Spanish) had the highest similarity score with the Twitter immigrants dataset, higher than Hatecheck test cases targeted at black people ("negros") or Jews ("judíos"), hence our decision to use indigenous people as a proxy.

only helped by NER. Sentiment analysis is globally helpful for many classes, but particularly for sentences involving *negated* positive or hateful statements.

functionality	NER	Sentiment	UD
threat	-8.23	-2.32	26.81
target	-3.54	4.70	-6.19
spell	-3.13	-5.72	12.59
slur	1.09	-6.30	14.42
ref	-6.80	2.17	7.77
profanity	-4.23	2.77	-0.44
phrase	-14.79	1.17	8.64
negate	4.19	3.57	1.98
ident	2.57	1.05	-14.42
derog	-1.60	2.02	18.58
counter	2.90	-11.83	-15.60

Table 4: Relative difference in macro-F1 score by class of functionality, between monolingual and zero-shot cross-lingual training (averaged across all language pairs), averaged across the two domains, for each auxiliary task.

7 Discussion

On the impact of each auxiliary task training, we experimented with jointly training hate speech detection and different auxiliary tasks: sentiment analysis, NER and UD tasks. In the immigrants domain, the NER and sentiment auxiliary tasks led to the best improvement on hate speech detection. The cross-lingual transferability of NER was facilitated by the fact that many named entities are the same across languages (e.g. person and organisation names); indeed, many successful unsupervised cross-lingual transfer systems for this task can be found in the literature (Rahimi et al., 2019; Bari et al., 2020).

Compared with the first two tasks, adding syntactic information had 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) showed 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, used POS-tagged text as input for classification. On the contrary, in a monolingual setting, Klemen et al. (2020) showed that morphological features added to LSTM and BERT-based hate speech detection models did not help with comment filtering. Similarly, using sequential auxiliary training of tasks such as POS tagging, Pruksachatkun et al. (2020) showed that the resulting additional low-level skills often led 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 style – short sentences, generally with low syntactic complexity – makes additional syntactic knowledge unhelpful (especially in a more difficult to parse user-generated content context) for a downstream task such as hate speech detection, which benefits more from semantic information.

Regarding the non-usage of POS taggers that could have been optimized for our User-Generated Content-based datasets, we investigated this possibility and conducted preliminary experiments for English – using the Tweebank (Jiang et al., 2022) as data source–, that showed that using a tagger trained on it did not bring much in terms of performance compared to "classic" UD POS taggers. Part of the reasons might come from the fact that our pre-processing step removes hashtags and normalized other Twitter's idiosyncrasies and hence make the data somewhat simpler to tag. Another reason to not investigate this further lies in the lack of availability of a UGC treebank for Spanish, breaking thus the symmetry of our experimental protocol. Last but not least, another reason we hypothesized for this lack of much improvement we noticed comes from the fact that the multilingual language model we used (XMLR and XMLR-T) were already providing strong results on UGC. This was corroborated by Riabi et al. (2021), who experimentally verified the robustness of language models when facing noisy UGC. Moreover Itzhak and Levy (2021) showed that subword-based language models were able to capture a significant amount of character-level alteration typical of UGC (Sanguinetti et al., 2020), explaining their surprising level of robustness when facing noisy content. However, we agree that better handling UGC content would be an interesting step, if not the next

step, especially if we can demonstrate that many idiosyncrasies align across languages in our target domains and hence are alleviated by the use of optimized tagging and parsing, eventually multilingual, models. This, in our minds, warrants another full-scale study with a thorough error analysis of cross-lingual syntactic transfer in noisy scenarios. We leave this for future work.

Cross-lingual zero-shot transfer on a domain with a gap between languages. In Section 3, we observed that using larger pre-trained multilingual language models, and if possible, multilingual models trained on corpora from the same source as the downstream task, improves cross-lingual zero-shot transfer. This adaptation has a significant and consistent positive impact. This is in line with the findings of Bose et al. (2021), who demonstrated the superiority of MLM over other tasks in a cross-corpora transfer setting. Similarly, van der Goot et al. (2021a) jointly trained auxiliary tasks with a downstream task (in their case, spoken language understanding) in a cross-lingual setting 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; 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 linguistic specificities and 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)). Furthermore, contrarily to Wikipedia where corpora are highly similar from one high-resource language to another in term of domains, Twitter data can significantly differ between languages due to cultural differences and events in the respective countries. Overall, when we used XLM-T, the model is only adapted 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 between corpus variations and the actual cultural and linguistic gap.

Discussions on computational costs and ethical considerations for this work can be found in Appendix 9.

8 Conclusion

In this work, we highlighted situations where zero-shot cross-lingual transfer of hate speech models fails because of the linguistic and cultural gap. We quantified the effect of the choice of multilingual language model and of auxiliary task training on these "failed" cases, showing the positive effect of NER and sentiment analysis multilingual training, but their limited improvement in the domain of hate speech against women. We performed a preliminary analysis on the effect of auxiliary tasks by *hate speech functionality* using the HateCheck test suite, hinting at which kind of hate speech benefits from transferring knowledge in both the source and the target languages for the three auxiliary tasks. Finally, we discussed limitations related to training data for language model pre-training, auxiliary tasks, and hate speech detection. All of our datasets with their new splits and models are freely available.[11], hoping that the sound experimental framework we designed will help strengthen future studies on cross-lingual hate-speech detection.

Acknowledgments

We warmly thank the reviewers for their very valuable feedback. This work received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101021607. The last author acknowledges the support of the French Research Agency via the ANR ParSiTi project (ANR16-CE33-0021). The authors are grateful to the OPAL infrastructure from Université Côte d'Azur for providing resources and support.

[11] https://github.com/ArijRB/Multilingual-Auxiliary-Tasks-Training-Bridging-the-Gap-between-Languages-for-Zero-Shot-Transfer-of-/

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 one 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 6 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 5.

Finally, the presence of bias in the pre-trained language models we use, due to the bias in the data they were trained on, may have an impact on hate speech detection, particularly on the topic of hate speech towards women. As a result, this area of research is currently under heavy scrutiny by the community.

Computational Costs. We conduct our experiments on RTX8000 GPUs. We test two models (XLM-R and XLM-T) on 7 different auxiliary tasks combinations, 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 5.

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

Table 5: Training time (in seconds) for one seed per model.

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.

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.

Cristina Bosco, Dell'Orletta Felice, Fabio Poletto, Manuela Sanguinetti, and Tesconi Maurizio. 2018. Overview of the evalita 2018 hate speech detection task. In *EVALITA 2018-Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, volume 2263, pages 1–9. CEUR.

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.

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.

Goran Glavaš, Mladen Karan, and Ivan Vulić. 2020. XHate-999: Analyzing and detecting abusive language across domains and languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6350–6365, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

Itay Itzhak and Omer Levy. 2021. Models in a spelling bee: Language models implicitly learn the character composition of tokens. *arXiv preprint arXiv:2108.11193*.

Hang Jiang, Yining Hua, Doug Beeferman, and Deb Roy. 2022. Annotating the Tweebank corpus on named entity recognition and building NLP models for social media analysis. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7199–7208, Marseille, France. European Language Resources Association.

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.

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.

Arij Riabi, Benoît Sagot, and Djamé Seddah. 2021. Can character-based language models improve downstream task performances in low-resource and noisy language scenarios? In *Proceedings of the Seventh Workshop on Noisy User-generated Text (W-NUT 2021)*, pages 423–436, Online. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, 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, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022. Multilingual HateCheck: Functional tests for multilingual hate speech detection models. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 154–169, Seattle, Washington (Hybrid). Association for Computational Linguistics.

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.

Manuela Sanguinetti, Lauren Cassidy, Cristina Bosco, Özlem Çetinoglu, Alessandra Teresa Cignarella, Teresa Lynn, Ines Rehbein, Josef Ruppenhofer, Djamé Seddah, and Amir Zeldes. 2020. Treebanking user-generated content: a UD based overview of guidelines, corpora and unified recommendations. *CoRR*, abs/2011.02063.

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.

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*.

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.

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.

Wenjie Yin and Arkaitz Zubiaga. 2021. Towards generalisable hate speech detection: a review on obstacles and solutions. *PeerJ Computer Science*, 7:e598.

A Datasets overview

A.1 Hate speech datasets overview

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.

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 6: Shared tasks used for the Hate speech corpora.

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 7: 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.

Language	immigrants	women
en	41.28	42.76
es	42	40.23
it	31.33	45.42

Table 8: Percentage of hateful examples in the train set for the comparable setting.

	immigrants			women		
	en	es	it	en	es	it
Nb of tokens per tweet						
avg	27.3	18.9	17.2	18.3	22.8	17.9
median	26.0	17.0	17.0	18.0	20.0	14.0
max	90	57	29	57	59	54
min	2	1	2	2	2	2
Nb of hashtags (avg per tweet, total unique nb)						
avg	2.0	0.2	0.6	0.2	0.2	0.2
unique	1162	214	491	211	292	228
Train/test OOV Ratio						
	0.4	0.6	0.5	0.5	0.5	0.5

Table 9: Descriptive statistics on hate speech detection training datasets.

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 10: 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 11: Twitter UD data overview.

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

Table 12: 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 13: Statistics of the WNUT 2016 NER shared task dataset.

Aux task	Src lang	immigrants			women		
		en	es	it	en	es	it
None	en	76.8	48.5	73.5	78.6	61.5	60.6
	es	65.9	84.2	60.7	72.5	80.3	51.9
	it	71.5	56.8	78.4	63.4	58.2	90.3
sent	en	-0.4	4.2†	-1.9	0.5	2.2	-0.2
	es	1.3	0.5	6.2	-2.6†	0.7	-9.6‡
	it	0.8	-1.8	-0.3	-5.1†	3.4	-0.3
NER	en	0.1	5.9‡	-4.7‡	-0.1	0.9	1.6
	es	-2.2	0.6	1.4	-5.9‡	1.5†	-6.0‡
	it	1.0	0.7	0.0	-2.7	2.2	0.5
UD	en	-0.4	2.9	-3.9	-0.1	-1.7†	-10.1‡
	es	-11.1‡	-0.7	-3.7	-2.4‡	0.4	-12.9‡
	it	-4.1‡	1.6	0.1	-8.7‡	-2.1	0.7†

(a) Detailed view.

Auxiliary Task	immigrants		women	
	mono	cross	mono	cross
None	79.8	62.8	83.1	61.3
Sent	-0.1	1.5	0.3	-2.0
NER	0.3	0.4	0.6	-1.7
UD	-0.3	-3.0	0.3	-6.3
Sent + NER	-0.2	1.3	0.6	-2.5

(b) Aggregated view.

Table 14: Effect (delta with hate speech detection baseline, averaged over 5 runs) of fine-tuning XLM-T on the three auxiliary tasks, on hate speech detection macro-F1 scores (%). Green values indicate an increase in score, red values a decrease. *Sent* stands for Sentiment and *Aux* for auxiliary.

B Complementary results

Aux. task	Src lang	en	es	it
None	en	*75.3*	51.9	70.1
	es	62.0	*83.4*	65.4
	it	69.2	51.3	*78.6*
MLM	en	1.1	-2.9	-1.4
	es	2.6	-2.9‡	0.3
	it	-1.6	-1.0	-0.1
NER	en	1.4†	1.0	-1.9
	es	3.1	0.4	-1.1
	it	3.3‡	4.5‡	-1.4†

Table 15: Effect (delta with XLM-R baseline) of MLM fine-tuning on sentences from NER datasets compared fine-tuning on NER as auxiliary tasks, on hate speech detection macro-F1 scores (%) for immigrants domain. Green values indicate an increase in score, red values a decrease.

Auxiliary task	Source lang	immigrants			women		
		en	es	it	en	es	it
None	en	*75.3*	51.9	70.1	*76.6*	51.6	49.9
	es	62.0	*83.4*	65.4	63.4	*77.8*	46.9
	it	69.2	51.3	*78.6*	60.3	57.3	*89.0*
UD	en	1.7†	-2.4	-1.2	0.7	-0.4	-10.6†
	es	-3.6	-1.1	-6.5†	-4.9	-0.4	-10.9‡
	it	-14.4‡	5.0‡	-1.6†	-14.7‡	-5.6	-0.3
UPOS	en	-0.6	-3.1	-1.4	0.9	-5.2	-1.2
	es	-4.0	-1.2	-3.9†	-0.9	1.9‡	-7.3†
	it	-4.7†	5.0‡	-1.0	-1.2	-3.4	-1.7

Table 16: **Ablation study**: Hate speech detection macro-F1 scores (%) of XLM-R fine-tuned on the UPOS task 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.