Evaluating the Effectiveness of Natural Language Inference for Hate Speech Detection in Languages with Limited Labeled Data

Janis Goldzycher **Moritz Preisig** **Chantal Amrhein** **Gerold Schneider**

Department of Computational Linguistics
University of Zurich
{goldzycher,amrhein,gschneid}@cl.uzh.ch, moritz.preisig@uzh.ch

Abstract

Most research on hate speech detection has focused on English where a sizeable amount of labeled training data is available. However, to expand hate speech detection into more languages, approaches that require minimal training data are needed. In this paper, we test whether natural language inference (NLI) models which perform well in zero- and few-shot settings can benefit hate speech detection performance in scenarios where only a limited amount of labeled data is available in the target language. Our evaluation on five languages demonstrates large performance improvements of NLI fine-tuning over direct fine-tuning in the target language. However, the effectiveness of previous work that proposed intermediate fine-tuning on English data is hard to match. Only in settings where the English training data does not match the test domain, can our customised NLI-formulation outperform intermediate fine-tuning on English. Based on our extensive experiments, we propose a set of recommendations for hate speech detection in languages where minimal labeled training data is available. [1]

1 Introduction

Hate speech is a global issue that transcends linguistic boundaries, but the majority of available datasets for hate speech detection are in English (Poletto et al., 2021; Yin and Zubiaga, 2021). This limits the capabilities of automatic content moderation and leaves most language communities around the world underserved. Creating labeled datasets is not only slow and expensive but also risks psychological impacts on the annotators (Kirk et al., 2022a). Although the number of non-English datasets is increasing, most languages still have limited or no datasets available (Poletto et al., 2021).
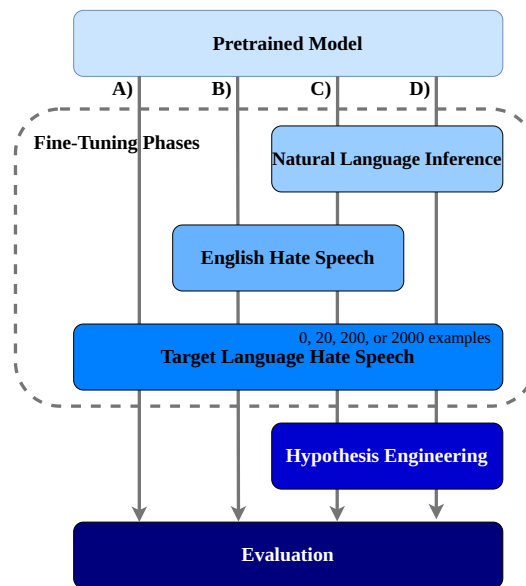

Figure 1: Approaches evaluated in this paper: A) Standard fine-tuning. B) Intermediate fine-tuning on English hate speech as proposed by Röttger et al. (2022a). Via C) and D), we explore natural language inference as an additional intermediate fine-tuning step. The natural language inference formulation further allows hypothesis engineering (Goldzycher and Schneider, 2022).

Consequently, there is a pressing need to develop methods that can efficiently expand hate speech detection into languages with less labeled data. Repurposing natural language inference models for text classification leads to well-performing zero-shot and few-shot classifiers (Yin et al., 2019). Recently, Goldzycher and Schneider (2022) showed that zero-shot NLI-based setups can outperform standard few-shot finetuning for English hate speech detection. This raises the question of whether natural language inference can be used to expand hate speech detection into more languages in a data-efficient manner.

In this paper, we aim to systematically evaluate the effectiveness of NLI fine-tuning for languages beyond English where we do not have an abun-

---

[1] We make our code publically available at https://github.com/jagol/xnli4xhsd.

dance of labeled training data. We give an overview of the experiment setup and compared approaches in Figure 1. Unlike Goldzycher and Schneider (2022), and inspired by Röttger et al. (2022a), we do not restrict ourselves to a zero-shot setup but further analyse the usefulness of an NLI-formulation when we have access to a limited amount of labeled examples in the target language as well as additional English data for intermediate fine-tuning. We believe that this mirrors a more realistic setup and allows us to offer clear recommendations for best practices for hate speech detection in languages with limited labeled data.

Our experiments with 0 up to 2000 labeled examples across five target languages (Arabic, Hindi, Italian, Portuguese, and Spanish) demonstrate clear benefits of an NLI-based formulation in zero-shot and few-shot settings compared to standard few-shot fine-tuning in the target language. While Röttger et al.'s (2022a) approach of fine-tuning on English data before standard few-shot learning on the target language proves to be a strong baseline, we reach similar performance when fine-tuning NLI-based models on intermediate English data. Building on the results by Goldzycher and Schneider (2022), who showed that targeted hypothesis engineering can help avoid common classification errors with NLI-based models, we find that such strategies offer an advantage in scenarios where we have expert knowledge about the domain but no in-domain English data for intermediate fine-tuning.

Overall, our contributions are the following:

1. We are able to reproduce the results of Röttger et al. (2022a), demonstrating the validity of their approach.

2. We evaluate NLI fine-tuning for expanding hate speech detection into more languages and find it to be beneficial if no English labeled data is available for intermediate fine-tuning.

3. We evaluate NLI-based models paired with hypothesis engineering and show that we can outperform previous work in settings where we have knowledge about the target domain but no domain-specific labeled English data.

2 Related Work

Hate speech is commonly defined as attacking, abusive or discriminatory language that targets protected groups or an individual for being a member of a protected group. A protected group is defined by characteristics such as gender, sexual orientation, disability, race, religion, national origin or similar (Fortuna and Nunes, 2018; Poletto et al., 2021; Vidgen et al., 2021; Yin and Zubiaga, 2021). The automatic detection of hate speech is typically formulated as a binary text classification task with short texts, usually social media posts and comments, as input (Founta et al., 2018). Despite most work being focused on English (Founta et al., 2018), in the last years there has been a growing trend to expand into more languages (Mandl et al., 2019, 2020; Röttger et al., 2022b; Yadav et al., 2023).

In what follows, we first review the relevant literature for multi- and cross-lingual hate speech detection, and then move on to the previous work in zero-shot and few-shot text classification with a specific focus on NLI-based methods, and finally focus on hypothesis engineering (Goldzycher and Schneider, 2022).

2.1 Multi- and Cross-lingual Hate Speech Detection

The scarcity of labeled datasets for hate speech detection in non-English languages has led to multiple approaches addressing this problem using meta-learning (Mozafari et al., 2022), active learning (Kirk et al., 2022b), label-bootstrapping (Bigoulaeva et al., 2023), pseudo-label fine-tuning (Zia et al., 2022), and multi-task learning using multilingual auxiliary tasks such as dependency parsing, named entity recognition and sentiment analysis (Montariol et al., 2022).

The most important research for our investigation is the study conducted by Röttger et al. (2022a). In their work, the authors train and evaluate models across five distinct languages: Arabic, Hindi, Italian, Portuguese, and Spanish. Their findings reveal that by initially fine-tuning multilingual models on English hate speech and subsequently fine-tuning them with labeled data in the target language, they achieve significant performance improvements in low-resource settings compared to only fine-tuning a monolingual model in the target language. In this study, we adopt their evaluation setup, reproduce their results and compare our results directly to their approach. For this reason, we will elaborate on and reference the specifics of their experimental setup throughout Section 3.

2.2 Zero-Shot and Few-Shot Classification

The development of language models which serve as a foundation for fine-tuning rather than training from scratch, has facilitated the implementation of zero-shot and few-shot text classification approaches, such as prompting (Liu et al., 2021) and task descriptions (Raffel et al., 2020). These techniques transform the target task into a format similar to the pre-training task and are typically employed in conjunction with large language models. Following this scheme, Chiu and Alexander (2021) leverage GPT-3 to detect hate speech with the prompts "Is this text racist?" and "Is this text sexist?".

In contrast to prompting, NLI-based prediction refers to reformulating the target task into an NLI task and thus into a (previous) fine-tuning task. In this setup, a model receives a premise and a hypothesis and is tasked with predicting whether the premise entails the hypothesis, contradicts it, or is neutral towards it. Yin et al. (2019) were the first to demonstrate the effectiveness of such an approach. They used an NLI model for zero-shot topic classification by inputting the text to be classified as the premise and constructing a hypothesis for each topic in the form of "This text is about <topic>". A prediction of *entailment* is to be interpreted as the input text belonging to the topic in the given hypothesis. Wang et al. (2021) demonstrated that this task reformulation benefits few-shot learning scenarios for various tasks, including offensive language identification.

2.3 Hypothesis Engineering

Goldzycher and Schneider (2022) pair an input text with multiple hypotheses in order to let an NLI model predict different aspects of the input text. They then use a rule-based approach to combine these predicted aspects into a final prediction for the hate speech label. More specifically, they distinguish between a main hypothesis and auxiliary hypotheses. The main hypothesis claims that the input text contains hate speech. The auxiliary hypotheses claim various relevant aspects such as that the input text is about a protected group in order to correct mispredictions of the main hypothesis. To find effective hypothesis combinations they conduct an error analysis on the English HateCheck (Röttger et al., 2021) dataset and propose four *strategies* based on this error analysis:

Filtering by target: Avoid false positives by predicting if any protected group is targeted in the input text.

Filtering reclaimed slurs: Avoid false positives by predicting indicators that a slur is used in a reclaimed fashion. Indicators used are: the speaker talks about themself or positive sentiment.

Filtering counterspeech: Avoid false positives by recognizing when another speech act is referenced, predicting if that speech act is hate speech and predicting the stance towards the referenced speech.

Catching dehumanizing comparisons: Avoid false negatives by checking if a protected group and negatively associated animals appear together in a sentence with a negative sentiment.

In our experiments, we will evaluate the effectiveness of the first three strategies for NLI-based hate speech detection. However, we will exclude the fourth strategy due to its lack of clear benefits in Goldzycher and Schneider (2022). More implementation details are provided in Section 3.3 and Appendix B.

3 Experiment Setup

Our experiment setup is largely based on the one created by Röttger et al. (2022a) and can be seen in Figure 2. In what follows, we first describe their setup, which we replicated as a baseline for our results. We then describe how we expanded their setup for the NLI-based experiments.

3.1 Reproducing Röttger et al. (2022a)

Data The authors use three English hate speech datasets (Founta et al., 2018; Kennedy et al., 2020; Vidgen et al., 2021, abbreviated as **FEN**, **KEN**, and **DEN**, respectively) which are all downsampled to 20,000 examples. **FEN** and **KEN** are sourced from Twitter and **DEN** consists of human-created adversarial examples. They further make use of five Twitter datasets in the respective target languages: Basile et al. (2019) for Spanish (**BAS19_ES**), Fortuna et al. (2019) for Portuguese (**FOR19_PT**), Modha et al. (2022) for Hindi (**HAS21_HI**), Ousidhoum et al. (2019) for Arabic (**OUS19_AR**), and Manuela et al. (2020) for Italian (**SAN20_IT**). The Multilingual HateCheck (MHC) (Röttger et al., 2022b) is used for additional, complementary evaluation. This suite of synthetic, evaluation-only
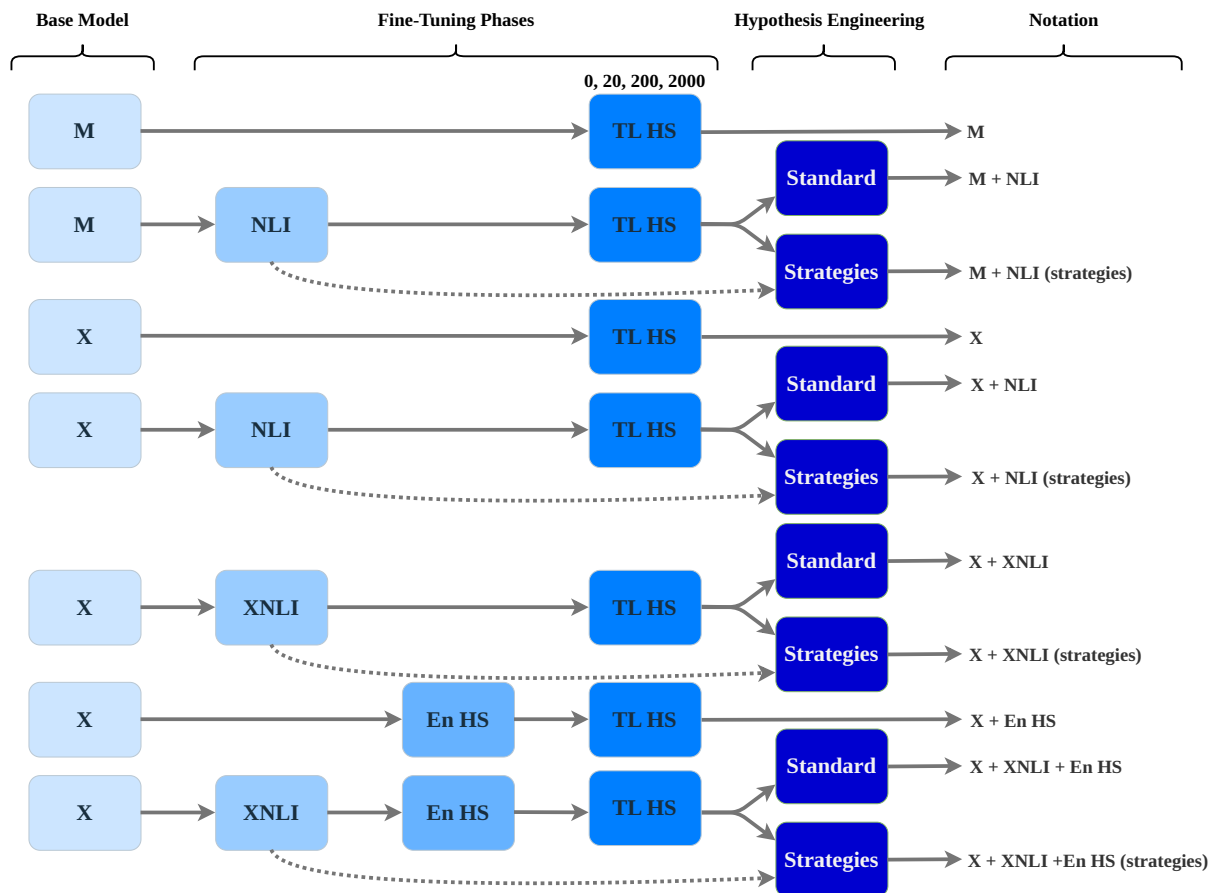
Figure 2: Experiment Setup. The base model is either monolingual **M** or multilingual **X**. In a first optional training phase, the model is fine-tuned either on the full XNLI dataset (**XNLI**) or on the subset of XNLI that is in the target language (**NLI**). A second optional training phase follows, in which a model is fine-tuned on an English hate speech dataset (**En HS**). Note that **En HS** is a stand-in for the specific English dataset the model is fine-tuned on, i.e. either **DEN**, if it is fine-tuned on Vidgen et al. (2021) or **FEN**, if it is fine-tuned on Founta et al. (2018). Finally, each model is fine-tuned on 0, 20, 200, or 2000 examples of a hate speech dataset in the target language (**TL HS**). If the model was fine-tuned on NLI, we either evaluate in a standard fashion with one hypothesis (**Standard**), or with hypothesis engineering strategies (**Strategies**). The dotted arrows show that for auxiliary hypotheses of the strategies a model version that was only fine-tuned on NLI is used. Further explanations of the dataset and model notation are provided in Section 3.

datasets, that covers a range of typical, but hard to classify, cases for hate speech detection is available in all five target languages. The datasets are further described in Appendix C.

Preprocessing All datasets are cleaned with the same preprocessing steps that have been used for training of XLM-T Barbieri et al. (2022). These consist of replacing all URLs with "https" and all usernames (strings starting with an "@") with "@user". Further, the authors downsampled the non-hate speech class in **FEN** and **KEN** such that the relative frequency of hate speech increased from 5.0% to 22% and from 29.3% to 50% respectively.

Models Röttger et al. (2022a) use Twitter-XLM-RoBERTa-base (Barbieri et al., 2022), typically abbreviated to XLM-T, as a multi-lingual base model. This model is derived from XLM-R (Conneau et al., 2020) and has been further pre-trained on a multilingual Twitter corpus. Further they use the following mono-lingual base-models: AraBERT-v2 (Antoun et al., 2020) for Arabic, Hindi BERT for Hindi, UmBERTo (Parisi et al., 2020) for Italian, BERTimbau (Souza et al., 2020) for Portuguese, and RoBERTuito (Pérez et al., 2022) for Spanish.

Training The training procedure consists of two fine-tuning phases: In the optional first phase, specifically proposed by Röttger et al. (2022a), a model is fine-tuned on English

	BAS19_ES			FOR19_PT			HAS21_HI			OUS19_AR			SAN20_IT			Avg. Diff.
N	20	200	2000	20	200	2000	20	200	2000	20	200	2000	20	200	2000	
M	0.48	0.67	**0.84**	0.46	0.62	0.71	0.46	0.49	0.56	0.45	0.55	0.68	0.40	0.70	**0.78**	-0.03
X	0.40	0.61	0.82	0.45	0.52	0.71	0.46	0.46	**0.59**	0.45	0.55	**0.69**	0.40	0.66	0.76	-0.03
X + DEN	**0.66**	**0.75**	0.83	**0.63**	0.68	0.71	0.51	0.53	0.58	0.51	0.64	0.67	**0.64**	**0.71**	0.76	-0.01
X + FEN	0.54	0.70	0.82	**0.63**	**0.69**	**0.72**	**0.54**	**0.55**	0.59	**0.59**	**0.66**	0.68	**0.64**	**0.71**	0.75	-0.01
X + KEN	0.63	0.72	0.82	**0.63**	0.68	0.71	0.52	**0.55**	0.59	**0.59**	0.65	**0.69**	**0.64**	**0.71**	0.75	0.00
Avg. Diff.	-0.03	-0.02	0.02	-0.01	-0.05	-0.01	-0.01	-0.02	-0.01	-0.01	-0.06	-0.02	0.00	-0.02	-0.01	-0.02

	HateCheck_ES			HateCheck_PT			HateCheck_Hi			HateCheck_Ar			HateCheck_It			Avg. Diff.
N	20	200	2000	20	200	2000	20	200	2000	20	200	2000	20	200	2000	
M	0.44	0.48	0.59	0.35	0.50	0.62	0.23	0.23	0.23	0.26	0.23	0.24	0.28	0.39	0.54	0.01
X	0.40	0.31	0.60	0.31	0.32	0.64	0.34	0.23	0.24	0.30	0.23	0.24	0.36	0.42	0.59	0.00
X + DEN	**0.82**	**0.82**	0.80	**0.79**	**0.81**	0.78	**0.57**	0.38	0.36	**0.61**	0.48	0.35	**0.82**	**0.81**	**0.79**	0.01
X + FEN	0.57	0.60	0.64	0.56	0.59	0.62	0.34	0.30	0.34	0.35	0.34	0.29	0.54	0.55	0.58	-0.01
X + KEN	0.57	0.58	0.63	0.62	0.64	0.64	0.40	0.30	0.31	0.31	0.30	0.29	0.53	0.59	0.62	0.02
Avg. Diff.	0.02	0.00	-0.01	0.02	-0.04	0.02	0.02	-0.04	0.03	-0.02	0.00	0.02	0.02	-0.01	0.06	0.01

Table 1: Reproduction of results on held-out test sets and the Multilingual HateCheck. "Avg. Diff." contains the average difference to the original results by Röttger et al. (2022a) per row and column respectively.

hate speech (**En HS**). In the second phase, the model is fine-tuned on N target language hate speech examples (**TL HS**), where $N \in \{10, 20, 30, 40, 50, 100, 200, 300, 400, 500, 1000, 2000\}$.[2] The training setup and the corresponding model notation are included in Figure 2 as **X + En HS**. Note that **En HS** is a stand-in for the specific English dataset the model is fine-tuned on, i.e. either **FEN**, **KEN**, or **DEN**. Additionally, Röttger et al. (2022a) compare against the baselines of fine-tuning a monolingual model directly on the target language (**M**) and fine-tuning a multilingual model directly on the target language (**X**). Training specifics, including hyperparameters, are provided in Appendix A.

3.2 NLI Fine-Tuning

In order to test the effectiveness of NLI fine-tuning, we add to this setup an additional optional phase which is placed before **En HS**. This optional first phase has three variants:

M + NLI: A monolingual model in the target language is fine-tuned on the subset of XNLI (Conneau et al., 2018)[3] that is also in the target language.

X + NLI: The multilingual model is fine-tuned on the subset of XNLI that is in the target language.

X + XNLI: The multilingual model is fine-tuned on the entire XNLI dataset, concatenated with the MNLI dataset (Williams et al., 2018)[4]. To encourage cross-lingual transfer learning, the translations of premises and hyptheses are shuffled such that for a given example the premise might be in Spanish and the hypothesis in Arabic.[5]

If a model has been trained on NLI examples, we continue to train that model in an NLI formulation, even when fine-tuning on hate speech data. The model is then presented with the input text as the premise and "This text is hate speech." as the hypothesis. The label "hate speech" then corresponds to "entailment" and "not hate speech" to "contradiction".

3.3 Hypothesis Engineering

We employ the combination of the three strategies "Filtering by Target", "Filtering Counterspeech", and "Filtering Reclaimed Slurs" since they led to the best results in Goldzycher and Schneider (2022). Further implementation details are provided in Appendix B. We evaluate the strategies in combination with all models that have been fine-tuned on XNLI or a subset of it. All models that were fine-tuned on a hate speech dataset are specific to one hypothesis claiming that there is hate

---

[2]As explained later in Section 4, we only train and evaluate at $N \in \{0, 20, 200, 2000\}$.

[3]More information on the XNLI dataset is given in Appendix C.

[4]More information on MNLI is given in Appendix C.

[5]This method of shuffling translations such that the premise and hypothesis are presented in different languages to the model has been employed for popular models such as joeddav/xlm-roberta-large-xnli.
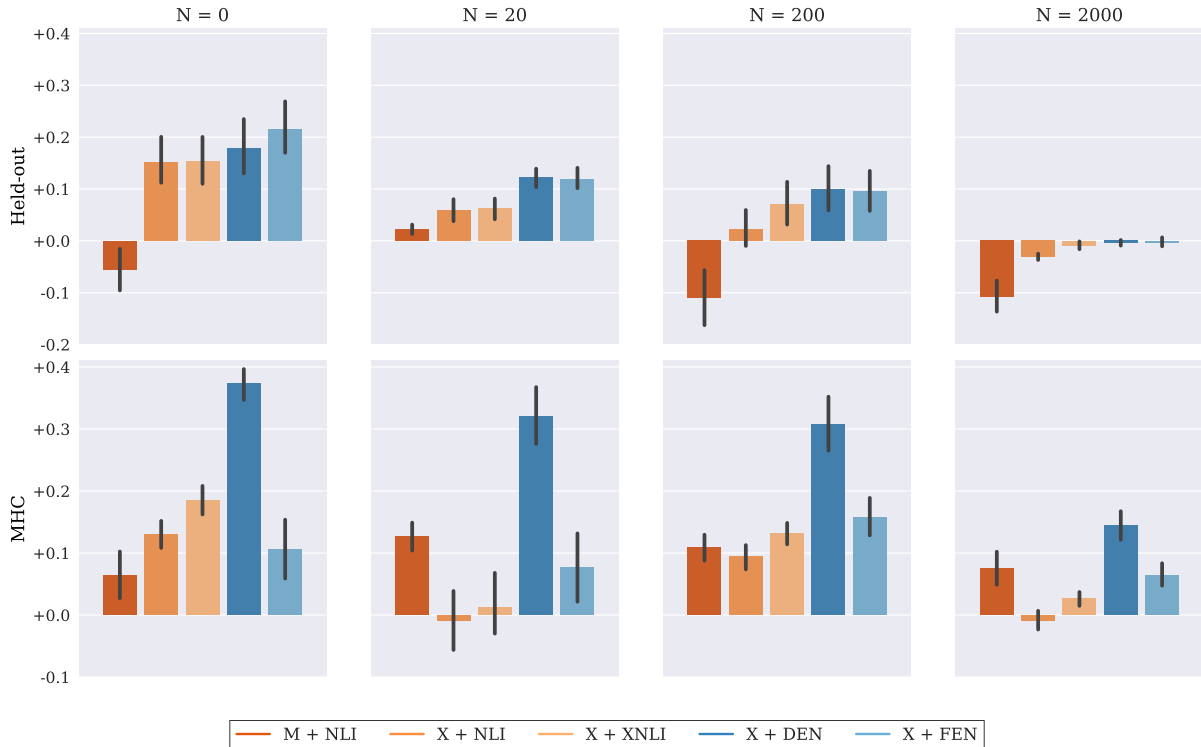
Figure 3: Evaluation of NLI fine-tuning on languages that are in XNLI, namely, Arabic, Hindi, and Spanish. The figure shows the absolute difference in macro-F_1 score when adding an intermediate fine-tuning step to fine-tuning in the target language (i.e. the difference to **M** and **X**, respectively).

speech in the input text. In such cases, we thus use the initial model, that has only been fine-tuned on an NLI dataset for all auxiliary hypothesis predictions.

4 Evaluation

Following Röttger et al. (2022a), we evaluate each setting displayed in Figure 2 on two test-sets: (1) the held-out test set in the target language and (2) the HateCheck dataset in the target language. But in contrast to their setup, we evaluate in $N \in \{0, 20, 200, 2000\}$ target language training examples. We thus evaluate at three scenarios (20, 200 and 2000) where our results are directly comparable to the results of Röttger et al. (2022a) and add a zero-shot scenario. The metric is macro-F_1 in order to account for imbalanced test sets. Like Röttger et al. (2022a), we train 10 models per setting, report the averaged results and the bootstrapped 95% confidence intervals, represented as errorbars in Figure 3 and shaded areas in Figures 4 to 7.

In what follows, we group the results according to research questions. The full results are given in Appendix D.

4.1 Can we Reproduce the Results of Röttger et al. (2022a)?

Table 1 contains the reproduction results and the average differences to the original results per language, number of examples N, and model. On average our results are lower by two percentage points on the held-out test sets and higher by one percentage point on the HateCheck test sets. We observe that: (1) Like Röttger et al. (2022a), our results follow a trend of diminishing returns. The larger performance increase often comes from increasing from 20 to 200 examples and not from increasing from 200 to 2000 examples, even though the absolute increase in examples is much larger in the second comparison. (2) Like Röttger et al. (2022a), we see that with an increasing number of examples, the benefit of fine-tuning on English hate speech decreases. At 2000 examples, the monolingual model that has directly been fine-tuned on target language examples has in most cases caught up with or even beats English fine-tuning.

Overall, we view our results as a confirmation of their findings. In order to simplify our evaluation, and since **X + FEN** and **X + KEN** have very similar results, we will only include **X + FEN** as a repre-
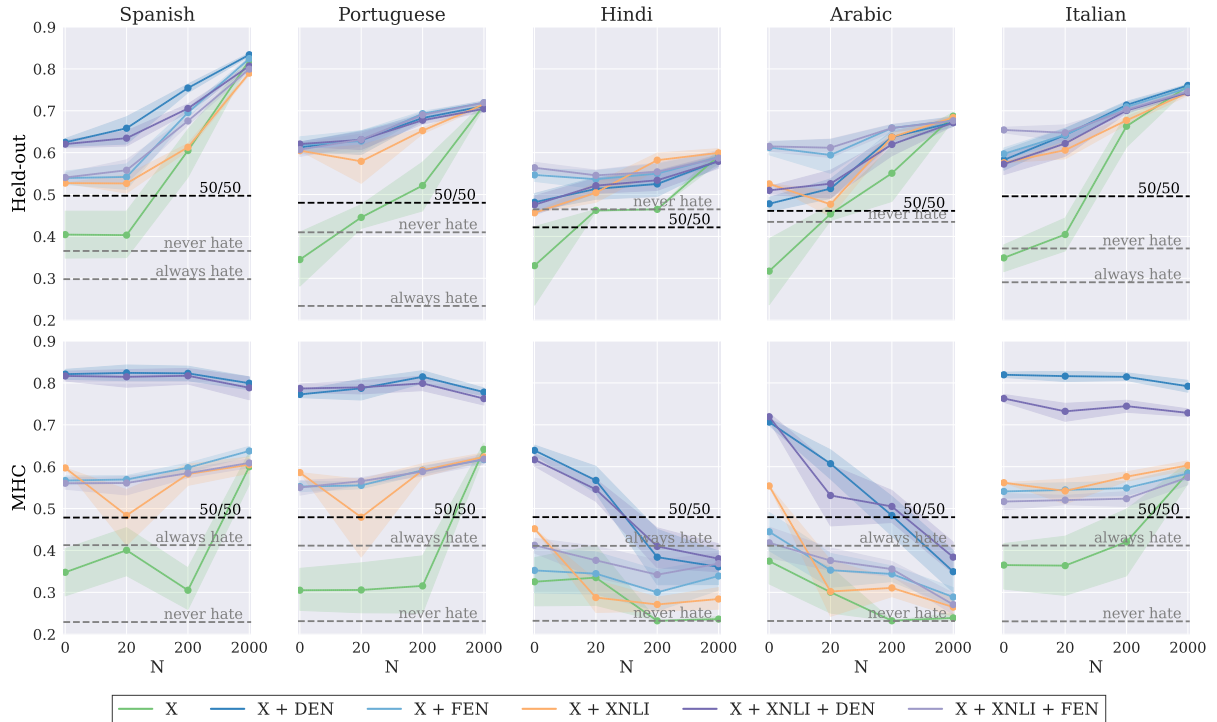
Figure 4: Evaluation of NLI fine-tuning on top of English hate speech fine-tuning. The results are given in macro-F_1.

sentative for natural data (as opposed to synthetic, adversarial data) in the following experiments.

4.2 How Does an NLI-Formulation Compare to Röttger et al. (2022a)?

We compare the baselines and settings proposed by Röttger et al. (2022a) with fine-tuning on monolingual NLI data in the target language (**M + NLI** and **X + NLI**) and fine-tuning on the entire XNLI dataset (**X + XNLI**). Only three languages (Arabic, Hindi, and Spanish) appear both in the evaluation setup and in XNLI. Since **M + NLI** and **X + NLI** training is only possible for languages in XNLI, we focus on the results in these three languages. The differences in performance averaged over the three languages are given in Figure 3.

Overall we see that introducing an intermediate fine-tuning step improves performance in most cases, but the benefits decrease with more target language examples available. While **X** is almost always improved by NLI fine-tuning, the stronger (recall from Table 1) baseline **M** only benefits from NLI fine-tuning on the HateCheck testset (second row of plots). When comparing **X + NLI** with **X + XNLI** we observe slightly more benefits from fine-tuning on the full XNLI dataset. Even though NLI-fine-tuning (orange) leads to clear benefits it is outperformed by fine-tuning on English hate

speech on **X** (blue) as proposed by Röttger et al. (2022a) in all setups. This finding raises the question if training on additional English labeled data is also beneficial in an XNLI-based setup which we aim to answer in the next section.

4.3 Are there Benefits to NLI-Finetuning when Given English Hate Speech Data?

To answer this question we compare the performance of intermediate fine-tuning on English hate speech (Röttger et al., 2022a) to first fine-tuning on XNLI and then fine-tuning on English hate speech. The results are displayed in Figure 4.

We make the following observations: (1) Since we now evaluate on all five languages, it is interesting to see how an NLI formulation performs for Portuguese and Italian which are not part of the XNLI dataset. For both testsets, we observe that even for unseen languages an NLI formulation has clear benefits over standard finetuning on target language examples (orange vs. green). (2) On held-out test sets (first row of plots), fine-tuning on both XNLI and English hate speech (purple) improves zero-shot performance in Hindi, Arabic and Italian and matches zero-shot performance of only fine-tuning on English hate speech (blue) in Spanish and Portuguese. (3) The combination of the two approaches (purple) can also lead to im-
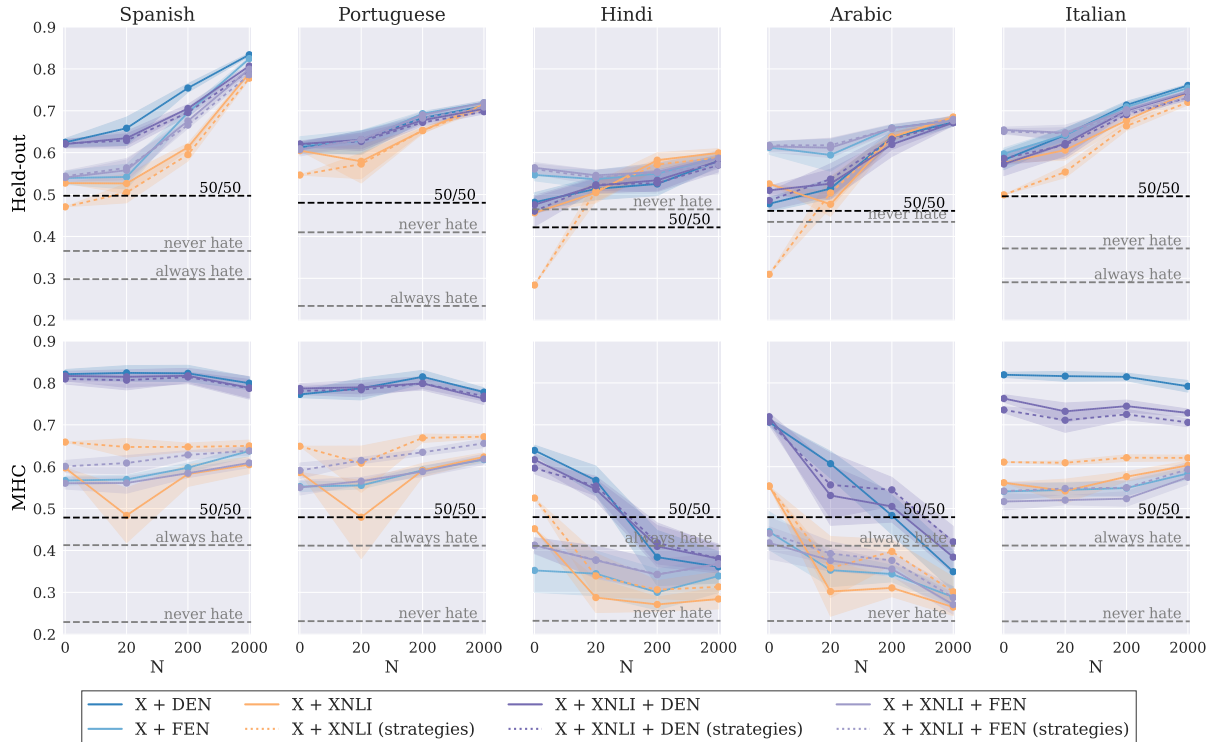
Figure 5: Evaluation of hypothesis engineering strategies.

proved few-shot results on Arabic and Hindi but reaches comparable performance or has a negative effect, when compared to only English hate speech fine-tuning (blue), in the other languages. (4) On Multilingual HateCheck (second row of plots) we observe mixed results, where additional intermediate XNLI fine-tuning tends to decrease results when fine-tuning on **DEN** but increase results when fine-tuning on **FEN**.

Overall, we find that performance improves more from intermediate fine-tuning on English (green vs. blue) than an NLI formulation (green vs. orange) and that there are negligible advantages to combining the two approaches (blue vs. purple). As discussed in the related work section, previous work by Goldzycher and Schneider (2022) showed that zero-shot hate speech detection for English could be improved with carefully engineered auxiliary hypotheses for an NLI setup. In the next section, we focus on the question whether hypothesis engineering is also beneficial for our zero-shot and few-shot setups with other languages.

4.4 Can Hypothesis Engineering further Improve Performance?

We evaluate if hypothesis engineering, specifically the three strategies proposed by Goldzycher and Schneider (2022) as described in Section 2.3, is

able to improve results. We take all models that have been fine-tuned on the full XNLI dataset and compare their performance with and without such hypothesis engineering strategies. The results are given in Figure 5.

On the held-out test sets (first row of plots), the strategies (dotted lines) show only in a few cases small positive effects but mostly negative effects. However, on HateCheck (second row of plots), we see that the strategies lead to clear performance improvements of **X + XNLI** and **X + XNLI + FEN**. This finding is not surprising since Goldzycher and Schneider (2022) specifically developed their hypothesis engineering strategies based on an error analysis on the English HateCheck data. For many languages, hypothesis engineering without intermediate English fine-tuning (orange dotted) even performs better than hypothesis engineering with fine-tuning on **FEN** (light purple dotted) which consists of Twitter data. Fine-tuning on English adversarial **DEN** hate speech examples (dark blue) remains the strongest approach for HateCheck since this fine-tuning data matches the testset conditions. However, our results show that if we have knowledge about the data the model will see at inference time (e.g. adversarial examples) but do not have matching English fine-tuning data (e.g. if only **FEN**

were available), hypothesis engineering geared towards the target domain (orange dotted line) is the best-performing approach.

5 Conclusion

In this work, we systematically evaluated the effectiveness of an NLI task formulation for hate speech detection in scenarios where only few labeled data in the target language are available. We were able to reproduce results by Röttger et al. (2022a), who showed that intermediate fine-tuning on English hate speech is beneficial in such scenarios.

Following their setup with our NLI-based experiments, we answered the following questions: (1) How does NLI fine-tuning compare to English hate speech fine-tuning? Our results showed that while NLI fine-tuning leads to strong improvements over only fine-tuning in the target language it is outperformed by English hate speech fine-tuning. (2) Are there benefits to combining NLI fine-tuning with English hate speech fine-tuning? We observed minor improvements when only 0 or 20 target language examples are available, but no benefits when more examples are available. (3) Can hypothesis engineering further improve performance on the previously best NLI-based setting? Our experiments demonstrated that hypothesis engineering can outperform other approaches only when the domain of the input data at inference time is known, but no matching training data is available.

Based on our results, we offer the following recommendations for hate speech detection in languages where little labeled data is available in the target language:

- **In general domain scenarios:** Follow Röttger et al. (2022a) and perform standard intermediate fine-tuning on English data before training on target language examples if any are available.

- **In scenarios where less English data is available, e.g. hate speech against a specific protected group:** An intermediate NLI fine-tuning step is likely to be strongly beneficial compared to only fine-tuning on limited English and target language examples.

- **In scenarios where we have knowledge about the target domain but no matching English fine-tuning data is available:** Here,

we suggest experimenting with targeted hypothesis engineering to reach the best possible performance. One exciting future avenue for this strategy is to focus on variation of protected groups across languages. Such culture and language-specific shifts will be hard to capture with English fine-tuning data but hypothesis engineering strategies offer more flexibility.

Finally, we want to highlight two areas for future work that arise from our experiments:

- **Fine-tuning phases:** We believe that a crucial circumstance limiting the effectiveness of NLI models for hate speech detection is the fact that NLI training datasets are typically from different domains, thus creating a domain gap between NLI fine-tuning and other training phases. We hypothesize that by mixing the fine-tuning phases the negative effects of the domain gap might be avoided or, at least, reduced.

- **Hypothesis engineering and model capacity:** In contrast to Goldzycher and Schneider (2022), we only observed positive results for hypothesis engineering in a few specific scenarios. One difference between the experiments in this paper and the ones in Goldzycher and Schneider (2022) that might explain the disparity in results is the model size: we use a base-sized multilingual RoBERTa model while they used an English-only BART-large model (Lewis et al., 2020), which we assume is generally more accurate in its NLI predictions and thus producing more reliable predictions for auxiliary hypotheses. Future work could thus evaluate the effect of model capacity on hypothesis engineering.

Acknowledgments

---

[6] https://www.digitalreligions.uzh.ch/en.html

Limitations

Even though our findings are backed by a large number of settings and experiments, the conclusions that can be drawn from this setup are limited in the following ways:

Datasets   We evaluated on five languages, a small fraction of the languages that could benefit from such models. Specifically, our results are limited to languages that appear in the pretraining dataset of XLM-T. Further, we showed that XNLI fine-tuning can lead to significant performance increases, even for languages that do not appear in XNLI. However, the two languages we tested on (Italian and Portuguese) are related to other languages in XNLI. There is a need for further research on how much of these benefits can be retained when the target language is less related to one of the languages in XNLI.

Models   For consistency and efficiency, we used the same multilingual model in all experiments. This means that our results are dependent on the specifics of the model, specifically the domain it has been pre-trained on and its model size. As already mentioned in Section 5, the small model size likely has had a negative impact on the results for hypothesis engineering.

Hypothesis Engineering   All results are limited to the specific strategies that we tested. Further, potential errors in the automatic translations of the hypotheses of the strategies might have impacted the results. A strength of the approach, that we did not evaluate, is that they can be adjusted to new languages and cultures simply by specifying e.g. a new set of protected groups or group characteristics via auxiliary hypotheses, thereby avoiding the criticism that zero-shot cross-lingual hate speech detection does not adjust to the specific circumstances of a language and culture (Nozza, 2022).

References

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 9–15, Marseille, France. European Language Resource Association.

Francesco Barbieri, Luis Espinosa Anke, and Jose Camacho-Collados. 2022. XLM-T: Multilingual language models in Twitter for sentiment analysis and beyond. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 258–266, Marseille, France. European Language Resources Association.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Irina Bigoulaeva, Viktor Hangya, Iryna Gurevych, and Alexander Fraser. 2023. Label modification and bootstrapping for zero-shot cross-lingual hate speech detection. *Language Resources and Evaluation (1574-0218)*.

Ke-Li Chiu and Rohan Alexander. 2021. Detecting hate speech with GPT-3. *arXiv:2103.12407 [cs]*. ArXiv: 2103.12407.

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.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. Association for Computational Linguistics.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in Text. *ACM Computing Surveys*, 51(4):85:1–85:30.

Paula Fortuna, João Rocha da Silva, Juan Soler-Company, Leo Wanner, and Sérgio Nunes. 2019. A hierarchically-labeled Portuguese hate speech dataset. In *Proceedings of the Third Workshop on Abusive*

Language Online, pages 94–104, Florence, Italy. Association for Computational Linguistics.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. *Proceedings of the International AAAI Conference on Web and Social Media*, 12(1).

Janis Goldzycher and Gerold Schneider. 2022. Hypothesis engineering for zero-shot hate speech detection. In *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, pages 75–90, Gyeongju, Republic of Korea. Association for Computational Linguistics.

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.

Hannah Kirk, Abeba Birhane, Bertie Vidgen, and Leon Derczynski. 2022a. Handling and presenting harmful text in NLP research. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 497–510, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hannah Kirk, Bertie Vidgen, and Scott Hale. 2022b. Is more data better? re-thinking the importance of efficiency in abusive language detection with transformers-based active learning. In *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, pages 52–61, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pretraining for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pretrain, Prompt, and Predict: A systematic survey of prompting methods in natural language processing. *arXiv:2107.13586 [cs]*. ArXiv: 2107.13586.

Thomas Mandl, Sandip Modha, Anand Kumar M, and Bharathi Raja Chakravarthi. 2020. Overview of the HASOC Track at FIRE 2020: Hate Speech and Offensive Language Identification in Tamil, Malayalam, Hindi, English and German. In *Forum for Information Retrieval Evaluation*, pages 29–32, Hyderabad India. ACM.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the HASOC Track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages. In *Proceedings of the 11th Forum for Information Retrieval Evaluation*, FIRE '19, pages 14–17, New York, NY, USA. Association for Computing Machinery. Eventplace: Kolkata, India.

Sanguinetti Manuela, Comandini Gloria, Elisa Di Nuovo, Simona Frenda, MARCO ANTONIO Stranisci, Cristina Bosco, Caselli Tommaso, Viviana Patti, Russo Irene, et al. 2020. HaSpeeDe 2 @ EVALITA2020: Overview of the EVALITA 2020 hate speech detection task. In *Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, pages 1–9. CEUR.

Sandip Modha, Thomas Mandl, Gautam Kishore Shahi, Hiren Madhu, Shrey Satapara, Tharindu Ranasinghe, and Marcos Zampieri. 2022. Overview of the HASOC subtrack at FIRE 2021: Hate speech and offensive content identification in english and indoaryan languages and conversational hate speech. In *Proceedings of the 13th Annual Meeting of the Forum for Information Retrieval Evaluation*, FIRE '21, page 1–3, New York, NY, USA. Association for Computing Machinery.

Syrielle Montariol, Arij Riabi, and Djamé Seddah. 2022. Multilingual auxiliary tasks training: Bridging the gap between languages for zero-shot transfer of hate speech detection models. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 347–363, Online only. Association for Computational Linguistics.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2022. Cross-Lingual Few-Shot Hate Speech and Offensive Language Detection Using Meta Learning. *IEEE Access*, 10:14880–14896. Conference Name: IEEE Access.

Debora Nozza. 2022. Nozza@LT-EDI-ACL2022: Ensemble modeling for homophobia and transphobia detection. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 258–264, Dublin, Ireland. Association for Computational Linguistics.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684, Hong Kong, China. Association for Computational Linguistics.

Loreto Parisi, Simone Francia, and Paolo Magnani. 2020. UmBERTo: An italian language model trained with whole word masking. `https://github.com/musixmatchresearch/umberto`.

Juan Manuel Pérez, Damián Ariel Furman, Laura Alonso Alemany, and Franco M. Luque. 2022. RoBERTuito: A pre-trained language model for social media text in Spanish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7235–7243, Marseille, France. European Language Resources Association.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: A systematic review. *Language Resources and Evaluation*, 55(2):477–523.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Paul Röttger, Debora Nozza, Federico Bianchi, and Dirk Hovy. 2022a. Data-efficient strategies for expanding hate speech detection into under-resourced languages. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5674–5691, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Paul Röttger, Haitham Seelawi, Debora Nozza, Zeerak Talat, and Bertie Vidgen. 2022b. 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.

Fábio Souza, Rodrigo Nogueira, and Roberto Lotufo. 2020. BERTimbau: Pretrained BERT models for Brazilian Portuguese. In *9th Brazilian Conference on Intelligent Systems, BRACIS, Rio Grande do Sul, Brazil, October 20-23 (to appear)*.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate 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 1: Long Papers)*, pages 1667–1682, Online. Association for Computational Linguistics.

Sinong Wang, Han Fang, Madian Khabsa, Hanzi Mao, and Hao Ma. 2021. Entailment as few-shot learner. *arXiv:2104.14690 [cs]*. ArXiv: 2104.14690.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Ankit Yadav, Shubham Chandel, Sushant Chatufale, and Anil Bandhakavi. 2023. LAHM : Large annotated dataset for multi-domain and multilingual hate speech identification. *CoRR*, abs/2304.00913.

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

Wenpeng Yin, Jamaal Hay, and Dan Roth. 2019. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3914–3923, Hong Kong, China. Association for Computational Linguistics.

Haris Bin Zia, Ignacio Castro, Arkaitz Zubiaga, and Gareth Tyson. 2022. Improving zero-shot cross-lingual hate speech detection with pseudo-label fine-tuning of transformer language models. *Proceedings of the International AAAI Conference on Web and Social Media*, 16:1435–1439.

A Training Details

All models were trained and evaluated using the `transformers` library (Wolf et al., 2020), version 4.26.1. The hyperparameters for the NLI-fine-tuning stage are provided in Table 2. We always

parameter	value
epochs	5
learning rate	2e-05
batch size	32
max sequence length	128

Table 2: Hyperparameters of NLI fine-tuning.

chose the best performing checkpoint on the validation set out the five epochs.

For intermediate fine-tuning on English hate speech and fine-tuning on target language examples, we used the same hyperparameters as Röttger et al. (2022a). They are given in Table 3 and Table 4.

parameter	value
epochs	3
learning rate	5e-05
batch size	16
max sequence length	128

Table 3: Hyperparameters for fine-tuning on English hate speech datasets.

parameter	value
epochs	5
learning rate	5e-05
batch size	16
max sequence length	128

Table 4: Hyperparameters for fine-tuning on hate speech datasets in the target language.

All other hyperparameters were kept at the default values of the huggingface Trainer-class.

B Hypothesis Engineering Details

Since we worked with monolingual models in the target language and with multilingual models that have been fine-tuned on English and other languages first, this raises the question in which language the hypotheses should be expressed. For monolingual models, we automatically translated all hypotheses with Google Translate to the target language. The model thus received the premise and hypothesis in the same language as input. For multilingual models, we kept the original English hypotheses. Since we shuffled the languages of premise and hypothesis in the NLI training-regime the models should be able to handle the differing languages well. Keeping the hypotheses in multilingual models in English also means that the hypothesis remains in the same language over multiple fine-tuning phases like intermediate fine-tuning on English hate speech data and target language fine-tuning.

Goldzycher and Schneider (2022) proposed two versions of the strategy "Filtering by Target": In the first version the model predicts if protected groups are targeted (e.g. "This text is about Muslims."). In the second version the model predicts if protected group characteristics are targeted (e.g. "This text is about religion."). Even though the second version performed worse in their experiments, we use this second version for our experiments, because its predictions are more neutral with respect to specific languages and cultures. This enabled us to use exactly the same strategies for each language. In a more sophisticated setup one could implement the first version predicting protected groups and adjust these groups for each language.

C Datasets

The key characteristics of all datasets we used in the experiments are described in Table 5. To create the MultiNLI dataset, sentences from diverse genres were collected and used as premises. Annotators then were tasked with creating artificial hypotheses for these premises. For XNLI, the test set was translated by human translators and the training set was translated automatically.

D Full Results

In order to highlight specific aspects of our results, we split them up over several Figures in the paper. The full results of all settings that we evaluated are provided in Figures 6 and 7.

code	paper	train	validation	test	% hate	source
		(X)NLI datasets				
MNLI	Williams et al. (2018)	40000	19650	-	-	diverse
XNLI	Conneau et al. (2018)	393000	24900	-	-	translation of MNLI
		English hate speech datasets (En HS)				
FEN	Founta et al. (2018)	20068	500	-	22.0	Twitter
KEN	Kennedy et al. (2020)	20692	500	-	50.0	Youtube, Twitter, Reddit
DEN	Vidgen et al. (2021)	38644	500	-	53.9	annotators, adversarial
		Target Language Hate Speech Datasets (TL HS)				
BAS19_ES	Basile et al. (2019)	4100	500	2000	41.5	Twitter
FOR19_PT	Fortuna et al. (2019)	3170	500	2000	31.5	Twitter
HAS21_HI	Modha et al. (2022)	3794	300	500	12.3	Twitter
OUS19_AR	Ousidhoum et al. (2019)	2053	300	1000	22.5	Twitter
SAN20_IT	Manuela et al. (2020)	5600	500	2000	41.8	Twitter
		Multilingual HateCheck (MHC)				
HateCheck_ES	Röttger et al. (2022b)	-	-	3745	70.3	annotators
HateCheck_PT	Röttger et al. (2022b)	-	-	3691	69.9	annotators
HateCheck_HI	Röttger et al. (2022b)	-	-	3565	69.8	annotators
HateCheck_AR	Röttger et al. (2022b)	-	-	3570	69.9	annotators
HateCheck_IT	Röttger et al. (2022b)	-	-	3690	70.0	annotators

Table 5: Statistics of training and evaluation datasets. In the case of FEN and KEN, we applied downsampling to the non-hate speech class. The table reflects the state after downsampling.
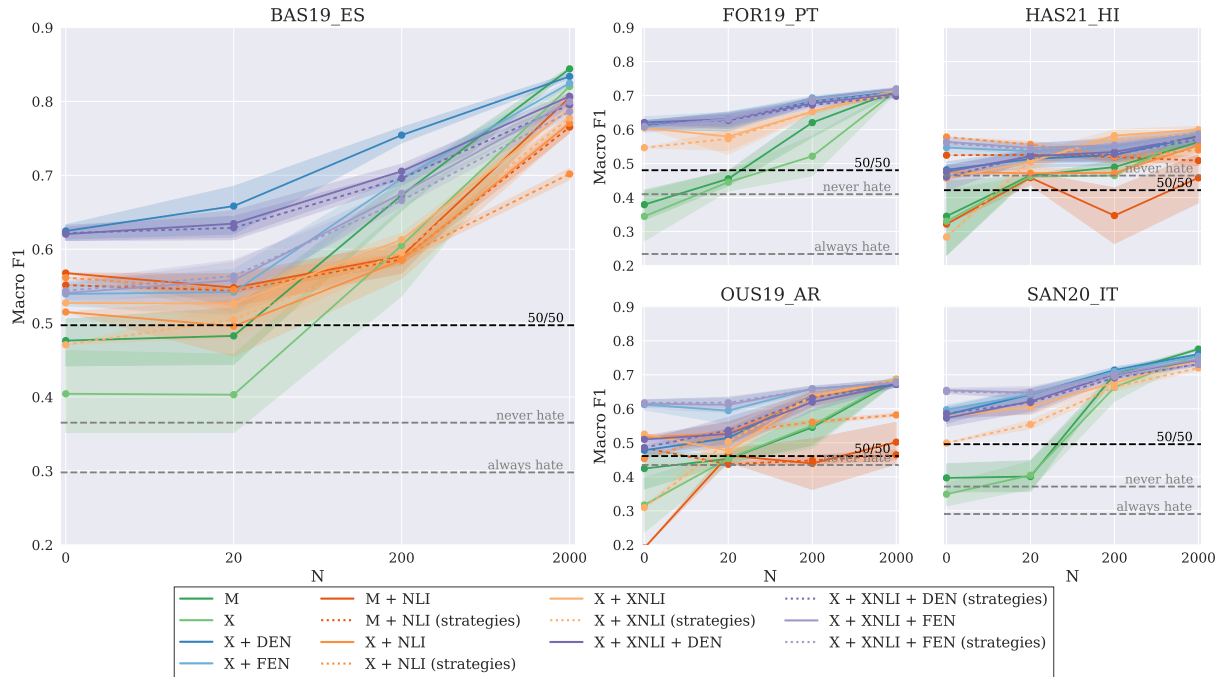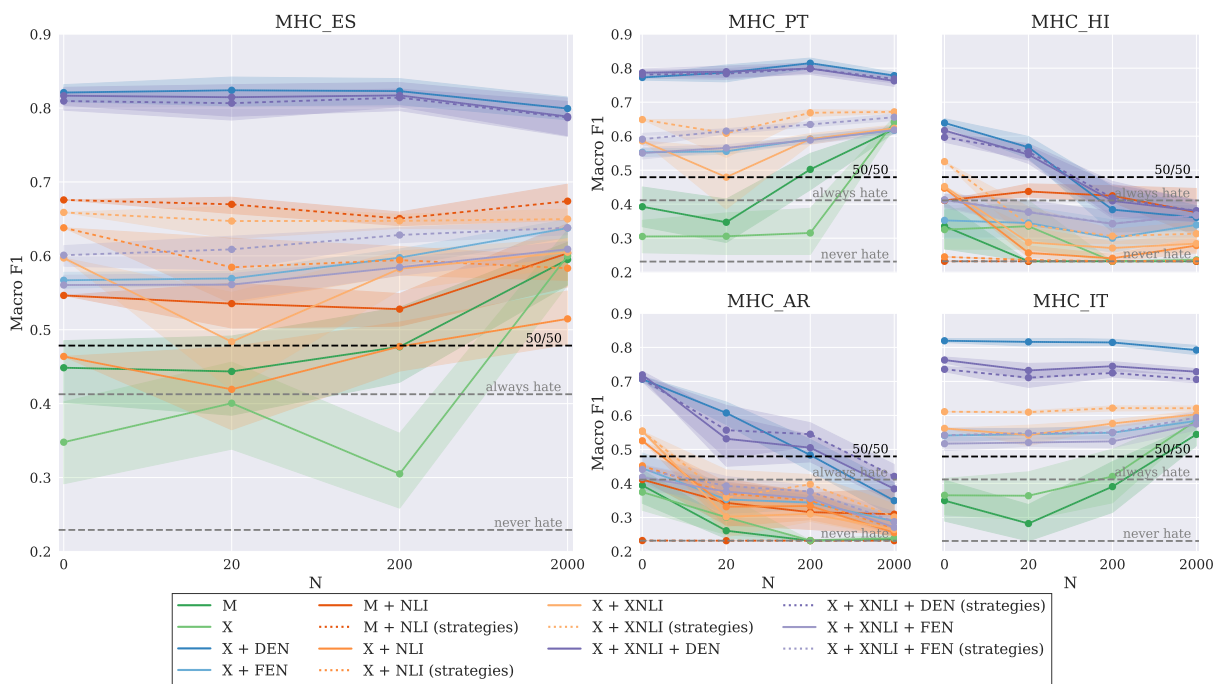

Figure 6: The full results on held-out test sets.

Figure 7: The full results on the multilingual HateCheck.