# How to Solve Few-Shot Abusive Content Detection Using the Data We Actually Have

**Viktor Hangya and Alexander Fraser**
Center for Information and Language Processing, LMU Munich,
Munich Center for Machine Learning,
{hangyav, fraser}@cis.lmu.de

**Abstract**

Due to the broad range of social media platforms, the requirements of abusive language detection systems are varied and ever-changing. Already a large set of annotated corpora with different properties and label sets were created, such as hate or misogyny detection, but the form and targets of abusive speech are constantly evolving. Since, the annotation of new corpora is expensive, in this work we leverage datasets we already have, covering a wide range of tasks related to abusive language detection. Our goal is to build models cheaply for a new target label set and/or language, using only a few training examples of the target domain. We propose a two-step approach: first we train our model in a multitask fashion. We then carry out few-shot adaptation to the target requirements. Our experiments show that using already existing datasets and only a few-shots of the target task the performance of models improve both monolingually and across languages. Our analysis also shows that our models acquire a general understanding of abusive language, since they improve the prediction of labels which are present only in the target dataset and can benefit from knowledge about labels which are not directly used for the target task.

**Keywords:** abusive content detection, transfer learning, few-shot training

## 1. Introduction

The wide spread of social media allowed us to communicate and share our opinions quickly and conveniently. However, it gives place to abusive content as well, which leaves some groups of people vulnerable. To push back abusive online content, various automated systems, and more importantly datasets (Poletto et al., 2021), were introduced covering various text genres such as forum (de Gibert et al., 2018), Twitter (Struß et al., 2019) or Instagram posts (Suryawanshi et al., 2020) of various languages (Vidgen and Derczynski, 2020), user groups such as women (Fersini et al., 2018) or LGBTQ+ (Leite et al., 2020) and tasks including hate speech (de Gibert et al., 2018), offensive language (Zampieri et al., 2019) or toxicity (Leite et al., 2020) detection, etc.

On the other hand, there is constantly a need to annotate new datasets supporting previously unseen target scenarios. To reduce annotation costs, related work leveraged transfer learning to build systems across languages (Ranasinghe and Zampieri, 2020) and domains (Glavaš et al., 2020). But finding the right source datasets is often challenging, since the label sets could differ. To alleviate the problem, previous work manually altered the label sets of the source datasets in order to match them to the target requirements. However, this approach is problematic, because it requires expertise in abusive language datasets, since i) the rules developed by other researchers for manual label matching are not reusable due to the rapid change in the application requirements

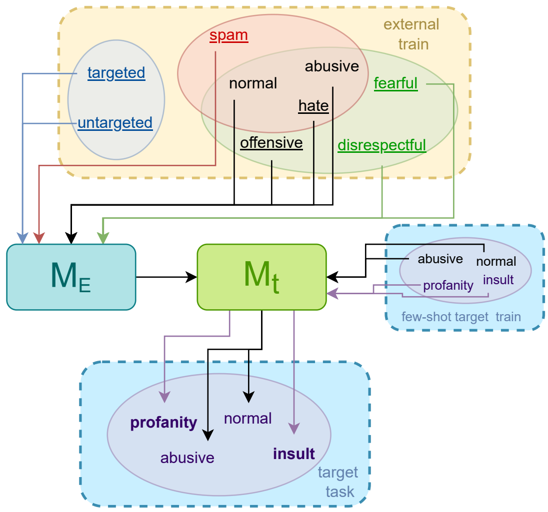


Figure 1: Two-step approach: $M_E$ is trained on the (external) datasets we already have, followed by its adaptation to the target task ($M_t$) with only a few-shots. Labels not directly used for the target task are underlined, target labels not contained in the external datasets are **bolded**.

and ii) the definition of the same label in some datasets could conflict, e.g., the offensive label of the `OLID` dataset includes profane language (Zampieri et al., 2019), while the same label does not in `HASOC` (Mandl et al., 2019). Thus, a precise understanding of abusive language phenomena is required. Additionally, iii) novel fine-grained labels do not have alternatives to be transferred from. Such fine-grained labels, e.g., related to a



*LREC-COLING 2024*, pages 8307–8322

specific view of a given community regarding an event, can be created on-the-fly as moderators or affected people face them. Thus, our goal is to eliminate the need for such rules and make information transfer more flexible with minimal target task annotations.

To this end, we introduce a method leveraging multiple already existing (external) datasets in order to understand general abusive language, allowing to build models cheaply for the target requirements without the need for manual dataset modifications. As shown in Figure 1, different datasets can inform the model about different types of abusive content. Some classes can directly be leveraged for the target task due to their matching label names and definitions. Others, such as labels with matching names but conflicting definitions or labels which are not contained in the target dataset at all, are leveraged only indirectly, by contributing to the general abusive language awareness of the model. Our approach consists of two steps: jointly training a language model on multiple external datasets using prompt-learning (Schick and Schütze, 2021). We then adapt the resulting model to the target requirements in the second step, using only a few samples per label from the target task (4-shots in the main experiments), which could even be created on-the-fly. Since the target task can contain unseen labels, i.e., labels which are not contained in any of the external datasets, or classes with conflicting label definitions, at least a few annotated samples are needed for model specialization.

We test our method on various tasks (e.g., hate, abuse and misogyny detection or target identification) in both monolingual and cross-lingual (English→German, →Italian, →Brazilian Portuguese, and →Hindi) setups. Additionally, our datasets cover multiple platforms including longer forum posts and shorter Twitter messages. Our experiments show improved performance when training using the external datasets compared to various baselines, including both monolingual and cross-lingual settings, on both binary and fine-grained test sets. We find that even unseen target labels are improved due to the better general abusive language understanding of our models. Our ablation study shows that external-only labels (labels which do not occur in the label set of the task we are carrying out) improve performance, showing that they contribute to general understanding as well. We experiment with different target data sizes and find that although our approach is more beneficial at lower sizes, when more data is available it is also effective. Finally, we perform model diagnostics using HateCheck (Röttger et al., 2021), further supporting our claim of better general abusive language understanding.

Our contributions are the following:[1]

- a multi-dataset training (MDT) approach, using prompt-learning based fine-tuning, for an efficient few-shot training which supports the ever-changing nature of abusive language detection,
- applicability across languages and text genres to support a wide range of target tasks cheaply,
- comprehensive analyses for a better understanding of model behavior.

## 2. Related Work

To alleviate the issue of missing datasets for a given target task, previous work leveraged transfer learning techniques. Ranasinghe and Zampieri (2020) built hate speech classifiers for Hindi, Spanish and Bengali by relying only on an English training dataset, while Glavaš et al. (2020) followed a similar approach for cross-domain experiments. They made the train and test corpora compatible using rule based label adaptation which, as discussed above, is often difficult. In this work, we eliminate this step and use external datasets without any modifications. Furthermore, Wiegand et al. (2018) showed that by adding seemingly similar English samples to a small amount of German training data the results decreased, while Nozza (2021) found that in zero-shot cross-lingual models language specific interjections are often misinterpreted leading to errors. These results indicate that selecting the right source dataset is not straight forward (and perhaps impossible in some cases). In this work, we leverage multiple external datasets for a robust abusive language understanding, and use a few-shots from the target dataset to specialize our models to the target domain. Similarly, Röttger et al. (2022) argues for the use of a small amount of target language training samples in order to extend hate detection to multiple languages. However, they focus on compatible binary datasets, while our approach is compatible with fine-grained tasks with unseen labels as well.

In contrast to transfer learning, the goal of multitask learning (MTL) is to build a shared model using various tasks in order to improve the performance on all of them, by exploiting common information in some tasks, and to perform multiple tasks with a single model (Caruana, 1997). Stickland and Murray (2019) proposed an MTL method based on pre-trained language models by introducing task specific parameters in each layer. Due to negative task interference however, single task

[1] Our code is publicly available at: `https://cistern.cis.lmu.de/multi_hs`

models perform best in many cases. To mitigate the issue of task interference, Pfeiffer et al. (2021) used adapters (Houlsby et al., 2019) in a multitask setting, showing that fusing information learned by task specific adapters can further boost the performance on a target task. However, MTL is not able to induce useful task dependencies when an imbalanced set, in terms of dataset sizes, is available. In contrast, a set of auxiliary tasks were used to improve the performance on a single target task in (Watanabe et al., 2022). Similarly, Mehmood et al. (2020) perform a final training step on the target biomedical NER task after MTL, while Kapil and Ekbal (2020) combine various abusive language related datasets. However, these methods rely on a large set of training data for the target task in order to improve the performance, which is often unavailable in the ever-changing field of abusive language detection. Magnossão de Paula et al. (2023) proposed to use task embeddings in order to reduce negative transfer in MTL, but they only consider binary abusive language detection tasks. In comparison, we use a wide range of tasks with heterogeneous label sets, including external-only labels, when just a little target data is available, and we show that this leads to a better general abusive language understanding which positively impacts even unseen target labels.

Our approach is also related to meta-learning (Hospedales et al., 2021), where the goal is to build a general model that is cheaply adaptable to a target task that is unknown at the time of meta-learning. In contrast, our goal is to build a classifier for a resource poor task that is known when training the full model, by leveraging already existing highly related datasets without strong expertise in abusive language phenomena. Additionally, Wang et al. (2021) showed that meta-learning has similar performance to MTL, thus we only use MTL as a baseline for simplicity.

## 3. Approach

We consider two sets of training corpora in our multi-dataset training (MDT) approach: external datasets ($D_E = \{D_{e_i} : i = 1..N\}$) which are not directly related to the target task and the target dataset ($D_t$) which is the target task for which we aim to build a classifier. The former are off-the-shelf datasets created for other tasks and/or languages containing a few thousands or sometimes tens of thousands of samples. In contrast, since our main goal is to reduce the costs of building systems for novel target tasks, $D_t$ contains only a few samples, 4-shots per label in our main experiments. MDT builds abusive language classifiers in two steps (Figure 1). First, we train a single model by fine-tuning a pre-trained LM ($M_0$) using only the external datasets in order to learn general abusive language understanding (resulting in $M_E$), which we adapt to the specificities of the target task in the second step (resulting in $M_t$). In contrast to MTL or meta-learning, where the final model supports multiple tasks, our final models ($M_t$) are built for a single target task. This imitates the use cases of social media platforms which need to build a specialized model supporting their own specific requirements. First, we discuss the used prompt-learning technique (Schick and Schütze, 2021), followed by the introduction of the two-step training.

### 3.1. Prompt-Learning

Prompt-learning was shown to be effective when only a small training set is available. Instead of using classification heads on top of pre-trained LMs, it aims to solve the task at hand using text generation. Depending on the used LM architecture, various techniques exist. We rely on encoder-only LMs in our experiments, thus use the method proposed by Schick and Schütze (2021), which employs the masked language modeling task (MLM) to perform text classification. Using pattern-verbalizer-pairs (PVPs) an input sentence is first transformed using the pattern, e.g., *I'll kill you.* $\rightarrow$ *I'll kill you. It was [MASK]*, and the task is to output the probability distribution over the vocabulary items at the masked position. Finally, the verbalizer maps the highest probability token, out of a set of valid tokens (see below), to labels of a given dataset, e.g., *threatening* $\rightarrow$ $\mathtt{threat}$ or *neutral* $\rightarrow$ $\mathtt{normal}$. During fine-tuning the model is trained to predict the token associated with the correct input label, using the MLM objective.

In our multi-dataset setup, we define PVPs for all external datasets and the target dataset separately ($PVP_E = \{PVP_{e_i} : i = 1..N\}$ and $PVP_t$). Having a dedicated pattern and verbalizer for each dataset makes our approach easy to be specialized for each dataset, and at the same time easy to use, since no single verbalizer handling all the labels is required. We defined only two different patterns: one for target detection datasets (`X` $\rightarrow$ `X` It was targeted at `[MASK]`, where `X` is the input text) and another for the rest (`X` $\rightarrow$ `X` It was `[MASK]`). On the other hand, the used verbalizers are specific for the label set of each dataset, which can be defined easily in general, i.e., we used 1-to-1 token to label mapping in most cases. We refer to Table 6 of the Appendix for more details about the exact patterns and verbalizers for each used dataset.

### 3.2. Multi-Dataset Training

**Step 1: General Model Training ($M_0 \rightarrow M_E$)** In each step of the training process we randomly select an external dataset $D_{e_i}$ and a batch of samples from it. Other than the shared model core, i.e., the pre-trained LM, we use the PVP related to $D_{e_i}$ for the forward-backward pass. For each dataset $D_{e_i}$ we use cross-entropy loss as the objective function $L_{e_i}$ to update the model. We run this process until convergence.

**Step 2: Model Specialization ($M_E \rightarrow M_t$)** In order to adapt $M_E$ to the target task, we simply continue training it on $D_t$ by using the $D_t$ specific PVP. Similarly to the above, we use cross-entropy loss $L_t$ to update the model until convergence. As shown by our experiments, the general abusive language understanding learned by $M_E$ helps this step to build a better model using just a few training samples.

## 4. Experimental Setup

### 4.1. Datasets

We selected a wide range of datasets for our experiments, covering various abusive language detection tasks, languages and text genres. We give a short overview in the following and further details, such as number of samples, exact PVPs, etc., in Table 6 of the Appendix.

**AMI** was created for the *Evalita 2018 shared task on Automatic Misogyny Identification* (Fersini et al., 2018), containing English and Italian tweets. We use both the binary and fine-grained misogyny labels as well as the target identification labels.

**GermEval** was introduced for the shared task on the *Identification of Offensive Language* in German tweets (Struß et al., 2019). We used both binary and fine-grained label sets.

**HASOC** The shared task on *Hate Speech and Offensive Content Identification* (Mandl et al., 2019) introduces datasets for English, German and Hindi containing Twitter and Facebook posts. We used its fine-grained abuse and target identification labels.

**HatEval** was built for *SemEval 2019 Task 5* about the detection of hate speech against immigrants and women in Spanish and English Twitter messages (Basile et al., 2019). We used its binary hate speech and target identification label sets.

**LSA** is a large scale fine-grained abusive dataset of English Tweets (Founta et al., 2018).

**MLMA** Ousidhoum et al. (2019) introduced a multilingual and multi-aspect hate speech dataset of English, French and Arabic Tweets. We leveraged the fine-grained hostility labels in English.

**OLID** The Offensive Language Identification Dataset contains English tweets annotated with offensive labels on three layers (Zampieri et al., 2019). We used its binary offensive text and target identification subsets.

| Dataset | Labels |
|---|---|
| External datasets ($D_E$) | |
| AMI binary misogyny En | misogyny, normal |
| AMI fine-grained misogyny En | stereotype, dominance, derailing, sexual_harassment, discredit |
| AMI binary target En | active, passive |
| HASOC fine-grained abusive En | hate, offensive, profanity |
| HASOC binary target En | targeted, untargeted |
| HatEval binary target En | individual, group |
| LSA fine-grained abusive En | abusive, hateful, spam, normal |
| MLMA fine-grained hostility En | abusive, hateful, offensive, disrespectful, fearful, normal |
| SRW fine-grained abusive En | sexism, racism, normal |
| Target datasets ($D_t$) | |
| HASOC fine-grained abusive En | hate, offensive, **profanity** |
| HASOC fine-grained abusive Hi | hate, offensive, **profanity** |
| HASOC fine-grained abusive De | hate, offensive, **profanity** |
| GermEval fine-grained offensive De | profanity, **insult**, abusive, normal |
| ToLD-Br fine-grained toxicity Pt-Br | **LGBTQ+phobia**, **obscene**, **insult**, racism, misogyny, **xenophobia**, normal |
| OLID fine-grained target En | individual, group, **other** |
| Stormfront binary hate En | hate, normal |
| HatEval binary hate En | hateful, normal |
| HatEval binary hate Es | hateful, normal |
| OLID binary offensive En | offensive, normal |
| GermEval binary offensive De | offensive, normal |
| AMI binary misogyny En | misogyny, normal |
| AMI binary misogyny It | misogyny, normal |

Table 1: Multi-dataset setup including external ($D_E$) and target ($D_t$) datasets. We consider similarly defined but differently named labels to be the same, such as hate and hateful, sexism and misogyny or individual and active. We bold **unseen** labels and underline labels which aren't used in any target dataset. We remove external datasets from $D_E$ which are from the same source as a given target dataset (in case of AMI, HASOC and HatEval).

**SRW** is an English Twitter set created for sexism and racism detection (Waseem and Hovy, 2016).

**Stormfront** was created for hate speech detection containing English forum posts from the *Stormfront* white supremacist forum (de Gibert et al., 2018). It is annotated with binary labels.

**ToLD-Br** is a Brazilian Portuguese Twitter dataset annotated for toxicity detection (Leite et al., 2020). We used its fine-grained label set containing a wide range of labels, including misogyny.

### 4.2. Multi-Dataset Setup

In the following we describe our multi-dataset setup, i.e., the 9 corpora in the external set ($D_E$) and the 13 target ($D_t$) datasets. For a high-level overview of the setup, including labels of the external and target datasets, we refer to Table 1. Note that although we list multiple target datasets, we build a dedicated model for each of them separately ($M_t$) in step 2 using $M_E$ which is trained on all the datasets in $D_E$ jointly in step 1. Although the goal of MDT is to allow for an easy external dataset selection specific to a target task, we only consider a single external set for all target datasets to save resources. However, this general setup shows that an external set can be selected without very careful data selection, and gives a lower bound of the achievable performance as we expect more benefits from a more specialized setup. The goal of the setup is to include a wide range of datasets related to abusive language detection, such as hate speech, offense, abuse, sexism, racism detection as well as target identification. Additionally, we include datasets from the same task category but with different label sets, e.g., `HASOC fine-grained abusive` (hate, offensive, profane) and `SRW fine-grained abusive` (sexism, racism, normal). We only include English datasets in $D_E$, while we used both English and non-English corpora (De, Hi, It, Pt-Br) as the target datasets to test cross-lingual transfer as well. Furthermore, we test on Stormfront which contains forum posts instead of Twitter and Facebook messages as the datasets in $D_E$ do. To avoid data leakage between the external train and the target test sets (in case of `AMI`, `HASOC` and `HatEval`), i.e., to filter samples which have the same input samples but with different labels or inputs from different languages with the same labeling methodology, we remove all datasets from $D_E$ which are from the same authors as the test set, e.g., we omit all `AMI` external datasets when training $M_E$ in step 1 in case we test on `AMI binary misogyny It`.[2]

### 4.3. Compared Systems

We compare MDT to four types of baseline systems. We use off-the-shelf pre-trained LMs and train them using the few-shot setup as in the second step of our proposed approach without training them on the external datasets (LM-base). As shown by Gururangan et al. (2020) fine-tuning LMs on the domain of the task of interest by further MLM training on unlabeled data can improve down-stream task performance. In order to test the effectiveness of this step in contrast to our approach which leverages labels instead, we run MLM on the external datasets of the above-mentioned setups for one epoch (MLM). To test the importance of the two separate steps of our approach we perform multitask learning, i.e., both the external datasets and the target dataset are used in a single step similarly as in step 1 in Section 3 (MTL). Additionally, we test adapter-fusion (Pfeiffer et al., 2021) which eliminates negative task interference by first, training independent adapters on each dataset, followed by combining them for the target task (Fusion).[3]

**Model parameters** We use *xlm-roberta-base* as our base LM (Conneau et al., 2020) for all the baselines and our MDT setups as well. In our early experiments we tested *bert-base-multilingual-cased* (Devlin et al., 2019) as well, which resulted in similar conclusions. However, XLM-R benefited slightly more from MDT which suggest that even larger models might be able to exploit general information from external datasets to a higher degree. For evaluation we used macro averaged $F_1$ score averaged over 5 different seeds[4] in order to reduce the high variance issue of few-shot classification (Zheng et al., 2022). We follow the standard n-shot setting for few-shot learning. Due to the high label bias of abusive language datasets, a large number of samples have to be considered for annotation in order to increase the training size of the minority classes with even a few examples. Thus, we selected $n = 4$ as a reasonable trade-off between costs and the amount of training data for the target datasets (step 2). We experiment with different $n$ values shown in Figure 2. We used the full training and validation sets of the external datasets (step 1). For all datasets we use the official train, validation and test splits if given, otherwise we take 80/20 train/test split of the full dataset and/or an additional 80/20 split of the train set for final training and validation if the latter is not given. For the implementation we used the *Huggingface transformers* (Wolf et al., 2020) and *OpenPrompt* (Ding et al., 2022) libraries for prompt-learning. The used hyperparameters are: batch size $1$,[5] gradient accumulation steps $16$, warm-up steps $10$, learning rate $5 \times 10^{-5}$ and dropout $0.1$. We ran a single epoch on each external dataset in step 1, since we found that longer training made our models biased towards some of the most frequent labels. In contrast, we used early stopping on the target $n$-shot validation set in step 2.

## 5. Results

First, we present our main results followed by the analysis of different few-shot sizes and the model performance on each label separately. Then we discuss an ablation study for a better understanding of how the external datasets affect the final performance. Finally, we briefly discuss our experiments on HateCheck.

[2]Note that this filtering step requires us to train separate $M_E$ models in these cases, resulting in 4 different $M_E$ models overall.

[3]We use prompt-training in contrast to the original work which uses classification heads.

[4]We use a single seed for MTL since it is trained on a large set of inputs jointly.

[5]Due to limited GPU memory, we could not test on larger batch sizes.

| | fine-grained | | | | | | |
|---|---|---|---|---|---|---|---|
| | abusive HASOC (1/3) | | | offensive GermEval | toxicity ToLD-Br | target OLID | |
| | En | Hi | De | De (1/4) | Pt-Br (4/7) | En (1/3) | avg. |
| LM-base | $32.76^{\pm5.94}$ | $33.59^{\pm5.68}$ | $32.55^{\pm11.71}$ | $21.48^{\pm2.22}$ | $8.25^{\pm3.54}$ | $36.94^{\pm3.80}$ | 27.59 |
| MLM | $35.98^{\pm7.70}$ | $\mathbf{35.81}^{\pm3.79}$ | $28.70^{\pm9.94}$ | $21.97^{\pm3.89}$ | $8.98^{\pm2.76}$ | $40.88^{\pm5.38}$ | 28.72 |
| MTL | $13.22^{\pm0.00}$ | $16.38^{\pm0.00}$ | $24.10^{\pm0.00}$ | $19.84^{\pm0.00}$ | $10.69^{\pm0.00}$ | $14.11^{\pm0.00}$ | 16.39 |
| Fusion | $37.35^{\pm8.06}$ | $34.29^{\pm4.14}$ | $29.18^{\pm8.09}$ | $21.33^{\pm3.88}$ | $8.26^{\pm1.20}$ | $42.45^{\pm4.03}$ | 28.81 |
| MDT | $\mathbf{40.48}^{\pm5.37}$ | $34.36^{\pm2.59}$ | $\mathbf{33.96}^{\pm7.83}$ | $\mathbf{27.70}^{\pm3.78}$ | $\mathbf{12.83}^{\pm1.78}$ | $\mathbf{49.55}^{\pm2.92}$ | **33.14** |

| | binary | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | hate | | | offensive | | misogyny | | |
| | Stormfront | HatEval | | OLID | GermEval | AMI | | |
| | En | En | Es | En | De | En | It | avg. |
| LM-base | $52.82^{\pm6.36}$ | $57.31^{\pm4.98}$ | $45.60^{\pm6.78}$ | $52.70^{\pm7.18}$ | $50.89^{\pm5.33}$ | $\mathbf{57.31}^{\pm4.25}$ | $60.33^{\pm10.45}$ | 53.85 |
| MLM | $54.88^{\pm4.89}$ | $58.91^{\pm2.53}$ | $46.37^{\pm9.00}$ | $53.63^{\pm7.09}$ | $52.81^{\pm1.16}$ | $48.68^{\pm8.20}$ | $64.24^{\pm5.82}$ | 54.22 |
| MTL | $54.22^{\pm0.00}$ | $54.40^{\pm0.00}$ | $48.85^{\pm0.00}$ | $56.00^{\pm0.00}$ | $47.05^{\pm0.00}$ | $49.73^{\pm0.00}$ | $45.32^{\pm0.00}$ | 50.80 |
| Fusion | $46.79^{\pm7.50}$ | $53.91^{\pm4.41}$ | $45.94^{\pm1.57}$ | $64.67^{\pm2.70}$ | $56.95^{\pm1.95}$ | $34.95^{\pm5.30}$ | $39.11^{\pm11.17}$ | 48.91 |
| MDT | $\mathbf{60.41}^{\pm5.75}$ | $\mathbf{60.20}^{\pm5.52}$ | $\mathbf{54.47}^{\pm0.79}$ | $\mathbf{64.81}^{\pm8.55}$ | $\mathbf{65.02}^{\pm4.17}$ | $47.77^{\pm1.36}$ | $\mathbf{66.98}^{\pm4.03}$ | **59.95** |

Table 2: Macro averaged $F_1$ scores and standard deviation (%) on the fine-grained and binary target datasets of our multi-dataset approach using 4-shot training. In case there are unseen labels in a given target dataset, we highlight them together with the overall number of labels in parentheses. The best result for each target dataset is in **bold**.

Our main results with 4-shot training are presented in Table 2. On a higher level it can be seen that MDT improved over all baselines in 11 out of 13 cases (in 12 cases compared to LM-base). MDT is significantly better ($\alpha = 0.05$) than LM-base on all datasets except on `AMI En` and `HASOC De`. We used the significance test proposed in (Dror et al., 2019), which accounts for the challenges of comparing deep neural networks, including the difficulties due to the use of multiple random seeds. The method compares the score distributions generated by different runs (random seeds) of a given model type using an approach based on *Almost Stochastic Dominance* relation of the distributions. The MLM, MTL and Fusion baselines also improve over the LM-base system, however, not as consistently and to a lesser extent than our approach. MTL and Fusion even achieve lower performance than LM-base when the averaged performance over all datasets is considered. This indicates that i) relying on the labels other than only the domain adaptation effect of MLM is beneficial, ii) the two-step approach of MDT is more effective, since the very low number of samples of the target dataset are suppressed by the external data samples when they are added directly into MTL and iii) the small number of target dataset samples is not enough to properly fuse dataset specific adapters (Fusion). Additionally, training on the external datasets makes the models more consistent over different runs, as shown by the decreased standard deviation values on most of the datasets.

MDT achieves comparable average improvements on the fine-grained and the binary target datasets. Looking at the former set, not only seen but unseen labels as well were improved (even on `ToLD-Br` with more than half of its labels unseen), suggesting that the general abusive language aware $M_E$ model helps learning the fine-grained label sets of these datasets even with only a few-shots being available. We discuss the improvements on the different labels in more details below in the per label analysis section. The only exception is the `HASOC fine-grained abusive Hi` dataset, where the MLM baseline achieved the best results, although MDT also improved over LM-base. Our conjecture is that this is partly due to the high ratio of English content in the dataset caused by its code-mixed nature, and as Table 2 suggests, MLM tends to perform better on English target datasets compared to non-English sets.

Although all labels of the binary target datasets are seen, as mentioned the definitions of some labels are different. For example, the offensive label of the `OLID binary offensive En` target dataset includes profanity, while the same label in the external `HASOC fine-grained abusive En` dataset does not. However, due to the inclusion of external training samples that are directly labeled as profane, the model is trained on all the necessary information. It only has to learn to combine them in the final model of a given target dataset, such as profanity and the more restrictive offensive label of `HASOC` into the general offensive label of `OLID`.

All the used external datasets are English. Comparing the improvements of the monolingual and cross-lingual setups of MDT, i.e., English and non-

| | hate | offensive | **profane** |
|---|---|---|---|
| LM-base | 30.64 | 24.39 | 43.24 |
| MDT | 52.94 | 26.64 | 41.85 |

(a) HASOC fine-grained abusive En

| | group | individual | **other** |
|---|---|---|---|
| LM-base | 41.85 | 51.06 | 17.92 |
| MDT | 60.07 | 65.69 | 22.89 |

(b) OLID fine-grained target En

| | misogyny | racism | **insult** | **xenophobia** | **LGBTQ+phobia** | **obscene** | normal |
|---|---|---|---|---|---|---|---|
| LM-base | 2.84 | 0.36 | 9.60 | 0.79 | 2.32 | 14.17 | 27.65 |
| MDT | 1.73 | 0.61 | 12.48 | 0.31 | 3.72 | 24.98 | 45.98 |

(c) ToLD-Br fine-grained toxic Pt-Br

| | hate | normal |
|---|---|---|
| LM-base | 30.28 | 75.36 |
| MDT | 36.86 | 83.95 |

(d) Stormfront binary hate En

| | misogyny | normal |
|---|---|---|
| LM-base | 55.26 | 65.40 |
| MDT | 69.33 | 64.64 |

(e) AMI binary misogyny It

Table 3: Per label 4-shot $F_1$ scores (%). Unseen labels are **bolded**.

English target datasets, we found that the external datasets are more beneficial monolingually. The average improvements are 9.51% (ignoring `AMI misogyny En`) and 6.09% respectively. This is not surprising given that cross-lingual transfer learning is almost always less effective. Still, it shows that the combination with English external datasets is beneficial to non-English test corpora as well. This is an important use case for reducing costs by dramatically reducing the need for human annotation. Additionally, all the external datasets contain Twitter or Facebook posts, while the Stormfront target dataset contains forum posts which tend to be longer and have different language use compared to microblog posts. Even in this case, the improvements of MDT are large compared with the baseline showing the generality of our model.

**N-shot analysis** Due to the high label bias of abusive language datasets, we consider 4-shots of the target datasets for efficiency. For example, out of the 12 833 training samples in the ToLD-Br dataset only 11 (0.08%) are labeled as racism, meaning that a given annotator has to look at more than a thousand text inputs to increase the number of this minority label with just one sample. For a more complete picture of the performance of our approach however, we present experiments with different $n$ values on a few selected datasets in Figure 2, while results on all the datasets are in Table 7 in the appendix. Similarly to Zhao et al. (2021) we find that although we average our results over 5 seeds, the performance can be unstable at lower $n$ values. It even decreases with the increase of training samples in same cases. However, MDT steadily outperforms LM-base, the gap only decreases at higher $n$ values. Still, even at $n = 64$ the baseline performs worse on 2 of the 3 datasets. In contrast, MDT has

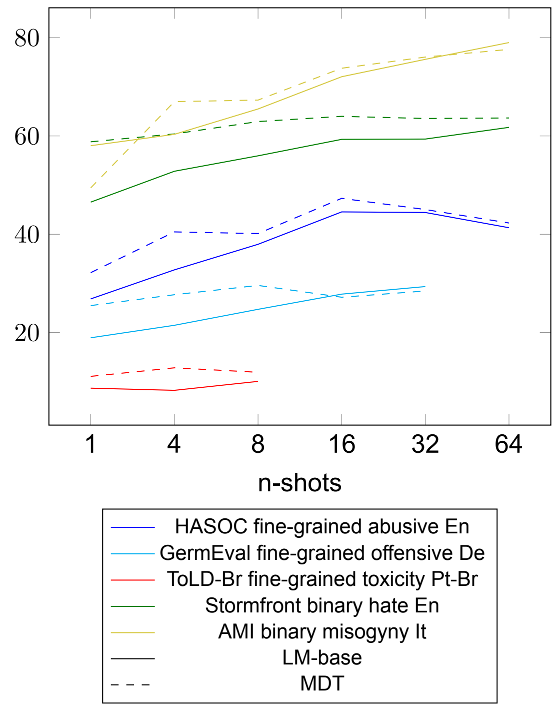


Figure 2: $F_1$ scores (%) at different number (n-shots) of target dataset samples comparing LM-base with our MDT approach. Some datasets have less than 64 samples available for the minority label.

the largest improvements compared to the baseline at lower $n$ values, which shows the strong advantage of using external datasets, especially for target datasets such as `ToLD-Br`, for which acquiring even one racism sample is expensive.

**Per label analysis** We present per label $F_1$ scores on a few selected datasets in Table 3. In case of the fine-grained `HASOC` dataset in subtable (a), the label *hate* was significantly improved, while the performance on *offensive* and *profane* were improved and decreased respectively with a similar margin. All labels were improved on the OLID dataset in subtable (b), even the unseen *other* label by almost 5 percentage points. Most interestingly, 3 out 4 unseen labels (5 out of 7 overall) were improved on the ToLD-Br dataset in subtable (c). Our conjecture is that the unseen *insult* label is related to the *fearful* label of the external MLMA dataset, since texts causing fear often involve insults as well, which leads to this improvement. Additionally, as stated by the authors of ToLD-Br (Leite et al., 2020), the unseen *insult* and *obscene* labels were often confused by the annotators, indicating their similarity, thus the latter could have also benefited from the *fearful* instances of MLMA. Similarly, the *LMBTQ+phobia* label is to some extent related to *sexism* external instances, thus MDT can leverage their similarity automatically without the need for manual label modifications. In contrast, *xenophobia*, which is somewhat related to *racism*, was not improved. However, since both models achieve less than $1$ percent $F1$, we believe that *xenophobia* is simply too hard to classify, which is the reason for no improvements in MDT. On the binary target datasets in subtables (d) and (e) the results are similar as in case of the fine-grained datasets, however all labels were improved in Stormfront, while only *misogyny* improved in AMI.

**Ablation study** We were interested in the additional value of labels which are not directly used in the target datasets. Thus, we removed the external-only labels from the external datasets in step 1, i.e., all labels which are not part of a given target dataset, and performed step 2 as normal ($\text{MDT}_{\text{abl}}$). We present experiments on datasets where MDT outperformed LM-base in Table 4. In the majority of the cases, especially for the binary datasets, we found that removing labels from the external datasets deteriorates the model's performance supporting our claim that by training first on a large set of diverse datasets the model can learn a general knowledge of abusive language which is then beneficial for the target task. Even though some labels in the external datasets aren't directly used for a given target task they are still beneficial.

**HateCheck** is a test suite containing 29 functional tests grouped into 11 classes (Röttger et al., 2021). It focuses on testing various aspects of hate speech detection models. It defines the following test classes: derogation (**F1-4**), threatening language (**F5-6**), slur usage (**F7-9**), profanity usage (**F10-11**), profane reference (**F12-13**), negation (**F14-15**), paraphrasing (**F16-17**), non-hate when mentioning protected groups (**F18-19**), counter speech (**F20-21**), abuse against non-protected targets (**F22-24**) and spelling variations (**F25-29**). We evaluated MDT using HatEval binary hate En as the target dataset compared to LM-base in Table 5. We found that MDT improves on cases testing a higher level of hate speech understanding, such as derogation and threatening, or slur and profanity usage, while the performance decreases on cases testing linguistics phenomena, such as negation or paraphrasing. These results further support our claim that MDT results in models with better abusive language understanding. On the other hand, they also highlight sensitivity to linguistic phenomena which should be improved in future work.

| | | MDT | $\text{MDT}_{\text{abl}}$ |
|---|---|---|---|
| fine-grained | HASOC abusive En | **40.48** | 36.96 |
| | HASOC abusive Hi | 34.36 | **38.38** |
| | HASOC abusive De | **33.96** | 32.55 |
| | GermEval offensive De | 27.70 | **28.04** |
| | ToLD-Br toxicity Pt-Br | **12.83** | 8.20 |
| | OLID target EN | **49.55** | 44.73 |
| binary | Stormfront hate En | **60.41** | 53.43 |
| | HatEval hate En | **60.20** | 56.91 |
| | HatEval hate Es | **54.47** | 52.69 |
| | OLID offensive En | **64.81** | 45.09 |
| | GermEval offensive De | **65.02** | 56.95 |
| | AMI misogyny It | **66.98** | 65.45 |
| | avg. | **47.56** | 43.28 |

Table 4: Macro averaged 4-shot $F_1$ scores of our ablation study which have labels removed from the external datasets that are not needed in the target datasets ($\text{MDT}_{\text{abl}}$).

| | LM-base | MDT |
|---|---|---|
| F1-4 | 45.23 | **61.95** |
| F5-6 | 48.59 | **49.72** |
| F7-9 | 28.88 | **30.36** |
| F10-11 | 50.00 | **52.38** |
| F12-13 | **74.91** | 74.73 |
| F14-15 | **29.15** | 24.89 |
| F16-17 | **73.18** | 49.53 |
| F18-19 | **27.14** | 9.57 |
| F20-21 | **8.30** | 4.44 |
| F22-24 | 14.51 | **21.82** |
| F25-29 | 41.46 | **59.44** |

Table 5: Macro averaged 4-shot $F_1$ scores on the HateCheck functional test classes using HatEval binary hate En as the target dataset. The results are averaged over test cases in the classes. The naming F$X-Y$ refers to the test class containing test cases between $X$ and $Y$.

## 6. Conclusions

Due to the large variety of the abusive content to be filtered, lack of resources is a major problem, larger than for many other NLP tasks. In order to eliminate the need for expensive dataset annotation for novel application scenarios, and thus reduce costs, we proposed a two-step multi-dataset approach (MDT) which exploits datasets we already have to learn general abusive language understanding and requires only a few annotated samples for the target task. Our experiments on various datasets showed that external datasets improve few-shot classification across tasks, text genres and languages. Additionally, our analysis reveals that external-only labels also contain useful information and consequently, that unseen labels can be improved as well, arguing for the utility of our approach.

## Limitations

The aim of our approach is that given the requirements (labels and their definitions) of a target task, train an abusive language classifier using external datasets. Although our approach supports selecting more closely related datasets, e.g., only hate speech or only misogyny and sexism datasets, we used a broad range of abusive language related external datasets in our experiments for simplicity, and to save computational resources. On the one hand, this setup shows the general applicability of MDT. However, our results do not show the potential of more specialized (but still easy to set up) configurations.

Additionally, we only used English corpora as external datasets. Although English has the most abusive language resources, datasets of other languages could also be used for this purpose. Our results show that MDT is more beneficial monolingually than cross-lingually, thus using resources from the same language as a non-English target dataset could be beneficial, which we did not test in this work.

Finally, our analysis on HateCheck reveals some weaknesses of our approach which indicate future directions in combining abusive language datasets.

## Acknowledgements

We thank the anonymous reviewers for their helpful feedback. The work was funded by the European Research Council (ERC; grant agreement No. 640550) and by the German Research Foundation (DFG; grant FR 2829/4-1 and FR 2829/7-1).

## 7. Bibliographical References

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.

Rich Caruana. 1997. Multitask learning. *Machine learning*, 28(1):41–75.

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.

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

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.

Ning Ding, Shengding Hu, Weilin Zhao, Yulin Chen, Zhiyuan Liu, Haitao Zheng, and Maosong Sun. 2022. OpenPrompt: An open-source framework for prompt-learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 105–113, Dublin, Ireland. Association for Computational Linguistics.

Rotem Dror, Segev Shlomov, and Roi Reichart. 2019. Deep dominance - how to properly compare deep neural models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2773–2785, Florence, Italy. Association for Computational Linguistics.

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2018. Overview of the Evalita 2018 Task on Automatic Misogyny Identification (AMI). *Proceedings of Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2018)*, 12:59–66.

Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontia dis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Mic hael Sirivianos, and Nicolas Kourtellis. 2018. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media*, pages 491–500.

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.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Timothy Hospedales, Antreas Antoniou, Paul Micaelli, and Amos Storkey. 2021. Meta-learning in neural networks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 44(9):5149–5169.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-Efficient Transfer Learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2790–2799.

Prashant Kapil and Asif Ekbal. 2020. A deep neural network based multi-task learning approach to hate speech detection. *Knowledge-Based Systems*, 210(October):106458.

João Augusto Leite, Diego Silva, Kalina Bontcheva, and Carolina Scarton. 2020. Toxic language detection in social media for Brazilian Portuguese: New dataset and multilingual analysis. 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 914–924, Suzhou, China. Association for Computational Linguistics.

Angel Felipe Magnossão de Paula, Paolo Rosso, and Damiano Spina. 2023. Mitigating negative transfer with task awareness for sexism, hate speech, and toxic language detection. In *2023 International Joint Conference on Neural Networks (IJCNN)*.

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, M ohana 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 *11th meeting of the Forum for Information Retrieval Evaluation*, pages 14–17.

Tahir Mehmood, Alfonso E. Gerevini, Alberto Lavelli, and Ivan Serina. 2020. Combining multi-task learning with transfer learning for biomedical named entity recognition. *Procedia Computer Science*, 176:848–857.

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.

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.

Jonas Pfeiffer, Aishwarya Kamath, Andreas Rücklé, Kyunghyun Cho, and Iryna Gurevych. 2021. AdapterFusion: Non-destructive task composition for transfer learning. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 487–503, Online. Association for Computational Linguistics.

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*, 55:477–523.

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 Proce ssing*, pages 5838–5844.

Paul Röttger, Debora Nozza, Federico Bianchi, and Dirk Hovy. 2022. 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, 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.

Timo Schick and Hinrich Schütze. 2021. Exploiting cloze-questions for few-shot text classification and natural language inference. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 255–269, Online. Association for Computational Linguistics.

Asa Cooper Stickland and Iain Murray. 2019. BERT and PALs: Projected Attention Layers for Efficient Adaptation in Multi-Task Lea rning. In *Proceedings of the 36th International Conference on Machine Learning*, pages 5986–5995.

Julia Maria Struß, Melanie Siegel, Josef Ruppenhofer, Michael Wiegand, and Manfred Klenner. 2019. Overview of Germeval Task 2, 2019 Shared Task on the Identification of Offensive Language. In *Proceedings of the 15th Conference on Natural Language Processing*, pages 354–365.

Shardul Suryawanshi, Bharathi Raja Chakravarthi, Mihael Arcan, and Paul Buitelaar. 2020. Multimodal meme dataset (MultiOFF) for identifying offensive content in image and text. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 32–41, Marseille, France. European Language Resources Association (ELRA).

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garba ge out. *PLoS ONE*, 15(12):1–32.

Haoxiang Wang, Han Zhao, and Bo Li. 2021. Bridging multi-task learning and meta-learning: Towards efficient training and effective adaptation. In *International Conference on Machine Learning*, pages 10991–11002.

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.

Taiki Watanabe, Tomoya Ichikawa, Akihiro Tamura, Tomoya Iwakura, Chunpeng Ma, and Tsuneo Kato. 2022. Auxiliary learning for named entity recognition with multiple auxiliary biomedical training data. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 130–139, Dublin, Ireland. Association for Computational Linguistics.

Michael Wiegand, Anastasija Amann, Tatiana Anikina, Aikaterini Azoidou, Anastasia Borisenkov, Kirstin Kolmorgen, Insa Kröger, and Chris tine Schäfer. 2018. Saarland University's Participation in the GermEval Task 2018 (UdSW) – Examining Different Types of Classifiers and Features. In *Proceedings of GermEval 2018, 14th Conference on Natural Language Processing (KONVENS 2018)*, pages 21–26.

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.

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

Mengjie Zhao and Hinrich Schütze. 2021. Discrete and soft prompting for multilingual models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8547–8555, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Mengjie Zhao, Yi Zhu, Ehsan Shareghi, Ivan Vulić, Roi Reichart, Anna Korhonen, and Hinrich Schütze. 2021. A closer look at few-shot crosslingual transfer: The choice of shots matters. 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 5751–5767, Online. Association for Computational Linguistics.

Yanan Zheng, Jing Zhou, Yujie Qian, Ming Ding, Chonghua Liao, Li Jian, Ruslan Salakhutdinov, Jie Tang, Sebastian Ruder, and Zhilin Yang. 2022. FewNLU: Benchmarking state-of-the-art methods for few-shot natural language understanding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 501–516, Dublin, Ireland. Association for Computational Linguistics.

# A. Additional Details

We present details of the used datasets in Table 6, such as source platform, number of samples (we only used 4 samples per label for training in the main experiments and an overall 16 samples following the original label distribution for validation in case of the target datasets) and used PVPs. We kept PVPs simple and uniform across datasets using English PVPs even for non-English datasets as well. We note however that in our initial experiments we tested machine translated PVPs which did not lead to significantly different results (Zhao and Schütze, 2021). Additionally, if a given token related to a label is split by the tokenizer, e.g., *dominance* $\rightarrow$ [*domina*, *#nce*], we take the averaged probabilities of the subwords at the `[MASK]` position as the probability of the related label. Finally, we show complete results of all setups in Table 7 (containing each target dataset in a separate subtable stretching over multiple pages).

| | source | #train | #valid | #test | verbalizer | Pattern |
|---|---|---|---|---|---|---|
| AMI fine-grained misogyny En | Twitter | 1,428 | 357 | 460 | stereotypical → stereotype<br>dominance → dominance<br>derailing → derailing<br>harassment → sexual_harassment<br>discrediting → discredit | X → X It was [MASK] |
| AMI binary misogyny En | Twitter | 3,200 | 800 | 1,000 | sexist → misogyny<br>neutral → normal | X → X It was [MASK] |
| AMI binary misogyny It | Twitter | 3,200 | 800 | 1,000 | sexist → misogyny<br>neutral → normal | X → X It was [MASK] |
| GermEval fine-grained offensive De | Twitter | 4,007 | 1,002 | 3,532 | profane → profanity<br>insulting → insult<br>abusive → abusive<br>neutral → normal | X → X It was [MASK] |
| GermEval binary offensive De | Twitter | 4,007 | 1,002 | 3,532 | offensive → offensive<br>neutral → normal | X → X It was [MASK] |
| HASOC fine-grained abusive En | Twitter, Facebook | 1,808 | 453 | 288 | hate → hate<br>offensive → offensive<br>profane → profanity | X → X It was [MASK] |
| HASOC fine-grained abusive De | Twitter, Facebook | 325 | 82 | 136 | hate → hate<br>offensive → offensive<br>profane → profanity | X → X It was [MASK] |
| HASOC fine-grained abusive Hi | Twitter, Facebook | 1,975 | 494 | 605 | hate → hate<br>offensive → offensive<br>profane → profanity | X → X It was [MASK] |
| HatEval binary hate En | Twitter | 3,055 | 764 | 850 | hate → hateful<br>neutral → normal | X → X It was [MASK] |
| HatEval binary hate Es | Twitter | 3,560 | 890 | 500 | hate → hateful<br>neutral → normal | X → X It was [MASK] |
| LSA fine-grained abusive En | Twitter | 29,728 | 7,433 | 9,291 | abusive → abusive<br>hate → hateful<br>spam → spam<br>neutral → normal | X → X It was [MASK] |
| MLMA fine-grained hostility En | Twitter | 5,549 | 1,388 | 1,735 | abusive → abusive<br>hate → hateful<br>offensive → offensive<br>disrespectful → disrespectful<br>fearful → fearful<br>neutral → normal | X → X It was [MASK] |
| OLID binary offensive En | Twitter | 10,592 | 2,648 | 860 | offensive → offensive<br>neutral → normal | X → X It was [MASK] |
| SRW fine-grained abusive En | Twitter | 6,504 | 1,626 | 2,033 | sexist → sexism<br>racist → racism<br>neutral → normal | X → X It was [MASK] |
| Stormfront binary hate En | Stormfront forum | 6,849 | 1,713 | 2,141 | hate → hate<br>neutral → normal | X → X It was [MASK] |
| ToLD-Br fine-grained toxicity Pt-Br | Twitter | 12,833 | 3,209 | 4,011 | homophobic → LGBTQ+phobia<br>obscene → obscene<br>insulting → insult<br>racist → racism<br>sexist → misogyny<br>xenophobic → xenophobia<br>neutral → normal | X → X It was [MASK] |
| HASOC binary target En | Twitter, Facebook | 4,681 | 1,171 | 1,153 | targeted → targeted<br>general → untargeted | X → X It was [MASK] |
| AMI binary target En | Twitter | 1,428 | 357 | 460 | individual → active<br>group → passive | X → X It was targeted at [MASK] |
| HatEval binary target En | Twitter | 3,732 | 933 | 1,318 | individual → individual<br>group → group | X → X It was targeted at [MASK] |
| OLID fine-grained target En | Twitter | 3,100 | 776 | 213 | individual → individual<br>group → group<br>other → other | X → X It was targeted at [MASK] |

Table 6: Dataset statistics for each (dataset, label configuration, language) triple. From left to right we indicate the source platform of the dataset, the number of total train, validation and test samples, used verbalizers (<predicted word> → <label>) which also indicates the labels of a given dataset, and patterns (where X is the input sentence). We kept our PVPs simple, i.e., most labels are mapped 1-to-1 to the same word, and we defined only two patterns. Note that we also used English PVPs for non-English datasets, since it was shown to perform well (Zhao and Schütze, 2021). Since different datasets often name the negative abuse class differently (e.g. no-hate, not-offensive, normal, etc.), we unified them by using the frequent *normal* label name. Additionally, similarly defined but differently named labels, such as hate and hateful or sexism and misogyny, are united by using the same verbalizers for them.

| | hate | offen. | profa. | avg. |
|---|---|---|---|---|
| **1-shot** | | | | |
| LM-base | $20.37^{\pm 16.46}$ | $34.26^{\pm 2.39}$ | $25.90^{\pm 12.28}$ | $26.85^{\pm 3.80}$ |
| MDT | $31.80^{\pm 21.12}$ | $30.37^{\pm 10.10}$ | $34.41^{\pm 20.77}$ | $32.19^{\pm 2.87}$ |
| **4-shot** | | | | |
| LM-base | $30.64^{\pm 17.57}$ | $24.39^{\pm 8.99}$ | $43.24^{\pm 5.11}$ | $32.76^{\pm 5.94}$ |
| MLM | $38.35^{\pm 21.08}$ | $31.91^{\pm 5.89}$ | $37.67^{\pm 11.63}$ | $35.98^{\pm 7.70}$ |
| MTL | $0.00^{\pm 0.00}$ | $39.66^{\pm 0.00}$ | $0.00^{\pm 0.00}$ | $13.22^{\pm 0.00}$ |
| Fusion | $50.72^{\pm 13.25}$ | $18.14^{\pm 11.70}$ | $43.20^{\pm 16.42}$ | $37.35^{\pm 8.06}$ |
| MDT | $52.94^{\pm 6.68}$ | $26.64^{\pm 13.34}$ | $41.85^{\pm 17.39}$ | $40.48^{\pm 5.37}$ |
| MDT-abl | $49.94^{\pm 10.26}$ | $19.01^{\pm 5.60}$ | $41.93^{\pm 13.72}$ | $36.96^{\pm 2.77}$ |
| **8-shot** | | | | |
| LM-base | $36.09^{\pm 18.50}$ | $28.12^{\pm 8.97}$ | $49.68^{\pm 5.98}$ | $37.96^{\pm 5.32}$ |
| MDT | $54.44^{\pm 8.98}$ | $23.55^{\pm 13.52}$ | $42.44^{\pm 8.61}$ | $40.14^{\pm 3.48}$ |
| **16-shot** | | | | |
| LM-base | $52.37^{\pm 12.34}$ | $26.66^{\pm 5.91}$ | $54.62^{\pm 7.43}$ | $44.55^{\pm 3.80}$ |
| MDT | $62.35^{\pm 5.45}$ | $23.21^{\pm 5.76}$ | $56.37^{\pm 8.98}$ | $47.31^{\pm 2.23}$ |
| **32-shot** | | | | |
| LM-base | $60.08^{\pm 11.25}$ | $18.43^{\pm 12.46}$ | $54.77^{\pm 9.13}$ | $44.43^{\pm 3.22}$ |
| MDT | $61.53^{\pm 8.39}$ | $18.26^{\pm 14.98}$ | $55.28^{\pm 6.11}$ | $45.03^{\pm 2.20}$ |
| **64-shot** | | | | |
| LM-base | $39.86^{\pm 23.03}$ | $30.31^{\pm 5.62}$ | $53.82^{\pm 9.36}$ | $41.33^{\pm 10.86}$ |
| MDT | $60.07^{\pm 4.17}$ | $25.59^{\pm 7.61}$ | $41.23^{\pm 23.92}$ | $42.29^{\pm 10.14}$ |

(a) HASOC fine-grained abusive En

| | hate | offen. | profa. | avg. |
|---|---|---|---|---|
| **1-shot** | | | | |
| LM-base | $42.46^{\pm 22.52}$ | $46.03^{\pm 20.92}$ | $10.56^{\pm 6.93}$ | $33.02^{\pm 12.44}$ |
| MDT | $24.35^{\pm 18.43}$ | $65.74^{\pm 10.60}$ | $18.16^{\pm 5.98}$ | $36.08^{\pm 7.51}$ |
| **4-shot** | | | | |
| LM-base | $28.81^{\pm 22.59}$ | $47.96^{\pm 24.69}$ | $20.88^{\pm 4.24}$ | $32.55^{\pm 11.71}$ |
| MLM | $29.85^{\pm 18.60}$ | $36.99^{\pm 21.05}$ | $19.26^{\pm 5.83}$ | $28.70^{\pm 9.94}$ |
| MTL | $0.00^{\pm 0.00}$ | $72.30^{\pm 0.00}$ | $0.00^{\pm 0.00}$ | $24.10^{\pm 0.00}$ |
| Fusion | $35.10^{\pm 13.09}$ | $35.15^{\pm 21.92}$ | $17.29^{\pm 6.50}$ | $29.18^{\pm 8.09}$ |
| MDT | $33.82^{\pm 19.38}$ | $49.80^{\pm 11.04}$ | $18.25^{\pm 4.09}$ | $33.96^{\pm 7.83}$ |
| MDT-abl | $28.59^{\pm 16.25}$ | $47.73^{\pm 27.24}$ | $21.33^{\pm 13.17}$ | $32.55^{\pm 11.16}$ |
| **8-shot** | | | | |
| LM-base | $16.45^{\pm 14.06}$ | $48.32^{\pm 18.31}$ | $18.15^{\pm 5.24}$ | $27.64^{\pm 4.79}$ |
| MDT | $36.94^{\pm 14.65}$ | $42.17^{\pm 18.52}$ | $20.55^{\pm 5.22}$ | $33.22^{\pm 6.65}$ |
| **16-shot** | | | | |
| LM-base | $26.37^{\pm 22.50}$ | $62.63^{\pm 10.75}$ | $16.98^{\pm 5.52}$ | $35.33^{\pm 8.64}$ |
| MDT | $43.96^{\pm 12.16}$ | $60.61^{\pm 9.98}$ | $23.63^{\pm 6.81}$ | $42.73^{\pm 7.51}$ |
| **32-shot** | | | | |
| LM-base | $5.95^{\pm 7.35}$ | $54.02^{\pm 8.79}$ | $20.67^{\pm 9.13}$ | $26.88^{\pm 2.44}$ |
| MDT | $34.71^{\pm 16.89}$ | $49.94^{\pm 17.30}$ | $15.99^{\pm 10.30}$ | $33.55^{\pm 7.46}$ |
| **64-shot** | | | | |
| LM-base | $25.98^{\pm 16.26}$ | $60.51^{\pm 9.13}$ | $19.28^{\pm 7.88}$ | $35.26^{\pm 5.96}$ |
| MDT | $27.59^{\pm 12.55}$ | $54.48^{\pm 12.94}$ | $23.66^{\pm 2.62}$ | $35.24^{\pm 5.26}$ |

(b) HASOC fine-grained abusive De

| | hate | offen. | profa. | avg. |
|---|---|---|---|---|
| **1-shot** | | | | |
| LM-base | $29.00^{\pm 15.58}$ | $24.21^{\pm 16.02}$ | $27.77^{\pm 13.25}$ | $26.99^{\pm 2.91}$ |
| MDT | $33.87^{\pm 13.10}$ | $29.49^{\pm 16.19}$ | $27.90^{\pm 22.33}$ | $30.42^{\pm 5.26}$ |
| **4-shot** | | | | |
| LM-base | $25.57^{\pm 18.46}$ | $33.84^{\pm 11.71}$ | $41.37^{\pm 12.38}$ | $33.59^{\pm 5.68}$ |
| MLM | $23.09^{\pm 16.30}$ | $34.63^{\pm 14.80}$ | $49.71^{\pm 5.74}$ | $35.81^{\pm 3.79}$ |
| MTL | $0.00^{\pm 0.00}$ | $49.13^{\pm 0.00}$ | $0.00^{\pm 0.00}$ | $16.38^{\pm 0.00}$ |
| Fusion | $45.92^{\pm 5.71}$ | $19.45^{\pm 11.77}$ | $37.51^{\pm 6.16}$ | $34.29^{\pm 4.14}$ |
| MDT | $40.01^{\pm 7.27}$ | $16.16^{\pm 17.45}$ | $46.89^{\pm 6.58}$ | $34.36^{\pm 2.59}$ |
| MDT-abl | $44.52^{\pm 6.07}$ | $18.81^{\pm 15.40}$ | $51.82^{\pm 8.43}$ | $38.38^{\pm 5.45}$ |
| **8-shot** | | | | |
| LM-base | $40.28^{\pm 14.55}$ | $31.14^{\pm 12.85}$ | $42.40^{\pm 14.03}$ | $37.94^{\pm 2.53}$ |
| MDT | $47.25^{\pm 3.31}$ | $19.72^{\pm 8.11}$ | $44.61^{\pm 8.93}$ | $37.19^{\pm 1.72}$ |

(c) HASOC fine-grained abusive Hi

| | abus. | insult | profa. | normal | avg. |
|---|---|---|---|---|---|
| **1-shot** | | | | | |
| LM-base | $15.82^{\pm 9.22}$ | $15.48^{\pm 6.61}$ | $3.87^{\pm 1.31}$ | $40.61^{\pm 16.10}$ | $18.94^{\pm 4.00}$ |
| MDT | $13.95^{\pm 5.56}$ | $5.82^{\pm 7.13}$ | $7.37^{\pm 2.28}$ | $74.81^{\pm 7.15}$ | $25.49^{\pm 2.90}$ |
| **4-shot** | | | | | |
| LM-base | $15.94^{\pm 9.67}$ | $18.61^{\pm 2.32}$ | $2.90^{\pm 0.66}$ | $48.46^{\pm 5.99}$ | $21.48^{\pm 2.22}$ |
| MLM | $19.41^{\pm 6.14}$ | $17.43^{\pm 3.81}$ | $3.47^{\pm 1.20}$ | $47.57^{\pm 11.07}$ | $21.97^{\pm 3.89}$ |
| MTL | $0.00^{\pm 0.00}$ | $0.00^{\pm 0.00}$ | $0.00^{\pm 0.00}$ | $79.38^{\pm 0.00}$ | $19.84^{\pm 0.00}$ |
| Fusion | $12.24^{\pm 4.51}$ | $13.35^{\pm 5.24}$ | $4.10^{\pm 0.79}$ | $55.63^{\pm 17.20}$ | $21.33^{\pm 3.88}$ |
| MDT | $23.30^{\pm 9.48}$ | $14.53^{\pm 4.17}$ | $8.60^{\pm 2.00}$ | $64.35^{\pm 7.62}$ | $27.70^{\pm 3.78}$ |
| MDT-abl | $30.52^{\pm 2.34}$ | $14.54^{\pm 3.43}$ | $0.00^{\pm 0.00}$ | $67.11^{\pm 9.65}$ | $28.04^{\pm 2.20}$ |
| **8-shot** | | | | | |
| LM-base | $23.02^{\pm 14.20}$ | $22.06^{\pm 2.11}$ | $4.01^{\pm 1.15}$ | $49.82^{\pm 12.13}$ | $24.73^{\pm 4.54}$ |
| MDT | $25.53^{\pm 8.68}$ | $13.73^{\pm 7.96}$ | $8.39^{\pm 2.13}$ | $70.72^{\pm 6.50}$ | $29.59^{\pm 3.17}$ |
| **16-shot** | | | | | |
| LM-base | $32.19^{\pm 8.24}$ | $15.98^{\pm 5.00}$ | $2.08^{\pm 1.92}$ | $61.01^{\pm 7.12}$ | $27.82^{\pm 2.01}$ |
| MDT | $20.99^{\pm 11.91}$ | $9.06^{\pm 7.93}$ | $9.50^{\pm 2.11}$ | $69.22^{\pm 15.18}$ | $27.19^{\pm 3.34}$ |
| **32-shot** | | | | | |
| LM-base | $18.74^{\pm 14.08}$ | $22.86^{\pm 5.56}$ | $6.95^{\pm 2.38}$ | $68.90^{\pm 11.86}$ | $29.36^{\pm 4.68}$ |
| MDT | $23.02^{\pm 15.61}$ | $9.21^{\pm 8.95}$ | $9.36^{\pm 2.63}$ | $72.27^{\pm 10.43}$ | $28.46^{\pm 6.16}$ |

(d) GermEval fine-grained offensive De

| | misogyny | racism | insult | xenophobia | LGBTQ+hobia | obscene | normal | avg. |
|---|---|---|---|---|---|---|---|---|
| **1-shot** | | | | | | | | |
| LM-base | $0.90^{\pm 0.95}$ | $0.58^{\pm 0.29}$ | $9.34^{\pm 2.72}$ | $0.09^{\pm 0.18}$ | $1.60^{\pm 1.51}$ | $9.86^{\pm 5.41}$ | $38.46^{\pm 11.40}$ | $8.69^{\pm 2.58}$ |
| MDT | $1.40^{\pm 1.04}$ | $0.11^{\pm 0.22}$ | $7.98^{\pm 7.22}$ | $0.39^{\pm 0.47}$ | $1.83^{\pm 1.77}$ | $22.57^{\pm 4.42}$ | $43.35^{\pm 11.39}$ | $11.09^{\pm 1.42}$ |
| **4-shot** | | | | | | | | |
| LM-base | $2.84^{\pm 4.14}$ | $0.36^{\pm 0.20}$ | $9.60^{\pm 2.68}$ | $0.79^{\pm 1.07}$ | $2.32^{\pm 1.25}$ | $14.17^{\pm 2.23}$ | $27.65^{\pm 17.87}$ | $8.25^{\pm 3.54}$ |
| MLM | $1.31^{\pm 0.96}$ | $0.88^{\pm 0.42}$ | $8.76^{\pm 5.02}$ | $0.35^{\pm 0.48}$ | $2.76^{\pm 2.04}$ | $15.01^{\pm 2.74}$ | $33.83^{\pm 16.75}$ | $8.98^{\pm 2.76}$ |
| MTL | $0.00^{\pm 0.00}$ | $0.00^{\pm 0.00}$ | $0.00^{\pm 0.00}$ | $0.00^{\pm 0.00}$ | $2.25^{\pm 0.00}$ | $15.77^{\pm 0.00}$ | $56.80^{\pm 0.00}$ | $10.69^{\pm 0.00}$ |
| Fusion | $0.75^{\pm 0.41}$ | $0.31^{\pm 0.10}$ | $8.71^{\pm 2.59}$ | $0.07^{\pm 0.13}$ | $2.73^{\pm 0.82}$ | $7.74^{\pm 3.00}$ | $37.48^{\pm 2.10}$ | $8.26^{\pm 1.20}$ |
| MDT | $1.73^{\pm 0.82}$ | $0.61^{\pm 0.89}$ | $12.48^{\pm 6.33}$ | $0.31^{\pm 0.38}$ | $3.72^{\pm 3.62}$ | $24.98^{\pm 4.50}$ | $45.98^{\pm 6.12}$ | $12.83^{\pm 1.78}$ |
| MDT-abl | $1.41^{\pm 0.50}$ | $1.11^{\pm 1.58}$ | $8.69^{\pm 1.84}$ | $2.78^{\pm 2.01}$ | $4.62^{\pm 2.60}$ | $13.82^{\pm 4.49}$ | $24.95^{\pm 11.60}$ | $8.20^{\pm 2.23}$ |
| **8-shot** | | | | | | | | |
| LM-base | $1.62^{\pm 0.99}$ | $0.69^{\pm 0.85}$ | $13.80^{\pm 3.92}$ | $0.46^{\pm 0.39}$ | $2.65^{\pm 0.47}$ | $16.39^{\pm 1.98}$ | $34.89^{\pm 12.98}$ | $10.07^{\pm 1.95}$ |
| MDT | $2.35^{\pm 0.83}$ | $0.67^{\pm 0.82}$ | $9.20^{\pm 7.54}$ | $0.94^{\pm 1.10}$ | $2.97^{\pm 1.95}$ | $23.45^{\pm 4.82}$ | $43.88^{\pm 8.39}$ | $11.92^{\pm 2.51}$ |

(e) ToLD-Br fine-grained toxic Pt-Br

| | group | indivi. | other | avg. |
|---|---|---|---|---|
| | | 1-shot | | |
| LM-base | $46.96^{\pm 9.06}$ | $43.39^{\pm 8.12}$ | $7.90^{\pm 9.76}$ | $32.75^{\pm 3.48}$ |
| MDT | $62.58^{\pm 3.50}$ | $57.42^{\pm 13.12}$ | $20.02^{\pm 7.92}$ | $46.67^{\pm 4.45}$ |
| | | 4-shot | | |
| LM-base | $41.85^{\pm 9.31}$ | $51.06^{\pm 8.90}$ | $17.92^{\pm 7.84}$ | $36.94^{\pm 3.80}$ |
| MLM | $47.17^{\pm 6.13}$ | $57.48^{\pm 6.38}$ | $17.99^{\pm 7.07}$ | $40.88^{\pm 5.38}$ |
| MTL | $14.12^{\pm 0.00}$ | $0.00^{\pm 0.00}$ | $28.22^{\pm 0.00}$ | $14.11^{\pm 0.00}$ |
| Fusion | $58.21^{\pm 3.19}$ | $51.48^{\pm 10.50}$ | $17.66^{\pm 8.32}$ | $42.45^{\pm 4.03}$ |
| MDT | $60.07^{\pm 11.50}$ | $65.69^{\pm 8.59}$ | $22.89^{\pm 6.60}$ | $49.55^{\pm 2.92}$ |
| MDT-abl | $60.80^{\pm 10.54}$ | $57.07^{\pm 14.66}$ | $16.33^{\pm 6.93}$ | $44.73^{\pm 4.15}$ |
| | | 8-shot | | |
| LM-base | $49.17^{\pm 5.58}$ | $49.02^{\pm 15.79}$ | $21.61^{\pm 6.07}$ | $39.93^{\pm 4.80}$ |
| MDT | $67.00^{\pm 3.35}$ | $65.98^{\pm 8.36}$ | $21.50^{\pm 7.27}$ | $51.49^{\pm 3.66}$ |

(f) OLID fine-grained target En

| | hate | normal | avg. |
|---|---|---|---|
| | | 1-shot | |
| LM-base | $26.31^{\pm 0.62}$ | $66.77^{\pm 6.47}$ | $46.54^{\pm 3.51}$ |
| MDT | $35.03^{\pm 5.29}$ | $82.58^{\pm 12.35}$ | $58.81^{\pm 7.55}$ |
| | | 4-shot | |
| LM-base | $30.28^{\pm 4.58}$ | $75.36^{\pm 9.38}$ | $52.82^{\pm 6.36}$ |
| MLM | $25.45^{\pm 12.15}$ | $84.31^{\pm 9.19}$ | $54.88^{\pm 4.89}$ |
| MTL | $16.30^{\pm 0.00}$ | $92.13^{\pm 0.00}$ | $54.22^{\pm 0.00}$ |
| Fusion | $30.18^{\pm 3.51}$ | $63.40^{\pm 11.55}$ | $46.79^{\pm 7.50}$ |
| MDT | $36.86^{\pm 2.87}$ | $83.95^{\pm 9.03}$ | $60.41^{\pm 5.75}$ |
| MDT-abl | $23.61^{\pm 8.76}$ | $83.26^{\pm 3.93}$ | $53.43^{\pm 4.44}$ |
| | | 8-shot | |
| LM-base | $31.14^{\pm 3.76}$ | $80.78^{\pm 4.96}$ | $55.96^{\pm 3.10}$ |
| MDT | $39.34^{\pm 1.68}$ | $86.52^{\pm 4.13}$ | $62.93^{\pm 2.70}$ |
| | | 16-shot | |
| LM-base | $34.80^{\pm 4.63}$ | $83.82^{\pm 7.12}$ | $59.31^{\pm 4.60}$ |
| MDT | $40.23^{\pm 2.50}$ | $87.74^{\pm 2.17}$ | $63.99^{\pm 1.31}$ |
| | | 32-shot | |
| LM-base | $36.45^{\pm 4.96}$ | $82.29^{\pm 5.44}$ | $59.37^{\pm 4.99}$ |
| MDT | $40.79^{\pm 1.42}$ | $86.31^{\pm 3.15}$ | $63.55^{\pm 1.29}$ |
| | | 64-shot | |
| LM-base | $38.79^{\pm 7.06}$ | $84.71^{\pm 4.90}$ | $61.75^{\pm 5.54}$ |
| MDT | $40.38^{\pm 1.97}$ | $86.93^{\pm 3.58}$ | $63.66^{\pm 2.05}$ |

(g) Stormfront binary hate En

| | hateful | normal | avg. |
|---|---|---|---|
| | | 1-shot | |
| LM-base | $46.18^{\pm 13.74}$ | $62.61^{\pm 8.12}$ | $54.40^{\pm 5.75}$ |
| MDT | $46.49^{\pm 11.06}$ | $54.76^{\pm 7.88}$ | $50.62^{\pm 3.70}$ |
| | | 4-shot | |
| LM-base | $55.20^{\pm 3.04}$ | $59.42^{\pm 10.52}$ | $57.31^{\pm 4.98}$ |
| MLM | $55.08^{\pm 5.70}$ | $62.73^{\pm 8.40}$ | $58.91^{\pm 2.53}$ |
| MTL | $38.72^{\pm 0.00}$ | $70.09^{\pm 0.00}$ | $54.40^{\pm 0.00}$ |
| Fusion | $57.43^{\pm 5.20}$ | $50.40^{\pm 8.22}$ | $53.91^{\pm 4.41}$ |
| MDT | $61.04^{\pm 4.19}$ | $59.35^{\pm 12.05}$ | $60.20^{\pm 5.52}$ |
| MDT-abl | $56.18^{\pm 4.34}$ | $57.63^{\pm 8.86}$ | $56.91^{\pm 2.82}$ |
| | | 8-shot | |
| LM-base | $48.29^{\pm 12.01}$ | $63.44^{\pm 11.82}$ | $55.86^{\pm 5.42}$ |
| MDT | $61.87^{\pm 1.41}$ | $65.62^{\pm 7.88}$ | $63.75^{\pm 3.83}$ |
| | | 16-shot | |
| LM-base | $53.38^{\pm 10.13}$ | $64.45^{\pm 3.70}$ | $58.91^{\pm 4.10}$ |
| MDT | $61.95^{\pm 4.45}$ | $65.02^{\pm 6.73}$ | $63.48^{\pm 3.62}$ |
| | | 32-shot | |
| LM-base | $62.72^{\pm 3.13}$ | $65.49^{\pm 6.95}$ | $64.10^{\pm 2.48}$ |
| MDT | $64.30^{\pm 5.29}$ | $64.97^{\pm 6.95}$ | $64.63^{\pm 3.71}$ |
| | | 64-shot | |
| LM-base | $61.90^{\pm 6.97}$ | $70.57^{\pm 4.09}$ | $66.23^{\pm 2.27}$ |
| MDT | $63.73^{\pm 4.54}$ | $69.86^{\pm 4.22}$ | $66.79^{\pm 2.40}$ |

(h) HatEval binary hate En

| | hateful | normal | avg. |
|---|---|---|---|
| | | 1-shot | |
| LM-base | $44.34^{\pm 14.02}$ | $46.96^{\pm 22.46}$ | $45.65^{\pm 7.07}$ |
| MDT | $46.49^{\pm 11.06}$ | $54.76^{\pm 7.88}$ | $50.62^{\pm 3.70}$ |
| | | 4-shot | |
| LM-base | $51.81^{\pm 6.45}$ | $39.39^{\pm 19.59}$ | $45.60^{\pm 6.78}$ |
| MLM | $42.27^{\pm 12.12}$ | $50.46^{\pm 25.12}$ | $46.37^{\pm 9.00}$ |
| MTL | $28.19^{\pm 0.00}$ | $69.52^{\pm 0.00}$ | $48.85^{\pm 0.00}$ |
| Fusion | $54.05^{\pm 6.81}$ | $37.84^{\pm 8.96}$ | $45.94^{\pm 1.57}$ |
| MDT | $50.25^{\pm 2.08}$ | $58.69^{\pm 3.00}$ | $54.47^{\pm 0.79}$ |
| MDT-abl | $44.02^{\pm 2.61}$ | $61.36^{\pm 3.56}$ | $52.69^{\pm 0.96}$ |
| | | 8-shot | |
| LM-base | $40.61^{\pm 6.04}$ | $63.78^{\pm 2.79}$ | $52.20^{\pm 3.70}$ |
| MDT | $54.02^{\pm 3.96}$ | $59.23^{\pm 3.47}$ | $56.62^{\pm 0.87}$ |

(i) HatEval binary hate Es

| | offensive | normal | avg. |
|---|---|---|---|
| **1-shot** | | | |
| LM-base | $38.57^{\pm 9.68}$ | $76.77^{\pm 8.12}$ | $57.67^{\pm 2.96}$ |
| MDT | $55.34^{\pm 3.21}$ | $85.90^{\pm 2.43}$ | $70.62^{\pm 0.70}$ |
| **4-shot** | | | |
| LM-base | $34.71^{\pm 6.52}$ | $70.69^{\pm 19.28}$ | $52.70^{\pm 7.18}$ |
| MLM | $39.93^{\pm 6.42}$ | $67.32^{\pm 17.26}$ | $53.63^{\pm 7.09}$ |
| MTL | $39.48^{\pm 0.00}$ | $72.53^{\pm 0.00}$ | $56.00^{\pm 0.00}$ |
| Fusion | $54.17^{\pm 2.20}$ | $75.17^{\pm 5.15}$ | $64.67^{\pm 2.70}$ |
| MDT | $55.55^{\pm 3.70}$ | $74.06^{\pm 14.25}$ | $64.81^{\pm 8.55}$ |
| MDT-abl | $37.94^{\pm 4.08}$ | $52.24^{\pm 17.81}$ | $45.09^{\pm 7.42}$ |
| **8-shot** | | | |
| LM-base | $40.74^{\pm 5.06}$ | $70.41^{\pm 12.03}$ | $55.58^{\pm 5.76}$ |
| MDT | $56.24^{\pm 1.27}$ | $83.67^{\pm 4.22}$ | $69.95^{\pm 1.61}$ |

(j) OLID binary offensive En

| | offensive | normal | avg. |
|---|---|---|---|
| **1-shot** | | | |
| LM-base | $38.25^{\pm 9.81}$ | $60.26^{\pm 10.28}$ | $49.25^{\pm 2.94}$ |
| MDT | $57.85^{\pm 5.75}$ | $75.16^{\pm 5.34}$ | $66.50^{\pm 1.79}$ |
| **4-shot** | | | |
| LM-base | $41.91^{\pm 5.26}$ | $59.86^{\pm 15.74}$ | $50.89^{\pm 5.33}$ |
| MLM | $43.45^{\pm 6.83}$ | $62.17^{\pm 5.94}$ | $52.81^{\pm 1.16}$ |
| MTL | $53.45^{\pm 0.00}$ | $40.64^{\pm 0.00}$ | $47.05^{\pm 0.00}$ |
| Fusion | $50.96^{\pm 3.03}$ | $62.95^{\pm 6.31}$ | $56.95^{\pm 1.95}$ |
| MDT | $55.74^{\pm 3.17}$ | $74.31^{\pm 9.90}$ | $65.02^{\pm 4.17}$ |
| MDT-abl | $33.83^{\pm 17.91}$ | $60.47^{\pm 19.37}$ | $47.15^{\pm 7.19}$ |
| **8-shot** | | | |
| LM-base | $40.43^{\pm 6.96}$ | $71.57^{\pm 2.23}$ | $56.00^{\pm 2.57}$ |
| MDT | $57.70^{\pm 1.33}$ | $71.64^{\pm 7.18}$ | $64.67^{\pm 3.35}$ |

(k) GermEval binary offensive De

| | misogyny | normal | avg. |
|---|---|---|---|
| **1-shot** | | | |
| LM-base | $45.47^{\pm 5.26}$ | $65.75^{\pm 2.71}$ | $55.61^{\pm 3.40}$ |
| MDT | $49.81^{\pm 14.77}$ | $39.48^{\pm 10.09}$ | $44.64^{\pm 3.18}$ |
| **4-shot** | | | |
| LM-base | $56.02^{\pm 3.10}$ | $58.61^{\pm 9.18}$ | $57.31^{\pm 4.25}$ |
| MLM | $59.22^{\pm 5.13}$ | $38.14^{\pm 19.71}$ | $48.68^{\pm 8.20}$ |
| MTL | $31.28^{\pm 0.00}$ | $68.18^{\pm 0.00}$ | $49.73^{\pm 0.00}$ |
| Fusion | $42.23^{\pm 26.40}$ | $27.68^{\pm 33.90}$ | $34.95^{\pm 5.30}$ |
| MDT | $53.57^{\pm 5.14}$ | $41.97^{\pm 5.13}$ | $47.77^{\pm 1.36}$ |
| MDT-abl | $57.54^{\pm 7.07}$ | $50.77^{\pm 4.11}$ | $54.16^{\pm 2.78}$ |
| **8-shot** | | | |
| LM-base | $58.15^{\pm 4.89}$ | $55.99^{\pm 11.84}$ | $57.07^{\pm 4.98}$ |
| MDT | $53.57^{\pm 3.48}$ | $48.07^{\pm 8.03}$ | $50.82^{\pm 5.02}$ |

(l) AMI binary sexism En

| | misogyny | normal | avg. |
|---|---|---|---|
| **1-shot** | | | |
| LM-base | $56.37^{\pm 17.70}$ | $59.66^{\pm 11.53}$ | $58.02^{\pm 8.47}$ |
| MDT | $53.10^{\pm 8.67}$ | $45.78^{\pm 13.44}$ | $49.44^{\pm 4.03}$ |
| **4-shot** | | | |
| LM-base | $55.26^{\pm 21.34}$ | $65.40^{\pm 4.06}$ | $60.33^{\pm 10.45}$ |
| MLM | $69.23^{\pm 5.45}$ | $59.25^{\pm 10.89}$ | $64.24^{\pm 5.82}$ |
| MTL | $40.79^{\pm 0.00}$ | $49.86^{\pm 0.00}$ | $45.32^{\pm 0.00}$ |
| Fusion | $53.51^{\pm 26.77}$ | $24.72^{\pm 30.39}$ | $39.11^{\pm 11.17}$ |
| MDT | $69.33^{\pm 5.39}$ | $64.64^{\pm 5.10}$ | $66.98^{\pm 4.03}$ |
| MDT-abl | $66.54^{\pm 7.19}$ | $64.35^{\pm 5.03}$ | $65.45^{\pm 4.02}$ |
| **8-shot** | | | |
| LM-base | $70.77^{\pm 3.42}$ | $60.21^{\pm 5.12}$ | $65.49^{\pm 3.17}$ |
| MDT | $68.95^{\pm 8.80}$ | $65.65^{\pm 6.41}$ | $67.30^{\pm 6.46}$ |
| **16-shot** | | | |
| LM-base | $74.11^{\pm 2.37}$ | $70.00^{\pm 10.28}$ | $72.05^{\pm 5.73}$ |
| MDT | $75.82^{\pm 3.34}$ | $71.74^{\pm 3.27}$ | $73.78^{\pm 2.98}$ |
| **32-shot** | | | |
| LM-base | $77.83^{\pm 6.63}$ | $73.35^{\pm 5.90}$ | $75.59^{\pm 6.12}$ |
| MDT | $78.33^{\pm 1.54}$ | $73.77^{\pm 4.60}$ | $76.05^{\pm 2.52}$ |
| **64-shot** | | | |
| LM-base | $81.23^{\pm 1.86}$ | $76.76^{\pm 3.94}$ | $79.00^{\pm 2.84}$ |
| MDT | $80.40^{\pm 1.22}$ | $74.81^{\pm 4.19}$ | $77.61^{\pm 2.65}$ |

(m) AMI binary misogyny It

Table 7: Per label and macro averaged $F_1$ scores for each target dataset. *LM-base*, *MLM*, *MTL* and *Fusion* rows indicate the model trained on the target task only, the masked language modeling, multitask learning and adapter fusion baselines, *MDT* our proposed approach, while *MDT-abl.* refers to the ablation studies where external only labels are removed from the external datasets.