LCT-1 at SemEval-2023 Task 10: Pre-training and Multi-task Learning for Sexism Detection and Classification

Konstantin Chernyshev[*] **Ekaterina Garanina**[*] **Duygu Bayram**
Qiankun Zheng **Lukas Edman**
University of Groningen
{k.chernyshev, e.garanina, d.bayram.1, q.zheng.9}@student.rug.nl
j.l.edman@rug.nl

Abstract

Misogyny and sexism are growing problems in social media. Advances have been made in online sexism detection but the systems are often uninterpretable. SemEval-2023 Task 10 on Explainable Detection of Online Sexism aims at increasing explainability of the sexism detection, and our team participated in all the proposed subtasks. Our system is based on further domain-adaptive pre-training (Gururangan et al., 2020). Building on the Transformer-based models with the domain adaptation, we compare fine-tuning with multi-task learning and show that each subtask requires a different system configuration. In our experiments, multi-task learning performs on par with standard fine-tuning for sexism detection and noticeably better for coarse-grained sexism classification, while fine-tuning is preferable for fine-grained classification[1].

1 Introduction

Sexism has been appearing frequently in online spaces in recent years, which not only makes online spaces unfriendly but also exacerbates social prejudice and causes serious harm to targeted groups. In order to control and mitigate it, considerable efforts have been made to detect online sexism (Fersini et al., 2018a,b; Bhattacharya et al., 2020; Fersini et al., 2020; Rodríguez-Sánchez et al., 2021). SemEval-2023 Task 10 on Explainable Detection of Online Sexism (EDOS) (Kirk et al., 2023) aims to improve interpretability via flagging sexist content (Task A) and further deciding on a particular type of sexism in the text, including 4-category (Task B) and 11-category (Task C) classification systems. In this paper, we present our participation in all EDOS subtasks.

Given the complexity of obtaining annotated data, the research on using additional data via different training techniques is quite intensive. Gururangan et al. (2020) illustrated the benefits of further pre-training and introduced two methods for it: only using the task data or only using the domain data. These approaches, used individually and in combination, are shown to perform well in domain-specific tasks with relatively low costs of computing resources.

Safi Samghabadi et al. (2020) developed a unified end-to-end neural model using a multi-task learning (MTL) approach to address the tasks of aggression and misogyny detection. Lees et al. (2020) pre-trained a BERT-based model on 1 billion comments and fine-tuned the model with multilingual toxicity data before fine-tuning it on the target dataset. They both demonstrated that using data from similar tasks and fine-tuning the model with it in a multi-task way could improve the model's performance.

Inspired by the above studies, we explore the impact of further pre-training as well as multi-task learning on the EDOS tasks in our work. Specifically, we collect several datasets that have been developed for the related tasks (Section 2.1) and include them with various annotation schemes, e.g., binary labeling of hate speech (HS), categorization of target groups, fine-grained misogyny and sexism classification.

First, we run experiments to choose the most suitable preprocessing for our models (Section 2.2) including emoji normalization (Koufakou et al., 2020; Bornheim et al., 2021), hashtag segmentation (Liu et al., 2019a), and masks for contents like usernames and links (Paraschiv and Cercel, 2019; Zeinert et al., 2021).

Next, we conduct further pre-training on the 2 million texts provided by the organizers and the other hate speech data (Section 2.3). We experiment with both domain-adaptive and task-adaptive pre-training strategies proposed by Gururangan et al. (2020). Finally, we investigate whether the

[*] Equal contribution.

[1] The source code is available at `https://github.com/lct-rug-2022/edos-2023`.

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1573–1581
July 13-14, 2023 ©2023 Association for Computational Linguistics

HS and related		Sexism and misogyny	
Dataset	**Entries**	**Dataset**	**Entries**
OLID	14100	AMI@EVALITA2018	5000
HatEval	13000	Call me sexist	11339
Measuring HS	39565	EXIST	5644
UB	448000	Online misogyny	6355

Table 1: Datasets in English used for pre-training and MTL. All statistics are reported after dataset processing specific for our purposes.

model can benefit from additional in-domain data (Section 2.4) using the MTL approach.

The contributions of this paper are as follows:

1. We find that further domain-adaptive pre-training can improve the performance on our tasks.

2. We conclude that the benefits of multi-task learning vary across the tasks. It performs on par with standard fine-tuning on Task A and surpasses it on Task B while being inferior to the performance of the fine-tuning on Task C.

2 System overview

We treat all tasks as classification problems. Specifically, we use pre-trained Transformer-based (Vaswani et al., 2017) models and focus on three approaches to enhance the model performance: data collection and preprocessing, additional pre-training, and MTL.

2.1 Data

The EDOS dataset consists of 14,000 messages collected from Reddit and Gab, the latter social media website known for its politically far-right user base. The messages are annotated into three different levels of granularity. The most general labeling scheme, Task A, is a binary classification task to detect whether a message is sexist. Task B details 4 categories of sexism within the 3,398 *sexist* messages of Task A, which are divided further into 11 categories in Task C. Moreover, the organizers include 2 million non-labeled texts from the same websites.

We split the EDOS dataset 80:20 into train and evaluation sets and use our evaluation set to rank the results of our experiments. Throughout this paper, our evaluation set is titled *eval*, whereas *dev* and *test* refer to the development and test sets of EDOS, respectively.

Apart from the data provided by the organizers, we select[2] several related datasets annotated for hate speech (HS), sexism, and other related concepts. We divide the datasets into two groups: HS (including related tasks), and sexism or misogyny datasets with further fine-grained annotation. Table 1 contains statistics for all datasets; their detailed descriptions, as well as shorter identifiers, are located in Appendix A. We have also collected the data in other languages to investigate the influence of multilingual training on the target tasks (Appendix D).

For pre-training on HS data, we use all collected datasets for HS and sexism. For MTL, we recompile the task datasets from the original data; the details are provided in Section 2.4.

2.2 Preprocessing

We consider the following preprocessing steps:

- Creating a uniform cleaning method across datasets. Since we use several datasets from different authors, the raw data does not always have usernames and URLs masked, and the existing masks differ. We use regular expressions to ensure that all usernames and URLs are masked and that all the masking tokens are the same.

- Normalizing hashtags. We use regular expressions to detect hashtags (#) and apply English word segmentation[3].

- Converting emojis to their natural language counterparts with the `emoji` Python library[4].

2.3 Further pre-training

One line of our research is exploring the further pre-training strategies following Gururangan et al. (2020). Specifically, we train the existing pre-trained model in an unsupervised manner using Masked Language Modeling (MLM) objective. We consider the following approaches: 1) domain-adaptive pre-training (DAPT): further pre-training of a model using domain-related data available, 2) task-adaptive pre-training (TAPT): utilizing only target task text data in an unsupervised manner, and 3) sequential application of the described techniques (DAPT+TAPT).

[2] For selecting the relevant datasets, we used `https://hatespeechdata.com` (Vidgen and Derczynski, 2021).

[3] `https://pypi.org/project/wordsegment/`

[4] `https://pypi.org/project/emoji/`

Task	Code	Labels	Entries	Datasets
Hate speech	hs	binary	33000	Measuring HS, HatEval
Offensive language	offensive	binary	14100	OLID
Toxicity	toxic	binary	40000	UB
Target (within HS)	target	IND, GRP, UNT, OTH	10110	HatEval, OLID, AMI@EVALITA2018
Gender mentioned	gender	binary	79565	Measuring HS, UB
Sexism	sexism	binary	28338	all sexism datasets

Table 2: Compiled hate speech tasks for MTL for Task A. The *Code* column provides the dataset codes for further reference in MTL discussion.

As the data for DAPT, we use 2 million texts from Reddit and Gab provided by the organizers and the collected HS data described above. For TAPT, we use the EDOS task data only, including both sexist and non-sexist texts.

2.4 Multi-task learning

To make use of the available annotated HS and sexism data, we apply the multi-task learning approach (Caruana, 1997). This approach assumes training on multiple tasks at once using a single model. Since the advent of BERT models, it is common to use a shared encoder and separate task-specific heads. In this setup, the loss is averaged among the heads. We consider MTL to be beneficial since it indirectly enriches relatively scarce target task data and provides the model with more information about hate speech and sexism.

For MTL we use the MaChAmp (van der Goot et al., 2021) toolkit. During multi-task learning on multiple datasets, it first splits the datasets into batches (each batch contains instances from one dataset only) and then concatenates and shuffles the split batches before training. During training, losses are averaged with pre-defined weights to represent the final loss. The best model is selected based on the aggregated metric.

We define two sets of tasks that we consider for Task A and Tasks B and C accordingly. MTL for Task A is based on HS datasets since we hypothesize that the variety of HS-related tasks can enhance the capabilities of the model in this domain and thus increase the performance on the target task. HS tasks are presented in Table 2. When compiling the task data, Measuring HS and UB datasets were cut with random sampling to avoid heavy imbalance inside the task dataset.

Task	Code	Entries
AMI@EVALITA2018	evalita	2245
Call me sexist	sexist	1241
EXIST	exist	2794
Online misogyny	online	448

Table 3: Sexism and misogyny classification tasks for MTL for Tasks B and C. The *Code* column provides the dataset codes for further reference in MTL discussion.

Since Tasks B and C are aimed at more precise sexism classification, we apply MTL on sexism datasets with different category systems. We consider each dataset to be a separate task; Table 3 contains the statistics per task. For details on classification in each dataset, we refer the reader to the corresponding papers. In addition to the external datasets listed in Table 3, we experiment with adding Task C to Task B MTL and vice versa.

3 Experiments

We conduct a series of fine-grained experiments on preprocessing, further pre-training and multi-task learning. We compare several preprocessing components, and the main focus of pre-training and MTL experiments is the input data. We do experiments sequentially, applying the findings from the previous step to the next ones.

3.1 Evaluation

Adhering to the official target metric of the shared task, we use the F1-macro score for the intermediate and final evaluation of all models.

While working on the submission version, we primarily used our eval set for evaluating the experiments and made final decisions via online submission to the dev leaderboard; the test set was not available. In this paper, we report all dev and test scores based on the data released by the organizers after the end of the shared task.

3.2 Baseline

For the baseline, we considered a variety of state-of-the-art pre-trained models, using the base-sized models. The performance of the best models is shown in Table 4. Despite the fact that HATEBERT (Caselli et al., 2021) is slightly better at Task A and DEBERTA-V3 (He et al., 2021) has the best performance in Task B, we opted for ROBERTA (Liu et al., 2019b) since it has a stable high score over all tasks. For most of the subsequent experiments, we used the ROBERTA-LARGE model, which is known to yield better performance.

Model	Task A	Task B	Task C
ROBERTA-BASE	0.8205	0.6034	**0.3599**
HATEBERT	**0.8295**	0.6052	0.2990
DEBERTA-V3-BASE	0.8088	**0.6217**	0.3192

Table 4: F1-macro score for top-3 baseline models on the eval set. The bold font indicates the best result for each task.

3.3 Preprocessing

Considering the preprocessing options described in Section 2.2, we tested various combinations of components by fine-tuning the ROBERTA-BASE model. Since usernames and links are not masked in some of the datasets, we do masking regardless of the resulting score, thus focusing on the effect of normalizing emojis and hashtags.

The results are displayed in Table 5. Based on the obtained results, we proceeded only with the unification of masks and normalization of emojis, leaving hashtags intact.

Masks	Emoji	Hashtags	Task A
+	-	-	0.8110
+	+	-	**0.8172**
+	-	+	0.8123
+	+	+	0.8169

Table 5: F1-macro eval score of ROBERTA-BASE model fine-tuned on Task A. The preprocessing is mask normalization, emoji normalization, and hashtags normalization respectively. The bold font indicates the best score.

3.4 Pre-training

We used the pre-trained ROBERTA-LARGE model for our experiments. We further pre-train the models using MLM task until convergence of validation loss, which we define as non-decreasing loss for 5 consecutive evaluation steps. The parameters are fixed across experiments (Appendix B).

We conducted the training on EDOS data only (TAPT), and on 2M texts alone or concatenated with the collected HS data (DAPT). We have also attempted sequential training on an extended dataset and EDOS data (DAPT+TAPT), but it did not perform well in our preliminary experiments. We trained the models using only mask normalization as preprocessing, as it proved to yield better results. The resulting language models were fine-tuned and tested on the target tasks. The final f1-macro scores are shown in Table 6.

For further experiments, including MTL, we used the obtained ROBERTA-LARGE model pre-trained on 2M texts for Task A and C, and ROBERTA-LARGE model pre-trained on 2M+HS texts for Task B. Models for Tasks A and B were selected based on eval score, while the model for Task C was selected by its dev score due to its exceptionally high value.

3.5 Multi-task learning

Using the MaChAmp toolkit, we fixed the training parameters for all tasks (Appendix C), used pre-trained models selected in the previous step (Section 3.4), and experimented on the dataset combinations. All combinations included the corresponding EDOS dataset.

As opposed to testing all dataset combinations, we did the experiments incrementally. First, we tested all datasets separately, i.e. MTL on each dataset paired with the target EDOS dataset. Afterwards, we formed triples and subsequently larger sets from the most prominent options. For the best dataset combinations based on our eval set, we conducted further fine-tuning on the target task only.

The results of MTL experiments are presented in Table 7. Moreover, we conducted several experiments on using multilingual data and models and observed a major drop in the performance. We discuss these experiments in Appendix D.

4 Discussion

For two approaches that we consider – fine-tuning and MTL – we generally selected the best models

Dataset	Task A			Task B			Task C		
	eval	**dev**	**test**	**eval**	**dev**	**test**	**eval**	**dev**	**test**
Baseline	0.8456	0.8463	0.8514	0.6544	0.6621	0.5997	0.5306	0.4504	0.4621
EDOS	0.8377	0.8508	0.8517	0.6394	**0.6723**	0.6338	0.4975	0.4573	0.4650
2M	**0.8474**	0.8456	**0.8534**	0.6602	0.6422	0.6056	0.5374	**0.5201**	0.4764
2M+HS	0.8408	**0.8612**	0.8476	**0.6756**	0.6655	**0.6359**	**0.5463**	0.4713	**0.4908**

Table 6: Further MLM pre-training of RoBERTa-large model using TAPT and DAPT approaches, with fine-tuning on tasks A, B, and C. The models are scored with F1-macro. Best scores for each task and set are in bold, submitted model is highlighted with grey.

Task	Datasets	eval	dev	test
A	offensive + FT	**0.8524**	0.8521	0.8446
	offensive, target	0.8363	**0.8553**	0.8470
	hs + FT	0.8460	0.8539	**0.8585**
B	edosC, evalita + FT	**0.6777**	0.7108	0.6277
	edosC, evalita, exist	0.6463	**0.7236**	**0.6575**
C	edosB, sexist + FT	**0.5489**	0.4515	**0.4854**
	edosB, sexist	0.5155	**0.4892**	0.4518

Table 7: F1-macro scores for best MTL models on eval, dev and test sets. Base models: RoBERTa-large further pre-trained on 2M for Tasks A and C, 2M+HS for Task B. +FT: further fine-tuning on the target task after MTL. Best scores are in bold, submission systems are highlighted with grey. We did not submit a MTL model for task C due to its inferior performance on the dev set compared to standard fine-tuning.

in each setup relying on eval scores and made the final choice between the two setups using the online development set. Evaluation of all our models on now available dev and test sets reveals the ranking mismatch among the partitions. This indicates both the complexity of the task and the partition unevenness, both of which greatly complicate the best model selection. Nevertheless, we can observe certain trends.

Our pre-training experiments show that the concept of further pre-training can be beneficial for the downstream task performance of sexism detection and classification. Different input data performed better for different tasks, although TAPT achieves lower scores compared to DAPT. Although such pre-training requires additional computational resources, the resulting language model can potentially be reused for other downstream tasks.

Considering MTL, the model for Task A benefits more from the tasks of a target, offensive language, and binary HS classification, which are only somewhat related to sexism. It can be due to the fact that the model gains a larger variety of information from more distant tasks. Performance on Tasks B and C is improved by joint training on these two tasks. Nevertheless, the effect of adding other datasets appears arbitrary, making it difficult to draw any conclusions. Another inconsistency is the impact of further fine-tuning, which varies quite significantly among the experiments.

Due to the above-mentioned partition unevenness, we observed the major score decrease of the submitted models on the test set, with the alternative system variations performing better by a large margin. The final test scores of the submitted models are 0.8446 for Task A, 0.6277 for Task B, and 0.4764 for Task C.

Going beyond submitted models, further domain adaptation pre-training improves the quality of sexism detection and classification compared to the baseline, and the selection of the further fine-tuning method depends greatly on the task. MTL outperforms standard fine-tuning for Task B and shows comparable results on Task A. For Task C, the results are consistently in favor of standard fine-tuning with carefully chosen hyperparameters.

5 Conclusion

Explainable models for online sexism detection are important for building a safe environment and mitigating interpretability problems. However, our results show that high performance becomes more difficult to achieve for such a complicated task as the labels become detailed and the task becomes fine-grained.

We found that domain-adaptive further pretraining of a language model improves its performance on a downstream task. Built on the domain-adapted models, MTL and standard fine-tuning behave differently depending on the task, which means that the task formulation has a heavy impact on the model selection even in the case of in-domain data.

Acknowledgements

We would like to thank Tommaso Caselli for the insightful supervision and multiple wonderful ideas.

We would like to thank the Center for Information Technology of the University of Groningen for their support and for providing access to the Peregrine high-performance computing cluster.

The first four authors are supported by the Erasmus Mundus Master's Programme "Language and Communication Technologies"[5].

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.

Shiladitya Bhattacharya, Siddharth Singh, Ritesh Kumar, Akanksha Bansal, Akash Bhagat, Yogesh Dawer, Bornini Lahiri, and Atul Kr. Ojha. 2020. Developing a multilingual annotated corpus of misogyny and aggression. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 158–168, Marseille, France. European Language Resources Association (ELRA).

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2019. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 491–500, New York, NY, USA. Association for Computing Machinery.

Tobias Bornheim, Niklas Grieger, and Stephan Bialonski. 2021. FHAC at GermEval 2021: Identifying German toxic, engaging, and fact-claiming comments with ensemble learning. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 105–111, Duesseldorf, Germany. Association for Computational Linguistics.

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

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

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

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

Elisabetta Fersini, Debora Nozza, and Paolo Rosso. 2020. AMI @ EVALITA2020: Automatic misogyny identification. In *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*, Online. CEUR.org.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018b. Overview of the task on automatic misogyny identification at IberEval 2018. *IberEval@SEPLN*, 2150:214–228.

Ella Guest, Bertie Vidgen, Alexandros Mittos, Nishanth Sastry, Gareth Tyson, and Helen Margetts. 2021. An expert annotated dataset for the detection of online misogyny. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1336–1350, Online. Association for Computational Linguistics.

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.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with

[5] `https://lct-master.org`

gradient-disentangled embedding sharing. In *International Conference on Learning Representations*.

Chris J Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. Constructing interval variables via faceted Rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277*.

Hannah Rose Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 Task 10: Explainable Detection of Online Sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, Toronto, Canada. Association for Computational Linguistics.

Anna Koufakou, Valerio Basile, and Viviana Patti. 2020. FlorUniTo@TRAC-2: Retrofitting word embeddings on an abusive lexicon for aggressive language detection. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 106–112, Marseille, France. European Language Resources Association (ELRA).

Alyssa Whitlock Lees, Ian Kivlichan, and Jeffrey Scott Sorensen. 2020. Jigsaw @ AMI and HaSpeeDe2: Fine-tuning a pre-trained comment-domain BERT model. In *Proceedings of Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020)*, Online.

Ping Liu, Wen Li, and Liang Zou. 2019a. NULI at SemEval-2019 task 6: Transfer learning for offensive language detection using bidirectional transformers. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 87–91, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

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

Andrei Paraschiv and Dumitru-Clementin Cercel. 2019. UPB at GermEval-2019 Task 2: BERT-based offensive language classification of German tweets. In *KONVENS*.

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, Laura Plaza, Julio Gonzalo, Paolo Rosso, Miriam Comet, and Trinidad Donoso. 2021. Overview of EXIST 2021: sEXism Identification in Social neTworks. *Procesamiento del Lenguaje Natural*, 67(0):195–207.

Niloofar Safi Samghabadi, Parth Patwa, Srinivas PYKL, Prerana Mukherjee, Amitava Das, and Thamar Solorio. 2020. Aggression and misogyny detection using BERT: A multi-task approach. In *Proceedings of the Second Workshop on Trolling, Aggression and Cyberbullying*, pages 126–131, Marseille, France. European Language Resources Association (ELRA).

Mattia Samory, Indira Sen, Julian Kohne, Fabian Flöck, and Claudia Wagner. 2021. “Call me sexist, but...” : Revisiting Sexism Detection Using Psychological Scales and Adversarial Samples. *Proceedings of the International AAAI Conference on Web and Social Media*, 15(1):573–584.

Manuela Sanguinetti, Gloria Comandini, Elisa Di Nuovo, Simona Frenda, Marco Antonio Stranisci, Cristina Bosco, Tommaso Caselli, Viviana Patti, and Irene Russo. 2020. HaSpeeDe 2 @ EVALITA2020: Overview of the EVALITA 2020 hate speech detection task. *EVALITA 2020 Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian*, pages 1–9.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Bertie Vidgen and Leon Derczynski. 2021. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):1–32.

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.

Philine Zeinert, Nanna Inie, and Leon Derczynski. 2021. Annotating Online Misogyny. 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 3181–3197, Online. Association for Computational Linguistics.

A Datasets

Here we list all datasets that we collected and used either in the main set of experiments (Table 1) or for exploring multilingual MTL (Table 8).

The hate speech (HS) datasets used include:

1. Offensive Language Identification Dataset (OLID) (Zampieri et al., 2019) – offensive language in tweets.

2. HatEval (Basile et al., 2019) – HS against immigrants and women in Twitter in English and Spanish.
3. Hate Speech Detection (HaSpeeDe) (Bosco et al., 2018; Sanguinetti et al., 2020) – HS in Italian social media, including Twitter, Facebook, and news.
4. Measuring Hate Speech (Measuring HS) (Kennedy et al., 2020) – fine-grained annotation of HS (including aggregated severity score) and target identity groups in social media.
5. Jigsaw Unintended Bias in Toxicity Classification (UB) (Borkan et al., 2019) – annotation for toxicity and target identity groups on the CivilComments platform. We worked with examples with annotated identities, which is roughly a quarter of the whole dataset.

The datasets on sexism and misogyny include:

1. Automatic Misogyny Identification at EVALITA 2018 evaluation campaign (AMI@EVALITA2018) (Fersini et al., 2018a) – misogyny in tweets in Italian and English.
2. Call me sexist, but... (Call me sexist) (Samory et al., 2021) – fine-grained sexism annotation on English social media data done by multiple workers. For our purposes, we use only messages where the class can be derived from the absolute majority of votes among the workers.
3. sEXism Identification in Social neTworks 2021 (EXIST) (Rodríguez-Sánchez et al., 2021) – sexism identification and classification in tweets in English and Spanish.
4. The Expert Annotated Online Misogyny Dataset (Online misogyny) (Guest et al., 2021) – misogyny classification of substrings in the text. In our work, we use full texts and consider only texts with all spans belonging to one category.

B Further pre-training parameters

Most of the pre-training parameters follow Gururangan et al. (2020). The parameters we updated are (TAPT / DAPT respectively):

- masking probability: 15%;
- batch size: 32 / 24;
- maximum number of epochs: 10 / 5;
- learning rate: 5e-6 for both pre-training and further fine-tuning.

Task	Lang	Entries	Datasets
hs	en	33000	HatEval, Measuring HS
	es	6600	HatEval
	it	12600	HaSpeeDe
evalita	en	2245	AMI@EVALITA2018
	it	2337	
exist	en	2794	EXIST
	es	2864	

Table 8: Task datasets for multilingual MTL. The codes en, es, it stand for English, Spanish, and Italian respectively.

C Multi-task learning parameters

We kept the original task dataset sizes, applied equal loss weights for all tasks, and made the following changes to the default MaChAmp v0.4 training parameters:

- batch size: 4;
- no discriminative fine-tuning and gradual layer unfreezing;
- learning rate: 5e-6 for multi-task and 1e-6 for further fine-tuning.

We did MTL for 20 epochs and further fine-tuning for 10 epochs.

D Multilingual MTL

For exploring the multilingual MTL, we used multilingual (e.g. EXIST) and entirely non-English datasets (e.g. HaSpeeDe). We conducted two experiments: MTL on Task A with multilingual version of *hs* dataset, which contains texts in English, Italian, and Spanish, and MTL on Task B with *evalita* and *exist* datasets, also comprised of these three languages. The multilingual dataset statistics are shown in Table 8. We used `roberta-large` model for English-only setup and `xlm-roberta-large` for the multilingual one.

	A, hs	**B, evalita, exist**
Multilingual	0.8175	0.5963
English	**0.8382**	**0.6259**

Table 9: F1-macro scores of models trained on multilingual and English-only data in the MTL setup.

The results are presented in Table 9. Since the multilingual setup performs noticeably worse, we decided not to pursue this direction of research.