CL-UZH at SemEval-2023 Task 10: Sexism Detection through Incremental Fine-Tuning and Multi-Task Learning with Label Descriptions

Janis Goldzycher
Department of Computational Linguistics
University of Zurich
goldzycher@cl.uzh.ch

Abstract

The widespread popularity of social media has led to an increase in hateful, abusive, and sexist language, motivating methods for the automatic detection of such phenomena. The goal of the SemEval shared task *Towards Explainable Detection of Online Sexism* (EDOS 2023) is to detect sexism in English social media posts (subtask A), and to categorize such posts into four coarse-grained sexism categories (subtask B), and eleven fine-grained subcategories (subtask C). In this paper, we present our submitted systems for all three subtasks, based on a multi-task model that has been fine-tuned on a range of related tasks and datasets before being fine-tuned on the specific EDOS subtasks. We implement multi-task learning by formulating each task as binary pairwise text classification, where the dataset and label descriptions are given along with the input text. The results show clear improvements over a fine-tuned DeBERTa-V3 serving as a baseline leading to F_1-scores of 85.9% in subtask A (rank 13/84), 64.8% in subtask B (rank 19/69), and 44.9% in subtask C (26/63).[1]

OFFENSIVE CONTENT WARNING: This report contains some examples of hateful content. This is strictly for the purposes of enabling this research, and we have sought to minimize the number of examples where possible. Please be aware that this content could be offensive and cause you distress.

1 Introduction

With social media's expanding influence, there has been a rising emphasis on addressing the widespread issue of harmful language, especially sexist language (Meyer and Cukier, 2006; Simons, 2015; Das et al., 2020). Automatic content moderation and monitoring methods have become indispensable due to the sheer amount of posts and comments on social media platforms. However, the deployment of automatic methods has led to a new problem: current approaches to sexism detection rely on transformer-based language models whose inner workings, in spite of model interpretability research, generally remain opaque (Sun et al., 2021). This stands in contrast with the need for explainable and transparent decision processes in content moderation.

The EDOS 2023 shared task (Kirk et al., 2023) focuses on the detection (subtask A), and coarse- (subtask B) and fine-grained (subtask C) categorization of sexism. The purpose of sexism categorization is to aid the explainability of sexism detection models, where categorization can serve as additional information for why a post was classified as sexist.

In this paper, we present our approach for all three subtasks. The annotated data for detecting sexism is scarce compared to other natural language processing tasks and is often not publicly available. In response to this, we adopt a multi-task learning approach, where we first train a general model for the detection of hateful and abusive language, and incrementally adapt it to the target task.

We implement multi-task learning via manipulation of the input, concretely by adding label descriptions, and dataset identifiers. This means that the model is presented with a pairwise text classification task where it gets a label description and a dataset identifier as a first sequence and the text to classify as the second sequence. The model then learns to predict if the label description presented in the first sequence, in the context of a dataset identifier, applies to the input text of the second sequence. Figure 1 demonstrates the approach. We collect data for a range of related tasks, including hate speech detection, offensive language detection, emotion detection, stance detection on feminism and abortion, and target group detection, leading to an auxiliary training set of over 560,000 annotated

[1] We make our code publicly available at https://github.com/jagol/CL-UZH-EDOS-2023.

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1562–1572

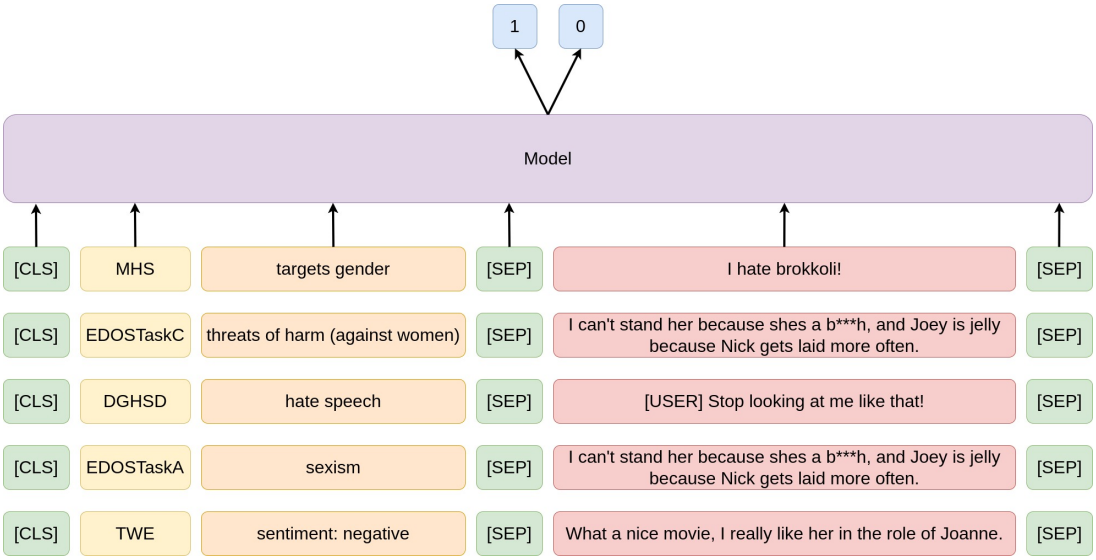

Figure 1: Task Formulation: The task is formulated as binary pairwise text classification where the model receives as input a dataset identifier, a label description, and an input text and predicts if the label, as learned for the given dataset, applies to the input text. Note that the same input text can appear with different annotations.

examples.

Our method involves a three-stage training process. As a first step, we train a general abusive language detection model using all available training data. In the second step, we further fine-tune this model on all three EDOS subtasks, and finally, in the third step, we fine-tune the model only on the target subtask.

Our models obtain strong results for subtask A, a macro-F_1 score of 0.859 achieving place 13 out of 84, but rank lower in subtasks B and C, indicating the proposed approach works comparatively well with few classes during inference time, but decreases in performance, relative to other approaches, with a higher number of classes. Our ablation study demonstrates that multi-task learning with label descriptions leads to clear performance improvements over a baseline consisting of DeBERTa-V3 (He et al., 2021) fine-tuned on each subtask. However, it remains unclear if there is a positive contribution from the additionally proposed dataset identifier.

2 Related Work

2.1 Sexism Detection

Sexism detection, sometimes also called sexism identification, is the task of predicting if a given text (typically a social media post) is sexist or not. Most research on the detection of harmful language has focused on more general phenomena such as offensive language (Pradhan et al., 2020), abusive language (Nakov et al., 2021), or hate speech (Fortuna and Nunes, 2018). Sexism intersects with these concepts but is not entirely covered by them since it also refers to subtle prejudices and stereotypes expressed against women. Accordingly, datasets for hate speech often include women as one of multiple target groups (Mollas et al., 2022; Vidgen et al., 2021; Waseem, 2016), and thus contain sexist texts, but are not exhaustive of sexism, since they do not cover its subtle forms. Recently, there has been increased attention on the detection of sexism and misogyny, leading to one shared task on sexism detection (Rodríguez-Sánchez et al., 2021) and three shared tasks on misogyny detection (Fersini et al., 2018a,b, 2020).

Harmful language detection tasks, such as sexism detection, are typically formulated as binary text classification tasks (Fortuna and Nunes, 2018). Categorizing sexism and misogyny is usually cast as a single-label multi-class classification task (Fersini et al., 2018a,b) with the exception of Parikh et al. (2019) who formulate the task as multi-label multi-task classification. Earlier approaches to sexism detection varied in their methods ranging from manual feature engineering using n-grams, lexical features, and word embeddings with statistical machine learning (Fersini et al., 2018a,b) to custom neural architectures (Fersini et al., 2020). Current approaches typically rely on fine-tuning pre-trained language models (Fersini et al., 2020; Rodríguez-Sánchez et al., 2021).

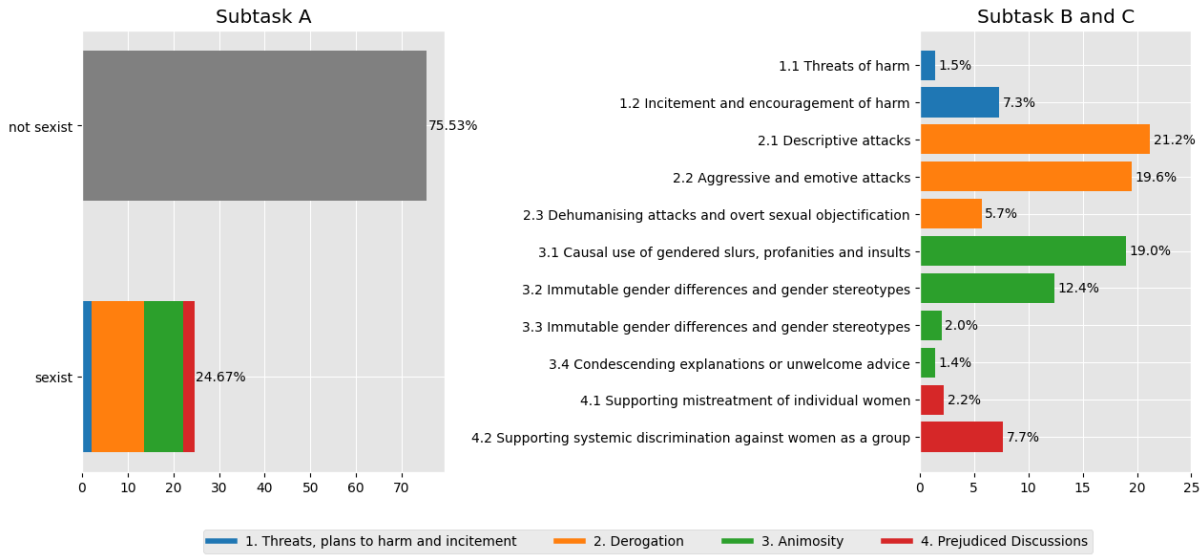

Figure 2: EDOS class distribution. Note that the category *not sexist* is absent from subtasks B and C. The percentages provided for these subtasks pertain solely to the *sexist* class, rather than the entire dataset.

2.2 Label descriptions and Multi-Task Learning

Prompts (Liu et al., 2021), and task descriptions (Raffel et al., 2020) have been used to condition generative models, while hypotheses (Wang et al., 2021) and label descriptions (Zhang et al., 2022) have been used to condition classification models in multi-task settings to produce a desired output. Multiple works have shown that multi-task learning (Caruana, 1998; Ruder, 2017) with auxiliary tasks such as polarity classification, aggression detection, emotion detection, and offensive language detection can benefit sexism detection (Abburi et al., 2020; Plaza-del Arco et al., 2021; Rodrıguez-Sánchez et al., 2021; Safi Samghabadi et al., 2020). However, to the best of our knowledge, our approach is the first to implement multi-task learning for sexism detection and categorization via label descriptions and without multiple model heads.

3 Data

3.1 EDOS Dataset

The EDOS 2023 dataset (Kirk et al., 2023) contains 20,000 posts from Gab and Reddit, labelled on three levels. On the top level, it is annotated with the binary labels *sexism* and *not-sexism*, where *sexism* is defined as "any abuse or negative sentiment that is directed towards women based on their gender or based on their gender combined with one or more other identity attributes (e.g. Black women, Muslim women, Trans women).".[2] The posts labelled as sexist are further classified into one of four categories, and eleven subcategories called vectors. The label taxonomy and the respective class distributions are displayed in Figure 2.

14,000 labelled examples are released as training data, and 2,000 examples and 4,000 examples are held back for validation and testing, respectively. Additionally, the shared task organizers provide one million unlabelled examples from Reddit and one million unlabelled examples from Gab. For our approach, we do not make use of the unlabelled data.

3.2 Auxiliary Datasets

However, we do make use of the following additional, labelled datasets as auxiliary training sets for multi-task learning:

DGHSD The "Dynamically Generated Hate Speech Dataset" (Vidgen et al., 2021) contains artificial adversarial examples aimed at tricking a binary hate speech detection model into predicting the wrong class.

MHS The "Measuring Hate Speech" dataset (Kennedy et al., 2020) contains comments sourced from Youtube, Twitter, and Reddit and is annotated for ten attributes related to hate speech. We only use the subset of labels listed in Table 1.

[2] https://codalab.lisn.upsaclay.fr/competitions/7124#learn_the_details-overview

dataset	label type	label value	size
DGHSD	hate speech	yes: 46.1% no: 53.9%	32,924
SBF	lewd	yes: 10.1% no: 89.9%	35,424
	offensive	yes: 47.1% no: 52.9%	35,424
MHS	hate speech	yes: 40.5% no: 59.5%	130,000
	targets gender	yes: 29.8% no: 70.2%	130,000
	targets women	yes: 21.9% no: 78.1%	130,000
TWE	offensive	yes: 33.1% no: 66.9%	11,916
	sentiment	negative: 15.5% neutral: 45.3% positive: 39.1%	45,615
	emotion	anger: 43.0% joy: 21.7% optimism: 9.0% sadness: 26.3%	3,257
	hate	yes: 42.0% no: 58.0%	9,000
	irony	yes: 50.5% no: 49.5%	2,862
	stance feminist	none: 18.9% against: 49.4% favor: 31.7%	597
	stance abortion	none: 27.1% against: 54.3% favor: 18.6%	587

Table 1: Label distributions of the auxiliary datasets.

SBF The "Social Bias Frames" dataset (Sap et al., 2020) is a combination of multiple Twitter datasets (Founta et al., 2018; Davidson et al., 2017; Waseem and Hovy, 2016) with newly collected data from Reddit, Gab, and Stormfront. We only use the subset of labels listed in Table 1.

TWE "TweetEval" (Barbieri et al., 2020) combines multiple datasets for different tasks into a single benchmark for detecting various aspects of tweets. We use the datasets for emotion classification (Mohammad et al., 2018), irony detection (Van Hee et al., 2018), hate speech detection (Basile et al., 2019), offensive language detection (Zampieri et al., 2019), sentiment detection (Rosenthal et al., 2017), and stance detection (Mohammad et al., 2016) for stances on the topics *feminism* and *abortion*.

3.3 Preprocessing

During preprocessing, we replaced all URLs in the input texts with the placeholder string "`[URL]`", all usernames (strings starting with an "`@`") with "`[USER]`", and all emojis with the respective textual description, also surrounded by brackets.

phase	Parameter	Value
general	loss function	cross-entropy loss
	optimizer	Adam (Kingma and Ba, 2015)
	β_1	0.9
	β_2	0.999
	learning rate	1e-6
	warmup steps	1,000
	effective batch size	32
	evaluation metric	macro-F_1
	early stopping	✓
1/3: AUX + EDOS	max epochs	1
2/3: EDOS	max epochs	20
	patience	5
3/3: EDOS A/B/C	max epochs	20
	patience	5

Table 2: Training hyperparameters. The left column refers to the different training phases. *general* applies to all training phases and all EDOS subtasks. *AUX+EDOS* refers to training on all auxiliary datasets, *EDOS* to training on all EDOS subtasks and *EDOS A/B/C* to training only on the target subtask.

4 System Description

We formulate each EDOS subtask as a binary pairwise classification task where the model predicts if a given label applies to the input text. This allows us to simultaneously train on multiple datasets with different labeling schemes and a different number of distinct labels without adjusting the model architecture or having to use multiple model heads.

Formally, our model receives as input (1) the concatenation of a dataset identifier $d_i \in D$ and a label description $l_j \in L$, and (2) the input text t. It predicts the probability distribution $y = \text{softmax}(\text{model}(\text{concat}(d_i, l_j), t))$ where $y \in \mathbb{R}^2$. y_1 then denotes the probability that l_j, given the context of d_i, does apply to t.

4.1 Model Details

We use DeBERTa (He et al., 2020), specifically DeBERTa-V3-large (He et al., 2021) fine-tuned on a range of natural language inference (NLI) datasets (Laurer et al., 2022)[3], since this model is already fine-tuned to classify and relate text pairs. We only change the output dimensionality from 3 to 2 for binary classification. In ablation tests, we also use the DeBERTa-V3-large without further fine-tuning.[4]

[3]The model is publicly available at https://huggingface.co/MoritzLaurer/DeBERTa-v3-large-mnli-fever-anli-ling-wanli.

[4]The model is publicly available at https://huggingface.co/microsoft/deberta-v3-large.

4.2 Label Descriptions

Where possible, we use the label names listed in Figure 2 and Table 1 as the label description. However, we make the following exceptions and adjustments: We strip the numbering from the label names for EDOS subtasks B, and C, and add the string " (against women)" at the end since this target group information is not yet in the label name. For multi-class classification in auxiliary datasets, we follow the format "`<label type>: <label value>`". For example, for sentiment classification, which has the three possible label values `negative`, `neutral`, and `positive`, we can generate a true example for positive sentiment with the label description "`sentiment: positive`".

4.3 Dataset Identifier

The same labels may have slightly different definitions in different datasets or may be differently applied due to different annotators. If no further information is given to the model, this could be a source of noise. Multiple datasets contain the label *hate speech* for our auxiliary datasets. To account for this, we introduce dataset identifiers, which are short dataset abbreviations of a few characters in length that are concatenated with the label description.

4.4 Training Procedure

We train the model in three phases: In the first phase, the model is trained with all available examples of all collected datasets. In the second phase, the best checkpoint from the previous phase is further fine-tuned on EDOS data from all three task levels (subtasks A, B, and C). Finally, in the third phase, the model is fine-tuned only on examples from the relevant subtask. We consider all three annotations from the three subtasks per example for validation during the first two phases, resulting in 6,000 annotations for each validation. In the last training phase, the model is only fine-tuned on one subtask. Thus, we only validated on the labels for that specific subtask. Further training details are provided in Table 2.

4.5 Random Negative Sampling

When converting a multi-class classification task (such as subtask B and C) to a binary pairwise text classification task, each positive example for class $c_k \in C$, where $k \in [0, ..., |C|]$, can be turned into $|C| - 1$ negative examples by choosing a label $c_k \in C \setminus \{c_k\}$. However, generating all possible negative examples for a positive example would result in an imbalanced training set. Therefore, in settings with more than two classes, we instead sample a random wrong class label during training to create one negative example for each positive example.[5] This means the model will be trained on different negative examples in each epoch while the positive examples stay the same.

4.6 Inference

During inference, we predict a probability p_i for each candidate class $c_i \in C$ and select the class with the highest probability. This means that we perform $|C|$ number of forward passes per prediction, except for binary classification (subtask A), where we can use just one forward pass to predict a probability for the label *sexism*.

Since our model produces just one probability for subtask A, we can select a probability threshold. For our official submission and for our ablation experiments, we test the thresholds $\{0.5, 0.6, 0.7, 0.8, 0.9\}$ on the validation set and use the highest performing threshold for the test set.

5 Experiments and Results

Table 4 contains the official evaluation scores showing strong results for subtask A and moderately good results for subtasks B and C.

5.1 Ablation Study

To illustrate the relative importance of the proposed methods, we systematically add components to a baseline model until we arrive at the submitted models and run each model version with three random seeds for our ablation tests. We evaluate the following settings: (1) "*single task EDOS*": We start with DeBERTa-V3-large models, fine-tuned on each subtask individually, serving as our baseline. (2) "*+ label description*": We add label descriptions while still only training each model on one subtask. (3) "*multi-task EDOS via label descriptions*": The models are trained on all three subtasks simultaneously using label descriptions. (4) "*+ NLI fine-tuning*": We repeat the setting but start training from the DeBERTa-V3 checkpoint fine-tuned on NLI datasets (see Section 4.1). (5)

[5] This applies to EDOS subtasks B and C, and the sentiment- emotion- and stance- detection tasks in TweetEval.

	A	B	C	AVG
single task EDOS	0.840	0.202	0.122	0.388
+ label description	0.851	0.160	0.098	0.370
multi-task EDOS via label descriptions	0.851	0.504	0.248	0.534
+ NLI fine-tuning	0.854	0.556	0.352	0.587
+ single task fine-tuning	0.850	0.623	0.412	0.629
+ fine-tuning on AUX	0.858	0.633	0.417	0.636
+ dataset identifier	0.858	0.629	0.431	0.640
+ class balancing	-	0.642	0.466	-

Table 3: Results of the ablation study on the test set. The metric is macro-F_1.

	F_1	Rank
Subtask A	0.859	13/84
Subtask B	0.648	19/69
Subtask C	0.449	26/63

Table 4: Results of the official evaluation on the test set.

"+ *single task fine-tuning*": We add a second fine-tuning phase in which the multi-task model is only fine-tuned on the target subtask. (6) "+ *fine-tuning on AUX* ": We add fine-tuning on auxiliary tasks and EDOS simultaneously as a first training phase. (7) "+ *dataset identifier*": We add the dataset identifier to the input. (8) *"class balancing"*: Finally, we perform upsampling to increase the relative frequency of scarce classes.[6] The upsampled version of the dataset is only used during the last fine-tuning phase.

Table 3 contains the test set results averaged over the three runs with different seeds. The full results for each run, including evaluations after intermediary training phases, are displayed in Appendix A. In what follows we analyze the effects of different system components and settings.

Baseline The baseline *single task EDOS* shows already a strong performance on subtask A, but leads to surprisingly low scores on subtasks B and C. We assume that this is due to underprediction and low performance of the very scarce classes (four classes of subtask C are below 3%), which can drastically reduce the macro-F_1 score.

Multi-Task Learning on all EDOS-Subtasks Comparing the baseline, with *multi-task EDOS via label descriptions*, shows a clear improvement of 1.1 percentage points (pp) from multi-task learning with label descriptions on subtask A, and drastic improvements, more than doubling performance, for subtasks B and C. Looking at single task models with label descriptions (+ *label description*) reveals that on subtask A the entire increase in performance is due to label descriptions while in subtasks B and C the dramatic performance increases are due to multi-task learning.

Starting with an NLI Model Starting training from a DeBERTa-V3 checkpoint that is already fine-tuned on NLI increases scores on all three subtasks. The more classes the task has, the larger is the increase.

Additional Single-Task Fine-Tuning Adding a second subtask-specific fine-tuning phase after training on all EDOS subtasks leads to increases of 6.7pp and 4.0pp for subtasks B and C. However, it reduces performance on subtask A by 0.4pp.

Multi-Task Learning on Auxiliary Tasks Inserting a first training phase that includes all auxiliary tasks and EDOS subtasks into the training process leads to further improvements of up to 1.0pp on all subtasks.

The Dataset Identifier Adding a dataset identifier to the input leads to mixed results. On subtask A the model does not change in performance, on subtask B it slightly decreases, and on subtask C we observe a clear increase of 1.4pp. Overall, we cannot draw a clear conclusion about the effects of the dataset identifier.

Class Balancing Finally, we observe that upsampling low-frequency classes in subtasks B and C has positive effects of 1.3pp and 3.5pp respectively.

[6] In subtask B, we increase classes with a frequency below 19% to ~19%. For subtask C, we upsample classes below 9% to ~9%.

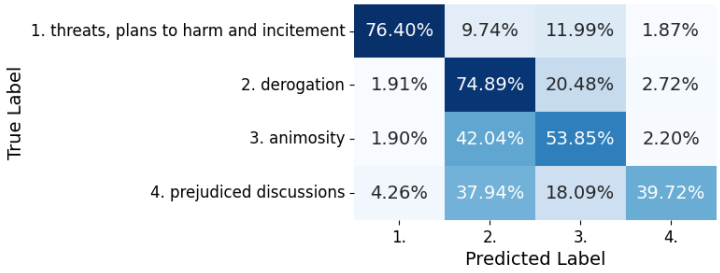

Figure 3: Visualized normalized confusion matrix for subtask B.

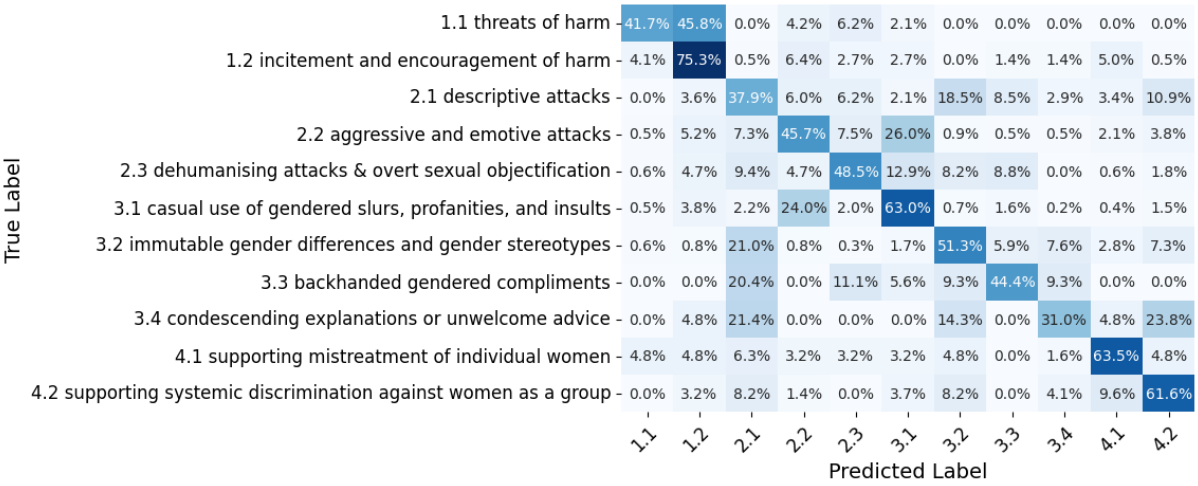

Figure 4: Visualized normalized confusion matrix for subtask C.

5.2 Error Analysis

Figures 3 and 4 display the confusion matrices averaged over three random seeds for the submitted model configurations for subtasks B and C.

In subtask B, we see that the category *threats, plans to harm, and incitement* is accurately predicted. The *derogation* class has a high recall, likely because it is the most common category. However, this also results in a significant number of false positives from the *animosity* and *prejudiced discussions* classes. As a consequence, these last two classes are strongly underpredicted.

In subtask C, it is evident that mispredictions generally stem from class confusions within the subtask B categories. Erroneous predictions outside of these categories are uncommon, except for *descriptive attacks*, which are frequently mistaken for various forms of *animosity*.

6 Conclusion

In this paper, we presented our approaches and results for all three subtasks of the shared task *Towards Explainable Sexism Detection*. We developed and evaluated a multi-task learning model that is trained in three phases: (1) training a general multi-task abusive language detection model, (2) fine-tuning the model on all three EDOS subtasks, thus specializing it in sexism detection, and (3) fine-tuning the model only on the target subtask. We implemented the multi-task capabilities only via input manipulation, i.e., label descriptions and dataset identifiers, without modifying the model architecture or using multiple model heads.

In the official shared task evaluation, our approach led to strong results on subtask A and moderately good results on subtask B and C, indicating that the method decreases more in performance with a higher number of classes than other approaches. Our ablation tests demonstrate that multi-task learning via label descriptions led to significant performance improvements on subtask A and large performance improvements on subtasks B and C. It remains unclear if the dataset identifier has any positive effect. Overall the results show that our model for binary sexism detection is reliable, but that there is still much room for improvement in sexism categorization.

Acknowledgments

We thank Chantal Amrhein and Simon Clematide, for the valuable conversations, and suggestions, and Jonathan Schaber, and Gerold Schneider for the helpful comments. We also thank the anonymous reviewers for their constructive feedback.

References

Harika Abburi, Pulkit Parikh, Niyati Chhaya, and Vasudeva Varma. 2020. Semi-supervised multi-task learning for multi-label fine-grained sexism classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5810–5820, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

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. 1998. *Multitask Learning. Learning to Learn*, pages 95–133. Springer US, Boston, MA.

Mithun Das, Binny Mathew, Punyajoy Saha, Pawan Goyal, and Animesh Mukherjee. 2020. Hate speech in online social media. *SIGWEB Newsl.*, Autumn 2020. New York, NY, USA.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. *Proceedings of the International AAAI Conference on Web and Social Media*, 11(1):512–515. Montréal, Québec, Canada.

Elisabetta Fersini, Debora Nozza, Paolo Rosso, et al. 2018a. Overview of the EVALITA 2018 Task on Automatic Misogyny Identification (ami). In *EVALITA Evaluation of NLP and Speech Tools for Italian Proceedings of the Final Workshop 12-13 December 2018, Naples*. Accademia University Press.

Elisabetta Fersini, Debora Nozza, Paolo Rosso, et al. 2020. AMI@ EVALITA2020: Automatic Misogyny Identification. In *Proceedings of the 7th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2020)*. CEUR Workshop Proceedings. Online.

Elisabetta Fersini, Paolo Rosso, and Maria Anzovino. 2018b. Overview of the Task on Automatic Misogyny Identification at IberEval 2018. *Ibereval@ sepln*, 2150:214–228. Seville (Spain).

Paula Fortuna and Sérgio Nunes. 2018. A Survey on Automatic Detection of Hate Speech in Text. *ACM Comput. Surv.*, 51(4). New York, NY, USA.

Antigoni Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. In *Twelfth International AAAI Conference on Web and Social Media*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

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:2009.10277 [cs].

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

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.

Moritz Laurer, Wouter van Atteveldt, Andreu Casas, and Kasper Welbers. 2022. Less Annotating, More Classifying – Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT - NLI. *online*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *arXiv:2107.13586 [cs]*. ArXiv: 2107.13586.

R. Meyer and M. Cukier. 2006. Assessing the attack threat due to irc channels. In *International Conference on Dependable Systems and Networks (DSN'06)*, pages 467–472.

Saif Mohammad, Felipe Bravo-Marquez, Mohammad Salameh, and Svetlana Kiritchenko. 2018. Semeval-2018 Task 1: Affect in Tweets. In *Proceedings of the*

12th International Workshop on Semantic Evaluation, pages 1–17. New Orleans, LA, USA.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. Semeval-2016 Task 6: Detecting Stance in Tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 31–41. San Diego, California, USA.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2022. ETHOS: A Multi-Label Hate Speech Detection Dataset. *Complex & Intelligent Systems*.

Preslav Nakov, Vibha Nayak, Kyle Dent, Ameya Bhatawdekar, Sheikh Muhammad Sarwar, Momchil Hardalov, Yoan Dinkov, Dimitrina Zlatkova, Guillaume Bouchard, and Isabelle Augenstein. 2021. Detecting Abusive Language on Online Platforms: A Critical Analysis. *arXiv:2103.00153 [cs]*. ArXiv: 2103.00153.

Pulkit Parikh, Harika Abburi, Pinkesh Badjatiya, Radhika Krishnan, Niyati Chhaya, Manish Gupta, and Vasudeva Varma. 2019. Multi-label categorization of accounts of sexism using a neural framework. 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 1642–1652, Hong Kong, China. Association for Computational Linguistics.

Flor Miriam Plaza-del Arco, M Dolores Molina-González, LAU López, and MT Martín-Valdivia. 2021. Sexism identification in social networks using a multi-task learning system. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021), XXXVII International Conference of the Spanish Society for Natural Language Processing., Málaga, Spain*, volume 2943, pages 491–499.

Rahul Pradhan, Ankur Chaturvedi, Aprna Tripathi, and Dilip Kumar Sharma. 2020. A Review on Offensive Language Detection. In *Advances in Data and Information Sciences*, Lecture Notes in Networks and Systems, pages 433–439, Singapore. Springer.

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

Francisco Rodrıguez-Sánchez, Jorge Carrillo-de Albornoz, and Laura Plaza. 2021. A multi-task and multilingual model for sexism identification in social networks. *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2021) co-located with the Conference of the Spanish Society for Natural Language Processing (SEPLN 2021)*.

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. *Proces. del Leng. Natural*, 69:229–240.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 502–518, Vancouver, Canada. Association for Computational Linguistics.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

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).

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Rachel Noelle Simons. 2015. Addressing Gender-Based Harassment in Social Media: A Call to Action. *iConference 2015 Proceedings*. Online.

Xiaofei Sun, Diyi Yang, Xiaoya Li, Tianwei Zhang, Yuxian Meng, Han Qiu, Guoyin Wang, Eduard Hovy, and Jiwei Li. 2021. Interpreting Deep Learning Models in Natural Language Processing: A Review. Arxiv:2110.10470.

Cynthia Van Hee, Els Lefever, and Véronique Hoste. 2018. SemEval-2018 task 3: Irony detection in English tweets. In *Proceedings of the 12th International Workshop on Semantic Evaluation*, pages 39–50, New Orleans, Louisiana. Association for Computational Linguistics.

Bertie Vidgen, Tristan Thrush, Zeerak Waseem, and Douwe Kiela. 2021. Learning from the worst: Dynamically generated datasets to improve online hate detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1667–1682, Online. Association for Computational Linguistics.

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

Zeerak Waseem. 2016. Are you a racist or am I seeing things? annotator influence on hate speech detection on Twitter. In *Proceedings of the First Workshop on NLP and Computational Social Science*, pages 138–142, Austin, Texas. Association for Computational Linguistics.

Zeerak Waseem and Dirk Hovy. 2016. Hateful Symbols or Hateful People? Predictive Features for Hate Speech Detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86.

Ruohong Zhang, Yau-Shian Wang, Yiming Yang, Donghan Yu, Tom Vu, and Li Lei. 2022. Long-tailed extreme multi-label text classification with generated pseudo label descriptions. *ArXiv*, abs/2204.00958.

A Full Results

Training sets and training phases	Run	LD	DI	Base model	Development set A	Development set ρ	Development set B	Development set C	Test set A	Test set B	Test set C
EDOS A	1	✗	✗	**DBV3**	0.840	0.7	-	-	0.837	-	-
EDOS A	2	✗	✗	**DBV3**	0.845	0.5	-	-	0.848	-	-
EDOS A	3	✗	✗	**DBV3**	0.837	0.5	-	-	0.836	-	-
EDOS B	1	✗	✗	**DBV3**	-	-	0.159	-	-	0.159	-
EDOS B	2	✗	✗	**DBV3**	-	-	0.159	-	-	0.159	-
EDOS B	3	✗	✗	**DBV3**	-	-	0.306	-	-	0.287	-
EDOS C	1	✗	✗	**DBV3**	-	-	-	0.114	-	-	0.115
EDOS C	2	✗	✗	**DBV3**	-	-	-	0.136	-	-	0.126
EDOS C	3	✗	✗	**DBV3**	-	-	-	0.110	-	-	0.124
EDOS A	1	✓	✗	**DBV3**	0.853	0.5	-	-	0.852	-	-
EDOS A	2	✓	✗	**DBV3**	0.845	0.5	-	-	0.852	-	-
EDOS A	3	✓	✗	**DBV3**	0.851	0.5	-	-	0.849	-	-
EDOS B	1	✓	✗	**DBV3**	-	-	0.162	-	-	0.159	-
EDOS B	2	✓	✗	**DBV3**	-	-	0.162	-	-	0.159	-
EDOS B	3	✓	✗	**DBV3**	-	-	0.159	-	-	0.161	-
EDOS C	1	✓	✗	**DBV3**	-	-	-	0.117	-	-	0.118
EDOS C	2	✓	✗	**DBV3**	-	-	-	0.094	-	-	0.086
EDOS C	3	✓	✗	**DBV3**	-	-	-	0.078	-	-	0.089
EDOS ABC	1	✓	✗	**DBV3**	0.865	0.5	0.556	0.225	0.851	0.530	0.253
EDOS ABC	2	✓	✗	**DBV3**	0.850	0.5	0.466	0.193	0.845	0.449	0.184
EDOS ABC	3	✓	✗	**DBV3**	0.860	0.7	0.570	0.329	0.857	0.533	0.309
EDOS ABC	1	✓	✗	**DBV3-NLI**	0.855	0.5	0.616	0.439	0.854	0.554	0.350
EDOS ABC	2	✓	✗	**DBV3-NLI**	0.855	0.5	0.617	0.431	0.852	0.555	0.355
EDOS ABC	3	✓	✗	**DBV3-NLI**	0.854	0.6	0.614	0.440	0.855	0.558	0.351
Ph1: EDOS ABC, Ph2: EDOS A	1	✓	✗	**DBV3-NLI**	0.853	0.5	-	-	0.848	-	-
Ph1: EDOS ABC, Ph2: EDOS A	2	✓	✗	**DBV3-NLI**	0.855	0.6	-	-	0.848	-	-
Ph1: EDOS ABC, Ph2: EDOS A	3	✓	✗	**DBV3-NLI**	0.854	0.9	-	-	0.855	-	-
Ph1: EDOS ABC, Ph2: EDOS B	1	✓	✗	**DBV3-NLI**	-	-	0.665	-	-	0.615	-
Ph1: EDOS ABC, Ph2: EDOS B	2	✓	✗	**DBV3-NLI**	-	-	0.683	-	-	0.637	-
Ph1: EDOS ABC, Ph2: EDOS B	3	✓	✗	**DBV3-NLI**	-	-	0.683	-	-	0.616	-
Ph1: EDOS ABC, Ph2: EDOS C	1	✓	✗	**DBV3-NLI**	-	-	-	0.496	-	-	0.412
Ph1: EDOS ABC, Ph2: EDOS C	2	✓	✗	**DBV3-NLI**	-	-	-	0.496	-	-	0.412
Ph1: EDOS ABC, Ph2: EDOS C	3	✓	✗	**DBV3-NLI**	-	-	-	0.496	-	-	0.412
Ph1: AUX + EDOS ABC	1	✓	✗	**DBV3-NLI**	0.825	0.5	0.283	0.247	0.831	0.263	0.228
Ph1: AUX + EDOS ABC	2	✓	✗	**DBV3-NLI**	0.825	0.5	0.291	0.230	0.828	0.269	0.237
Ph1: AUX + EDOS ABC	3	✓	✗	**DBV3-NLI**	0.824	0.5	0.302	0.245	0.827	0.284	0.239
Ph1: AUX + EDOS ABC Ph2: EDOS ABC	1	✓	✗	**DBV3-NLI**	0.850	0.6	0.601	0.422	0.860	0.541	0.382
Ph1: AUX + EDOS ABC Ph2: EDOS ABC	2	✓	✗	**DBV3-NLI**	0.851	0.6	0.608	0.421	0.859	0.549	0.379
Ph1: AUX + EDOS ABC Ph2: EDOS ABC	3	✓	✗	**DBV3-NLI**	0.853	0.5	0.602	0.424	0.857	0.538	0.380
Ph1: AUX + EDOS ABC Ph2: EDOS ABC, Ph3: EDOS A	1	✓	✗	**DBV3-NLI**	0.855	0.6	-	-	0.860	-	-
Ph1: AUX + EDOS ABC Ph2: EDOS ABC, Ph3: EDOS A	2	✓	✗	**DBV3-NLI**	0.854	0.7	-	-	0.858	-	-
Ph1: AUX + EDOS ABC Ph2: EDOS ABC, Ph3: EDOS A	3	✓	✗	**DBV3-NLI**	0.855	0.7	-	-	0.857	-	-
Ph1: AUX + EDOS ABC Ph2: EDOS ABC, Ph3: EDOS B	1	✓	✗	**DBV3-NLI**	-	-	0.663	-	-	0.594	-
Ph1: AUX + EDOS ABC Ph2: EDOS ABC, Ph3: EDOS B	2	✓	✗	**DBV3-NLI**	-	-	0.693	-	-	0.657	-
Ph1: AUX + EDOS ABC Ph2: EDOS ABC, Ph3: EDOS B	3	✓	✗	**DBV3-NLI**	-	-	0.689	-	-	0.649	-
Ph1: AUX + EDOS ABC Ph2: EDOS ABC, Ph3: EDOS C	1	✓	✗	**DBV3-NLI**	-	-	-	0.507	-	-	0.423
Ph1: AUX + EDOS ABC Ph2: EDOS ABC, Ph3: EDOS C	2	✓	✗	**DBV3-NLI**	-	-	-	0.495	-	-	0.425
Ph1: AUX + EDOS ABC Ph2: EDOS ABC, Ph3: EDOS C	3	✓	✗	**DBV3-NLI**	-	-	-	0.478	-	-	0.401
Ph1: AUX + EDOS ABC	1	✓	✓	**DBV3-NLI**	0.796	0.5	0.260	0.196	0.805	0.236	0.214
Ph1: AUX + EDOS ABC	2	✓	✓	**DBV3-NLI**	0.798	0.5	0.259	0.206	0.801	0.237	0.207
Ph1: AUX + EDOS ABC	3	✓	✓	**DBV3-NLI**	0.793	0.5	0.257	0.226	0.803	0.253	0.230
Ph1: AUX + EDOS ABC Ph2: EDOS ABC	1	✓	✓	**DBV3-NLI**	0.862	0.6	0.637	0.466	0.859	0.605	0.395
Ph1: AUX + EDOS ABC Ph2: EDOS ABC	2	✓	✓	**DBV3-NLI**	0.852	0.5	0.565	0.411	0.857	0.533	0.370
Ph1: AUX + EDOS ABC Ph2: EDOS ABC	3	✓	✓	**DBV3-NLI**	0.849	0.6	0.569	0.424	0.859	0.534	0.366
Ph1: AUX + EDOS ABC Ph2: EDOS ABC, Ph3: EDOS A	1	✓	✓	**DBV3-NLI**	0.858	0.7	-	-	0.859	-	-
Ph1: AUX + EDOS ABC Ph2: EDOS ABC, Ph3: EDOS A	2	✓	✓	**DBV3-NLI**	0.855	0.6	-	-	0.856	-	-
Ph1: AUX + EDOS ABC Ph2: EDOS ABC, Ph3: EDOS A	3	✓	✓	**DBV3-NLI**	0.862	0.5	-	-	0.861	-	-
Ph1: AUX + EDOS ABC Ph2: EDOS ABC, Ph3: EDOS B	1	✓	✓	**DBV3-NLI**	-	-	0.674	-	-	0.633	-
Ph1: AUX + EDOS ABC Ph2: EDOS ABC, Ph3: EDOS B	2	✓	✓	**DBV3-NLI**	-	-	0.665	-	-	0.642	-
Ph1: AUX + EDOS ABC Ph2: EDOS ABC, Ph3: EDOS B	3	✓	✓	**DBV3-NLI**	-	-	0.664	-	-	0.613	-
Ph1: AUX + EDOS ABC Ph2: EDOS ABC, Ph3: EDOS C	1	✓	✓	**DBV3-NLI**	-	-	-	0.522	-	-	0.455
Ph1: AUX + EDOS ABC Ph2: EDOS ABC, Ph3: EDOS C	2	✓	✓	**DBV3-NLI**	-	-	-	0.464	-	-	0.419
Ph1: AUX + EDOS ABC Ph2: EDOS ABC, Ph3: EDOS C	3	✓	✓	**DBV3-NLI**	-	-	-	0.473	-	-	0.419
Ph1: AUX + EDOS ABC Ph2: EDOS ABC, Ph3: EDOS B (up to ~19%)	1	✓	✓	**DBV3-NLI**	-	-	0.679	-	-	0.653	-
Ph1: AUX + EDOS ABC Ph2: EDOS ABC, Ph3: EDOS B (up to ~19%)	2	✓	✓	**DBV3-NLI**	-	-	0.677	-	-	0.642	-
Ph1: AUX + EDOS ABC Ph2: EDOS ABC, Ph3: EDOS B (up to ~19%)	3	✓	✓	**DBV3-NLI**	-	-	0.661	-	-	0.632	-
Ph1: AUX + EDOS ABC Ph2: EDOS ABC, Ph3: EDOS C (up to ~9%)	1	✓	✓	**DBV3-NLI**	-	-	-	0.473	-	-	0.462
Ph1: AUX + EDOS ABC Ph2: EDOS ABC, Ph3: EDOS C (up to ~9%)	2	✓	✓	**DBV3-NLI**	-	-	-	0.497	-	-	0.470
Ph1: AUX + EDOS ABC Ph2: EDOS ABC, Ph3: EDOS C (up to ~9%)	3	✓	✓	**DBV3-NLI**	-	-	-	0.516	-	-	0.466

Table 6: Full results of the ablation study on the development and test set. *LD* refers to label descriptions, and *DI* refers to dataset identifiers. *DBV3* refers to DeBERTa-V3-large and *DBV3-NLI* refers to DeBERTa-V3-large fine-tuned on NLI datasets. ρ refers to the threshold applied for subtask A. The settings containing the settings of the models submitted to the official evaluation are marked in grey. *Ph1*, *Ph2*, and *Ph3* stand for training phase 1, 2, and 3 respectively.