DH-FBK at SemEval-2023 Task 10: Multi-Task Learning with Classifier Ensemble Agreement for Sexism Detection

Elisa Leonardelli
Fondazione Bruno Kessler
Trento, Italy
eleonardelli@fbk.eu

Camilla Casula
Fondazione Bruno Kessler
University of Trento
Trento, Italy
ccasula@fbk.eu

Abstract

This paper presents the submissions of the DH-FBK team for the three tasks of Task 10 at SemEval 2023. The Explainable Detection of Online Sexism (EDOS) task aims at detecting sexism in English text in an accurate and explainable way, thanks to a fine-grained annotation that follows a three-level schema: sexist or not (Task A), category of sexism (Task B) and vector of sexism (Task C) exhibited. We use a multi-task learning approach in which models share representations from all three tasks, allowing for knowledge to be shared across them. Notably, with our approach a single model can solve all three tasks. In addition, motivated by the subjective nature of the task, we incorporate inter-annotator agreement information in our multi-task architecture. Although disaggregated annotations are not available, we artificially estimate them using a 5-classifier ensemble, and show that ensemble agreement can be a good approximation of crowd agreement. Our approach achieves competitive results, ranking 32nd out of 84, 24th out of 69 and 11th out of 63 for Tasks A, B and C respectively. We finally show that low inter-annotator agreement levels are associated with more challenging examples for models, making agreement information useful for this kind of task.

⚠ **Warning**: *this paper contains examples that may be offensive or upsetting.*

1 Introduction

The prevalence of sexism on the Internet can be damaging to the women it targets and make online spaces hostile. In order to address this, automated methods are often used to identify sexist content on a large scale, though typically only high-level categories are identified, without further explanation. Improved interpretability and understanding of the decisions made by artificial intelligence when flagging what is sexist and why it is sexist could empower both users and moderators. Recently, the SemEval-2023 Task 10 (Kirk et al., 2023) "Explainable Detection of Online Sexism" (EDOS) has been proposed, to encourage research on sexist language detection in a more accurate and explainable way. This shared task focuses on exploring methods for identifying explicit sexism and classifying the type of behavior expressed into fine-grained categories in English data from Gab and Reddit.

In this paper, we present the **DH-FBK** entries for the three tasks of EDOS. We propose a multi-task learning approach, a paradigm that leverages training signals of related tasks at the same time by exploiting a shared representation in the model. In particular, we simultaneously train models on all three tasks, corresponding to the three levels of the annotation schema utilized. The hypothesis is that the annotation levels contain mutually relevant information, making knowledge sharing beneficial for the execution of each task.

Furthermore, motivated by the subjectivity of the task at hand and in connection to our research interests, we use an additional auxiliary task in the multi-task configuration. Previous research (Leonardelli et al., 2021; Reidsma and op den Akker, 2008; Jamison and Gurevych, 2015) has showed how in an offensive language detection task, training on data with low levels of inter-annotator agreement can be detrimental for model performance. However, recent research has also showed that this depends on the reason for disagreement, and that it can be helpful to include information about agreement level when training (Sandri et al., 2023). As disaggregated annotations are not available for participants to use, we artificially estimate a measure of agreement by using an ensemble of five classifiers to mimic annotator judgments on train set items. We then use the level of agreement between classifiers as a proxy for inter-annotator agreement, and exploit this information in one of the tasks we train our models on.

Moreover, in the analysis section of the paper,

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1894–1905

we discuss the impact of the auxiliary agreement task on training, by showing a number of additional experiments in which we compare the effect of using synthetic agreement information with that of using actual gold crowd agreement information (released by organizers after competition end) and the effect of removing data with low agreement from training. Furthermore, we discuss impact of agreement on test set, by showing how selecting test data according to different levels of annotators' agreement has a strong effect on classifiers performance.

Our method achieves competitive results on the SemEval-2023 Task 10, ranking well in all the tasks, especially the ones concerning the more fine-grained classifications. With our approach, a single model would suffice for achieving reasonably good performance on all three tasks. However, we used two models for our submission, given that the task was designed so that the fine-grained tasks had smaller test sets.

2 Related work

Currently, the definition of sexism and misogyny is under discussion (see for example Richardson-Self (2018)). One of the most widely accepted definitions of misogyny involves the expression of hostility and hatred towards women. In contrast, sexism comprises any form of oppression or prejudice against women, ranging from more subtle language to overtly hostile (as in the case of misogyny). Research on online misogyny and sexism is relatively recent and in fast development.

One of the first attempts is by Waseem and Hovy (2016). In the context of a broader investigation on online hate speech, they categorised misogyny as a sub-branch of hate speech ("hate towards women") and dedicated a part of their hate-speech dataset to this. Interestingly, authors claim that annotating misogyny would require more than a binary present/absent label.

In the same year, Hewitt et al. (2016) collected the first dataset annotated for misogynistic content. They aimed to highlight the challenges of identifying misogynist abuse, rather than finding examples of abusive tweets and analysed the usage of the keywords such as *cunt*, *slut* and *bitch*, and revealed how 'they have crept into general use as a form of address', highlighting how the problem of defining what constitutes misogynist language remains.

Research on automatic misogyny identification further attracted attention when Anzovino et al. (2018) proposed the first taxonomy of misogynistic language behaviors in social media (five categories: *Discredit*, *Harassment & Threats of Violence*, *Derailing*, *Stereotype & Objectification*, and *Dominance*). In the same year, Anzovino and colleagues organised two shared tasks for the evaluation of systems performing automatic misogyny identification (AMI) in English, Italian, and Spanish (Fersini et al., 2018a,b) with benchmarks annotated on two levels, using the proposed taxonomy. After this first attempt to categorise misogyny into different types, other research works have proposed solutions along the same line, creating new datasets, extending taxonomies and adding new languages (e.g. Rodríguez-Sánchez et al. (2020); Zeinert et al. (2021); Almanea and Poesio (2022)).

3 Data and task description

The Explainable Detection of Online Sexism (EDOS) task is focused on the detection of sexism beyond high level categories. Sexism, defined for the annotators of this corpus 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*" (Kirk et al., 2023), is in fact often identified by automatic systems in a binary fashion, with tools simply flagging online content as sexist or not sexist, providing no further details or explanations. Content moderation based on this kind of tools, however, can often result in poor explainability and interpretability (Kirk et al., 2023). The goal of the EDOS task is that of developing models for detecting sexism in English that are both more accurate and more explainable, thanks to more fine-grained annotations in the training data.

3.1 The "Explainable Detection of Online Sexism" (EDOS) dataset

The organizers of SemEval-2023 Task 10 provide participant teams with the "Explainable Detection of Online Sexism" (EDOS) annotated dataset, a dataset containing social media posts from Gab and Reddit introduced in Kirk et al. (2023).

The EDOS dataset contains 20,000 posts annotated according to a three-level annotation schema, illustrated in Figure 1:

1. Posts are annotated as *sexist* or *not sexist*, following the binary categorization of most sexism detection tools. Sexist posts are the mi-

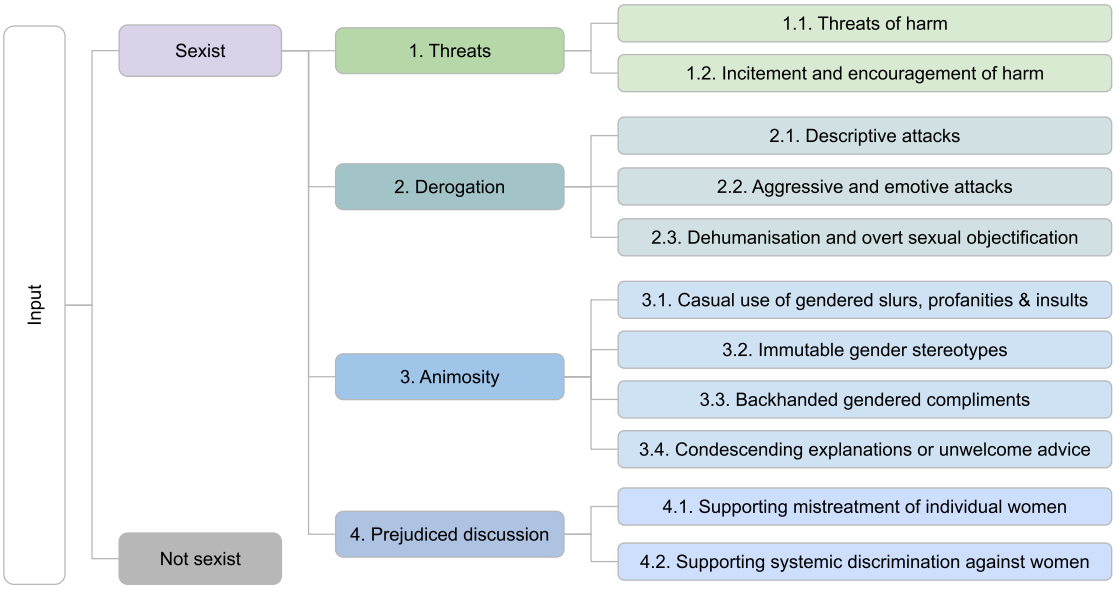

Figure 1: Taxonomy of the EDOS dataset. The figure is adapted from the figure provided by the task organizers (Kirk et al., 2023).

nority, and constitute around 24% of the entire dataset[1]. Sexist posts are then annotated according to the other two levels of annotation, while posts that are not sexist are only assigned the binary label.

2. Posts annotated as sexist are annotated with the *category* of sexism they exhibit. The EDOS dataset is annotated according to 4 possible categories: (1) *threats*, (2) *derogation*, (3) *animosity*, and (4) *prejudiced discussions*. Categories are also distributed unevenly, with *threats* constituting 9% of the posts, *derogation* 47%, *animosity* 34%, and *prejudiced discussions* 10%.

3. Posts labeled as sexist are also annotated with the *vector* used for conveying sexist content. The taxonomy used by EDOS includes 11 possible vectors of sexism, each of them associated with one of the 4 categories in the second level of annotation. The list of vectors can be seen in the right part of Figure 1.

Each item has been annotated by a crowd of three. A more detailed description of the annotation schema, along with definitions and examples, is provided in Appendix B.

3.2 Task setup

The EDOS task is divided into three subtasks, organized hierarchically reflecting the taxonomy of the dataset (Section 3.1).

1. **TASK A - Binary Sexism Detection**: a two-class (or binary) classification where systems have to predict whether a post is sexist or not sexist;

2. **TASK B - Category of Sexism**: for posts which are sexist, a four-class classification where systems have to predict one of four categories: (1) threats, (2) derogation, (3) animosity, (4) prejudiced discussions;

3. **TASK C - Fine-grained Vector of Sexism**: for posts which are sexist, an 11-class classification where systems have to predict one of 11 fine-grained vectors.

For all three tasks, the metric used by the organizers for evaluation is macro-averaged F_1 score, since it gives the same weight to all classes, and all tasks are imbalanced classification tasks.

[1]Abusive language identification datasets in general tend to have imbalanced classes. The average percentage of abusive posts in abusive language detection datasets found in the survey by Vidgen and Derczynski (2021) is 36.7%.

4 Methods

Given the hierarchical nature of the taxonomy of the EDOS task, there are strong correlations among the three subtasks. For instance, identifying the presence of a vector in a post (task C) necessarily implies both identifying a specific category of sexism (task B) and identifying a post as sexist (task A). Because of this, we aim at leveraging the information shared across the annotation levels with multi-task learning, a paradigm that aims at exploiting training signals of related tasks at the same time by using a shared representation in the model (Caruana, 1997).

In addition to exploiting the information shared across the levels of annotation for the three subtasks, we aim at leveraging information on how potentially ambiguous or challenging a post can be. Given that during the participation in the EDOS task we had no information on inter-annotator agreement level or disaggregated annotation for each post, we "artificially" created an *agreement* label for each post using an ensemble of 5 classifiers, which we refer to as *ensemble agreement*. The idea of exploiting classifier agreement follows the finding by Leonardelli et al. (2021) that classifier ensembles can be useful for identifying ambiguous or challenging posts in the context of offensive language detection. We describe the process we followed for assigning ensemble agreement labels in Section 4.1.

4.1 Ensemble agreement

Our goal is that of estimating a level of agreement for each post, which can then be exploited as additional information in our multi-task training setup, in addition to the three EDOS subtasks. In order to do this, we employ an ensemble of classifiers.

To obtain an approximation of agreement level for each post, we divide the available training data X using 5 folds, creating 5 separate 80/20 train/validation splits so that each example in X is in the validation set of one fold. This is necessary for us to have (*i*) an ensemble prediction for each example and (*ii*) enough data for training the ensemble classifier we use to estimate agreement.

More specifically, we divide the training data X into 5 subsets $X_1, X_2, ..., X_5$. We then create 5 folds using these subsets so that each subset is used as validation data once. Then, for each fold, the steps we follow for labeling training instances with *ensemble agreement* are:

1. We train 5 multi-task classifiers $C_1, ..., C_5$ on the training data for the current fold. The details of the classifiers are discussed in Sec. 4.1.1.

2. We use the multi-task classifiers $C_1, ..., C_5$ to predict the annotations of the examples in the validation set of the current fold.

3. Based on the predictions of the classifiers, we assign ensemble agreement labels to the validation set of the current fold based on how many classifiers agree with the actual gold annotation. The ensemble agreement label is thus a number between 0 and 5. This procedure is further explained in Sec. 4.1.2.

In the end, we use 5 classifier ensembles to estimate agreement levels, each constituted by 5 multi-task classifiers in turn.

4.1.1 Multi-task classifiers

The five classifiers we use in each of our ensembles are multi-task, so that they can learn from all three levels of annotation in EDOS at the same time. The rationale is that all annotation levels can improve model performance as they contain useful and (partially) non redundant information.

Specifically, we use a pre-trained model (in our case, RoBERTa base (Liu et al., 2019)) as the shared encoder for all tasks. We fine-tune the encoder using the implementation by van der Goot et al. (2021) of the inverse square root learning rate decay used in Howard and Ruder (2018), while a separate decoder is utilized by each task. Through this method, all tasks benefit from the of mutual signals encoded via a shared representation that is jointly fine-tuned during training.

The tasks we use for our multi-task models correspond to the three annotation levels in the EDOS dataset, as well as the three subtasks of the EDOS shared task, and are weighted equally:

- Binary classification of sexism (corresponding to Task A in the EDOS shared task).

- Category of sexism classification (EDOS Task B), which in this case is remapped to 5 classes (0-4) instead of 4 (1-4), so that the 0 label can be used for examples annotated as *not sexist* in the first level of annotation, which means they have no category of sexism assigned to them. This is necessary in order for the multi-task models to still have an output for each

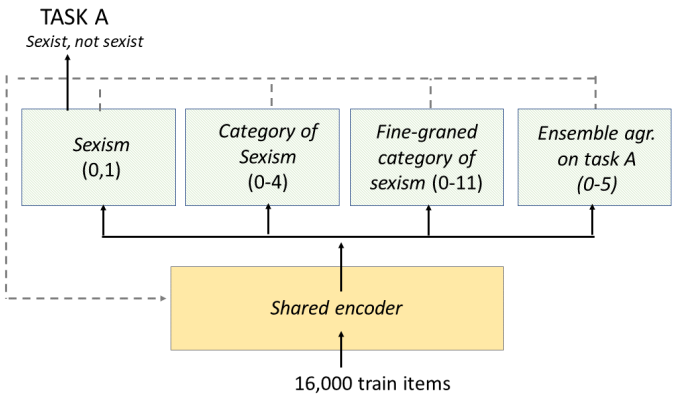

Figure 2: The design of the multitask model used for Task A predictions (MT_TASKA).

task regardless of the predictions for the other tasks.

- Vector of sexism classification (EDOS Task C), which is remapped so as to allow for a "no vector" label ("0"). The labels used for this task are therefore 12 (0-11).

In addition to being able to leverage information across levels of annotation, multi-task models also have the advantage of having multiple outputs. As a result, our multi-task models can predict all three levels of annotation simultaneously.

4.1.2 Agreement estimation

After training 5 multi-task classifiers for each fold, we use them to predict the annotation of the texts in the validation split for the current fold on all three annotation levels, i.e. binary annotation of sexism, category of sexism, and sexism vector.

We then compare the predictions of the 5 classifiers in the current fold's ensemble with the gold labels in the EDOS dataset, and for each annotation level we assign an *ensemble agreement* label corresponding to the number of classifiers in the current ensemble that predicted the correct gold label, between 0 and 5. In practice, this results in the addition of three labels to our data: (*i*) the ensemble agreement for binary classification (task A), (*ii*) the ensemble agreement for category classification (task B), and (*iii*) the ensemble agreement for vector classification (task C).

The estimations of agreement level we obtain using the ensemble models are then used as auxiliary tasks in the models we use for our final submissions, in addition to the multi-task framework described in Sec. 4.1.1. We describe our models in Section 4.2.

4.2 Models

We participated in each of the three tasks proposed within the EDOS challenge. Using the multi-task paradigm, where all subtasks are simultaneously solved, a single model could suffice to have a predictions for all the subtasks. However, the EDOS task was organized so that in the evaluation phase first the test set for Task A was released. Given that the Tasks B and C depend on the binary annotation of sexism - as category and vector annotations are only available for posts annotated as *sexist* - a subset of the Task A test set with only *sexist* posts was released at a later date and constituted the test set for Tasks B and C. As a result, using only one model to solve all the three tasks would result in error propagation from Task A to Tasks B and C, especially with regards to false negatives, i.e. posts that were incorrectly identified as *not sexist* in Task A, which would have no Task B or Task C predictions. To solve this issue, we build two separate models, one for Task A and one for Tasks B and C.

4.2.1 Task A model

For our Task A submission, we use a multi-task setup similar to that of the classifiers in the ensembles (Section 4.1.1), with the additional task of predicting *ensemble agreement* on Task A. The model, from now on MT_TASKA, is therefore trained on one main task and three auxiliary tasks, as illustrated in Figure 2:

- Binary sexism classification (Task A, 2 classes), which we use as the main task, mean-

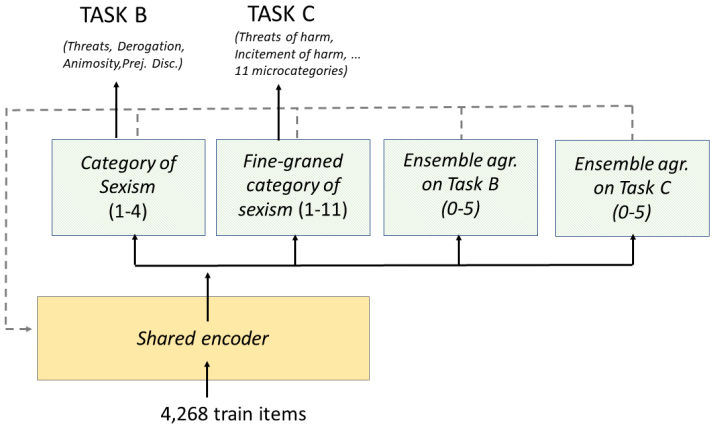

Figure 3: The design of the multitask model used for Task B and Task C predictions (MT_TASKBC).

ing that the predictions of this task are the output we consider for our predictions,

- Sexism category classification (Task B, 4 classes + 1), remapped so that there is an additional 0 label for posts that are considered *not sexist*,

- Sexism vector classification (Task C, 11 classes + 1), remapped again so that there is an additional 0 label for posts that are considered *not sexist*,

- *Ensemble agreement* prediction of binary sexism labels from Task A (6 classes).

4.2.2 Tasks B and C model

The MT_TASKBC model, which we use for our Task B and Task C submissions, is built similarly to the MT_TASKA model. Given that for tasks B and C the test set is composed exclusively of sexist posts, Task A information (binary classification of sexism) is unnecessary. We therefore discard Task A information for this model, using the sexism category (Task B) and sexism vector (Task C) annotations along with their respective *ensemble agreement* measures. This results in four tasks for our multi-task setup, illustrated in Figure 3:

- Sexism category classification (Task B, 4 classes), main task whose output is used for our Task B predictions,

- Sexism vector classification (Task C, 11 classes), main task whose output is used for our Task C predictions,

- *Ensemble agreement* prediction of sexism categories from Task B (6 classes),

- *Ensemble agreement* prediction of sexism vectors from Task C (6 classes).

5 Experiments

In this section, we first outline the experimental setup, then we present the results of our models on the three tasks. We finally show some additional analyses and discussion.

5.1 Experimental Setup

We employ the MaChAmp v0.2 toolkit (van der Goot et al., 2021) and RoBERTa-base (Liu et al., 2019) as our shared encoder. All tasks are addressed as classification tasks (with different number of classes according to tasks). We fine-tune every model (around 110M trainable parameters) on one GPU[2] for 10 epochs using default MaChAmp hyperparameter values (see Table 4 in Appendix). In addition, during training we let each class receive equal weight so that minority classes are not underrepresented, and we also introduce loss weights. The multi-task learning loss is computed as $L = \sum_t \lambda_t L_t$, where L_t is the loss for task t and λ_t the corresponding weighting parameter, and we provide a different loss weight for the auxiliary tasks. For model MT_TASKA, we empirically set $\lambda_t = 1$ for the main task, and $\lambda_t = 0.5$ for the other tasks. For model MT_TASKBC, we empirically set $\lambda_t = 0.7$ for the two main tasks, and $\lambda_t = 0.25$ for the two auxiliary tasks.

[2]NVIDIA Titan Xp

We only use the (labelled) data supplied by the organizers and split 90% of it into a training set and 10% into a development set. In order to maximize the effectiveness of our method, we generate 5 models starting from 5 pseudo-random seeds, and label the test according to the majority (3/5). As for multi-class tasks (B and C), a 3/5 majority is not always present. In this case, we randomly select among the prediction of the 5 models.

5.2 Results

The official results for our submissions for Tasks A, B and C are shown in Table 1. We report macro-averaged F_1 score and overall rank of our systems, as well as those of the best performing team for comparison.

Task A As shown in Table 1, the predictions of our MT_TASKA model obtain a macro F_1 score of 0.8402, and rank 32nd out of 84 submissions, in the highest half of the leaderboard. The difference with the best system is relatively little in terms of F_1 (Δ 0.036), but the rank achieved is quite low. This is likely caused by the high number of teams that participated in this task and that scored very close to one another.

Task B Our MT_TASKBC model ranked 24th out of 69 participating teams on the official leaderboard, scoring an F_1 of 0.7326. With respect to Task A, we ranked higher, though now the score difference from the best system is greater and reaches almost 10 points.

Task C Our MT_TASKBC model ranked 11th out of 63 participating teams on the official leaderboard, reaching an F_1 of 0.5606. This is the task where we ranked best among the submitted tasks.

As mentioned in Section 5.1, in order to enhance our competitiveness, for each of the two configurations used to solve the tasks, we run 5 restarts and submit predictions as the majority aggregation. For Task A, the average F_1 of the 5 models is 0.8355, thus the usage of five restarts advanced our performance on average 0.005, corresponding (hypothetically) to 10 positions on the leaderboard. For task B, the mean result of the single models is 62.792±0.601 (F_1±std), which gives an advantage on average of 4 positions on the leaderboard. For task C, the result of the single models is 46.484±0.611 (F_1±std), which gives an advantage on average of 11 positions on the leaderboard.

TASK A	**F_1**	**rank**
Task Best System	0.8746	1/84
MT_TASKA	0.8402	32/84
TASK B	**F_1**	**rank**
Task Best System	0.7326	1/69
MT_TASKBC	0.6385	24/69
TASK C	**F_1**	**rank**
Task Best System	0.5606	1/63
MT_TASKBC	0.4935	11/63

Table 1: Official test set results of our models compared to the system that best performed in the shared task

5.3 Analysis: the role of agreement

In this section, focusing on Task A, we explore the correlation between ensemble agreement and crowd agreement and the role of agreement in both training and testing. Besides the 5-classifier ensemble agreement used in our multi-task system, in this section we also consider crowd-annotations, released by organizers after the competition ended. As the crowd of annotators is composed by three, it maps into two levels of agreement: full agreement (3/3) and disagreement (2/3). To statistically compare the results of sections 5.3.2 and 5.3.3, we run 12 reruns and report the average and standard deviation of the predictions.

5.3.1 Synthetic and crowd agreement correlation

To evaluate the validity of approximating annotator agreement with a classifier ensemble, we computed Pearson's correlation coefficient between the agreement of the classifiers and that of annotators (for ensemble agreement, values of 4 and 5 for classifier agreement were considered agreement, while lower values disagreement). A similarity between the two patterns of agreement is detected as they present a moderate correlation (r = 0.33). However, the 5 classifiers are trained on identical experimental setups, just varying the composition of the training set from the same dataset. Introducing more variety in the classifiers constituting the ensemble could increase ability of the ensemble to represent disagreements.

5.3.2 Role of agreement in train set

Table 2 shows the additional experiments we conduct to examine the role of agreement in training data, especially in function of the type and level of agreement (crowd-annotators or artificial

Exp.	Config.	Aux. agr. task on	Training on	Tr. size	F_1±std
1.	MT_taskA	class. ens.-agr.	all	16,000	**0.836*** ±0.003
2.	MT_taskA	class. ens.-agr.	class. ens.-agr. > 1	14,527	**0.836*** ±0.002
3.	MT_taskA	crowd-agr.	all	16,000	**0.837*** ±0.003
4.	MT_taskA	-	all	16,000	0.831 ±0.004
5.	taskA only	-	all	16,000	0.826 ±0.007

Table 2: Classifier performance (F_1) with different subsets of the training set, and on different types of agreement (artificial, crowd). Statistically significant results (compared to the lowest F_1 among MT config.) marked with ($*$)

5-classifier ensemble). For this set of experiments, we follow the same experimental setup we used for Task A predictions, except for Exp.4, where we omit the auxiliary task on agreement and Exp.5, where we trained the system only on the principal task (Task A) without any multi-task. For each experiment, we present F_1 scores averaged across 12 reruns and the standard deviation.

Exp.1 configuration reproduces exactly the submitted system for Task A (see Sec. 4.2.1 and 5). Exp. 2 also uses the same configuration, but it is trained using a smaller train set, i.e.data with low artificial agreement are removed from the training (all items that are correctly classified by less than 2 of the 5-classifier ensemble). This is inspired by results in Leonardelli et al. (2021), showing how training on data with low agreement is not beneficial. Indeed in Exp.2, even by removing up to almost 10% of the training data (with low agreement), performance is not decreasing. In Exp.3, the artificial agreement used for the auxiliary task on agreement is replaced by the crowd-agreement values (agreement/disagreement). Results show how indeed crowd-agreement and artificial agreement can be both used for this task. Moreover in Exp.4, to assess the impact of agreement task from the multitask configuration, the task on agreement is removed. Indeed, this setup scores the lowest F_1 among the results of Table 2. Finally, Exp.5 shows how the performance for Task A, when removing multi-task learning, decreases.

To reliably assess the differences in performance between the models, we run significance statistics.[3] Exp. 1, 2 and 3 all take agreement into consideration, either by removing data with low agreement or by using an the agreement auxiliary task on crowd or artificial agreement. Interestingly these three tasks do not statistically differ among them, but all of them statistically differ from Exp. 4, where agreement is not considered.

[3]We implement Almost Stochastic Order (Dror et al., 2019; Del Barrio et al., 2018) as implemented by Ulmer et al. (2022). For the 4 experiments utilizing multi-task paradigm, we compare the models' scores across the 12 restarts and consider a threshold of $\tau = 0.2$ to measure statistical significance. This threshold is equivalent to a Type I error rate of p-value .05 based on Ulmer et al. (2022).

5.3.3 Role of agreement in test set

We assess the variations in performance based on the agreement of crowd-annotators on evaluating the test set. To this end, we divide our test set into two subsets based on agreement level and calculate F_1 separately for each one. The results of Table 3 suggests a drastic reduction in classification performance when handling low crowd-agreement data, showcasing that ambiguous data are the most difficult to classify. This finding replicates findings reported by Leonardelli et al. (2021), which observed similar outcomes on offensive language detection task, in their dataset and also and is systems submitted for a popular hate speech benchmark. Interestingly, when separating the offensive and not offensive classes, 3 shows how performance is lower for the offensive one. Though, the drop in performance when classifying data not unanimously annotated is present for both classes. However, for the offensive one, the drop in performance when data present disagreement is worse, as F_1 diminishes and reaches the worst performance overall, indicating these are the most difficult cases to classify.

Test on	Test size	F_1	F_1	
			Not off.	Off.
crowd-agr. 3/3	3,115	0.896	0.96	0.827
crowd-agr. 2/3	885	0.676	0.796	0.55

Table 3: Performance of submitted model for Task A on subsets with different crowd-agreement level. We report the average macro-F_1 obtained from 12 restarts for the overall results

6 Conclusion

In this paper, we described the systems submitted for EDOS SemEval-2023 Task 10. We utilize

a multi-task learning approach which allows our models to share representations across all three tasks of EDOS. Additionally, due to the subjective nature of this task, we include inter-annotator agreement information into our multi-task framework. As disaggregated annotations are not available, we artificially estimate them with a 5-classifier ensemble. Our system achieved competitive results on the official leaderboard ranking in upper half in all three tasks, without relying on any external data.

Moreover, we conducted a thorough analysis on the role and impact of agreement on Task A (Table 2). We showed how, while using an ensemble to construct artificial agreement for training models can be a complex and expensive task, it can be replaced by the crowd-annotations agreement if available, making the method more viable. We showed how taking agreement into consideration during the training phase can be done in different ways while still being beneficial. Furthermore, by considering F_1s relatively to test data with different crowd-agreement levels, we show how the performance of our system drastically decreases when evaluating more "difficult" items, i.e. items where annotators disagreed, for which labels thus reflect only a partial consensus towards one or the other judgments.

We hypothesise that agreement could play an important role on classification of finer-grained labels for sexism (Task B and C), given that for higher levels of accuracy (and higher number of options available for annotation), more personal sensitivity comes into play. We leave this aspect for the near future investigation. Finally, data for which annotators failed to reach unanimous consensus are around 22% of the current dataset, and represent an important challenge in terms of classification performance, in part due to the fact that they represent subjective points of view in judgements that are legitimate in this type of tasks. Our hope is that our work can contribute to the debate about the importance of having different points of view, rather than disregarding them, when compiling datasets and designing classifiers.

Limitations

The current investigation only applies to English and needs to be expanded to other languages, with the option of adapting the used taxonomy to fit different languages. Furthermore, the current data does not take into account more subtle forms of sexism, which can be overt or implicit. Additionally, the investigation into the role of agreement that has been conducted for Task A is missing for tasks B and C and requires further exploration. Finally, while our method for estimating agreement appears to reliably estimate actual crowd agreement levels, it could not work equally well in circumstances different from ours, so the use of artificial agreement measures should be investigated further.

Acknowledgements

Part of this work was funded by the StandByMe European project (REC-RDAP-GBV-AG-2020) on "Stop online violence against women and girls by changing attitudes and behaviour of young people through human rights education" (GA 101005641). This research was also supported by the StandByMe 2.0 project (CERV-2021-DAPHNE) on "Stop gender-based violence by addressing masculinities and changing behaviour of young people through human rights education" (GA 101049386).

References

Dina Almanea and Massimo Poesio. 2022. Armis-the arabic misogyny and sexism corpus with annotator subjective disagreements. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2282–2291.

Maria Anzovino, Elisabetta Fersini, and Paolo Rosso. 2018. Automatic identification and classification of misogynistic language on twitter. In *Natural Language Processing and Information Systems: 23rd International Conference on Applications of Natural Language to Information Systems, NLDB 2018, Paris, France, June 13-15, 2018, Proceedings 23*, pages 57–64. Springer.

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

Eustasio Del Barrio, Juan A Cuesta-Albertos, and Carlos Matrán. 2018. An optimal transportation approach for assessing almost stochastic order. In *The Mathematics of the Uncertain*, pages 33–44. Springer.

Rotem Dror, Segev Shlomov, and Roi Reichart. 2019. Deep dominance - how to properly compare deep neural models. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2773–2785. Association for Computational Linguistics.

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, Paolo Rosso, and Maria Anzovino. 2018b. Overview of the task on automatic misogyny identification at ibereval 2018. *Ibereval@ sepln*, 2150:214–228.

Sarah Hewitt, Thanassis Tiropanis, and Christian Bokhove. 2016. The problem of identifying misogynist language on twitter (and other online social spaces). In *Proceedings of the 8th ACM Conference on Web Science*, pages 333–335.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia. Association for Computational Linguistics.

Emily Jamison and Iryna Gurevych. 2015. Noise or additional information? Leveraging crowdsource annotation item agreement for natural language tasks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 291–297.

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*. Association for Computational Linguistics.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic. 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. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Dennis Reidsma and Rieks op den Akker. 2008. Exploiting 'subjective' annotations. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pages 8–16, Manchester, UK. Coling 2008 Organizing Committee.

Louise Richardson-Self. 2018. Woman-hating: On misogyny, sexism, and hate speech. *Hypatia*, 33(2):256–272.

Francisco Rodríguez-Sánchez, Jorge Carrillo-de Albornoz, and Laura Plaza. 2020. Automatic classification of sexism in social networks: An empirical study on twitter data. *IEEE Access*, 8:219563–219576.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. Why don't you do it right? analysing annotators' disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2420–2433, Dubrovnik, Croatia. Association for Computational Linguistics.

Dennis Ulmer, Christian Hardmeier, and Jes Frellsen. 2022. deep-significance: Easy and meaningful signifcance testing in the age of neural networks. In *ML Evaluation Standards Workshop at the Tenth International Conference on Learning Representations*.

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

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.

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.

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.

A Hyperparameters

Hyperparameter	Value
Optimizer	AdamW
β_1, β_2	0.9, 0.99
Dropout	0.3
Epochs	10
Batch size	32
Learning rate (LR)	0.0001
LR scheduler	Slanted triangular
Decay factor	0.38
Cut fraction	0.2

Table 4: Default MaChAmp hyperparameter settings (van der Goot et al., 2021) used for all our experiments.

B Label definitions and examples

The descriptions of the vectors and relative examples provided by the task organizers are presented in table 5.

Category	Vector	Definition	Example
1. Threats, plans to harm and incitement	1.1 Threats of harm	Expressing intent, willingness or desire to harm a woman, group of women or individual women. This could include, but is not limited to: physical, sexual, emotional or privacy-based forms of harm.	I'll kill any women that talks back to me
	1.2 Incitement and encouragement of harm	Inciting or encouraging an individual, group, or general audience to harm a woman or group of women. It includes language where the author seeks to rationalise and/or justify harming women to another person.	Raping her would put her in her place
2. Derogation	2.1 Descriptive attacks	Characterising or describing women in a derogatory manner. This could include, but not limited to: negative generalisations about women's abilities, appearance, sexual behaviour, intellect, character, or morals.	Women's football is so shit, they're so slow and clumsy
	2.2 Aggressive and emotive attacks	Expressing strong negative sentiment against women, such as dislike, disgust, or hatred. This can be through direct description of the speaker's subjective emotions, baseless accusations, or the use of gendered slurs, gender-based profanities and gender-based insults.	I hate women
	2.3 Dehumanising attacks and overt sexual objectification	Derogating women by comparing them to non-human entities such as vermin, disease or refuse, or overtly reducing them to sexual objects.	Women are pigs
3. Animosity	3.1 Causal use of gendered slurs, profanities and insults	Using gendered slurs, gender-based profanities and insults, but not to intentionally attack women. Only terms that traditionally describe women are in scope (e.g. 'bitch', 'slut').	Stop being such a little bitch
	3.2 Immutable gender differences and gender stereotypes	Asserting immutable, natural or otherwise essential differences between men and women. In some cases, this could be in the form of using women's traits to attack men. Most sexist jokes will fall into this category.	Men and women's brains are wired different bro, that's just how it is
	3.3 Backhanded gendered compliments	Ostensibly complimenting women, but actually belittling or implying their inferiority. This could include, but is not limited to: reduction of women's value to their attractiveness or sexual desirability, implication that women are innately frail, helpless or weak.	Women are delicate flowers who need to be cherished
	3.4 Condescending explanations or unwelcome advice	Offering unsolicited or patronising advice to women on topics and issues they know more about (known as 'mansplaining')	My gf always complains about period pains but she just doesn't understand the medical science for eliminating them!
4. Prejudiced Discussions	4.1 Supporting mistreatment of individual women	Expressing support for mistreatment of women as individuals. Support can be shown by denying, understating, or seeking to justify such mistreatment.	Women shouldnt show that much skin, it's their own fault if they get raped
	4.2 Supporting systemic discrimination against women as a group	Expressing support for systemic discrimination of women as a group. Support can be shown by denying, understating, or seeking to justify such discrimination.	The leadership of men in boardrooms is a necessary evil—corporations need to be efficiently run

Table 5: Detailed definitions of vectors and examples by Kirk et al. (2023).