DH-FBK at HODI: Multi-Task Learning with Classifier Ensemble Agreement, Oversampling and Synthetic Data

Elisa Leonardelli[1], Camilla Casula[1,2]

[1]Fondazione Bruno Kessler, Trento, Italy
[2]University of Trento, Trento, Italy

Abstract
We describe the systems submitted by the DH-FBK team to the HODI shared task, dealing with Homotransphobia detection in Italian tweets (Subtask A) and prediction of the textual spans carrying the homotransphobic content (Explainability - Subtask B). We adopt a multi-task approach, developing a model able to solve both tasks at once and learn from different types of information. In our architecture, we fine-tuned an Italian BERT-model for detecting homotransphobic content as a classification task and, simultaneously, for locating the homotransphobic spans as a sequence labeling task. We also took into account the subjective nature of the task by artificially estimating the level of agreement among the annotators using a 5-classifier ensemble and incorporating this information in the multi-task setup. Moreover, we experimented by extending the initial training data with oversampling (Run 1) and via generation of synthetic data (Run2). Our runs achieve competitive results in both tasks. Finally, we conducted a series of additional experiments and a qualitative error analysis.

Keywords
Multi-task learning, data augmentation, agreement, subjective tasks

***Warning**: This paper contains examples of potentially offensive content.*[1]

1. Introduction

In recent years, social media use has increased globally, with platforms enabling users to post, share and comment about any topic at any time. With the increase of online communication, proliferation of online hateful comments has become a major problem. Natural language processing (NLP) research is essential for the mitigation of online hate speech, as it can help in understanding the phenomenon and assisting in automating the process at a large scale.

The NLP community has been tackling this problem through the creation of datasets and models, especially focusing on some of the most vulnerable communities, such as migrants [2] or women [3]. The application of automatic methods for detecting hate speech targeting LGBTQIA+ people specifically is a recent development, having been addressed for the first time in English and Tamil [4] and more recentely in Locatelli et al. [5].

The evaluation task for Homotransphobia Detection in Italian (HODI) [6] proposed at Evalita 2023 [7], aims to explore homotransphobia on Twitter in Italian, taking a deeper look into an issue that has not been adequately addressed in either the global or Italian NLP communities. To this end, the task organizers released a dataset of 6,000 tweets annotated for homophobic and transphobic content (Subtask A) and highlighting the span range expressing it within the sentence (Subtask B), encouraging the developing of models able to detect homotransphobic content in an accurate and explainable way.

In this paper, we present the submitted systems by the DH-FBK team for the two HODI subtasks. Based on the hypothesis that the two layers of annotations provided are highly correlated and thus knowledge sharing will help with the completion of each task, we implemented a multi-task architecture, similarily to the ones proposed in Ramponi and Leonardelli [8] and Leonardelli and Casula [9]. This setup allows leveraging training signals of related tasks at the same time by exploiting a shared representation in the model. Specifically, we simultaneously train a model on the two HODI subtasks, addressing Subtask A as a classification task, and the extraction of the spans containing homotransphobic language (Subtask B) as a Sequential Labeling (SL) problem, locating the spans by BIO tags [10]. Importantly, this multi-task approach allows to develop a unique model for addressing both tasks, and we are one of the two teams who participated in both tasks. Moreover, given the subjectivity of the task, we add an auxiliary task to the multi-task configuration to incorporate information related to annotator agreement. Previous studies have shown that training on data with low agreement between annotators can lead to a decrease in model performance [11]. However, more recent research has shown that this depends on the source of the disagreement and that the level of agreement should still be taken into account when training

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
eleonardelli@fbk.eu (E. Leonardelli); ccasula@fbk.eu (C. Casula)

[1] Profanities have been obfuscated with PrOf (https://github.com/dnozza/profanity-obfuscation) [1]

[12]. Since disaggregated annotations are not accessible to participants in the task, we estimate agreement levels through the use of an ensemble of 5 classifiers, to imitate annotator judgments, similarly to the work conducted in Leonardelli and Casula [9]. Additionally, following the organizers' suggestion to increase the train size of the data, we experimented with different methods for augmenting the training size, i.e. oversampling [13] and data generation [14].

Our best performing run (Run 1) achieved competitive results, ranking 4th for Subtask A and 2nd for Subtask B.

Finally, we discuss the impact of the different elements we combined in our models by conducting a series of additional experiments in Section 5.2, showing the benefits of augmenting training data, especially using oversampling, and showing the relative beneficial impact of the auxiliary task on agreement, which is effective only in combination with oversampling and not with the synthetic data. We then also conduct a qualitative analysis to discover the most difficult cases.

2. The HODI dataset

The HODI dataset is composed of 6,000 Italian tweets. The tweets have been collected from the 1st of May 2022 to the 31st of August 2022 using a set of 21 keywords associated with language that might potentially target minority groups victims of homotransphobia. Entries are annotated following a two-layer scheme:

1. Homotransphobia detection: a tweet contains homotransphobic language or not (binary).
2. Rationales detection (explainability): when a tweet is considered homotransphobic, the span of text that contains the homotranshphobic part is highlighted (list of character positions).

3. Task description

The organizers provided participants with the HODI dataset, described in Section 2. 5,000 annotated tweets were released during the first phase of the competition, out of which 2,008 labeled as homotransphobic. In a second phase, the remaining 1,000 tweets were released unlabeled as test data. The task is divided into two subtasks, reflecting the layers of annotations of the dataset:

- Subtask A - Homotransphobia detection: binary classification, the goal is to predict whether a tweet is homotransphobic or not.
- Subtask B - Explainability: participants are required to predict the spans of homotransphobic tweet that were responsible for the homotransphobic label of the tweet.

The metric used for evaluation of Subtask A is macro-F_1, while character-based F_1 is used for evaluating Subtask B, similarly to Pavlopoulos et al. [15].

4. Methods

4.1. Multi-task setup

To exploit the strong correlations between the annotations of Subtasks A and B, we used a multi-task learning setup [16], showed in Figure 1. Our model is trained simultaneously on tasks relative to both levels of annotation and, by utilizing a shared representation, all the available information is available to the model. Moreover, the tasks under scrutiny are highly subjective. For instance, we observed some inconsistencies across sentences (for example articles being included/excluded in spans). To leverage the uncertainty around words that are potentially ambiguous, and given that no information about agreement among annotators was released, we 'artificially' created an *agreement* label by using the agreement of an ensemble of 5 classifiers. This procedure is described in more detail in Appendix A. In summary, we use three tasks for our multi-task model: two main tasks corresponding to the two annotation levels (and subtasks) of HODI, and an additional auxiliary task relative to synthetic agreement. The three tasks can be summarized as:

- Homotransphobia (Subtask A): binary classification of homotransphobia.
- Explainability (Subtask B): annotations for this task are released at character level. We convert each sentence from character to word-level annotation, and associate each word to a label for whether it belongs to the homotransphobic span (Explainibility). Moreover, since often spans are comprised of entire phrases, annotations followed sequence labelling, using a BIO tagging scheme [17] in which each word can be at the beginning, inside or outside of a span. After converting the data into this format, Subtask B can be carried out as a sequence labelling prediction task.
- Agreement on Subtask B: it is addressed as sequence labelling at word-level. It can assume values between [0-5], reflecting how many of the 5-classifier ensemble, described in Appendix A, predict a specific word in agreement with the gold label of Subtask B.

4.2. Synthetic Data

The use of synthetic data has been proposed as a method to increase the amount of available training data for hate speech and offensive language detection tasks, especially

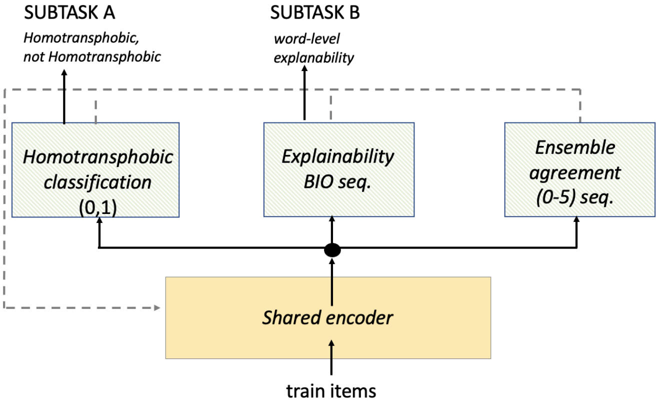
SUBTASK A
Homotransphobic, not Homotransphobic
SUBTASK B
word-level explanability
Homotransphobic classification (0,1)
Explainability BIO seq.
Ensemble agreement (0-5) seq.
Shared encoder
train items

Figure 1: The multi-task configuration of the model we used for Subtask A and B predictions.

when relying on machine-generated data [18, 19]. Although data augmentation using generative models has been found to not always be reliable in improving models [14], we aim at exploring whether it can help the performance of our models for the HODI task.

A widely used method of generating synthetic data consists in fine-tuning a generative model on annotated data and then using it for generating new sequences. These generated sequences are then passed through a classifier in order to confirm the label assignment made by the generator, since generative models are not always reliable in their label assignment [20].

While the majority of works that exploit model-generated data for the detection of offensive language have no particular focus on any target category or phenomenon, our experiments are focused on specifically detecting homotransphobia. Because of this, the generated texts should be correct regarding both the label and the focus on the phenomenon. In part due to this, and in part to the limited availability of generative large language models for Italian, we decide to generate new data using an encoder-decoder model trained on Italian, IT5 [21], in its 738M-parameter (`large`) configuration. The details of our data augmentation process can be found in Appendix B.

Given that the augmentation process provides us with synthetic examples annotated for Subtask A (Homotransphobia detection), but not for Subtask B (detection of rationales), we additionally estimate Subtask B labels for the generated instances, using the model of the first submitted run (generated data were used only in run2), while the agreement for the auxiliary task was estimated using the ensemble classifier described in Appendix A.

4.3. Experimental Setup

The models developed for the two runs submitted by our team, are both based on a pre-trained Italian BERT [2]. For fine-tuning the models in the multi-task setup described in Figure 1, we employed the MaChAmp v0.2 toolkit [22], a tool that supports a variety of standard NLP tasks out-of-the-box, also in a multi-task setup. We employed the pre-trained BERT as our shared encoder for all tasks, while a separate decoder is utilized by each task. We fine-tuned the model (110M parameters) for 15 epochs on a single GPU [3], using default MaChAmp hyperparameters[4]. For the training process, we assign each class equal weight to guarantee minority classes are not underrepresented. We introduced also loss weights for the multi-task learning loss, calculated as $L = \sum_t \lambda_t L_t$, where L_t accounts for the loss for task t and λ_t being the respective weighting parameter. We set $\lambda_t = 0.8$ for the primary tasks, and $\lambda_t = 0.5$ for the auxiliary tasks.

4.4. Submitted Systems Description

For the competition, we submitted two different runs with predictions by models created using the same setup described in Section 4.3 and Figure 1, but trained on different sets of data. Starting from the suggestion from the organizers to augment the size of the training set, we experimented with oversampling and data generation in the following way:

- **Run 1:** the data made available from the organizers are oversampled by repeating them twice.

[2] dbmdz/bert-base-italian-cased

[3] NVIDIA Titan Xp

[4] Defaults MaChAmp hyperparameter settings used for all our experiments; Optimizer: AdamW; β_1, β_2 = 0.9; Dropout = 0.2; Batch Size = 32; Learning rate = 0.0001

- **Run 2:** In addition to oversampling the HODI data, similarly to run1, we add 4,000 synthethically generated examples (see Section 4.2).

We split HODI data into 90% training set and 10% development set. For Run2, the synthetic examples were added to the training set.

5. Evaluation

5.1. Results

Table 1 shows the official results of our submissions for Subtasks A and B. All runs for both tasks beat the organizers' baseline.

For Subtask A we report macro-averaged F_1 score and overall rank of our runs, as well as those of the teams who performed better than us and the baseline. Our best performance (Run 1) obtained a macro F_1 score of 0.795, ranking 4th out of 18 submitted runs (3rd out of 8 teams), while run 2 ranked 7th out of 18 submissions (4th out of 8 teams).

For Subtask B we report the overall ranking, given that the leaderboard is short and only another team participated in Subtask B. One run of the other team participating in this task beat our result, while our best scoring run (Run 1) ranked 2nd.

Table 1
Overview of the results

Subtask A	Run	F1	Team
1	Run 3	0.8108	metzi
2	Run 2	0.8	metzi
3	Run 1	0.7959	odang4hodi
4	**Run 1**	**0.7950**	**DH-FBK**
5	Run 2	0.7942	extremITA
6	Run 2	0.7920	odang4hodi
7	**Run 2**	**0.7837**	**DH-FBK**
		...	
13	-	0.6691	baseline_a
19		...	
Subtask B			
1	Run 2	0.723	extremITA
2	**Run 1**	**0.705**	**DH-FBK**
3	**Run 2**	**0.701**	**DH-FBK**
4	Run 1	0.66	extremITA
5	-	0.205	baseline_b

5.2. Additional experiments

Regarding the impact of generated data, when adding the synthetic data in the training (Run 1) performance decreases in both tasks, showing that the augmentation with generated data does not improve the generalization of models compared to oversampling. In fact, we hypothesize that the addition of synthetic data might push models to be over-reliant on specific identity terms or profanities, hurting its generalization capabilities, a phenomenon that has been observed in data augmentation using generative models [14]. Moreover, to dissect the impact of oversampling and the impact of the auxiliary task, we run a series of additional experiments[5]. Results are shown in Table 2. To evaluate the role of oversampling we replicate the setup of the two submitted runs but omitting the oversampling of the HODI data from the training (Exp 1 and Exp 2). By comparing results (Run 1 vs Exp 1 and Run 2 vs Exp 2), we can observe how oversampling the data is generally beneficial, especially if no synthetic data are used. Moreover, in Exp 3 and Exp 4 we replicate the submitted experiments but exclude the auxiliary task. By comparing Run1 and Exp 3, we can observe how in this case the task is indeed beneficial, while it is not when comparing Run 2 and Exp 4, where synthetic examples are part of the training data. This suggests that the estimation of agreement for generated data might not be informative.

5.3. Qualitative Error Analysis - Subtask A

To perform a qualitative analysis on the most problematic tweets, we isolated the tweets that were incorrectly classified by all models in Table 2. The most consistent false negative regards the missed detection of tweets containing a specific offensive slang word (*f*mminiello*). One possible reason is that this word is not generally common (as it belongs to a local language variety), and it was not present in the training set. Observing the posts incorrectly classified as homotransphobic by the models, we identified (doubtful) sense of humour or metaphorical expressions (*andare a fare in culo, essere fr*cio col culo degli altri*) as possible reasons. Another possible reason could also be over-reliance on specific terms.

6. Conclusions

We described our participation to the HODI evaluation task at Evalita 2023. We used a multi-task learning approach to share representations between the two tasks involved and, additionally, considering the subjectivity of the task, we incorporated inter-annotator agreement information into our framework, estimating them with a 5-classifier ensemble. We experimented augmentation of the training data available by oversampling them and via generated data. We were one of the few teams who participated in both tasks, and our systems performed competitively.

Moreover, we conducted an analysis on the role and impact of the various aspects we combined. Our results show oversampling is generally beneficial, especially

[5] The organizers released the labels for the test set after the closing of the evaluation phase

Table 2
Classifier performance by varying the subsets of the training set and the auxiliary task

Exp.	Multi-task setup		Training		F_1 subtasks	
	main	auxiliary	on	size	A	B
Run 1 (sub)	subtask A & B	ens.-agr.	2 x data	10,000	**0.795**	**0.705**
Run 2 (sub)	subtask A & B	ens.-agr.	2 x data + synth. data	14,000	0.784	0.701
Exp 1 (post)	subtask A & B	ens.-agr.	data	5,000	0.787	0.692
Exp 2 (post)	subtask A & B	ens.-agr.	data + synth. data	9,000	0.789	0.691
Exp 3 (post)	subtask A & B	-	2 x data	10,000	0.785	0.694
Exp 4 (post)	subtask A & B	-	2 x data + synth. data	14,000	0.788	0.698

when combined with the auxiliary task on agreement. The usage of generated data instead has limited benefits, compared to oversampling or additional auxiliary tasks. Finally, performing a qualitative analysis on the most frequent causes of error, we identified specific homotransphobic slang terms that were problematic to be identified by our models.

Acknowledgments

The work of Elisa Leonardelli was partially 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). Her 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

[1] D. Nozza, D. Hovy, The state of profanity obfuscation in natural language processing scientific publications, in: Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, 2023.

[2] T. Bourgeade, A. T. Cignarella, S. Frenda, M. Laurent, W. Schmeisser-Nieto, F. Benamara, C. Bosco, V. Moriceau, V. Patti, M. Taulé, A multilingual dataset of racial stereotypes in social media conversational threads, in: Findings of the Association for Computational Linguistics: EACL 2023, 2023, pp. 674–684.

[3] E. Fersini, P. Rosso, M. Anzovino, Overview of the task on automatic misogyny identification at ibereval 2018., Ibereval@ sepln 2150 (2018) 214–228.

[4] B. R. Chakravarthi, R. Priyadharshini, R. Ponnusamy, P. K. Kumaresan, K. Sampath, D. Thenmozhi, S. Thangasamy, R. Nallathambi, J. P. McCrae, Dataset for identification of homophobia and transophobia in multilingual youtube comments, arXiv preprint arXiv:2109.00227 (2021).

[5] D. Locatelli, G. Damo, D. Nozza, A cross-lingual study of homotransphobia on twitter, in: Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP), 2023, pp. 16–24.

[6] D. Nozza, A. T. Cignarella, G. Damo, T. Caselli, V. Patti, HODI at EVALITA 2023: Overview of the Homotransphobia Detection in Italian Task, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[7] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[8] A. Ramponi, E. Leonardelli, Dh-fbk at semeval-2022 task 4: leveraging annotators' disagreement and multiple data views for patronizing language detection, in: Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022), 2022, pp. 324–334.

[9] E. Leonardelli, C. Casula, Dh-fbk at semeval-2023 task 10: Multi-task learning with classifier ensemble agreement for sexism detection, in: Proceedings of the 17th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 1894–1905. URL: https://aclanthology.org/2023.semeval-1.261.

[10] Q. Zhu, Z. Lin, Y. Zhang, J. Sun, X. Li, Q. Lin, Y. Dang, R. Xu, Hitsz-hlt at semeval-2021 task 5: Ensemble sequence labeling and span boundary detection for toxic span detection, in: Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021), 2021, pp. 521–526.

[11] E. Leonardelli, S. Menini, A. Palmero Aprosio, M. Guerini, S. Tonelli, Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 10528–10539. URL: https://aclanthology.org/2021.emnlp-main.822. doi:10.18653/v1/2021.emnlp-main.822.

[12] M. Sandri, E. Leonardelli, S. Tonelli, E. Jezek, Why don't you do it right? analysing annotators' disagreement in subjective tasks, in: Proceedings of the 2023 Conference of the European Chapter of the Association for Computational Linguistics, 2023.

[13] E. Leonardelli, S. Menini, S. Tonelli, Dh-fbk@ haspeede2: Italian hate speech detection via self-training and oversampling., in: Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), volume 2765, 2020.

[14] C. Casula, S. Tonelli, Generation-based data augmentation for offensive language detection: Is it worth it?, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 3359–3377. URL: https://aclanthology.org/2023.eacl-main.244.

[15] J. Pavlopoulos, J. Sorensen, L. Laugier, I. Androutsopoulos, Semeval-2021 task 5: Toxic spans detection, in: Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021), 2021, pp. 59–69.

[16] R. Caruana, Multitask learning, Machine learning 28 (1997) 41–75.

[17] J. Lafferty, A. McCallum, F. C. Pereira, Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001).

[18] T. Wullach, A. Adler, E. Minkov, Fight Fire with Fire: Fine-tuning Hate Detectors using Large Samples of Generated Hate Speech, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 4699–4705. URL: https://aclanthology.org/2021.findings-emnlp.402.

[19] T. Hartvigsen, S. Gabriel, H. Palangi, M. Sap, D. Ray, E. Kamar, ToxiGen: A Large-Scale Machine-Generated Dataset for Adversarial and Implicit Hate Speech Detection, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 3309–3326. URL: https://aclanthology.org/2022.acl-long.234. doi:10.18653/v1/2022.acl-long.234.

[20] A. Anaby-Tavor, B. Carmeli, E. Goldbraich, A. Kantor, G. Kour, S. Shlomov, N. Tepper, N. Zwerdling, Do Not Have Enough Data? Deep Learning to the Rescue!, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 7383–7390. doi:10.1609/aaai.v34i05.6233.

[21] G. Sarti, M. Nissim, IT5: Large-scale text-to-text pretraining for italian language understanding and generation, ArXiv preprint 2203.03759 (2022). URL: https://arxiv.org/abs/2203.03759.

[22] R. van der Goot, A. Üstün, A. Ramponi, I. Sharaf, B. Plank, 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, Association for Computational Linguistics, Online, 2021, pp. 176–197. URL: https://aclanthology.org/2021.eacl-demos.22. doi:10.18653/v1/2021.eacl-demos.22.

Appendix

A. Ensemble agreement

For posts (of the training set) that were annotated as homotransphobic, we aim at obtaining an approximation of the agreement level on each word of the post, as being considered part of the span is correlated with labeling the post as homotranspophobic. This information is then exploited as additional information in our multi-task training setup, specifically as an extension to the sequence labelling prediction of Subtask B.

We split the training data X provided by the HODI organizers in 5 folds $X_1, X_2, ..., X_5$, creating 5 separate train/validation splits, being careful that each item of the train appears in the validation set of one fold. We employ an ensemble of classifiers, a method first suggested by Leonardelli et al. 2021 [11], where each classifier of the ensemble is trained using slightly different configurations by varying the initial conditions such as the initial seed and the number of epochs, so that the 5 classifiers produce similar but not identical predictions. The classifiers are produced in the multi-task setup showed in Figure 1, but without the Auxiliary Task on agreement. In this manner, we have ensemble predictions for each of the entries of the training data. Based on the predictions of the classifiers, we assign ensemble agreement labels to the validation set (at a word-level) 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. We consider this information as proxy for item's difficulty and annotators' disagreement.

B. Data augmentation pipeline

The pipeline we follow for augmenting the available data for the task is as follows:

1. We fine-tune a classifier (in our case a BERT-base model trained on Italian) [2] on the HODI training data.
2. We fine-tune IT5-Large on the same training data, formatting the task so that the input is ‘*Scrivi un tweet:*’ or ´*Scrivi un tweet omotransfobico:*’ (‘*Write a tweet:*’ or ´*Write a homotransphobic tweet:*’) depending on the gold label of each example, and the output is the actual post.
3. We use the fine-tuned IT5 model to generate new data, using the same type of input we use in Step 2.
4. We filter the generated data using the fine-tuned classifier from Step 1, keeping only the examples for which the label assignment is the same for the classifier and the generator [20, 14]. We additionally remove duplicates and normalize URLs as URL.
5. We rank generated examples based on the confidence of the classification model we used for filtering, retaining the top 2,000 examples for each class. This number is chosen in order to ideally double the size of the dataset, and we use generated examples that are equally split among the labels so as to artificially mitigate the class imbalance.