HODI at EVALITA 2023: Overview of the first Shared Task on Homotransphobia Detection in Italian

Debora Nozza[1], Alessandra Teresa Cignarella[2,3], Greta Damo[1], Tommaso Caselli[4] and Viviana Patti[2]

[1]Department of Computing Sciences, Bocconi University, Milan, Italy
[2]Department of Computer Science, University of Turin, Turin, Italy
[3]aequa-tech, Turin, Italy
[4]Center for Language and Cognition, University of Groningen, Groningen, The Netherlands

Abstract
HODI is a new shared task for the automatic detection of homotransphobia in Italian presented at EVALITA 2023. The challenge is organized into two subtasks: Subtask A focuses on the binary textual classification of homotransphobic tweets, while Subtask B is concerned with the identification of "rationales" for explainability in the form of textual spans of text. We have received a total of 19 runs for Subtask A and 5 runs for Subtask B from a total of 8 participating teams from 6 different countries. We present here an overview of the HODI shared task, the datasets, the evaluation methodology, the results obtained by the participants, and a discussion of the methodology adopted by the teams.

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

Keywords
Natural Language Processing, Hate Speech, Homotransphobia

1. Introduction

Despite advancements in human and civil rights, the Internet remains a hostile environment for LGBTQIA+ individuals. The increasing frequency, severity, and complexity of online hate crimes are mirrored in the real world. In a recent ISTAT-UNAR survey[1] on discrimination on work places suffered by LGBTQIA+ people, 43.9% of the participants has been target of insults, 61.8% suffered a micro-aggression (including hate speech), and 1.1% has been physically assaulted. In addition to this, anti-LGBTQIA+ hate crimes have risen drastically in the past three years.[2] Natural Language Processing (NLP) is a key subject of research for combating online hate speech since it can automate the process at scale while reducing online moderators' labor and mental stress [2]. Despite the NLP community's interest in hate speech detection datasets and models [3], very few studies covered hate speech against the LGBTQIA+ community [4, 5, 6]. Given the *target-oriented* nature of hate speech and the ineffectiveness of transferring hate speech detection models to different unseen hate speech targets [7, 8, 9, 10], the lack of a dedicated benchmark was a pending issue for homotransphobia in Italian, which we addressed with this work.

The Homotransphobia Detection in Italian (HODI)[3] shared task at EVALITA 2023 [11] **identifies Italian hate speech directed at the LGBTQIA+ community**. This will allow us to investigate a phenomenon that has received little attention from the worldwide NLP community and has never been investigated for Italian.

Being able to automatically determine whether a message is hateful or not is an important contribution to the fight against homotransphobia, yet this is not sufficient. Systems are always prone to errors, even the most accurate ones. This means that it could be the case that a perfectly normal message gets labeled as hateful simply because it contains an identity term (e.g., the word "gay"). Or the opposite can happen. Flagging a message for this kind of content should always be accompanied by an explanation that will shed (some) light on the way the system has taken its decision. Furthermore, recent Euro-

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
debora.nozza@unibocconi.it (D. Nozza); alessandrateresa.cignarella@unito.it (A. T. Cignarella); greta.damo@studbocconi.it (G. Damo); t.caselli@rug.nl (T. Caselli); viviana.patti@unito.it (V. Patti)
https://deboranozza.com/ (D. Nozza); https://www.unito.it/persone/acignare (A. T. Cignarella); https://www.rug.nl/staff/t.caselli/ (T. Caselli); https://www.unito.it/persone/vpatti (V. Patti)
0000-0002-7998-2267 (D. Nozza); 0000-0002-4409-6679 (A. T. Cignarella); 0000-0003-2936-0256 (T. Caselli); 0000-0001-5991-370X (V. Patti)

CEUR Workshop Proceedings (CEUR-WS.org)

[1]https://www.istat.it/it/files//2022/03/REPORTDISCRIMINAZIONILGBT_2022_rev.pdf

[2]https://www.theguardian.com/world/2021/dec/03/recorded-homophobic-hate-crimes-soared-in-pandemic-figures-show

[3]Task website: https://hodi-evalita.github.io/ and task repository: https://github.com/HODI-EVALITA/HODI_2023

Text	Subtasks	
	A	B
Odio i fr*ci	1	[0,1,2,3,7,8,9,10,11]
Morte ai gay torinesi	1	[0,1,2,3,4,9,10,11]
Divento fr*cio per te	0	0
Gay ed etero, stessi diritti	0	0

Table 1
Examples of the annotated data.

	Subtask A		Subtask B	
Split	Hate	Not	Single Token	Multi Token
Train	2,008	2,992	48	1,960
Test	511	489	16	495

Table 2
HODI data statistic overview.

pean legislation (General Data Protection Regulation – GDPR [12]) has introduced a "right to explanation". This necessitates a paradigm change from performance-based models to interpretable models [13]. This shared task will also contribute towards this need by assessing the **models' explanation abilities** to recognize the terms relevant for hate speech. This will allow, in the future, to control for possible biases of models overfitting to specific terms (e.g., *gay*) [14, 6], as well as use the explanations to generate counternarratives.

2. Task Description

HODI is structured on two subtasks (see examples in Table 1):

- **Subtask A - Homotransphobia detection**: this is a binary classification task where systems must classify a message as hateful or not against LGBTQIA+ community.
- **Subtask B - Explainability**: once a message is classified as hateful, the objective is to identify the rationales of the classification model, i.e., those tokens in the sequence that contributed to the flagging of the message.

The two tasks are strictly interconnected, but they have been run independently.

3. Training and Testing Data

Data Collection Data have been collected from Twitter using a keyword-based approach from May 1st, 2022 until August 31st, 2022. The selection is influenced by the observation that the summer months coincide with the pride celebrations, leading to increased discussions and engagement on social media regarding the subjects relevant to our objective. Additionally, May 17th is recognized globally as the International Day Against Homophobia, Biphobia, and Transphobia, further emphasizing the significance of this time frame for our task. We focused both on keywords that are commonly used in hateful contexts (e.g., *fr*cio*) and on others related to specific events that directly involve or affect the LGBTQIA+ community (e.g., *Pride*, *DDL Zan*). The complete list of keywords can be found in Appendix A. The decision to use keywords identifying events has been done because of a tendency to observe a surge in homotransphobic messages around them. In this way, we limited the presence of only explicit profanity-driven keywords that may introduce biases in the data and, consequently, in the trained models. As a result, the final dataset does not correspond to the natural distribution of hate on social media, which is lower.

Data Annotation Our annotation guidelines[4] have been developed by re-using previous guidelines for similar shared tasks, namely HatEval [15] and AMI [16]. In particular, we define a message as being hateful by applying the following definition:

> *any communication that disparages a person or a group on the basis of some characteristics, such as color, race, ethnicity, gender, sexual orientation, religion, nationality, or other aspects.*

Following the proposals in [17], our definition of hate speech and annotation guidelines have benefited from a series of interactions with some members of the Italian LGBTQIA+ community. In addition to this, we managed to have the data manually labeled by three members of the Italian LGBTQIA+ community (two males and one female). Each message has been annotated in parallel by each annotator for both subtasks. The annotators labeled whether the text is hateful or not and targets the LGBTQIA+ community. Then, the annotation for Subtask B targeting explainability is performed following the approach in [13]. In particular, our annotators have been asked to highlight the span of text that could support their labeling decision, the so-called *rationales*. We asked annotators to provide rationales only for the tweets considered hateful. These span annotations help us to investigate deeper the manifestations of hateful speech.

[4]Available for consultation here: https://github.com/HODI-EVALITA/HODI_2023

	step1	step2	step3	*avg*
Subtask A	0.543	0.658	0.547	0.583
Subtask B	0.617	0.627	0.700	0.648

Table 3
Inter-annotator agreement calculated with Fleiss' kappa coefficient (Subtask A) and % observed agreement (Subtask B) in the three steps.

The annotation campaign has been conducted in three different steps by giving the annotators 2,000 tweets each for each step. The inter-annotator agreement (IAA) has been calculated at the end of every step. In Table 3, we display the measures of the IAA on both subtasks, calculated with Fleiss' kappa coefficient (Subtask A) and % observed agreement (Subtask B). The average of the IAA obtained in both subtasks is *substantial* according to the interpretation of [18]. It is particularly impressive how the three annotators reached an IAA of 0.648 on the selection of homotransphobic spans of text, considering the difficulty and subjectivity of the task.

Extracting Gold Labels In this shared task, we decided to provide the participants with aggregated gold labels for both tasks rather than releasing the annotations separately. The aggregation process has been implemented as follows: for Subtask A, the gold label was chosen through a majority voting strategy. Since the annotators were three, and they could select only between two labels (0/1), there was always a clear prevalence for one or the other. On the other hand, for Subtask B, the gold span of text has been established by merging the three spans selected by the three annotators. Finally, in the fashion proposed in the SemEval 2021 shared task of toxic spans detection [25], we released the annotation of spans as a list of indices referring to the position of characters in the text (see Table 1).

Data Statistics Table 2 presents a summary of the annotated data for both subtasks. We provided 5,000 training and 1,000 testing tweets. The data we provided are roughly balanced (40% hateful tweets in training and 51% in the test set). For Subtask B, we report the number of messages with a single-token rationale and those with multi-token rationales. It can be seen how in both train and test, the majority of spans containing homophobic expressions are composed of more than one token. On the other hand, in the train set, there are 48 tweets where the hateful span contains only one word. In the test set, those cases are even fewer, i.e., only 16. Table 1 shows examples of data annotations for both Subtask A and B, with the rationales highlighted in yellow for better understanding.

4. Evaluation Measures and Baseline

Systems have been evaluated using the following metrics per task:

Subtask A. We use standard evaluation metrics for text classification, namely Precision, Recall, and F1-score per class. The ranking of the systems is based on the macro-averaged F1-score of the hateful and non-hateful messages.

Subtask B. Systems are evaluated using Intersection-Over-Union (IOU) [26], an *agreement* metrics. Token-level IOU is the size of the overlap of the character of the tokens they cover divided by the size of their union. We count a prediction as a match if it overlaps with any of the ground truth rationales by more than some threshold. We use these partial matches to calculate an F1 score and subsequently rank the systems.

Two different methods have been implemented to compare models to baselines:

Subtask A. Logistic Regression classifier based on TF-ID using unigrams and bigrams only.

Subtask B. A random classifier following the implementation of the organizers of the SemEval-2021 Task 5, Toxic Spans Detection [25].

The HODI GitHub repository[5] contains the code for calculating evaluation metrics and producing predictions using the baselines.

5. Participants and Results

We have received submissions from eight teams, for a total of 18 runs for Subtask A and four for Subtask B. Only two teams participated in Subtask B. Two teams used the same approach and system architecture for participating in other EVALITA 2023 tasks, namely `O-Dang` for HaSpeeDe and `extremITA` for all tasks. The majority of the teams were from academia, with only one industrial participant.

Participants were allowed to submit a maximum number of three runs for each subtask. Note that, in the case of submissions for both tasks, participants were asked to submit their predictions for Subtask A and Subtask B at the same time, i.e., in the same evaluation window. Table 4 provides a summary of the teams, illustrating their country and the subtasks they addressed.

[5] https://github.com/HODI-EVALITA/HODI_2023

Team	Country	Task	dbmdz-BERT-Italian	AlBERTO	Open AI Davinci	IT5	Camoscio	UmBERTO - Oscar Corpus	Twitter XML-R Sentiment	Fine-tuning	Knowledge injection	Data augmentation	Multi-task Learning	Few-shot Learning	Feature Extraction	Prompting
DH-FBK [19]	IT	A, B	✓							✓		✓	✓			
CHILab [20]	IT	A		✓											✓	
extremITA [21]	IT	A, B				✓	✓									✓
O-Dang [22]	IT,UK	A		✓	✓					✓	✓					
LCTs [23]	ES,NL	A						✓		✓						
Team_Tamil [24]	IE,IN	A							✓					✓		

Table 4
Overview of the participating systems who submitted the report. We list each team's tasks, pre-trained language models, and methods investigated.

Subtask A - Homotransphobia detection The homotransphobia detection task received 19 submissions from 8 teams, as shown in Table 5. The best result has been obtained by LCTs, where the team fine-tuned an Italian pretrained RoBERTa model named UmBERTo[6] for 10 epochs. Thus, this underscores the fact that relying solely on domain-specific approaches is still insufficient when it comes to effectively utilizing large models and extensive training. 6 out of 8 teams provide better results than the baseline. Due to a code error in the official submission that was not ranked in the shared task's official results, the team CHILab resubmitted amended runs (**) after the deadline.

Subtask B - Explainability The subtask related to the identification of the rationales behind prediction decisions received 5 runs from 2 teams. Table 6 shows the results in terms of F1. Considering the task's inherently complex and unique nature, teams had to invest additional effort beyond what is typically required for a binary prediction task, leading to an anticipated decrease in participation. Both teams outperformed the random baseline. The best performing submission by extremITA obtained the homophobic rationales interrogating an instruction-tuned decoder-only model (i.e., LLaMA) with the natural language instruction "*Con quali parole l'autore del testo precedente esprime odio omotransfobico? Separa le sequenze di parole con [gap]*" (*en: In what words does the author of the previous text express homotransphobic hatred? Separate the word sequences with [gap].*). While the ability to prompt such models has already been demonstrated to be effective by [27], the Subtask B results further highlight the power of large language models to perform even more difficult subjective tasks, such as explaining homophobic hatred.

6. Discussion

In Table 4, we present an overview of the participating systems for which we have received a system description paper. This section delves into the team's varied approaches from different perspectives.

Language Models Following a trend already seen in other evaluation campaigns [15, 16], all of the proposed systems make use of pre-trained language models (PTLMs) based on encoders only (dbmdz-BERT-italian[7], AlBERTO [28], UmBERTO[8], and Twitter-XML-R-sentiment[9]), or decoders only (Open AI Davinci [29], Camoscio[10]), or using a full Transformer architecture (IT5 [30]). Only two teams used multilingual models (Twitter-XML-R-sentiment and Open AI Davinci), while all the others used Italian monolingual PTLMs. For the Italian PTLMs, only AlBERTO [28] has been trained with a language variety compatible with the task's data, i.e., social media data. It is remarkable that pure fine-tuning of PTLMs has been done only by one team (LCTs). Another team, Team_Tamil, proposes

[6] https://huggingface.co/Musixmatch/umberto-commoncrawl-cased-v1

[7] https://huggingface.co/dbmdz/bert-base-italian-cased

[8] https://huggingface.co/Musixmatch/umberto-commoncrawl-cased-v1

[9] https://huggingface.co/citizenlab/twitter-xlm-roberta-base-sentiment-finetunned

[10] https://github.com/teelinsan/camoscio

Team	Macro F1	Rank
LCTs$_3$	**0.8108**	1
LCTs$_2$	0.8000	2
O-Dang$_1$	0.7959	3
DH-FBK$_1$	0.7950	4
extremITA$_2$	0.7942	5
O-Dang$_2$	0.7920	6
DH-FBK$_2$	0.7837	7
O-Dang$_3$	0.7804	8
LCTs$_1$	0.7709	9
CHILab$_2$**	0.7525	-
CHILab$_3$**	0.7454	-
extremITA$_1$	0.7431	10
CHILab$_1$**	0.7248	-
INGEOTEC$_1$	0.7153	11
Team_Tamil$_1$	0.6735	12
Baseline	0.6691	13
SOVRAG$_3$	0.6634	14
SOVRAG$_2$	0.6334	15
SOVRAG$_1$	0.6108	16
CHILab$_3$	0.5528	17
CHILab$_1$	0.5205	18
CHILab$_2$	0.5199	19

Table 5
Results for Subtask A - Homotransphobia detection. Numbers in subscripts for the team names correspond to the submitted run.

zero and few-shot learning of fine-tuned classification language models aiming at solving hate speech detection (e.g., [31]) or emotion-related tasks (e.g., [32]) in Italian and multilingual settings. For all other participants, fine-tuning represents just one component of other architectures and solutions.

Features and Additional Data No system has used external features from specialized lexical resources. Only one participant, DH-FBK, has extended the available training materials for both subtasks using synthetic data obtained with IT5. The authors have retained only the top 2,000 examples for each class as a strategy to double the size of the HODI training set per class as well as to mitigate class imbalance.

Prompting Following recent advancements in generative language models, two teams, O-Dang and extremITA, made use of prompting engineering techniques. In the case of O-Dang, prompts have been used to query the Open AI Davinci model to extract additional data concerning the names of entities of type "PERSON" that are present in the training set. The information thus obtained is concatenated to the original message as a form of knowledge injection. The extremITA team took a more radical path by addressing all EVALITA 2023

Team	F1 score	Rank
extremITA$_2$	**0.7228**	1
DH-FBK$_1$	0.7051	2
DH-FBK$_2$	0.7008	3
extremITA$_1$	0.6598	4
Baseline	0.2050	5

Table 6
Results for Subtask B - Explainability. Numbers in subscripts for the team names correspond to the submitted run.

tasks by means of prompting. They apply two different prompting approaches, compliant with the models they use (IT5 and Camoscio). The authors exploited zero-shot prompting, which means they did not give the models any examples from the training data. They only specialized the natural language instruction for the different tasks.

Interaction between Subtask A and Subtask B The only team that exploited as much as possible the interaction between the two subtasks in the design of their system is DH-FBK. The authors developed a multi-task learning architecture using the MaChAmp v2.0 toolkit [33].

7. Conclusion and Future Work

This paper introduces HODI, the first shared task on homotransphobia detection in Italian. The task aims to not only identify homotransphobic messages but also investigate the underlying reasons behind them. We have analyzed the submissions from participating teams and concluded that satisfactory results have been achieved in detecting homotransphobia in Italian. Furthermore, notable progress has been made in the explainability task, although further work is required in this area. To continue advancing in this field, future efforts should focus on constructing larger and more diverse datasets. Additionally, there is a need to enhance the detection models and improve their ability to explain the specific words or features that contribute to a hateful classification.

Acknowledgments

The work of A.T. Cignarella and V. Patti was partially funded by the International project *STERHEOTYPES - Studying European Racial Hoaxes and sterEOTYPES*, funded by the Compagnia di San Paolo and VolksWagen Stiftung under the 'Challenges for Europe' Call for Projects (CUP: B99C20000640007). The work of D. Nozza was partially funded by Fondazione Cariplo (grant No. 2020-4288, MONICA). Debora Nozza is a member of the MilaNLP group, and the Data and Marketing Insights

Unit of the Bocconi Institute for Data Science and Analysis.

A special mention also to the people who helped us with the annotation of the dataset and the assessment of guidelines: Davide, Greta, and Mauro, thank you very much for your great help.

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, Toronto, Canada, 2023, pp. 3897–3909. URL: https://aclanthology.org/2023.findings-acl.240.

[2] M. Chaudhary, C. Saxena, H. Meng, Countering online hate speech: An NLP perspective, arXiv preprint arXiv:2109.02941 (2021). URL: https://arxiv.org/abs/2109.02941.

[3] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti, Resources and benchmark corpora for hate speech detection: a systematic review, Language Resources & Evaluation 55 (2021) 477–523.

[4] B. R. Chakravarthi, R. Priyadharshini, T. Durairaj, J. P. McCrae, P. Buitaleer, P. K. Kumaresan, R. Ponnusamy, Findings of the shared task on Homophobia Transphobia Detection in Social Media Comments, in: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, Association for Computational Linguistics, 2022.

[5] 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). URL: https://arxiv.org/abs/2109.00227.

[6] D. Nozza, F. Bianchi, A. Lauscher, D. Hovy, Measuring harmful sentence completion in language models for LGBTQIA+ individuals, in: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 26–34. URL: https://aclanthology.org/2022.ltedi-1.4. doi:10.18653/v1/2022.ltedi-1.4.

[7] D. Nozza, Exposing the limits of zero-shot cross-lingual hate speech detection, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Association for Computational Linguistics, Online, 2021, pp. 907–914. URL: https://aclanthology.org/2021.acl-short.114. doi:10.18653/v1/2021.acl-short.114.

[8] E. W. Pamungkas, V. Patti, Cross-domain and cross-lingual abusive language detection: A hybrid approach with deep learning and a multilingual lexicon, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, Association for Computational Linguistics, Florence, Italy, 2019, pp. 363–370. URL: https://aclanthology.org/P19-2051. doi:10.18653/v1/P19-2051.

[9] E. W. Pamungkas, V. Basile, V. Patti, Misogyny detection in twitter: a multilingual and cross-domain study, Information Processing & Management 57 (2020) 102360. URL: https://www.sciencedirect.com/science/article/pii/S0306457320308554. doi:https://doi.org/10.1016/j.ipm.2020.102360.

[10] P. Chiril, E. W. Pamungkas, F. Benamara, V. Moriceau, V. Patti, Emotionally informed hate speech detection: A multi-target perspective, Cognitive Computation 14 (2022) 322–352. URL: https://doi.org/10.1007/s12559-021-09862-5. doi:10.1007/s12559-021-09862-5.

[11] 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.

[12] Eu regulation 2016/679 general data protection regulation (GDPR), https://eur-lex.europa.eu/eli/reg/2016/679/oj, 2016. URL: https://eur-lex.europa.eu/eli/reg/2016/679/oj, accessed:2022-10-09.

[13] B. Mathew, P. Saha, S. M. Yimam, C. Biemann, P. Goyal, A. Mukherjee, Hatexplain: A benchmark dataset for explainable hate speech detection, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 35:17, 2021, pp. 14867–14875.

[14] L. Dixon, J. Li, J. Sorensen, N. Thain, L. Vasserman, Measuring and mitigating unintended bias in text classification, in: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 67–73. URL: https://doi.org/10.1145/3278721.3278729. doi:10.1145/3278721.3278729.

[15] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, M. Sanguinetti, 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, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp.

54–63. URL: https://aclanthology.org/S19-2007. doi:10.18653/v1/S19-2007.

[16] E. Fersini, D. Nozza, P. Rosso, AMI @ EVALITA2020: Automatic Misogyny Identification, in: V. Basile, D. Croce, M. Di Maro, L. C. Passaro (Eds.), Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020), CEUR.org, Online, 2020.

[17] T. Caselli, R. Cibin, C. Conforti, E. Encinas, M. Teli, Guiding principles for participatory design-inspired natural language processing, in: Proceedings of the 1st Workshop on NLP for Positive Impact, Association for Computational Linguistics, Online, 2021, pp. 27–35. URL: https://aclanthology.org/2021.nlp4posimpact-1.4. doi:10.18653/v1/2021.nlp4posimpact-1.4.

[18] J. R. Landis, G. G. Koch, An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers, Biometrics (1977) 363–374.

[19] E. Leonardelli, C. Casula, DH-FBK at HODI: Multi-Task Learning with Classifier Ensemble Agreement, Oversampling and Synthetic Data, 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.

[20] I. Siragusa, R. Pirrone, CHILab at HODI: A minimalist approach, 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.

[21] C. D. Hromei, D. Croce, V. Basile, R. Basili, Extremita at EVALITA: Multi-Task Sustainable Scaling to Large Language Models at its Extreme, 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.

[22] C. Di Bonaventura, A. Muti, M. A. Stranisci, O-Dang at HODI and HaSpeeDe3: A Knowledge-Enhanced Approach to Homotransphobia and Hate Speech Detection in 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.

[23] D. Locatelli, L. Locatelli, LCTs at HODI: Homotransphobic Speech Detection on Italian Tweets, 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.

[24] R. Ponnusamy, P. K. Kumaresan, K. K. Ponnusamy, C. Rajkumar, R. Priyadharshini, B. Raja Chakravarthi, Team_Tamil at HODI: Few-Shot Learning for Detecting Homotransphobia in Italian Language, 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.

[25] 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), ACL, 2021, pp. 59–69.

[26] J. DeYoung, S. Jain, N. F. Rajani, E. Lehman, C. Xiong, R. Socher, B. C. Wallace, ERASER: A benchmark to evaluate rationalized NLP models, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online, 2020, pp. 4443–4458. URL: https://aclanthology.org/2020.acl-main.408. doi:10.18653/v1/2020.acl-main.408.

[27] F. M. Plaza-del arco, D. Nozza, D. Hovy, Respectful or toxic? using zero-shot learning with language models to detect hate speech, in: The 7th Workshop on Online Abuse and Harms (WOAH), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 60–68. URL: https://aclanthology.org/2023.woah-1.6.

[28] M. Polignano, P. Basile, M. de Gemmis, G. Semeraro, V. Basile, ALBERTO: Italian BERT language understanding model for NLP challenging tasks based on tweets, in: Proceedings of the 6th Italian Conference on Computational Linguistics, CLiC-it 2019, volume 2481, CEUR Workshop Proceedings (CEUR-WS. org), CEUR-WS.org, 2019. URL: http://ceur-ws.org/Vol-2481/paper57.pdf.

[29] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al., Training language models to follow instructions with human feedback, Advances in Neural Information Processing Systems 35 (2022) 27730–27744.

[30] 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.

[31] D. Nozza, F. Bianchi, G. Attanasio, HATE-ITA: Hate speech detection in Italian social media text, in: Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH), Association for Computational Linguistics, Seattle, Washington (Hybrid), 2022, pp. 252–260. URL: https://aclanthology.org/2022.woah-1.24. doi:10.18653/v1/2022.woah-1.24.

[32] F. Bianchi, D. Nozza, D. Hovy, FEEL-IT: Emotion and sentiment classification for the Italian language,

in: Proceedings of the Eleventh Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Association for Computational Linguistics, Online, 2021, pp. 76–83. URL: https://aclanthology.org/2021.wassa-1.8.

[33] 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.

A. Appendix

Keywords used for tweet collection.

gay, pride, lesbica, f*nocchio, fr*cio, fr*cia, b*cchinaro, c*lattone, rottinc*lo, piglianc*lo, succhiac*zzo, ciucciac*zzo, c*landa, leccaf*ga, b*liccio, b*sone, f*mminiello, p*mpinaro, effeminato, c*chino, b*caiolo, ch*cca, r*cchione, invertito, travestito, passivo, deviato, ddl zan