DisaggregHate It Corpus: A Disaggregated Italian Dataset of Hate Speech

Marco Madeddu[1], Simona Frenda[1,2], Mirko Lai[1,2], Viviana Patti[1] and Valerio Basile[1]

[1] *Università di Torino, Italy*
[2] *Aequa-tech srl, Turin, Italy*

Abstract

Recent studies in Machine Learning advocate for the exploitation of disagreement between annotators to train models in line with the different opinions of humans about a specific phenomenon. This means that datasets where the annotations are aggregated by majority voting are not enough. In this paper, we present an Italian disaggregated dataset concerning hate speech and encoding some information about the annotators: the DisaggregHate It Corpus. The corpus contains Italian tweets that focus on the topic of racism and has been annotated by native Italian university students. We explain how the dataset was gathered by following the recommendation of the *perspectivist* approach [1], encouraging the annotators to give some socio-demographic information about them. To exploit the disagreement in the learning process, we proposed two types of *soft labels*: softmax and standard normalization. We investigated the benefit of using disagreement by creating a baseline binary model and two regression models that were respectively trained on the 'hard' (aggregated label by majority voting) and the two types of 'soft' labels. We tested the models in an in-domain and out-of-domain setting, evaluating their performance using the cross-entropy as a metric, and showing that the models trained on the soft labels performed better.

Keywords

hate speech, perspectivism, disagreement

1. Introduction

The rise of the Internet and social media platforms has given many users the opportunity to express their opinion online. Unfortunately, this leads to the diffusion of a new online phenomenon: the hate speech. To prevent the viral spread of this kind of expressions on social media, hate speech detection became a popular task in Natural Language Processing (NLP). A lot of tools have been created to detect and counter hate speech[2, 3, 4].

Recently, there have been studies that suggest trying to shift away from the golden standard approach in Machine Learning, especially in tasks partly subjective and influenced by the social and cultural context, like hate speech [5, 1]. These works advocate that different opinions given in the annotation process are not a noise factor but can be used to make better systems [6]. This shift inspires scholars to try different techniques to train models using datasets where the target label is not simply determined by majority voting on the annotations. In this line, two theoretical paradigms have been established, both looking for the inclusion of different perspectives: the *learning from disagreement* and *perspectivism.* The former could be considered like a 'soft perspectivist approach' because it takes into account the presence of disagreement in the annotated data, while the latter, overcomes the idea of "ground truth" in the construction of datasets and on the creation and evaluation of NLP models, focusing more on who the annotators are.

Our work could be considered a tentative to approach hate speech detection, exploiting the possible disagreement among the annotators. Usually, models are trained on data associated to a 'hard' label. In the case of binary classification, each item is assigned a label whose value is either 0 or 1. The hard label value is commonly obtained through majority voting, therefore this implies that controversial instances have the same label as the ones that saw all annotators in agreement. This may be thought of like a loss of valuable information that can be used in the training phase of the models [7]. On the other hand, 'soft' labels approaches try to avoid this waste of data by assigning a real number to the label. Different functions can be used in the process of determining the value of the soft label, such as standard normalization or a softmax function [8].

In this context, we created the DisaggregHate It Corpus, a new disaggregated dataset about hate speech in the Italian language that incorporates some socio-demographic information about annotators[1]. A corpus like this could be beneficial in exploring how different segments of population are sensitive to certain social issues like hate speech, and how this information can be used to create better systems.

After explaining the different characteristics of the

CLiC-it 2023: 9th Italian Conference on Computational Linguistics, Nov 30 — Dec 02, 2023, Venice, Italy

marco.madeddu@edu.unito.it (M. Madeddu); simona.frenda@unito.it (S. Frenda); mirko.lai@unito.it (M. Lai); viviana.patti@unito.it (V. Patti); valerio.basile@unito.it (V. Basile)

CEUR Workshop Proceedings (CEUR-WS.org)

[1] The corpus is available here: https://github.com/madeddumarco/DisaggregHateIt

dataset in section 3 we will validate the corpus by using it as the training set of different models in section 4. The performed experiments show that training models on a soft label rather than a hard label leads to better results. As suggested by [7], we used the cross entropy metric for evaluating the models.

2. Related Work

The past years have seen an increase in using different paradigms that try to model the different opinions of human annotators, especially in cases of recognition of subjective phenomena, like hate speech. Adopting a *soft perspectivist approach*, recent challenges like Le.Wi.Di (*Learning with disagreement*) shared task were proposed at SemEval 2021 and 2023 [8, 9]. In particular, this shared task asked participants to model various phenomena, such as humor and hate speech detection, exploiting the *soft labels*. These, contrary to the *hard labels* (simple labels), are obtained computing a sort of distribution of the labels chosen by annotators. Modelling this distribution, the systems are able to approximate the probability distribution of the opinions about the specific phenomenon. A *strong perspectivist approach*, instead, looks at whom the annotators are and how to model their opinion [1].

In the experimental part of this work, we focused especially on the use of soft labels to model the different labels without considering the information available about annotators, on the example of Uma et al. [7]. In this study, the authors experimented the application of the soft labels to detect various phenomena, employing a standard and a softmax normalization of the labels. They proved that in both hard and soft evaluation settings, respectively using accuracy and cross-entropy metrics, the use of soft labels in the modelling leads to better results.

Following their example, we evaluated the new disaggregated dataset on hate speech, DisaggregHate It Corpus, composed of Italian tweets, and enriched with some socio-demographic information about the annotators. Our idea was to create a dataset according perspectivist recommendations provided by Cabitza et al. [1] to ensure the transparency of the created perspective dataset. Among these recommendations, the authors mention the involvement of enough and heterogeneous annotators, and the collection of information about them. Moreover, with our work we meet also other their recommendations such as the report about the annotation process, the use of hard labels (computed by majority voting) and the soft labels (to represent the distribution of the decisions provided by annotators), and finally, we validated our models in an out-of-domain setting.

The works on hate speech that comply to some of these recommendations and release disaggregated datasets, are few, and to our knowledge, are only in other languages. One of the most famous is the Measuring of Hate Speech corpus [10][2] available only in English, that encodes various dimensions of hate speech (with disaggregated labels) and also different information about annotators. Follow: the HS-Brexit disaggregated dataset created by Akhtar et al. [11], ToxCR dataset [12] and JSRPData [13], on hate speech and toxic language. All of these datasets are in English and contain little information about the annotators. About Italian language, to our knowledge, only IMSyPP-IT dataset [14] have been released with disaggregated labels but without information about annotators.

In this context, a dataset like DisItaggragated released with disaggregated labels about hate speech, and that encodes also some information about annotators, contributes to enrich the resources for Italian community and to encourage the modeling of perspectives and different opinions in a very subjective phenomenon like hate speech.

3. Dataset

In this section, we first introduce our dataset by illustrating the context of the annotation process and secondly the general statistics about the corpus. Further, we will analyze the distribution of the positive and negative label for both the hard and soft label.

3.1. Corpus Creation

The DisaggregHate It Corpus used for this work is composed of 1100 tweets extracted from Contro L'Odio [15], an Italian corpus that focuses on racist hate and in particular on discrimination towards immigrants. The annotation process carried out as part of a master degree course, so the participants are all university students aged between 21 and 30, and native of the Italian language. A specific educational web platform has been realized on the example of the one developed by [16], for allowing the annotation process and the collection of some basic information about the annotators. For each tweet, annotators have been asked to decide the presence hate speech (yes or no), irony (yes or no) and the stance of the author of the message towards immigration issues (positive, neutral, or negative)[3]. For our experiment, we only considered the hate speech annotations, so from this point forward when we will talk about the target label we are referring to the **hate speech** one.

[2] This dataset is released on HuggingFace: https://huggingface.co/datasets/ucberkeley-dlab/measuring-hate-speech

[3] The used guidelines are the ones adopted to annotate data in the HaSpeeDe context [17] (for hate speech), in the context of IronIta [18] (for irony), and in the context of SardiStance [19] (for stance). Especially the last guidelines have been adapted to the context of immigration.

Profile	Annotators	Tweets	Krippendorff's α
City <50k	11	300	0.32
City >50k	8	300	0.40
TSCI	4	100	0.19
Humanistic	1	100	-
Men	12	403	0.28
Women	11	400	0.24
Low SM	30	700	0.32
High SM	36	700	0.34

Table 1
Information about annotators

Annotators provided basic information about their gender, how many social media platforms they use, if they live in a city with more than 50 thousands residents and their school background (TSCI or Humanistic). The participants could choose to give one or more information about them.

In order to collect as many annotations as possible for each tweet, students have been grouped in teams of minimum 5 components, and each annotator was asked to annotate at least 100 tweets per group. However, some sets of data have been annotated by more than 1 group of students and others only by few annotators. Therefore, every tweet has a number of annotation in a range from 1 to 13.

In this context, we computed the agreement among the annotators, taking into account the information that they provided, using Krippendorff's Alpha [20]. This metric, indeed, allows evaluating agreement when the matrix of annotations is sparse (i.e., the number of annotators is not constant for each tweet and, thus, some values could miss). We did not report the value of Krippendorff's Alpha for the 'humanistic' profile as the function requires at least two annotators (see Table 1). The value of Krippendorff's Alpha for the whole dataset is 0.34.

In table 1 we can observe that the agreement intra-group is quite low as Krippendorff's alpha values that are equal to 0 indicate absence of reliability meanwhile values that are equal to 1 show perfect agreement [20]. It means that annotators have different perception of hate speech even if they share the same socio-demographic trait. The only profile that shows a fair agreement is the 'City >50K' (living in a city with more than 50 thousands residences). However, the scores are low, motivating an approach based on *learning with disagreement.*

3.2. Hard and Soft Label Distribution

We assigned each tweet three different labels: a hard label and two soft labels. The hard label matches the majority vote of annotations, while the two soft labels, respectively, employ a standard and a softmax normalisation. Using the generalization of Uma et al. [7], given an instance i and C classes , we can determine a vector $[d_i^1, d_i^2, .., d_i^C,]$ where d_i^j is the number of votes given by the annotators for class j to the instance i. Softmax normalization determines the value of the soft label l_i^j for each example i and class j with the following formula:

$$l_i^j = \frac{exp(d_i^j)}{\sum_a exp(d_i^a)}$$

Meanwhile standard normalization is obtained by applying:

$$l_i^j = \frac{d_i^j}{\sum_a d_i^a}$$

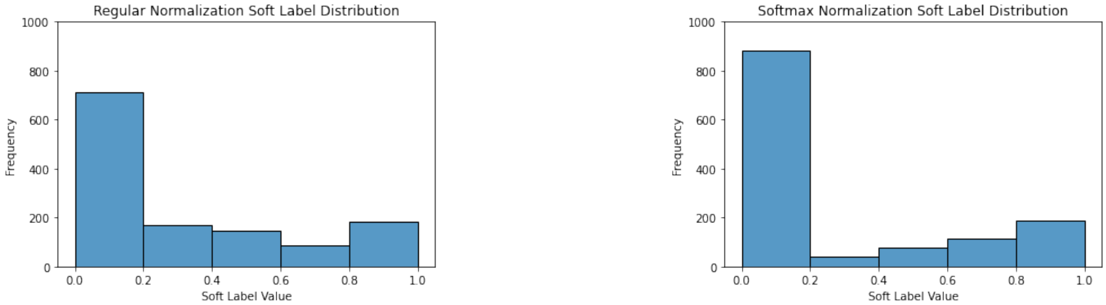
This case study addresses the hate speech annotation as a binary problem: $j \in [HS, \neg HS]$. We computed standard and a softmax normalisations l_i^{HS} for the sole positive class.

Addressing the data labelling with a soft label approach prevents discarding annotations and allows for the creation of a more informative annotated corpus. As Uma et al. [7] pointed out the softmax normalization, unlike the standard one, assigns to an instance a non zero value even where a class received zero votes. Therefore, the softmax normalisation could be seen as a way to smooth the label distribution, but it could also cause some side effects. Indeed, whenever $d_i^c \simeq \sum_a d_i^a$, i.e. there is complete agreement among annotators but the there is only a very small number of annotators, $l_i^j \forall j \neq c$ will be however sensibly larger than 0. Therefore, the use of standard normalization would be preferable in the presence of many classes and few annotators.

In table 2 we can observe how many ties, positive and negative instances are present in our dataset when we apply a majority voting to obtain a hard label. We can also observe how many tweets had an even number of annotators resulting in possible ties. We can see that there is different percentages of positive instances in some demographic division criteria like gender (Men and Women). Other category distinctions, like the one based on social media usage, show little difference between the two groups (Low SM and Hight SM). The number of ties is very different between the various categories

Category	Examples	Ties %	Pos. %	Neg. %	Tie Chance %
Whole Dataset	1100	3.7%	15.3%	80.9%	66%
City <50k	300	0%	9.6%	90.3%	0%
City >50k	300	6.6%	11.6%	81.6%	33.3%
TSCI	100	15%	9%	76%	100%
Humanistic	100	0%	19%	81%	0%
Men	403	18.11%	10.17%	71.71%	74%
Women	400	9.25%	23.25%	67.5%	31%
Low SM	700	6%	12.14%	81.85%	71%
High SM	700	5.85%	12%	82.14%	42%

Table 2
Dataset Composition



(a) Distribution for standard normalization

(b) Distribution for softmax normalization

Figure 1: Histograms for distributions of soft labels

ranging from 0% to more than 18% of the total instances. A very high number of ties indicates the presence of controversial instances that could be very important in the training phase of a model. The Krippedndorff's alpha values paired with the number of ties show that the DisaggregHate It Corpus contains a not neglectable level of disagreement between the annotators. Overall, we can see that the DisaggregHate It Corpus is unbalanced towards the negative class; therefore, in Section 4, we proposed to train the models using weighted labels.

In Figure 1, we observe the label distribution using standard or softmax normalization. We can observe that there are more negative instances than positive ones as the most represented bin is the one with $l_i^{HS} < 0.2$. We can observe a mostly similar tendency comparing the Figures 1a and 1b even if the standard normalization has more examples in the bins for the middle values. Overall, we can observe that annotators usually tend to be in agreement when there is a clear signal of hate speech, indeed the bin with values $l_i^{HS} > 0.8$ has more instances compared to other ones.

4. Experiments

The DisaggregHate It Corpus has been used to carry out two main settings of experiments: in-domain (test set of DisaggregHate It Corpus corpus) and out-of-domain (two test sets of two new shared tasks at EVALITA 2023). The tested models are: a standard model trained on aggregated labels (called here *Binary*), and two new models trained on soft labels (called here *Regression*) computed in two different manners. The former trained to detect the presence or absence of hate speech in the tweets, the latter trained to give a probability about the presence of hate speech in the tweets in line with the distribution of labels provided by annotators.

4.1. Models Description

We built all of our models by fine-tuning an already existing BERT (Bidirectional Encoder Representations from Transformers) based model for Italian. BERT is the state-of-the-art family of Large Language Models based on the transformer architecture [21]. There are a lot of BERT models that have been trained on large amount of data, thus they can be easily fine-tuned to perform in other tasks by fine-tuning them with smaller data sets. The model we chose to use is the uncased Italian BERT model with the Huggingface identifier: dbmdz/bert-base-italian-uncased created by the MDZ Digital Library team [22]. We accessed it through the Huggingface platform and the Python library Transformers which offers easy to use functions to design a simple architecture for fine-tuning the pre-trained models for specific tasks like the one of classification (i.e., *BertForSequenceClassification*). Considering the characteristics of our dataset and the kind of experiments that we wanted to perform, we designed some specific techniques.

The first regards the output of the network. We created three different models: one trained for binary classification with the hard label of the dataset, and two regression models respectively trained on the soft label computed with standard normalization and the softmax normalization. Taking into account the need of using a *soft metric* (cross-entropy) to compare the performance of our models, as suggest by [7, 8, 9], for the binary classifier we obtained soft label predictions by applying the softmax function to the logit outputs. The probabilities from the regression models are simply obtainable thanks to the Transformers library by setting the number of labels parameter to 1 of a classification model. As the outputs of the regression models are not bounded, we applied the clip function to limit their value between 0 and 1.

The second is about the different balance of the classes in our dataset. The DisaggregHate It Corpus contained, indeed, more negative label examples than positive ones (see Table 2). To deal with this, we experimented by assigning different weights to the positive and negative label. We obtained these different weights through the *compute_class_weight* function present in the scikit learn Python library. These weights were used in the calculation of the loss function for each model. The binary model was trained with a weighed cross-entropy loss function and given that the training set contained hard labels, we easily assigned different weights to each label. The regression models were trained with a weighed Mean Squared Error loss function and as the label values were real number, we assigned the postive binary label weight to examples with a soft label value ≥ 0.5 and we assigned the negative binary label weight to the rest. The models were trained for 5 epochs, each with a learning rate parameter equal to $2e^{-5}$.

4.2. In and Out-of-Domain Testing

The **in-domain test set** has been extracted from the DisaggregHate It Corpus, selecting 20% of the entire dataset, while the rest was used for the training and validation sets. As **out-of-domain test sets**, we used two datasets in the Italian language that have been released in the occasion of the 2023 edition of the EVALITA campaign. The first one is the corpus regarding the second task of the HaSpeeDe3 (Hate Speech Detection) shared task [23] annotated in regard to political and religious hate. The used test set from HaSpeeDe3 is composed of 5600 tweets, containing 2144 positive examples. The second dataset is the corpus from the HoDI (Homotransphobia Detection in Italian) shared task [24] containing 5000 tweets about homophobia. The test set of HoDI is composed of 5000, containing 2008 positive examples. So after training our models with the in-domain training sets we tested them on the in-domain tests, the entire HaSpeeDe 3 and HoDI training sets.

4.3. Results

In table 3 we report all the results in terms of cross-entropy (CE) for the in-domain and out-of-domain experiments. We decided to only report the CE scores with certain test sets to avoid an unfair comparison. Therefore, we excluded testing the regression model trained on the standard normalization soft labels with the softmax normalized test set, and vice versa. As the binary model soft label predictions are obtained by applying the softmax function, thus we decided it is adequate to calculate the CE with the softmax normalized test set. About the out-of-domain testing, we calculated the CE between the soft label predictions and the hard label versions of the test sets, as the disaggregated annotations are not available.

We can observe in table 3 that both regression models report better scores than the binary models in all tests both in-domain and out-of-domain. When we compare the CE score obtained with the binary model with the ones obtained with regression models, we can see a very significant difference in favor of the regression model in both scenarios. Observing in details the standard normalization and softmax normalization regression models, we notice that the softmax normalization seems works better in general, in both experimental settings. However, if in the in-domain setting, the scores report a difference of 5% in terms of Δ, in the out-of-domain setting, the results from both regression models are similar. These results are in line with the ones obtained in the study of Uma et al. [7].

Moreover, we can observe that both regression models score slightly worse when compared to the in-domain setting, and this could have been expected as the cross-domain task is difficult. Another factor of this drop in performance could be that the target label of the cross-domain datasets was binary and not a real number. This encourages the releasing of datasets with disaggregated labels.

Model Type	Train Set	In-Domain Test		HaSpeeDe 3	HoDI
		CE Std.	CE Softmax	CE	CE
Binary	Hard Label	-	1.084	0.814	0.851
Regression	Standard Norm. Label	**0.616**	-	0.668	**0.674**
Regression	Softmax Norm. Label	-	**0.588**	**0.662**	0.678

Table 3
Cross Entropy Test Results

5. Conclusion

In this work, we presented the DisaggregHate It Corpus, a new disaggregated dataset in the Italian language of hate speech. To our knowledge, it is the first dataset released with disaggregated labels and some socio-demographic information about the annotators. Computing the agreement among annotators with the same profile, we noticed that the Krippendolf'α is very low. Moreover, this information, paired with the number ties obtained by majority voting, showed us how disagreement is a real factor in corpora. That motivates the need to approach the hate speech detection task with models that encode the different opinions of humans annotators. To this purpose, we experimented with the use of a soft label, exploring two different computation of soft labels: standard and softmax normalization.

To continue our study on the usage of disagreement as a factor in learning we carried out different experiments testing the performance of our models in two specific settings: in-domain and out-of-domain. We created a binary model based on the hard labels and two regression models trained on the soft labels (computed with the two different normalization, regular and softmax). Inspired by previous works [7, 8, 9], we evaluated the models, employing the cross-entropy between the soft labels of annotations and the model predictions. Observing the results, we noticed that the regression models perform better both when considering in-domain and out-of-domain test sets. This implies that a soft label is helpful to integrate annotators disagreement inside our models in order to be more in line with the distribution of the opinions of human annotators.

Taking into account these results, we plan to use the same DisaggregHate It Corpus, to explore a stronger perspectivist approach modelling the perspectives of different groups of annotators on the basis of their socio-demographic traits or other commonalities.

Ethics Statement

The annotation process involved students of the Politecnico di Torino, who performed this task in an educational environment. The guidelines and the information about the annotation task have been shared via the educational platform exploited for implementing the annotation process, and discussed during the lessons. The efforts required to the students has been limited to their time and oriented to complete a project work being part of the exam of *Internet e social media: tecnologie e derive della comunicazione in rete*. This annotation task has been used, first of all, to give the students the opportunity to discuss the disagreement, encouraging a deep reflection on the importance of developing high quality annotated resources, to train and evaluate machine learning models.

References

[1] F. Cabitza, A. Campagner, V. Basile, Toward a perspectivist turn in ground truthing for predictive computing, Proceedings of the AAAI Conference on Artificial Intelligence 37 (2023) 6860–6868. URL: https://ojs.aaai.org/index.php/AAAI/article/view/25840. doi:10.1609/aaai.v37i6.25840.

[2] A. Schmidt, M. Wiegand, A survey on hate speech detection using natural language processing, in: Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media, Association for Computational Linguistics, Valencia, Spain, 2017, pp. 1–10. URL: https://aclanthology.org/W17-1101.

[3] P. Fortuna, S. Nunes, A survey on automatic detection of hate speech in text, ACM Computing Surveys 51 (2018) 85:1–85:30. URL: https://doi.org/10.1145/3232676.

[4] F. Poletto, V. Basile, M. Sanguinetti, C. Bosco, V. Patti, Resources and benchmark corpora for hate speech detection: A systematic review, Language Resources and Evaluation 55 (2021) 477–523. URL: https://rdcu.be/cCdaB.

[5] B. Plank, The "problem" of human label variation: On ground truth in data, modeling and evaluation, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 10671–10682. URL: https://aclanthology.org/2022.emnlp-main.731.

[6] V. Basile, M. Fell, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, A. Uma, We need to con-

sider disagreement in evaluation, in: Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future, Association for Computational Linguistics, Online, 2021, pp. 15–21. URL: https://aclanthology.org/2021.bppf-1.3. doi:10.18653/v1/2021.bppf-1.3.

[7] A. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, M. Poesio, A case for soft loss functions, Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 8 (2020) 173–177. URL: https://ojs.aaai.org/index.php/HCOMP/article/view/7478. doi:10.1609/hcomp.v8i1.7478.

[8] A. Uma, T. Fornaciari, A. Dumitrache, T. Miller, J. Chamberlain, B. Plank, E. Simpson, M. Poesio, SemEval-2021 task 12: Learning with disagreements, in: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), Association for Computational Linguistics, Online, 2021, pp. 338–347. URL: https://aclanthology.org/2021.semeval-1.41. doi:10.18653/v1/2021.semeval-1.41.

[9] E. Leonardelli, A. Uma, G. Abercrombie, D. Almanea, V. Basile, T. Fornaciari, B. Plank, V. Rieser, M. Poesio, Semeval-2023 task 11: Learning with disagreements (lewidi), 2023. arXiv:2304.14803.

[10] C. J. Kennedy, G. Bacon, A. Sahn, C. von Vacano, Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application, ArXiv abs/2009.10277 (2020). URL: https://api.semanticscholar.org/CorpusID:221836648.

[11] S. Akhtar, V. Basile, V. Patti, Modeling annotator perspective and polarized opinions to improve hate speech detection, Proceedings of the AAAI Conference on Human Computation and Crowdsourcing 8 (2020) 151–154. URL: https://ojs.aaai.org/index.php/HCOMP/article/view/7473. doi:10.1609/hcomp.v8i1.7473.

[12] D. Kumar, P. G. Kelley, S. Consolvo, J. Mason, E. Bursztein, Z. Durumeric, K. Thomas, M. Bailey, Designing toxic content classification for a diversity of perspectives, in: Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021), 2021, pp. 299–318.

[13] N. Goyal, I. D. Kivlichan, R. Rosen, L. Vasserman, Is your toxicity my toxicity? exploring the impact of rater identity on toxicity annotation, Proceedings of the ACM on Human-Computer Interaction 6 (2022) 1–28.

[14] M. Cinelli, A. Pelicon, I. Mozetič, W. Quattrociocchi, P. Kralj Novak, F. Zollo, Italian YouTube Hate Speech corpus, 2021. URL: http://hdl.handle.net/11356/1450, slovenian language resource repository CLARIN.SI.

[15] A. T. Capozzi, M. Lai, V. Basile, F. Poletto, M. Sanguinetti, C. Bosco, V. Patti, G. Ruffo, C. Musto, M. Polignano, et al., Computational linguistics against hate: Hate speech detection and visualization on social media in the" contro l'odio" project, in: CEUR Workshop Proceedings, volume 2481, CEUR-WS, 2019, pp. 1–6.

[16] S. Frenda, A. T. Cignarella, M. A. Stranisci, M. Lai, C. Bosco, V. Patti, et al., Recognizing hate with nlp: The teaching experience of the# deactivhate lab in italian high schools, in: CEUR WORKSHOP PROCEEDINGS, volume 3033, CEUR-WS. org, 2021, pp. 1–7.

[17] C. Bosco, D. Felice, F. Poletto, M. Sanguinetti, T. Maurizio, et al., Overview of the evalita 2018 hate speech detection task, in: Ceur workshop proceedings, volume 2263, CEUR, 2018, pp. 1–9.

[18] A. T. Cignarella, S. Frenda, V. Basile, C. Bosco, V. Patti, P. Rosso, Overview of the EVALITA 2018 task on irony detection in Italian tweets (IronITA), in: Proceedings of the Sixth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian (EVALITA 2018) co-located with the Fifth CLiC-it, volume 2263, 2018, pp. 1–6.

[19] A. T. Cignarella, M. Lai, C. Bosco, V. Patti, R. Paolo, et al., Sardistance@ evalita2020: Overview of the task on stance detection in italian tweets, in: Proceedings of the Seventh Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2020), Ceur, 2020, pp. 1–10.

[20] K. Krippendorff, Computing Krippendorff's Alpha-Reliability, Technical Report, University of PennSylvania, 2011. URL: https://www.asc.upenn.edu/sites/default/files/2021-03/Computing%20Krippendorff%27s%20Alpha-Reliability.pdf.

[21] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[22] K. Clark, M.-T. Luong, Q. V. Le, C. D. Manning, Electra: Pre-training text encoders as discriminators rather than generators, in: International Conference on Learning Representations, 2020, pp. 1–14. URL: https://openreview.net/forum?id=r1xMH1BtvB.

[23] M. Lai, F. Celli, A. Ramponi, S. Tonelli, C. Bosco, V. Patti, Haspeede3 at evalita 2023: Overview of the political and religious hate speech detection 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, pp. 1–8. URL: https://ceur-ws.org/Vol-3473/paper22.pdf.

[24] 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, pp. 1–8. URL: https://ceur-ws.org/Vol-3473/paper26.pdf.