Handling Disagreement in Hate Speech Modelling

Petra Kralj Novak[1,2], Teresa Scantamburlo[3], Andraž Pelicon[2,4], Matteo Cinelli[5], Igor Mozetič[2], and Fabiana Zollo[3(✉)]

[1] Central European University, Vienna, Austria
novakpe@ceu.edu
[2] Jožef Stefan Institute, Ljubljana, Slovenia
{andraz.pelicon,igor.mozetic}@ijs.si
[3] Ca' Foscari University, Venice, Italy
{teresa.scantamburlo,fabiana.zollo}@unive.it
[4] Jožef Stefan International Postgraduate School, Ljubljana, Slovenia
[5] Sapienza University, Rome, Italy
matteo.cinelli@uniroma1.it

Abstract. Hate speech annotation for training machine learning models is an inherently ambiguous and subjective task. In this paper, we adopt a perspectivist approach to data annotation, model training and evaluation for hate speech classification. We first focus on the annotation process and argue that it drastically influences the final data quality. We then present three large hate speech datasets that incorporate annotator disagreement and use them to train and evaluate machine learning models. As the main point, we propose to evaluate machine learning models through the lens of disagreement by applying proper performance measures to evaluate both annotators' agreement and models' quality. We further argue that annotator agreement poses intrinsic limits to the performance achievable by models. When comparing models and annotators, we observed that they achieve consistent levels of agreement across datasets. We reflect upon our results and propose some methodological and ethical considerations that can stimulate the ongoing discussion on hate speech modelling and classification with disagreement.

Keywords: Hate speech · Annotator agreement · Diamond standard evaluation

1 Introduction

Modern research in machine learning (ML) is driven by large datasets annotated by humans via crowdsourcing platforms or spontaneous online interactions [5].

The authors acknowledge financial support from the EU REC Programme (2014–2020) project IMSyPP (grant no. 875263), the Slovenian Research Agency (research core funding no. P2-103), and from the project "IRIS: Global Health Security Academic Research Coalition".

D. Ciucci et al. (Eds.): IPMU 2022, CCIS 1602, pp. 681–695, 2022.
https://doi.org/10.1007/978-3-031-08974-9_54

Most annotation projects assume that a single preferred or even correct annotation exists for each item—the so-called "gold standard". However, this reflects an idealisation of how humans perceive and categorize the world. Virtually, all annotation projects encounter numerous cases in which humans disagree. The reasons behind disagreement can be various. For example, people can disagree because of accidental mistakes or misunderstandings experienced during the annotation process. In other cases, disagreement can originate from the inherent ambiguity of the annotation task or the annotators' subjective beliefs.

When labels represent different (subjective) views, ignoring this diversity creates an arbitrary target for training and evaluating models: If humans cannot agree, why would we expect the correct answer from a machine to be any different [7]? And, if the machine is able to learn an artificial gold standard, would it make it a perfect (infallible) predictor? The acknowledgement of multiple perspectives in the production of ground truth stimulated a reconsideration of the classical gold standard and the growth of a new research field developing alternative approaches. A recent work proposed a data perspectivist approach to ground truthing and suggested a spectrum of possibilities ranging from the traditional gold standard to the so-called "diamond standard", in which multiple labels are kept throughout the whole ML pipeline [3]. It has also been observed that training directly from *soft labels* (i.e., distributions over classes) can achieve higher performance than training from aggregated labels under certain conditions (e.g., large datasets and high quality annotators) [24]. Studies in hate speech classification came to similar conclusions and showed that supervised models informed by different perspectives on the target phenomena outperform a baseline represented by models trained on fully aggregated data [1].

In this paper, we focus on hate and offensive speech detection, which, similarly to other tasks like sentiment analysis, is inherently subjective. Thus, a disagreement between human annotators is not surprising. In sentiment analysis, disagreement ranges between 40–60% for low quality annotations, and between 25–35% even for high quality annotations [13,17]. Until recently, the subjectivity factor has been largely ignored in favor of a gold standard [26,27]. This led to a dramatic overestimation of model performance on human-facing ML tasks [12]. Here we investigate the specifics of hate speech annotation and modelling through the development of three large hate speech datasets and respective ML models. We present the process for data collection and annotation, the training of state-of-the-art ML models and the results achieved during the evaluation step. Our approach is characterized by two elements. First, we embrace disagreement among annotators in all phases of the ML pipeline and use a diamond standard for model training and evaluation. Second, we evaluate annotators' and models' performance through the lens of disagreement by applying the same performance measures to different comparisons (inter-annotator, self-agreement, and annotator vs model). Our experience led us to reflect and discuss a variety of methodological and ethical implications of handling multiple (conflicting) perspectives in hate speech classification. We conclude that disagreement is a genuine and crucial component of hate speech modelling and needs greater consideration within the ML community. A carefully designed annotation procedure

supports the study of annotators' disagreement, discerns authentic dissent from spurious differences, and collects additional information that could possibly justify or contextualize the annotators' opinion. Moreover, a greater awareness of disagreement in hate speech datasets can generate more realistic expectations on the performance and limits of the ML models used to make decisions about the toxicity of online contents.

The paper is structured as follows. Section 2 presents the annotation process resulting in three large diamond standard hate speech datasets. Section 3 describes our training and evaluation of neural network-based models from diamond standard data, and reports the results by comparing the models' performance to the annotators' agreement. Finally, in Sect. 4, starting from our own results and experience, we discuss some implications of addressing disagreement in hate speech.

2 Data Selection and Annotation

Annotation campaign design and management drastically influences the quality of the annotated data. In this section, we first introduce the annotation schema used for annotating over 180,000 social media items in three different languages (English, Italian, and Slovenian). Then, we describe our annotation campaign and describe the procedure used to monitor and evaluate the annotation progress.

2.1 Annotation Schema

A simple and intuitive annotation schema facilitates the annotation efforts, and reduces possible errors and misunderstandings. However, since the definition of hate speech is a subtle issue there are other possible categorizations—see [18] for a systematic review. The annotation schema presented in this paper is adapted from the OLID [26] and FRENK [16] schemas, yet it is simpler, while retaining most of their expressiveness. The annotation procedure consists of two steps: first, the type of hate speech is determined, then the target of hate speech, when relevant, is identified. We distinguish between the following four **speech types**:

- **Acceptable**: does not present inappropriate, offensive or violent elements.
- **Inappropriate**: contains terms that are obscene or vulgar; but the text is not directed at any specific target.
- **Offensive**: includes offensive generalizations, contempt, dehumanization, or indirect offensive remarks.
- **Violent**: threatens, indulges, desires or calls for physical violence against a target; it also includes calling for, denying or glorifying war crimes and crimes against humanity.

In the case of offensive or violent speech, the annotation schema also includes a target. There are ten pre-specified targets: Racism, Migrants, Islamophobia,

Antisemitism, Religion (other), Homophobia, Sexism, Ideology, Media, Politics, Individual, and Other. For Italian, an additional "North vs. South" target was included (see Sect. 4.1.). We used the same schema to annotate three datasets: English YouTube, Italian YouTube, and Slovenian Twitter (see Table 1).

Table 1. Description of the datasets used for model training and evaluation. There are data sources, topics covered, timeframe, and the number of annotated items in the training and evaluation sets.

Language	Source	Topic	Period	Training set	Evaluation set
English	**YouTube**	Covid-19	Feb 2020 – May 2020	51,655	10,759
Italian	**YouTube**	Covid-19	Jan 2020 – May 2020	59,870	10,536
Slovenian	**Twitter**	General	Dec 2017 – Oct 2020	50,000	10,000

2.2 Data Selection and Annotation Setup

For each language, we selected two separate sets of data for annotation to be used for training and evaluating machine learning models. To overcome the class-imbalance problem (most hate speech datasets are highly unbalanced [20], see also Table 2), the training data selection was optimized to get hate speech-rich training datasets. This was achieved by selecting the data from large collections based on simple classifiers trained on publicly available hate speech data: we used the FRENK data [16] for Slovenian and English, and a dataset of hate speech against immigrants for Italian [22]. This led to training datasets with about two times more violent hate speech (the minority class) than we would get from a random sample. The evaluation dataset was randomly sampled from a period strictly following the training data time-span.

Table 2. Distribution of hate speech classes across the three application datasets. There is the total size of the collected data, and the classes assigned by the hate speech classification models.

Dataset	No. of tweets/ YT comments	Acceptable	Inappropriate	Offensive	Violent
English YouTube	20,227,765	13,670,748 (67.58%)	226,774 (1.12%)	6,222,405 (30.76%)	107,838 (0.53%)
Italian YouTube	1,273,936	1,047,056 (82.19%)	50,949 (4.00%)	164,600 (12.92%)	11,331 (0.89%)
Slovenian Twitter	12,961,136	9,721,259 (75.00%)	109,348 (0.84%)	3,115,207 (24.03%)	15,322 (0.12%)

Annotators were recruited and selected in Slovenia and Italy. Excellent knowledge of the target language (native speakers of Slovenian and Italian and proficient users of English) as well as an interest in social media and hate speech problems were required. Annotators were provided with written annotations guidelines[1] in their mother tongue. Guidelines included a description of the labels and the instructions on how to select them. They also provided practical information about the annotation interface and contact information to be used in case of doubts or requests. We provided continuous support to the annotators through online meetings and a dedicated group on Facebook.

Based on the number of annotators, we distributed the data according to the following constraints:

- Each social media item should be annotated twice.
- Each annotator gets roughly the same number of items.
- All pairs of annotators have approximately the same overlap (in the number of items) for pair-wise annotator agreement computation.
- For Twitter, each annotator is assigned some items (tweets) twice to compute self-agreement.
- For YouTube: a) Threads (all comments to a video) are kept intact; b) Each annotator is assigned both long and short threads.

Such a careful distribution of work enables continuous monitoring and evaluation of the annotation progress and quality. The annotators were working remotely on their own schedule. Internal deadlines were set to discourage procrastination. We monitored the annotation progress by keeping track of the number of completed annotations and evaluating the self- and inter-annotator agreement measures (see Sect. 3.1). Agreement between (pairs of) annotators

Table 3. The annotator agreement and overall model performance. Two measures are used: Krippendorff's (ordinal) *Alpha* and accuracy (*Acc*). The first column is the self-agreement of individual annotators (available for Twitter data only), and the second column is the aggregated inter-annotator agreement between different annotators. The last two columns are the model evaluation results, on the training and the out-of-sample evaluation sets, respectively. Note that the overall model performance is comparable to the inter-annotator agreement.

	Agreement				Classification model			
Dataset	Self-agreement		Inter-annotator		Training set		Evaluation set	
	Alpha	*Acc*	*Alpha*	*Acc*	*Alpha*	*Acc*	*Alpha*	*Acc*
English YouTube	–	–	0.60	0.78	0.55	0.75	0.60	0.83
Italian YouTube	–	–	0.59	0.78	0.60	0.79	0.58	0.84
Slovenian Twitter	0.79	0.88	0.60	0.79	0.61	0.80	0.57	0.80

[1] Hate speech annotation guidelines in English are available as part of IMSyPP D2.1: http://imsypp.ijs.si/wp-content/uploads/IMSyPP-D2.1-Hate-speech-DB-2.pdf, starting from page 16.

(see Table 3) was regularly computed during the process, enabling early detection of poorly-performing annotators, i.e., annotators disagreeing systematically with other annotators, either due to misunderstanding of the task, not following the guidelines or not devoting enough attention.

We used the described schema and protocol for developing three diamond standard datasets, and made them available on the Clarin repository: English YouTube[2], Italian YouTube[3], and Slovenian Twitter[4], summarized in Table 1. In the Slovenian dataset, the tweets are annotated independently, while the English and Italian datasets include contextual information in the form of threads of YouTube comments: Every comment is annotated for hate speech, yet the annotators were also given the context of discussion threads. Furthermore, the YouTube datasets are focused on the COVID-19 pandemic topic.

3 Model Training and Evaluation

We used the three diamond standard datasets to train and evaluate machine learning hate speech models. For each dataset, a state-of-the-art neural model based on a Transformer language model was trained end-to-end [6] to distinguish between the four speech classes. The models were trained directly on the diamond standard data, i.e., the training examples were repeated with several equal or disagreeing labels. For Italian, we used AlBERTo [19], a BERT-based language model pre-trained on a collection of tweets in the Italian language. For English, the base version of English BERT with 12 Transformer blocks [6] was used. For Slovenian, a trilingual CroSloEng-BERT [23], which was jointly pretrained on Slovenian, Croatian and English languages, was used. All three models are available at the IMSyPP project HuggingFace repository[5].

We used the Italian and Slovenian models in two previous analytical studies on hate speech in social media. The Italian model was used in a work investigating relationships between hate speech and misinformation sources on the Italian YouTube [4]. The Slovenian model was used to perform an analysis on the evolution of retweet communities, hate speech and topics on the Slovenian Twitter during 2018–2020 [8–10].

3.1 Evaluation Measures

A distinctive aspect of our approach is to apply the same measures a) to estimate the agreement between the human annotators and b) to estimate the agreement between the results of model classification and the manually annotated data. There are several measures of agreement, and to get robust estimates from different problem perspectives, we apply three well-known measures from the fields

[2] English dataset: https://www.clarin.si/repository/xmlui/handle/11356/1454.
[3] Italian dataset: https://www.clarin.si/repository/xmlui/handle/11356/1450.
[4] Slovenian dataset: https://www.clarin.si/repository/xmlui/handle/11356/1398.
[5] IMSyPP HuggingFace model repository: https://huggingface.co/IMSyPP.

of inter-rater agreement and machine learning: Krippendorff's *Alpha*, accuracy (*Acc*) and F_1 score.

There are several properties of hate speech modelling that require special treatment: i) The four speech types are ordered, from normal to the most hateful, violent speech, and therefore disagreements have very different magnitudes, thus we use ordinal Krippendorff's *Alpha*; ii) The four speech classes are severely imbalanced, a further reason to use Krippendorff's *Alpha*; iii) Since we also need a class-specific measure of (dis)agreement, F_1 is used.

The speech types are modelled by a discrete, ordered 4-valued variable $c \in C$, where $C = \{A, I, O, V\}$, and $A \prec I \prec O \prec V$. The values of c denote acceptable speech (abbreviated A), inappropriate (I), offensive (O) or violent (V) hate speech. The data items that are labelled by speech types are either individual YouTube comments or Twitter posts. The data labeled by different annotators is represented in a reliability data matrix. The data matrix is a n-by-m matrix, where n is the number of items labeled, and m is the number of annotators. An entry in the matrix is a label $c_{iu} \in C$, assigned by the annotator $i \in \{1, \ldots, m\}$ to the item $u \in \{1, \ldots, n\}$. The data matrix does not have to be full, i.e., some items might not be labelled by all the annotators.

A **coincidence matrix** is constructed from the reliability data matrix. It tabulates all the combined values of c from two different annotators. The coincidence matrix is a k-by-k square matrix, where $k = |C|$, the number of possible values of C, and has the following form:

	c'			$\sum$
	.	.	.	.
c	.	$N(c, c')$	.	$N(c)$
	.	.	.	.
$\sum$	.	$N(c')$	.	N

An entry $N(c, c')$ accounts for all coincidences from all pairs of annotators for all the items, where one annotator has assigned a label c and the other c'. $N(c)$ and $N(c')$ are the totals for each label, and N is the grand total. The coincidences $N(c, c')$ are computed as:

$$N(c, c') = \sum_u \frac{N_u(c, c')}{m_u - 1} \quad c, c' \in C$$

where $N_u(c, c')$ is the number of (c, c') pairs for the item u, and m_u is the number of labels assigned to the item u. When computing $N_u(c, c')$, each pair of annotations is considered twice, once as a (c, c') pair, and once as a (c', c) pair. The coincidence matrix is therefore symmetrical around the diagonal, and the diagonal contains all the matching labelling.

We can now define the three evaluation measures that we use to quantify the agreement between the annotators, as well as the agreement between the model and the annotators. Since the annotators might disagree on the labels, there is no "gold standard". The performance of the model can thus only be compared to a (possibly inconsistent) labelling by the annotators.

Krippendorff's *Alpha*[14] is defined as follows:

$$Alpha = 1 - \frac{D_o}{D_e},$$

where D_o is the actual disagreement between the annotators, and D_e is disagreement expected by chance. When annotators agree perfectly, $Alpha = 1$, when there is a baseline agreement as expected by chance, $Alpha = 0$, and when the annotators disagree systematically, $Alpha < 0$. The two disagreement measures, D_o and D_e, are defined as:

$$D_o = \frac{1}{N}\sum_{c,c'} N(c,c') \cdot \delta^2(c,c'), \qquad D_e = \frac{1}{N(N-1)}\sum_{c,c'} N(c) \cdot N(c') \cdot \delta^2(c,c').$$

The arguments $N(c,c'), N(c), N(c')$ and N refer to the values in the coincidence matrix, constructed from the labeled data.

$\delta(c,c')$ is a difference function between the values of c and c', and depends on the type of decision variable c (nominal, ordinal, interval, etc.). In our case, c is an ordinal variable, and δ is defined as:

$$\delta(c,c') = \sum_{i=c}^{i=c'} N(i) - \frac{N(c) + N(c')}{2} \quad e.g., \quad c, c', i \in \{1^{st}, 2^{nd}, 3^{rd}, 4^{th}\}.$$

Accuracy (*Acc*) is a common, and the simplest, measure of performance of the model which measures the agreement between the model and the "gold standard". However, it can be also used as a measure of agreement between two annotators. *Acc* is defined in terms of the observed disagreement D_o:

$$Acc = 1 - D_o = \frac{1}{N}\sum_{c} N(c,c).$$

Accuracy does not account for the (dis)agreement by chance, nor for the ordering of hate speech classes. Furthermore, it can be deceiving in the case of unbalanced class distribution.

F-score (F_1) is an instance of a well-known effectiveness measure in information retrieval [25] and is useful for binary classification. In the case of multi-class problems, it can be used to measure the performance of the model to identify individual classes. $F_1(c)$ is the harmonic mean of precision (*Pre*) and recall (*Rec*) for class c:

$$F_1(c) = 2 * \frac{Pre(c) * Rec(c)}{Pre(c) + Rec(c)}.$$

In the case of a coincidence matrix, which is symmetric, the 'precision' and 'recall' are equal, since false positives and false negatives are both cases of disagreement. $F_1(c)$ thus degenerates into:

$$F_1(c) = \frac{N(c,c)}{N(c)}.$$

In terms of the annotator agreement, $F_1(c)$ is the fraction of equally labelled items out of all the items with label c.

3.2 Annotator Agreement and Model Performance

For the evaluation, we use the same measures to estimate the agreement between the human annotators, and the agreement between the model classification and the manually annotated diamond standard data. Table 3 summarizes the overall annotator agreement and the models' performance in terms of Krippendorff's (ordinal) *Alpha* and accuracy (*Acc*) on all three datasets.

The annotators agree on the hate speech label on nearly 80% of the data points ($Acc = 0.78$–0.79). Our models agree with at least one annotator in over 80% of the cases ($Acc = 0.80$–0.84). Considering the high class imbalance and the ordering of the hate speech classes, a comparison in terms of Krippendorff's (ordinal) *Alpha* is more appropriate: Table 3 shows a very consistent agreement of about 0.6 ($Alpha = 0.55$–0.60) both between the annotators and the models on all three datasets.

The very misleading performance estimates as computed by accuracy are evident from Table 4. We consider two cases of binary classification. In the first case, all three types of speech which are not acceptable (e.g., inappropriate, offensive, or violent) are merged into a single, unacceptable class. In the second case, all types of speech which are not violent (e.g., acceptable, inappropriate, or offensive) are merged into a non-violent class. The performance of such binary classification is then estimated by *Alpha* and *Acc*. The estimates in the first case are comparable to the results in Table 3. In the second case, however, the *Alpha* values drop considerably, while the *Acc* scores rise to almost 100% ($Acc = 0.97$–0.99). This is due to a high imbalance of the non-violent vs. violent items, with a respective ratio of more than 99:1. The *Alpha* score, on the other hand, indicates that the model performance is low, barely above the level of classification by chance ($Alpha = 0.26$–0.39 on the evaluation set).

Class-specific results comparing the model and the annotator agreement in terms of F_1 are available in Table 5. The F_1 scores of the models would in absolute sense not be considered high. Yet they are comparable and in many cases even higher than the F_1 scores between the annotators. The only exception (still consistent in all three datasets) is the relatively low models' performance for the violent class. This is consistent with the binary classification results (Non-violent vs. Violent) in Table 4. We hypothesise, with high degree of confidence, that a poor identification of the violent class is due to the scarcity of training examples.

Table 4. The annotator agreement and model performance for two cases of binary classification: Acceptable (A) vs. Unacceptable class (either I, O, or V), and Violent (V) vs. Non-violent class (either A, I, or O). The performance is measured by the *Alpha* and accuracy (*Acc*) scores. Note the very high and misleading *Acc* scores for the second case, where the class distribution between the Violent and Non-violent classes is highly imbalanced. The *Alpha* scores, on the other hand, are very low, barely above the level of classification by chance.

Agreement		Acceptable vs. Unacceptable		Non-violent vs. Violent	
Dataset	Model	*Alpha*	*Acc*	*Alpha*	*Acc*
Inter-annotator		0.59	0.80	0.54	0.98
English	Train. set	0.55	0.77	0.45	0.98
YouTube	Eval. set	0.60	0.84	0.29	0.99
Inter-annotator		0.60	0.82	0.61	0.98
Italian	Train. set	0.62	0.83	0.51	0.97
YouTube	Eval. set	0.59	0.87	0.39	0.99
Self-agreement		0.79	0.90	0.69	0.99
Inter-annotator		0.60	0.81	0.61	0.99
Slovenian	Train. set	0.62	0.82	0.24	0.99
Twitter	Eval. set	0.57	0.81	0.26	0.99

Table 5. Class-specific annotator agreement and model performance. The classification is done into four hate speech classes (A, I, O, V), and the performance is measured by the F_1 score for each class individually. Note a relatively low model performance for the Violent class (F_1(V)).

Agreement		Acceptable	Inappropriate	Offensive	Violent
Dataset	Model	F_1(A)	F_1(I)	F_1(O)	F_1(V)
Inter-annotator		0.82	0.32	0.75	0.55
English	Train. set	0.78	0.39	0.74	0.46
YouTube	Eval. set	0.89	0.25	0.69	0.30
Inter-annotator		0.86	0.52	0.63	0.62
Italian	Train. set	0.87	0.53	0.65	0.53
YouTube	Eval. set	0.92	0.59	0.58	0.39
Self-agreement		0.92	0.62	0.85	0.69
Inter-annotator		0.85	0.48	0.71	0.62
Slovenian	Train. set	0.85	0.52	0.73	0.25
Twitter	Eval. set	0.86	0.46	0.69	0.26

4 Discussion

Given the intrinsically subjective nature of judging offensive and violent content, it might be argued that a diamond standard should be preferred in this and other

similar contexts through all the phases of the machine learning pipeline. In the following, we discuss methodological and ethical implications of this approach.

4.1 Methodological Implications

Working with diamond standard data influences the data annotation, machine learning training and evaluation. We argue that selecting the data to be annotated, setting up the annotation campaign, monitoring its execution and evaluating the quality of annotations during and after the annotation campaign, are crucial steps that influence the final quality of the annotated data. Yet, the importance of annotation campaigns is often neglected in machine learning pipelines. An important practical dilemma when building diamond standard datasets is still to be investigated: when faced with an intrinsically subjective task (e.g., hate speech detection, sentiment analysis) how should one decide upon how many facets should a diamond have vs. how large should it be? The more diamond faces (i.e., the number of labels per item) ensure better data quality and enable the identification of ambiguous cases. Yet, when limited with the number of labels an annotation campaign can afford, is it better to have more data items labeled (thus a larger dataset with more variety) or more labels to the same items? Is this trade-off the same for the training as well as for the evaluation set?

Our second focus is on model evaluation: we propose a perspectivist view, as we evaluate model performance through the lens of disagreement by applying the same, proper performance measures to evaluate the annotator agreement and the model quality. Standard metrics assume a different meaning in a context where the same object can be assigned to multiple legitimate labels. For example, precision and recall lose the asymmetry that is implicitly assumed between the outcome retrieved from direct observation (also called 'real' outcome) and the prediction provided by the ML models, as we show in Sect. 3.1. In the case of ordered labels (e.g., our speech labels), mutual information, proposed by [24] as a good evaluation measure when learning with disagreement, is not appropriate as it neglects the labels' ordering. Proper performance measures in our case include ordinal Krippendorff's *Alpha*, which accommodates both the ordered nature of the labels (from normal to the most hateful, violent speech, and consequently a varying magnitude of disagreements), and class imbalance (where the Violent class is underrepresented). Furthermore, we use F_1 for the estimation of class-specific disagreement and misclassification, but not macro-F_1. Macro-F_1 is not an appropriate measure to aggregate individual F_1 scores to estimate the overall model performance [11].

In our perspectivist view on model evaluation, model performance is closely tied to the agreement between annotators. This means that annotator agreement poses intrinsic limits to the performance achievable by the ML models. This is implemented by the use of the same measures for all comparisons (e.g., between the annotators and between the annotators and the model). We observed that the level of agreement between our models and the annotators reaches the inter-annotator agreement when applying the overall performance measure (ordinal

Krippendorff's *Alpha*). This indicates that the model is limited by the annotator agreement and can not be drastically improved. However, when considering the class-specific F_1 values, the model reaches the inter-annotator agreement in all classes except for the minority class (i.e., Violent). Without a comparison to the F_1 scores of the annotators, or binary classification Non-violent vs. Violent, this shortcoming of the classification model would not have been detected.

4.2 Ethical Implications

The problem of ground truthing in hate speech modelling has also some ethical and legal implications. Even though the perception and interpretation of offensive and violent speech can vary among people and cultures, it is also true that the lack of respect is a moral violation and can have tangible negative effects on subjects. Some people, for example, can suffer from depression or even physical injuries after being largely exposed to violent and offensive communication [21]. In this regard, many countries impose restrictions to protect individuals from discriminatory and threatening content and digital platforms strive for the limitation of hate speech.

Defining hate speech subsumes important decisions about the ethical and legal boundaries of public debates and bears responsibility for limiting the right of freedom of expression, thereby including or excluding people from democratic participation. Not surprisingly, the introduction of legal boundaries to remove hate speech from the public sphere has raised various criticisms. For example, some consider hate speech bans as a form of paternalism, incompatible with the assumption that humans are responsible and autonomous individuals, while others fear that the power of judging hate speech would put the state in a position to decide what can or cannot be said [2].

The tension between the right to safety and the right to freedom of expression becomes even more controversial when one deals with ML models for hate speech detection and removal. In this context, the decision as to whether accepting or rejecting a potentially harmful content leverages the capacity of ML algorithms to make accurate predictions. However, our results and other studies (e.g. [12]) suggest that measuring hate speech classification in terms of prediction accuracy can be elusive when annotators disagree: a classifier cannot be accurate when the data is inconsistent due to many conflicting views. Deliberating upon items that cannot be classified in a clear-cut way is a questionable practice and requires greater scrutiny among ML developers, managers and policy makers. Achieving a consensus in predictive tasks might not necessarily be an ideal outcome. On the contrary, diversity can improve collective predictions [15]. Moreover, if predictions are accompanied by additional information including the reasons behind the predictions, cultivating a positive disagreement can foster more fruitful judgments.

5 Conclusions and Future Work

In this paper, we adopt a perspectivist approach to data annotation, model training and evaluation of hate speech classification. Our first emphasis is on the annotation process leading to the diamond standard data, as we argue that it influences the final data quality, and thereof the machine learning model quality. As the main point, we propose a perspectivist view on model evaluation, as we evaluate model performance through the lens of disagreement by applying the same, proper performance measures to evaluate the annotator agreement and the model quality. We argue that annotator agreement poses intrinsic limits to the performance achievable by models. By following the same annotation protocol, model training and evaluation, we developed three large scale hate speech datasets and the corresponding machine learning models. All our results are consistent across the three datasets: Trained and reliable annotators disagree in about 20% of the cases, model performance reaches the annotator agreement in the overall evaluation, while for the minority class (Violent) there is still some room for improvement. A broad reflection on the role of disagreement in hate speech detection leads us to consider some methodological and ethical implications that could stimulate the ongoing debate, not limited to hate speech modelling but to subjective classification tasks where disagreement is likely to arise and make a difference.

References

1. Akhtar, S., Basile, V., Patti, V.: Modeling annotator perspective and polarized opinions to improve hate speech detection. In: Proceedings AAAI Conference on Human Computation and Crowdsourcing, vol. 8, pp. 151–154 (2020)
2. Anderson, L., Barnes, M.: Hate speech. In: Zalta, E.N. (ed.) The Stanford Encyclopedia of Philosophy. Metaphysics Research Lab Stanford University (2022)
3. Basile, V., Cabitza, F., Campagner, A., Fell, M.: Toward a perspectivist turn in ground truthing for predictive computing. arXiv:2109.04270 (2021)
4. Cinelli, M., Pelicon, A., Mozetič, I., Quattrociocchi, W., Novak, P.K., Zollo, F.: Dynamics of online hate and misinformation. Sci. Rep. **11**(1), 1–12 (2021). https://doi.org/10.1038/s41598-021-01487-w
5. Cristianini, N., Scantamburlo, T., Ladyman, J.: The social turn of artificial intelligence. AI Soc. 1–8 (2021). https://doi.org/10.1007/s00146-021-01289-8
6. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv:1810.04805 (2018)
7. Dumitrache, A., Aroyo, L.,Welty, C.: A crowdsourced frame disambiguation corpus with ambiguity. In: Proceedings of NAACL (2019)
8. Evkoski, B., Ljubešić, N., Pelicon, A., Mozetič, I., Kralj Novak, P.: Evolution of topics and hate speech in retweet network communities. Appl. Netw. Sci. **6**(1), 1–20 (2021). https://doi.org/10.1007/s41109-021-00439-7
9. Evkoski, B., Mozetič, I., Ljubešić, N., Novak, P.K.: Community evolution in retweet networks. PLoS One **16**(9), e0256175 (2021). https://doi.org/10.1371/journal.pone.0256175,Non-anonymized version available at arXiv:2105.06214

10. Evkoski, B., Pelicon, A., Mozetič, I., Ljubešić, N., Novak, P.K.: Retweet communities reveal the main sources of hate speech. PLoS ONE **17**(3), e0265602 (2022). https://doi.org/10.1371/journal.pone.0265602
11. Flach, P., Kull, M.: Precision-recall-gain curves: PR analysis done right. In: Cortes, C., Lawrence, N.D., Lee, D.D., Sugiyama, M., Garnett, R. (eds.) Advances in Neural Information Processing Systems, pp. 838–846. Curran Associates (2015)
12. Gordon, M.L., Zhou, K., Patel, K., Hashimoto, T., Bernstein, M.S.: The disagreement deconvolution: bringing machine learning performance metrics in line with reality. In: Proceedings CHI Conference on Human Factors in Computing Systems, pp. 1–14 (2021)
13. Kenyon-Dean, K., et al.: Sentiment analysis: It's complicated! In: Proceedings of NAACL, pp. 1886–1895 (2018)
14. Krippendorff, K.: Content Analysis, An Introduction to its Methodology. Sage Publications, 4th edn. (2018)
15. Landemore, H., Page, S.E.: Deliberation and disagreement: problem solving, prediction, and positive dissensus. Politics Philos. Econ. **14**(3), 229–254 (2015)
16. Ljubešić, N., Fišer, D., Erjavec, T.: The FRENK datasets of socially unacceptable discourse in Slovene and English (2019), arXiv:1906.02045
17. Mozetič, I., Grčar, M., Smailović, J.: Multilingual Twitter sentiment classification: the role of human annotators. PLoS One**11**(5), e0155036 (2016). https://doi.org/10.1371/journal.pone.0155036
18. Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., Patti, V.: Resources and benchmark corpora for hate speech detection: a systematic review. Lang. Res. Eval. **55**(2), 477–523 (2020). https://doi.org/10.1007/s10579-020-09502-8
19. Polignano, M., Basile, P., De Gemmis, M., Semeraro, G., Basile, V.: AlBERTo: Italian BERT language understanding model for NLP challenging tasks based on tweets. In: Italian Conference on Computational Linguistics, vol. 2481, pp. 1–6 (2019)
20. Rathpisey, H., Adji, T.B.: Handling imbalance issue in hate speech classification using sampling-based methods. In: IEEE International Conference on Science in Information Technology), pp. 193–198 (2019)
21. Saha, K., Chandrasekharan, E., De Choudhury, M.: Prevalence and psychological effects of hateful speech in online college communities. In: Proceedings 10th ACM Conference on Web Science, pp. 255–264 (2019)
22. Sanguinetti, M., Poletto, F., Bosco, C., Patti, V., Stranisci, M.: An Italian Twitter corpus of hate speech against immigrants. In: Proceedings of 11th International Conference on Language Resources and Evaluation (2018)
23. Sojka, P., Kopeček, I., Pala, K., Horák, A. (eds.): TSD 2020. LNCS (LNAI), vol. 12284. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58323-1
24. Uma, A.N., Fornaciari, T., Hovy, D., Paun, S., Plank, B., Poesio, M.: Learning from disagreement: a survey. Artif. Intell. Res. **72**, 1385–1470 (2021)
25. Van Rijsbergen, C.: Information Retrieval. Butterworth, 2nd edn. (1979)
26. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R.: Predicting the type and target of offensive posts in social media. In: Proceedings of NAACL-HLT, pp. 1415–1420 (2019)
27. Zampieri, M., et al.: SemEval-2020 task 12: Multilingual offensive language identification in social media. arXiv:2006.07235 (2020)