Agreeing to Disagree: Annotating Offensive Language Datasets with Annotators' Disagreement

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio
Marco Guerini, Sara Tonelli
Fondazione Bruno Kessler, Trento, Italy
`{eleonardelli,menini,aprosio,guerini,satonelli}@fbk.eu`

Abstract

Since state-of-the-art approaches to offensive language detection rely on supervised learning, it is crucial to quickly adapt them to the continuously evolving scenario of social media. While several approaches have been proposed to tackle the problem from an algorithmic perspective, so to reduce the need for annotated data, less attention has been paid to the quality of these data. Following a trend that has emerged recently, we focus on the level of agreement among annotators while selecting data to create offensive language datasets, a task involving a high level of subjectivity. Our study comprises the creation of three novel datasets of English tweets covering different topics and having five crowd-sourced judgments each. We also present an extensive set of experiments showing that selecting training and test data according to different levels of annotators' agreement has a strong effect on classifiers performance and robustness. Our findings are further validated in cross-domain experiments and studied using a popular benchmark dataset. We show that such hard cases, where low agreement is present, are not necessarily due to poor-quality annotation and we advocate for a higher presence of ambiguous cases in future datasets, particularly in test sets, to better account for the different points of view expressed online.

1 Introduction

When creating benchmarks for NLP tasks through crowd-sourcing platforms, it is important to consider possible issues with inter-annotator agreement. Indeed, crowd-workers do not necessarily have a linguistic background and are not trained to perform complex tasks, thus jeopardizing benchmark quality. Furthermore, some crowd-workers try to maximize their pay by supplying quick answers that have nothing to do with the correct label. This issue has been tackled in the past by proposing approaches to control for annotators' expertise and reliability (Hovy et al., 2013), trying to identify spammers and mitigate their effect on annotation, or by repeating labeling on targeted examples (Sheng et al., 2008). However, not all tasks are the same: while in some cases, like for instance PoS-tagging or parsing, disagreement among annotators is more likely due to unclear annotation guidelines and can usually be reconciled through adjudication, full annotators' agreement should not be necessarily enforced in *social computing tasks*, whose goal is to study and manage social behavior and organizational dynamics, especially in virtual worlds built over the Internet (Wang, 2007). In these tasks – which include offensive language detection among others – subjectivity, bias and text ambiguity play an important role (Aroyo et al., 2019), and being an inherent component of the task they should be measured and analysed rather than discarded (Klenner et al., 2020; Basile, 2020). Indeed, instead of aiming for a global consensus on what constitutes verbal abuse on social media, we investigate the impact of different degrees of disagreement, how classifiers behave with ambiguous training and test data, and the role of disagreement in current shared tasks. More specifically, we first collect and annotate three datasets of English tweets covering different domains, to test if agreement among a pool of generic classifiers can be considered a proxy for annotator agreement. We then focus on how annotator agreement (both in training and test set) impacts classifiers' performance, considering domain-specific and generic classifiers as well as in-domain and out-of-domain experiments. We also show that low agreement examples – no matter how difficult they can be – still provide useful signal for training offensive language detection systems and do not represent random annotations. So "coin-flipping" or example removal seems not to be the right strategy to solve these disagreement cases. Then, we measure disagreement in the English test set of the last Offenseval shared task (Zampieri et al., 2020),

and analyse to what extent the high performance achieved by most participating systems is related to high agreement in annotation.

We release the new annotated datasets upon request,[1] including more than 10k tweets covering three domains. The messages have been labeled with 50k crowd-worker judgements and annotated with agreement levels. To our knowledge, this represents the first dataset explicitly created to cover different agreement levels in a balanced way.We also advocate for the release of more datasets like the one we propose, especially for highly subjective tasks, where the need to include different points of view should be accounted for.

NOTE: This paper contains examples of language which may be offensive to some readers. They do not represent the views of the authors.

2   Related Work

While there has been an extensive discussion on minimal standards for inter-annotator agreement to ensure data quality (Di Eugenio and Glass, 2004; Passonneau, 2004; Artstein and Poesio, 2008), recently an increasing number of works argue that disagreement is unavoidable because language is inherently ambiguous (Aroyo and Welty, 2015), proposing ways to tackle annotators' disagreement when building training sets (Dumitrache et al., 2019). Hsueh et al. (2009), for example, identify a set of criteria to select informative yet unambiguous examples for predictive modeling in a sentiment classification task. Rehbein and Ruppenhofer (2011) analyse the impact that annotation noise can have on active learning approaches. Other works along this line investigate the impact of uncertain or difficult instances on supervised classification (Peterson et al., 2019), while Beigman Klebanov and Beigman (2014) show that including hard cases in training data results in poorer classification of easy data in a word classification task. Along the same lines, Jamison and Gurevych (2015) show that filtering instances with low agreement improve classifier performance in four out of five tasks. Both works observe that the presence of such instances lead to misclassifications.

Several approaches have been presented that implement strategies to deal with disagreement when training classifiers for diverse tasks. In most cases, disagreement has been treated as a consequence of low annotation quality, and addressed through methodologies aimed at minimising the effects of noisy crowdsourced data. Simpson et al. (2020), for example, present a Bayesian sequence combination approach to train a model directly from crowdsourced labels rather than aggregating them. They test their approach on tasks such as NER where disagreement is mainly due to poor annotation quality. Other works have focused instead on uncertainty in PoS-tagging, integrating annotators' agreement in the modified loss function of a structured perceptron (Plank et al., 2014). Also Rodrigues and Pereira (2018) propose an approach to automatically distinguish the good and the unreliable annotators and capture their individual biases. They propose a novel crowd layer in deep learning classifiers to train neural networks directly from the noisy labels of multiple annotators, using only backpropagation.

Other researchers have suggested to remove hard cases from the training set (Beigman Klebanov and Beigman, 2009) because they may potentially lead to poor classification of easy cases in the test set. We argue instead that disagreement is inherent to the kind of task we are going to address (i.e. offensive language detection) and, in line with recent works, we advocate against forced harmonisation of annotators' judgements for tasks involving high levels of subjectivity (Klenner et al., 2020; Basile, 2020). Among recent proposals to embrace the uncertainty exhibited by human annotators, Gordon et al. (2021) propose a novel metric to evaluate social computing tasks that disentangles stable opinions from noise in crowd-sourced datasets. Akhtar et al. (2020), instead, divide the annotators into groups based on their polarization, so that different gold standard datasets are compiled and each used to train a different classifier.

Compared to existing works, our contribution is different in that we are interested mainly in the process of dataset creation rather in evaluation metrics or classification strategies. Indeed, our research is guided mainly by research questions concerning the data selection process, the composition of datasets and the evaluation using controlled levels of agreement. To this purpose, we create the first dataset for offensive language detection with three levels of agreement and balanced classes, encompassing three domains. This allows us to run comparative in-domain and out-of-domain evaluations, as well as to analyse existing benchmarks like the Offenseval dataset (Zampieri et al., 2020) using the

---

[1] See Ethics Statement section for further details

same approach. While few crowd-sourced datasets for toxic and abusive language detection have been released with disaggregated labels (Davidson et al., 2017), they have not been created with the goal of analysing disagreement, therefore no attention has been paid to balance the number of judgments across different dimensions, like in our case.

3   Data Selection and Annotation

In our study, we focus on three different domains, which have been very popular in online conversations in 2020: Covid-19, US Presidential elections and Black Lives Matter (BLM) movement. After an empirical analysis of online discussions, a set of hashtags and keywords for each domain are defined (e.g. *#covid19*, *#election202*, *#blm*). Then, using Twitter public APIs, tweets in English containing at least one of the above keywords are collected in a time span between January and November 2020 (for more details about data collection see Appendix D). From this data collection, we randomly select 400,000 tweets (around 130,000 for each domain), which we then pre-process by splitting hashtags into words using the Ekphrasis tool (Gimpel et al., 2010) and then replacing all mentions to users and urls with ⟨user⟩ and ⟨url⟩ respectively.

3.1   Ensemble of classifiers to select data for annotation

Since we do not know the real distribution of agreement levels in the data we collected, random sampling for annotation might be a sub-optimal choice. Thus, we developed a strategy to pre-evaluate the tweets, trying to optimize annotators' effort by having a balanced dataset (in fact data might be very skewed leading to over-annotation of some classes and under-annotation of others). To pre-evaluate the tweets we use a heuristic approach by creating an ensemble of 5 different classifiers, all based on the same BERT configuration and fine-tuned starting from the same abusive language dataset (Founta et al., 2018). Since the original dataset contains four classes (`Spam`, `Normal`, `Abusive` and `Hateful`), we first remove the tweets from the `Spam` class and map the remaining ones into a binary `offensive` or `non-offensive` label, by merging `Abusive` and `Hateful` tweets into the `offensive` class and mapping the `Normal` class into the `non-offensive` one. We then select 15k tweets from the Founta dataset (~100k tweets) for speeding up the process, as we are not interested in the overall performance of the different classifiers, but rather in their relative performances. Each classifier of the ensemble is trained using a different balance for the training and the evaluation set, so to yield slightly different predictions. In particular, all five classifiers are trained with the BERT-Base uncased model[2], a max seq length of 64, a batch size of 16 and 15 epochs. One classifier has been trained using 12k tweets in the training and 3k in the validation set, a second classifier was trained using the same training instances but repeated twice (24k), while the validation set remained the same. In a third and fourth configuration, we repeat twice the offensive and the non-offensive training instances respectively. Finally, in a fifth configuration we change the proportion between training and validation set (10k for training, 5k for validation).

The rationale for this choice is twofold: (i) since we will collect 5 crowd-annotations for each tweet, we want to have an intuitive and possible direct comparison between ensemble agreement and annotators' agreement (i.e. five votes per tweet coming from the classifiers and five from crowd-workers). (ii) The dataset in Founta et al. (2018) has been specifically created to encompass several types of offensive language. We can therefore consider it as a general prior knowledge about verbal abuse online before adapting our systems to the 3 domains of interest.

In the following sections we will denote *unanimous agreement* with A^{++} (i.e. agreement between 5 annotators or classifiers), *mild agreement* with A^{+} (i.e. 4 out of 5 annotations agreeing on the same label), and *weak agreement* with A^{0} (i.e. the 5 annotations include 3 of them in agreement and 2 in disagreement). When focusing also on the label we will use the same notation, representing offensive tweets as $O^{++/+/0}$ and non offensive ones as $N^{++/+/0}$ respectively.

The pre-evaluation through the classifier ensemble resulted in the following agreement distribution: about 92% of the data was classified as A^{++}. For about 5% of the data, agreement among the classifiers was A^{+}, while for the remaining 3% of the data, they fell in the A^{0} situation.

---

[2]12-layer, 768-hidden, 12-heads, 110M parameters https://github.com/google-research/bert

3.2 Data Annotation with AMT

In order to analyse the relation between automated and manual annotation with respect to agreement and disagreement, we select an equal number of tweets from each class of agreement of the ensemble (A^{++}, A^+, A^0) to be manually annotated. For each domain and each agreement class we select 1,300 tweets – equally divided between `offensive` and `non-offensive` predictions – for a total of 3,900 tweets per domain.

Every tweet is annotated by 5 native speakers from the US, who we expect to be familiar with the topics, using Amazon Mechanical Turk. We follow for all domains the same annotation guidelines, aimed at collecting crowd-workers' judgements on the offensiveness of the messages using the binary labels `offensive` and `not offensive` (see Guidelines included in Appendix A).

To ensure high-quality annotations, we select a pool of tweets from the three domains of interest and ask three expert linguists to annotate them. The tweets with perfect agreement are used as gold standard. We then include a gold standard tweet in every HIT (group of 5 tweets to be annotated). If a crowd-worker fails to evaluate the gold tweet, the HIT is discarded. Moreover, after the task completion we remove all the annotations done by workers who did not reach a minimum overall accuracy of 70% with respect to the gold standard. As a consequence of this quality control, for some tweets we could not collect five annotations, and they had to be removed from the final dataset. On the other hand, it was a crucial process to minimise the possible impact of spam and low-quality annotations on disagreement – which is the focus of our analysis. The total number of tweets annotated using AMT is 10,753, including 3,472 for Covid-19, 3,490 for US elections and 3,791 for BLM. Some (slightly modified) examples of tweets judged with different levels of agreement by crowd-annotators are reported in Table 1.

3.3 Annotators and Ensemble Agreement

If we use the majority vote for crowd-annotated data, the datasets have an average distribution of 31% of offensive and 69% non-offensive tweets, while it is 50% each according to ensemble annotation we used for sampling. This means that our classifiers tend to label more tweets as offensive compared to human annotators, as shown in the confusion matrix in Fig. 1. It is interesting to note

that, although the tweets to be annotated were selected evenly across classifiers' agreement classes, the agreement between annotators is not uniformly distributed.

As regards annotators' agreement, for about 43% of the tweets annotated we have full consensus between annotators (A^{++}). The vast majority of these tweets were judged unanimously as non-offensive (34,12% N^{++}), and only 8,05% of the data were judged unanimously offensive (O^{++}), the less represented type of agreement. For the remaining data, 29,35% has mild agreement (A^+, 4 out of 5 annotators agreed) with 19% N^+ and 10,35% O^+, and another 28,28% of the data in the class A^0 (3 vs 2 annotators) with 15,56% N^0 and 12,92% O^0.

	Classifiers ensemble						
	N^{++}	N^+	N^0	O^0	O^+	O^{++}	Total
N^{++}	1434 13.34%	695 6.46%	542 5.04%	485 4.51%	369 3.43%	144 1.34%	3669 34.12%
N^+	299 2.78%	390 3.63%	395 3.67%	387 3.60%	373 3.47%	199 1.85%	2043 19.00%
N^0	118 1.10%	284 2.64%	325 3.02%	328 3.05%	313 2.91%	305 2.84%	1673 15.56%
O^0	40 0.37%	173 1.61%	226 2.10%	262 2.44%	294 2.73%	394 3.66%	1389 12.92%
O^+	19 0.18%	101 0.94%	133 1.24%	139 1.29%	212 1.97%	509 4.73%	1113 10.35%
O^{++}	15 0.14%	37 0.34%	68 0.63%	81 0.75%	119 1.11%	546 5.08%	866 8.05%
Total	1925 17.90%	1680 15.62%	1689 15.71%	1682 15.64%	1680 15.62%	2097 19.50%	10753 100.00%

Figure 1: Confusion matrix (raw number of tweets and percentage) between classifiers ensemble agreement (x-axis) and crowd-annotators agreement (y-axis) on offensive ("O") and non-offensive ("N") labels.

We also compute Pearsons' correlation coefficient between the agreement of the ensemble classifiers and that of annotators. It achieves a moderate correlation ($r = 0.51$), showing that training an ensemble of classifiers on generic data to pre-screen domain-specific tweets before manual annotation could help identifying tweets that are either unambiguous or more challenging. A similar correlation ($r = 0.50$) was obtained on an ensemble of BiL-STM classifiers trained with the same training and development sets of the five BERT-based classifiers, suggesting that the pre-screening approach could be used also with other classifiers.

3.4 Qualitative analysis of (dis)agreement

Through a manual analysis of the tweets belonging to the A^0 class, we can identify few phenomena that lead to disagreement in annotation. In many cases,

N^{++}	Stand for something or else fall for anything. #BlackLivesMatter
	Hello world! What a great day to be alive #Trump2020 #MAGA
N^+	Come on man! Lock'em up!!! #maga
	Not the first time. You all misspelled #blacklivesmatter. Speak up! @user
N^0	Set fire to Fox News (metaphorically)
	@user is outing #BLACK_LIVES_MATTER as a cult! HE IS CORRECT!
O^0	#DISGUSTING #Democrats terrorize old folks just before #elections2020
	I love this shit! #BlackLivesMatter
O^+	@user You're a bumbling fool #elections2020
	Elections 2020: Red Rapist v. Blue Racist
O^{++}	Y'all trending about kpop stans instead of #BlackLivesMatter big fack you
	Crazy idiots. This is batshit bullshit. #elections2020

Table 1: Examples of tweets with different degrees of crowd-workers' agreement. The messages have been created starting from real examples by slightly changing their wording, so to make it impossible to retrieve the original ones on Twitter. N=Not offensive, O=Offensive. $++/+/0$ correspond to high, medium and low agreement respectively.

tweets are ambiguous and more context would be needed to fully understand whether the user wanted to offend someone or not. These cases include the presence of deictic expressions or pronouns referring to previous tweets, see for example:

(1) *Shoulda thrown this clowns bike off the bridge!*
(2) *Won't work. Gangs will terrorize the city. Murder at will and maybe they'll shoot the Mayor.*

Other cases include generic expressions of anger that are not targeted against a specific person or group, or expressions of negative feelings, see for example:

(3) *Amen ! Enough of this crap !*

Finally, questions, and in particular rhetorical questions, are very frequent in the A^0 class and their interpretation seems to represent a challenging task for crowd-workers:

(4) *if George Floyd was white would the cop have acted in the same violent, murderous way?*
(5) *What is it with these kids of leftist politicians?*

Overall, disagreement does not seem to stem from poor annotation of some crowd-workers, but rather from genuine differences in the interpretation of the tweets. Additionally, BLM and US American elections are recent events and annotators may have been biased by their personal opinion on the topic during annotation, an effect that has already been highlighted in Sap et al. (2019, 2020).

4 Classification experiments

After collecting information on human agreement on tweets covering three different domains, we aim at assessing the impact of (dis)agreement on classifier behaviour.

To this end, we create several balanced configurations of the datasets, so to control for the effect of agreement level, label distribution and domain topic. We first split the data into a training and test set of 75% and 25% for each domain. Then, to control for the effect of training data size, we further downsample all sets to the smallest one, so that each agreement sample is equally represented (A^{++}, A^+, A^0). In this way, we obtain 3 sets of training data – one per ambiguity level – containing 900 tweets each. Every set further contains 300 tweets from each domain, half for `offensive` label and half for `non-offensive` label so to control also for the effect of label distribution across domains and agreement levels.

4.1 Impact of (dis)agreement in training data

To assess the impact of agreement level in training data, we run a series of experiments by comparing two different classifiers: the first one relies on BERT directly fine-tuned on domain data, while the second foresees also an intermediate fine-tuning step using the entire dataset in Founta et al. (2018), inspired by the supplementary training approach from Phang et al. (2018). BERT is used with the same parameters of the ensemble classifiers, reported in Section 3.1. The domain data used for fine-tuning are built starting from the training data described above divided into different agreement levels (A^{++}, A^+, A^0 and their combinations).

Training split	Training size	All domains	Founta + all domains
A^{++}	900	0.746	**0.757**
A^{+}	900	0.734	0.753
A^{0}	900	0.639	0.683
$A^{++/+}$	1800	**0.755**	**0.756**
$A^{+/0}$	1800	0.728	0.724
$A^{++/0}$	1800	0.723	0.730
$A^{++/+/0}$	2700	0.745	0.752
Baseline - training only on Founta et al. data: 0.667 F1			

Table 2: Performance (F1) when training on data with different levels of human agreement (rows), fine-tuned either on domain data or using the dataset from Founta et al. (2018) and domain data.

Results are reported in Table 2. Note that, for training, the tweets in a given partition for all domains are merged, while they are tested on each domain separately. The reported F1 is an average of the three results (results for each domain can be found in the Appendix and are consistent with the ones reported here). We observe that, if we consider only one level of agreement, data with total agreement are the best for prediction (A^{++}), up to the point that A^{++} data alone provide better results than using all data available in the three splits (*all*), despite the different size (900 vs. 2700 instances). Additionally, the combination of high and mild agreement data ($A^{++/+}$) yields results that are in line with the best configuration obtained with two fine-tuning steps (0.755 vs 0.757). This result clearly indicates that for this kind of task it is not necessary to collect huge datasets for fine-tuning, since few data from the target domain may suffice if properly selected. Finally, the effect of using low agreement data for training is detrimental, in line with findings reported in past works (Reidsma and op den Akker, 2008; Jamison and Gurevych, 2015). This can be spotted in two results: the use of generic data alone as in our baseline is better than using low agreement in-domain data (0.667 vs. 0.639) and all configurations where A^{0} is added to mild and high agreement data perform worse than without A^{0} (0.734 vs 0.728 and 0.746 vs 0.723).

4.2 Impact of (dis)agreement in test data

As a next step, we investigate how classifier's performance varies as a function of annotators' agreement in the test data. To this end, we divide also our test set into subsets according to the same agreement levels (A^{++}, A^{+}, A^{0}) and calculate separate F1s on each of these splits. We run the classifier for

'all domains' described in Section 4.1, i.e. trained on the three domains and tested on one of them. Results, reported in Table 3, are obtained by averaging the F1 for each domain.

We observe a dramatic drop in performance when agreement decreases in the test set, indicating that ambiguous data are the most challenging to classify. These results highlight the need to control for ambiguity also in the test set when creating offensive language benchmarks (for example in shared tasks), in order to avoid high system performance being due to a lack of challenging examples. The best performance on ambiguous data is obtained when training on unambiguous and mildly ambiguous data ($A^{++/+}$). Interestingly, adding A^{+} data to A^{++} data leads to the highest increase in performance exactly for A^{0} test data (from 0.552 to 0.574). This rules out the possibility that a certain level of disagreement in the training set is more effective in classifying the same type of ambiguity in the test set (e.g. train and test on A^{0} data), and suggests that high agreement or mild agreement training sets perform better in all cases.

Training split	Training size	Tested on	F1
$A^{++/+}$	1800	A^{++}	0.860
$A^{++/+}$	1800	A^{+}	0.768
$A^{++/+}$	1800	A^{0}	0.574
A^{++}	900	A^{++}	0.847
A^{++}	900	A^{+}	0.763
A^{++}	900	A^{0}	0.552
A^{0}	900	A^{++}	0.662
A^{0}	900	A^{+}	0.639
A^{0}	900	A^{0}	0.567

Table 3: Performance on $A^{++/+}$, A^{++}, A^{0} data, classified with "all domains" configuration in Table 2.

4.3 Impact of (dis)agreement on out-of-domain data

We then test the effect of cross-domain classification according to agreement levels, so to minimise the impact of possible in-domain overfitting. We repeat the experiments described in the previous section by using two domains for training and the third for testing. As an example, a classifier model was trained using A^{++} data from Covid 19 and US Presidential campaign, and tested on A^{++} data on BLM. This has been repeated for each domain and each agreement level. For conciseness of presentation, we report in Table 4 the F1 obtained by averaging F1 on each domain (results for each domain can be found in the Appendix and also in

this case they are consistent with the one reported here). Results confirm that (i) the classifier yields a good performance when the training data have high agreement, even in an out-of-domain scenario, and (ii) adding A^0 data to the training set has a detrimental effect on performance. Finally, if we compare these results with Table 2, we observe that the effect of overfitting on in-domain data is very limited.

Training split	Training size	Out of domain	Founta + out of domain
A^{++}	600	0.719	0.747
A^{+}	600	0.677	0.716
A^{0}	600	0.567	0.658
$A^{++/+}$	1,200	**0.732**	**0.748**
$A^{+/0}$	1,200	0.659	0.715
$A^{++/0}$	1,200	0.714	0.722
$A^{++/+/0}$	1,800	0.722	0.737

Table 4: Performance (F1) in the out-of-domain setting. Results are the average F1 obtained by the classifier on each domain when trained on the other two.

4.4 (Dis)agreement versus Randomness

An additional question we want to address is whether low agreement data provide some useful information for training offensive language detection systems or if the effect of such data is no more that of random annotation.

We therefore replicate the experiments of Table 2 by replacing the label of A^0 data with a random one. Since we want to obtain the same controlled distribution we assign the same probability to N and O labels. Results are reported in Table 5. As can be seen, when using A^{0rand} data the results worsen as compared to A^0, indicating that the label in A^0 are not assigned by chance and they can contain useful signal for the classifier, albeit challenging. Consistently with previous results, the more gold and high agreement data is added to the training, the smaller the effect of A^{0rand}. These results show also that coin-flipping, which has been suggested in past works to resolve hard disagreement cases (Beigman Klebanov and Beigman, 2009), may not be ideal because it leads to a loss of information.

5 Experiments on Offenseval dataset

Our experiments show that when training and test data include tweets with different agreement levels, classification of offensive language is still a challenging task. Indeed, our classification results reported in Table 2 and 4 suggest that on this kind

Training split	Training size	All domains	Founta + all domains
A^{0}	900	0.639	**0.683**
A^{0rand}	900	0.505	0.576
$A^{+/0}$	1800	**0.728**	0.724
$A^{+/0rand}$	1800	0.657	0.689
$A^{++/0}$	1800	0.723	**0.730**
$A^{++/0rand}$	1800	0.684	0.703
$A^{++/+/0}$	2700	0.745	**0.752**
$A^{++/+/0rand}$	2700	0.719	0.730

Table 5: Performance (F1) when training on data with different levels of human agreement (rows) and replacing A^0 labels with random ones (A^{0rand}). First line of each group is reported from Table 2 for comparison.

of balanced data, F1 with Transformer-based models is ≈0.75. However, system results reported for the last Offenseval shared task on offensive language identification in English tweets (Zampieri et al., 2020) show that the majority of submissions achieved an F1 score > 0.90 on the binary classification task.

We hypothesize that this delta in performance may depend on a limited presence of low agreement instances in the Offenseval dataset used for evaluation (Zampieri et al., 2019). We therefore randomly sample 1,173 tweets from the task test data (30% of the test set) and annotate them with Amazon Mechanical Turk using the same process described in the previous sections (5 annotations per tweet). We slightly modify our annotation guidelines by including the cases of profanities, which were explicitly considered offensive in Offenseval guidelines.

Results, reported in Table 6 (left column) show that the outcome of the annotation is clear-cut: more than 90% of the tweets in the sample have either a high (A^+) or very high (A^{++}) agreement level. Furthermore, only 6.4% of the annotations (75) have a different label from the original Offenseval dataset, 50% of which are accounted for by the A^0 class alone. So our annotation is very consistent with the official one and the distribution is very skewed towards high agreement levels, as initially hypothesized.

To understand whether this skewness can be generalised, i.e. if this sample distribution might be representative of a population distribution, we also estimate the distribution of agreement levels in the initial pool of data (around 400k tweets) we collected using US Election, BLM and Covid-related hashtags (Section 3).[3] The estimate of the distri-

---

[3]Since we pre-selected the tweets to be annotated through

Agreement	Offenseval		400k Tweets	
A^{++}	75.62%	(887)	68.52%	(274,514)
A^+	14.75%	(173)	19.08%	(76,457)
A^0	9.63%	(113)	12.40%	(49,694)
N^{++}	64.36%	(755)	65.12%	(260,925)
N^+	5.80%	(68)	15.50%	(62,085)
N^0	4.60%	(54)	7.94%	(31,813)
O^0	5.03%	(59)	4.46%	(17,882)
O^+	8.95%	(105)	3.59%	(14,372)
O^{++}	11.25%	(132)	3.39%	(13,589)

Table 6: Comparison of agreement distribution in Offenseval sample and projection on 400k tweets.

bution for class A^+, A^{++} and A^0 is reported in Table 6 (right column). A comparison between the two columns shows that disagreement distribution in the Offenseval sample is in line with the distribution in the data we initially collected before balancing, providing initial evidence that this distribution – with few disagreement cases – might be a 'natural' one for online conversations on Twitter.

Differences emerge when considering the ratio of offensive tweets. In Offenseval data, the percentage of offensive tweets is more than double the percentage in our data (25.23% vs. 11.44%), because the authors adopted several strategies to overrepresent offensive tweets (Zampieri et al., 2019).

As a final analysis, we collect the runs submitted to Offenseval and compute the F1 score of each of these systems over the three levels of agreement separately. Overall, we consider all runs that in the task obtained F1 > 0.75, i.e. 81 runs out of 85. Results are reported in Table 7 as the average of the F1 obtained by the different systems. This last evaluation confirms our previous findings, since F1 increases when agreement level increases in test data. This finding, together with the distribution of agreement levels, shows that the high performance obtained by the best systems in the shared task is most probably influenced by the prevalence of tweets with total agreement.

Offenseval 2020 - test subsets	F1	StDev
A^{++} (887 tweets)	0.915	± 0.055
A^+ (173 tweets)	0.817	± 0.075
A^0 (113 tweets)	0.656	± 0.067

Table 7: Average F1 obtained by the best systems at Offenseval 2020 ± StDev.

the classifier ensemble, to estimate the real distribution of agreement levels in our data we classified with the ensemble all of them (400k tweets). Then, to determine the proportion of each class of agreement, we projected the distribution of annotators' agreement level for each ensemble class, using the confusion matrix reported in Figure 1.

6   Discussion and Conclusions

We have presented a data annotation process and a thorough set of experiments for assessing the effect of (dis)agreement in training and test data for offensive language detection. We showed that an ensemble of classifiers can be employed to preliminarily select potentially unambiguous or challenging tweets. By analysing these tweets we found that they represent real cases of difficult decisions, deriving from interesting phenomena, and are usually not due to low-quality annotations. We also found that these challenging data are minimally present in a popular benchmark dataset, accounting for higher system performance. We believe that such hard cases should be more represented in benchmark datasets used for evaluation of hate speech detection systems, especially in the test sets, so to develop more robust systems and avoid overestimating classification performance. This goal can be achieved by integrating the common practice of oversampling the minority *offensive* class with the oversampling of minority *agreement* classes.

From a multilingual perspective, we also noted that at Offenseval 2020 the best performing systems on Arabic scored 0.90 F1 with a training set of 8k tweets, 0.85 on Greek with less than 9k tweets, and 0.82 on Turkish despite having more than 32k examples for training. This shows that the amount of training data is not sufficient to ensure good classification quality, and that also in this case a study on disagreement levels could partly explain these differences (this is further corroborated by the fact that for Turkish the lowest overall inter-annotator agreement score was reported).

As future work, we plan to develop better approaches to classify (dis)agreement, in order to ease oversampling of low agreement classes. Preliminary experiments (not reported in this paper) show that the task is not trivial, since supervised learning with LMs such as BERT does not work properly when trying to discriminate between ambiguous and not ambiguous tweets. Indeed, BERT-based classification performed poorly both in the binary task (ambiguous vs. not ambiguous) and in the three-way one (offensive vs. not offensive vs. ambiguous). This suggests that ambiguity is a complex phenomenon where lexical, semantic and pragmatic aspects are involved, which are difficult to capture through a language model.

This corpus, together with the experiments presented in this paper, will hopefully shed light onto

the important role played by annotators' disagreement, something that we need to understand better and to see as a novel perspective on data. Indeed, if we want to include diversity in the process of data creation and reduce both the exclusion of minorities' voices and demographic misrepresentation (Hovy and Spruit, 2016), disagreement should be seen as a signal and not as noise.

7 Ethics Statement

The tweets in this dataset have been annotated by crowd-workers using Amazon Mechanical Turk. All requirements introduced by the platform for tasks containing adult content were implemented, for example adding a warning in the task title. We further avoid to put any constraints on the minimum length of sessions or on the minimum amount of data to be labeled by each crowd-worker, therefore they were not forced to prolonged exposure to offensive content. Indeed, we observed that crowd-workers tended to annotate for short sessions, on average 20 minutes, which suggests that annotating was not their main occupation. Crowd-workers were compensated on average with 6 US$ per hour.

Although we put in place strict quality control during data collection, we compensated the completed hits also when annotations were finally discarded because they did not reach the minimum accuracy threshold of 70% w.r.t. the gold standard. We also engaged in email conversations with crowd-workers when they were blocked because of mismatches with the gold standard tweets. In several cases, we clarified with them the issue and subsequently unlocked the task.

Concerning the annotated dataset, we support scientific reproducibility and we would like to encourage other researchers to build upon our findings. However, we are aware that ethical issues may arise related to the complexity and delicacy of judgments of offensiveness in case they are made public. Therefore, in compliance with Twitter policy, we want to make sure that our dataset will be reused for non-commercial research only[4] avoiding any discriminatory purpose, event monitoring, profiling or targeting of individuals. The dataset, in the form of tweet IDs with accompanying annotation, can be obtained upon request following the process described at this link: `https://github.com/dhfbk/annotators-agreement-dataset`. Re-

questors will be asked to prove their compliance with Twitter policy concerning user protection and non-commercial purposes, as well as to declare that they will not use our dataset to collect any sensitive category of personal information. Also, releasing the tweet IDs instead of the text will enforce users' right to be forgotten, since it will make it impossible to retrieve tweets if their authors delete them or close their account. Although we are aware of the risks related to developing and releasing hate speech datasets, this research was carried out with the goal of improving conversational health on social media, and even exposing the limitations of binary offensive language detection. We believe that our findings confirm the context- and perspective-dependent offensiveness of a message, and we therefore avoid binary labels, stressing the importance of taking multiple points of view (in our case, five raters) into account. Following the same principle of avoiding profiling, crowd-workers' IDs are not included in the dataset, so that it will not be possible to infer annotator-based preferences or biases.

References

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2020. Modeling annotator perspective and polarized opinions to improve hate speech detection. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, 8(1):151–154.

Lora Aroyo, Lucas Dixon, Nithum Thain, Olivia Redfield, and Rachel Rosen. 2019. Crowdsourcing subjective tasks: The case study of understanding toxicity in online discussions. In *Companion Proceedings of The 2019 World Wide Web Conference*, WWW '19, page 1100–1105, New York, NY, USA. Association for Computing Machinery.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine*, 36(1):15–24.

Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.

---

[4]https://developer.twitter.com/en/developer-terms/policy

Valerio Basile. 2020. It's the end of the gold standard as we know it. on the impact of pre-aggregation on the evaluation of highly subjective tasks. In *Proceedings of the AIxIA 2020 Discussion Papers Workshop co-located with the the 19th International Conference of the Italian Association for Artificial Intelligence (AIxIA2020), Anywhere, November 27th, 2020*, volume 2776 of *CEUR Workshop Proceedings*, pages 31–40. CEUR-WS.org.

Beata Beigman Klebanov and Eyal Beigman. 2009. From annotator agreement to noise models. *Computational Linguistics*, 35(4):495–503.

Beata Beigman Klebanov and Eyal Beigman. 2014. Difficult cases: From data to learning, and back. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 390–396, Baltimore, Maryland. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the 11th International AAAI Conference on Web and Social Media*, ICWSM '17.

Barbara Di Eugenio and Michael Glass. 2004. Squibs and discussions: The kappa statistic: A second look. *Computational Linguistics*, 30(1):95–101.

Anca Dumitrache, Lora Aroyo, and Chris Welty. 2019. A crowdsourced frame disambiguation corpus with ambiguity. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2164–2170, Minneapolis, Minnesota. Association for Computational Linguistics.

Antigoni-Maria Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *11th International Conference on Web and Social Media, ICWSM 2018*.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A Smith. 2010. Part-of-speech tagging for twitter: Annotation, features, and experiments. Technical report, Carnegie-Mellon University, School of Computer Science.

Mitchell L. Gordon, Kaitlyn Zhou, Kayur Patel, Tatsunori Hashimoto, and Michael S. Bernstein. 2021. The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, New York, NY, USA. Association for Computing Machinery.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1120–1130, Atlanta, Georgia. Association for Computational Linguistics.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Pei-Yun Hsueh, Prem Melville, and Vikas Sindhwani. 2009. Data quality from crowdsourcing: A study of annotation selection criteria. In *Proceedings of the NAACL HLT 2009 Workshop on Active Learning for Natural Language Processing*, pages 27–35, Boulder, Colorado. Association for Computational Linguistics.

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

Manfred Klenner, Anne Göhring, and Michael Amsler. 2020. Harmonization sometimes harms. In *Proceedings of the 5th Swiss Text Analytics Conference and the 16th Conference on Natural Language Processing, SwissText/KONVENS 2020, Zurich, Switzerland, June 23-25, 2020 [online only]*, volume 2624 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Rebecca J. Passonneau. 2004. Computing reliability for coreference annotation. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

J. Peterson, R. Battleday, T. Griffiths, and O. Russakovsky. 2019. Human uncertainty makes classification more robust. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9616–9625.

Jason Phang, Thibault Févry, and Samuel R Bowman. 2018. Sentence encoders on stilts: Supplementary training on intermediate labeled-data tasks. *arXiv preprint arXiv:1811.01088*.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Learning part-of-speech taggers with inter-annotator agreement loss. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 742–751, Gothenburg, Sweden. Association for Computational Linguistics.

Ines Rehbein and Josef Ruppenhofer. 2011. Evaluating the impact of coder errors on active learning. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 43–51, Portland, Oregon, USA. Association for Computational Linguistics.

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

Filipe Rodrigues and Francisco C. Pereira. 2018. Deep learning from crowds. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 1611–1618. AAAI Press.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5477–5490. Association for Computational Linguistics.

Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 614–622, New York, NY, USA. Association for Computing Machinery.

Edwin Simpson, Jonas Pfeiffer, and Iryna Gurevych. 2020. Low resource sequence tagging with weak labels. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8862–8869.

Fei-Yue Wang. 2007. Toward a paradigm shift in social computing: The acp approach. *IEEE Intelligent Systems*, 22(5):65–67.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the Type and Target of Offensive Posts in Social Media. In *Proceedings of NAACL*.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 Task 12: Multilingual Offensive Language Identification in Social Media (OffensEval 2020). In *Proceedings of SemEval*.

A Annotation Guidelines for AMT

This section contains the instructions provided to annotators on Amazon Mechanical Turk. The first part changes according to the domain:

Covid-19: *The tweets in this task have been collected during the pandemic. Would you find the content of the messages offensive? Try to judge the offensiveness of the tweets independently from your opinion but solely based on the abusive content that you may find.*

US Presidential campaign: *The tweets in this task have been collected during the last US Presidential campaign. Would you find the content of the messages offensive? Try to judge the offensiveness of the tweets independently from your political orientation but solely based on the abusive content that you may find.*

Black Lives Matter: *These tweets are related to the Black Lives Matter protests. Would you find the content of the messages offensive? Try to judge the offensiveness of the tweets independently from your opinion but solely based on the abusive content that you may find.*

The second part of the task description, instead, is the same for all the domains, containing a definition of what is offensive and informing the workers that there is a quality check on the answers:

Offensive: *Profanity, strongly impolite, rude, violent or vulgar language expressed with angry, fighting or hurtful words in order to insult or debase a targeted individual or group. This language can be derogatory on the basis of attributes such as race, religion, ethnic origin, sexual orientation, disability, or gender. Also sarcastic or humorous expressions, if they are meant to offend or hurt one or more persons, are included in this category.*

Normal: *tweets that do not fall in the previous category.*

Quality Check: *the HIT may contain a gold standard sentence, manually annotated by three different researchers, whose outcome is in agreement. If that sentence is wrongly annotated by a worker, the HIT is automatically rejected.*

Asking annotators to label the tweets independently from their views, opinions or political orientation was inspired by recent works, showing that making explicit possible biases in the annotators contributes to reduce such bias (Sap et al., 2019).

Training	Training Size	All domains			Founta + all domains		
		BLM	Covid	Election	BLM	Covid	Election
A^{++}	900	**0.756**	**0.752**	0.730	**0.768**	0.752	**0.752**
A^{+}	900	0.745	0.724	**0.734**	0.774	0.736	0.748
A^{0}	900	0.647	0.644	0.626	0.689	0.652	0.707
$A^{++/+}$	1800	**0.776**	**0.756**	**0.732**	**0.779**	**0.738**	**0.750**
$A^{+/0}$	1800	0.738	0.738	0.707	0.744	0.698	0.729
$A^{++/0}$	1800	0.732	0.733	0.704	0.746	0.723	0.721
$A^{++/+/0}$	2700	0.758	0.736	0.742	0.766	0.748	0.742

Table 8: Test results on single domains using a model trained on all domains.

Training	Training Size	Out of domain			Founta + out of domain		
		BLM	Covid	Election	BLM	Covid	Election
A^{++}	600	**0.699**	**0.734**	**0.723**	**0.760**	0.736	**0.746**
A^{+}	600	0.681	0.720	0.631	0.718	0.706	0.725
A^{0}	600	0.557	0.603	0.542	0.674	0.629	0.672
$A^{++/+}$	1800	**0.696**	**0.758**	**0.742**	**0.740**	**0.771**	**0.734**
$A^{+/0}$	1800	0.641	0.686	0.649	0.733	0.696	0.716
$A^{++/0}$	1800	0.695	0.726	0.720	0.737	0.706	0.722
$A^{++/+/0}$	2700	0.737	0.736	0.692	0.734	0.756	0.720

Table 9: Results in the out-of-domain setting, testing the classifier on each domain when trained on the other two.

B Impact of (dis)agreement on classification - results in detail

Table 8 displays domain-specific results related to the analysis shown in Section 4.1 of the main document, where for sake of brevity we have shown only an average between the three domains. The table confirms that also on single domains, training data with higher level of agreement improve predictions, while training data with low level of agreement are detrimental. Classification took about 2 minutes on a Titan X for the runs using only domain-specific data. Adding the intermediate fine-tuning on data from Founta et al. (2018) increases the time to 1.5 hours.

C Impact of (dis)agreement on out-of-domain data - results in detail

Similar to the previous table, Table 9 displays out-of-domain results related to the analysis shown in Section 4.3 of the main document, where we report only an average between the three domains. The results are consistent with the average scores reported in the main document, i.e. that training data with high agreement improve prediction, while training data with low agreement are detrimental. Classification took about the same time of the runs in the single domain configuration.

D Twitter data collection

Through its application programming interface (API), Twitter provides access to publicly available messages upon specific request. For each of the domains analysed, a set of hashtags and keywords was identified that unequivocally characterizes the domain and is collectively used. During a specific period of observation, all the tweets containing at least an item of this hashtags/keywords seed list were retrieved in real time (using "filter" as query). The most relevant entries from the covid-19 seed list are: *covid-19*, *coronavirus*, *ncov*, *#Wuhan*, *covid19*, *sarscov2* and *covid*. Data were collected in the time span between 25 January and 09 November 2020. The most relevant entries from the blm seed list are: *george floyd*, *#blm*, *black lives matter*. Tweets were collected between 24 May 2020 and 16 June 2020. The most relevant entries from the US Elections seed list are: *#maga*, *#elections2020*, *Trump*, *Biden*, *Harris*, *Pence*. The tweets were collected between 30 September 2020 and 04 November 2020.

For each domain, a big bulk of data was collected in real time for each specific time span. From these about 400,000 tweets were randomly selected and evaluated with the ensemble method as described in Section 3 of the main paper.