Why Don't You Do It Right? Analysing Annotators' Disagreement in Subjective Tasks

Marta Sandri
Dept. of Humanities
University of Pavia, Italy
sandri.marta97@gmail.com

Elisa Leonardelli
Fondazione Bruno Kessler
Trento, Italy
eleonardelli@fbk.eu

Sara Tonelli
Fondazione Bruno Kessler
Trento, Italy
satonelli@fbk.eu

Elisabetta Jezek
Dept. of Humanities
University of Pavia, Italy
jezek@unipv.it

Abstract

Annotators' disagreement in linguistic data has been recently the focus of multiple initiatives aimed at raising awareness on issues related to 'majority voting' when aggregating diverging annotations. Disagreement can indeed reflect different aspects of linguistic annotation, from annotators' subjectivity to sloppiness or lack of enough context to interpret a text.

In this work we first propose a taxonomy of possible reasons leading to annotators' disagreement in subjective tasks. Then, we manually label part of a Twitter dataset for offensive language detection in English following this taxonomy, identifying how the different categories are distributed. Finally we run a set of experiments aimed at assessing the impact of the different types of disagreement on classification performance. In particular, we investigate how accurately tweets belonging to different categories of disagreement can be classified as offensive or not, and how injecting data with different types of disagreement in the training set affects performance. We also perform offensive language detection as a multi-task framework, using disagreement classification as an auxiliary task.

Warning: *This paper contains examples that may be offensive or upsetting.*

1 Introduction

The development of benchmark datasets based on the reconciliation of annotators' disagreement has been a standard practice within the NLP community for decades. However, in the last few years this paradigm has been questioned (Basile, 2020; Abercrombie et al., 2022), since aggregating discording labels is based upon the assumption that texts have a single interpretation and that annotators' disagreement is something that should be corrected. On the contrary, it may convey useful information on the task, the data and the annotators themselves (Aroyo and Welty, 2015). For example, annotators may disagree because of poorly described guidelines, or because they interpret in different ways a text based on their background and beliefs. The text may be ambiguous because of a lack of context, or annotators may simply be working poorly, without paying much attention to the task they should perform on the text (Uma et al., 2021).

Having a better understanding of the reasons behind annotators' disagreement could lead to better annotation guidelines and provide insights into the advantages and limits of developing NLP systems able to deal with disagreement (Davani et al., 2022). It would also contribute to a better understanding of how disagreement interferes with systems' performance. Finally, automatically detecting instances that are likely to obtain discording labels could help researchers in creating linguistic datasets with more or less challenging examples (Lehmann et al., 1996).

In this work, we contribute to the research line on annotators' disagreement by first presenting a taxonomy, where we classify possible reasons behind discording annotations in subjective tasks, i.e. tasks admitting diverse valid beliefs about what the correct data labels should be (Rottger et al., 2022). We also annotate part of an existing dataset for offensive language detection (Leonardelli et al., 2021) in order to assess the validity of our categorisation. We further perform several experiments aimed at addressing the following research questions: *i)* What are the most challenging categories of disagreement to classify in a task of abusive language detection? *ii)* What is the effect of including specific categories of disagreement in the training set? *iii)* Can multi-task learning be effectively used for offensive language detection using disagreement classification as auxiliary task? Are these

Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, pages 2428–2441

results consistent across disagreement categories?

Evaluation results can contribute to ongoing studies investigating the negative impact of disagreement on systems performance (Schwartz et al., 2011; Beigman Klebanov and Beigman, 2014; Leonardelli et al., 2021; Uma et al., 2021) by providing a more fine-grained view on disagreement types.

The annotated data used in our experiments can be obtained as an extension of Leonardelli et al. (2021)'s dataset, at this link: `https://github.com/dhfbk/annotators-agreement-dataset`.

2 Related work

Despite all the efforts to minimize inter-annotator disagreement, every large-scale project involving linguistic annotation has to deal with cases of diverging perceptions among human workers, especially in tasks where human sensibility is at play. However, disagreement due to interpretation divergences is also found in text annotation tasks usually perceived as objective, such as Part-of-Speech tagging (Plank et al., 2014), semantic role labeling (Dumitrache, 2019) or word sense disambiguation (Martínez Alonso et al., 2015).

In subjective tasks, the hypothesis that a single truth exists for all instances has been debated in several past works (Basile et al., 2021; Uma et al., 2021; Basile et al., 2022). Indeed, many researchers suggested that disagreement has to be treated as a signal, and not as noise (Aroyo and Welty, 2015; Plank et al., 2014; Basile et al., 2019). For this reason, Plank (2022) has recently proposed to use *human label variation* rather than the term *disagreement* to capture the fact that two or more views may sometimes be plausible in text annotation. Along the same line, different approaches have been proposed with the primary aim of amending the traditional way in which disagreement is dealt with (i.e., to treat it as noise and discard it from the training set), thus demonstrating that disagreement provides insights into human perception, linguistic data as well as classification systems.

Concerning the integration of disagreement in NLP classifiers, few approaches have been proposed so far (for an overview, see Uma et al. (2021)). Beside aggregating judgments based on majority voting, which has the clear shortcoming of reducing different voices and points of view in favour of the dominant one, some approaches consider disagreement as an indicator of the difficulty of the instances to be annotated (Reidsma and op den Akker, 2008). Past works have proposed to train separate classifiers, one for each annotator, and build an ensemble classifier that makes a prediction when all classifiers agree on the class label (Basile, 2020) or to adopt a multi-task architecture using a shared representation to model annotator disagreements (Davani et al., 2022). Similarly, Fornaciari et al. (2021) use soft-labels (i.e., probability distributions over the annotator labels) as an auxiliary task in a multi-task neural model. As an alternative, disagreement can also be excluded from the data by training the model only on items that show high agreement (Reidsma and op den Akker, 2008) or, conversely, models can be trained and evaluated only on disagreement-raising items (Beigman Klebanov and Beigman, 2014).

Concerning the analysis of annotators' behaviour on subjective tasks, past works showed that disagreement is to be expected (Basile, 2020; Davani et al., 2022). This is confirmed by Kenyon-Dean et al. (2018), showing that around 30% of the instances in a Twitter corpus for sentiment analysis are controversial, thus likely to lead to disagreement. They also propose to merge them in a new class of sentiment called "complicated", so that they are not excluded from the data. Sang and Stanton (2022) show that age and personality are factors that greatly influence annotators' perception of offensive content. Kocoń et al. (2021) show that disagreement can be reduced and annotation quality increased by combining text representations of labeled annotators' opinions with their demographic traits. Akhtar et al. (2019) deal with disagreement in hate speech annotation by partitioning the annotators into clusters reflecting more uniform subjective judgments.

As regards the categorisation of disagreement in subjective tasks, one of the first studies on this topic was presented in Beigman Klebanov et al. (2008), where in a metaphor annotation task a distinction is made between annotators' attentions slips and genuine disagreement. Basile et al. (2021) identify three main sources of disagreement: individual differences related to annotators' background and beliefs, stimulus characteristics, i.e. inherent ambiguity of texts, and context. Uma et al. (2021) further extend this categorisation by listing the following reasons behind disagreement: annotation errors, imprecise or vague annotation scheme, context-dependent text ambiguity, introduced in

Poesio et al. (2019), item difficulty and annotators' subjectivity. In our proposed taxonomy, we rely on previous works in the choice of categories, keeping the distinction between inherent text ambiguity and context-dependent one, as well as annotators' subjectivity. We further specify these categories by proposing subtypes that take into account different linguistic phenomena.

A recent work introducing a taxonomy of disagreement has also been presented in Jiang and de Marneffe (2022), which focuses on the task of natural language inference. Interestingly, the authors propose a three-layered taxonomy which has some high-level overlaps with ours, for example our *Ambiguity* class can be roughly mapped onto *Uncertainty in sentence meaning*, and *Subjectivity* onto *Annotator behavior*. The lower categories, however, are in some cases very task-specific. The analysis of task-specific and task-independent categories of disagreement may be an interesting research direction to explore in the future.

3 A Taxonomy of Disagreement

As a first step towards a better understanding of the reasons behind annotators' disagreement, we propose a taxonomy of disagreement for subjective tasks, which is built starting from past categorisations presented in the literature and iteratively adding subtypes based on the analysis of examples in existing datasets. An overview of the taxonomy is reported in Figure 1. The taxonomy includes four macro-categories, which are further specified through more fine-grained subtypes. For each category we report also selected examples, all taken from the dataset of disagreement presented in Leonardelli et al. (2021) (see a more detailed description of the dataset in Section 4). In the following, we illustrate each category in detail.

3.1 Sloppy Annotation

This category covers errors in the annotation due to annotators' carelessness. This can happen in particular with crowd-sourcing platforms, when annotators are recruited without a proper training and their annotation quality is not monitored. In general, this type of disagreement can be minimised by adopting tools that identify which annotators are not trustworthy (Hovy et al., 2013).

The only specification for this category is *Noise* (Figure 1), corresponding to messages clearly labeled with the wrong category, see for instance the tweet below annotated as "offensive":

(1) *In a singular voice, art across the world..*

3.2 Ambiguity

The second source of disagreement we include in the taxonomy is *Ambiguity*, that is a much debated topic in linguistics. It is generally referred to as a property of words or phrases to allow more than one interpretation (Tuggy, 1993; Cruse, 2004). This category comprises mainly cases of figurative language, a phenomenon related to the use of words with a diverging meaning from their literal use in order to express colorful images and emotions (Dancygier and Sweetser, 2014). This category includes six subtypes:

Analogy. This label includes comparison mechanisms, such as simile and metaphor, along with the figure of speech of analogy. Both these phenomena involve the cognitive concept of mapping between two conceptual domains (Dancygier and Sweetser, 2014). Below we report an example of metaphor:

(2) *Trump is a walking petri dish. His goal is to spread the virus to as many people as possible.*

False Assertion. Under this label we group all the cases that are characterized by an assertion that is false if compared to the reality of facts and that can therefore trigger irony (Cignarella et al., 2018). In other words, users express the opposite of what they think or something false and exaggerated in relation to the context. This figure of speech needs knowledge of the world to be understood and this is the main reason behind disagreement. An example is reported below:

(3) *Another attempt backfired on them, George Floyd cured Covid-19 and opened up the economy!*

Rhetorical Question. We group under this label all the instances containing a question asked not to obtain an answer but with the purpose of rhetorically pointing out a concept (Stivers and Enfield, 2010), see example below:

(4) *why do we treat our prisoners like this? Is it really because we decided once you commit a crime you're worthless non-human?*

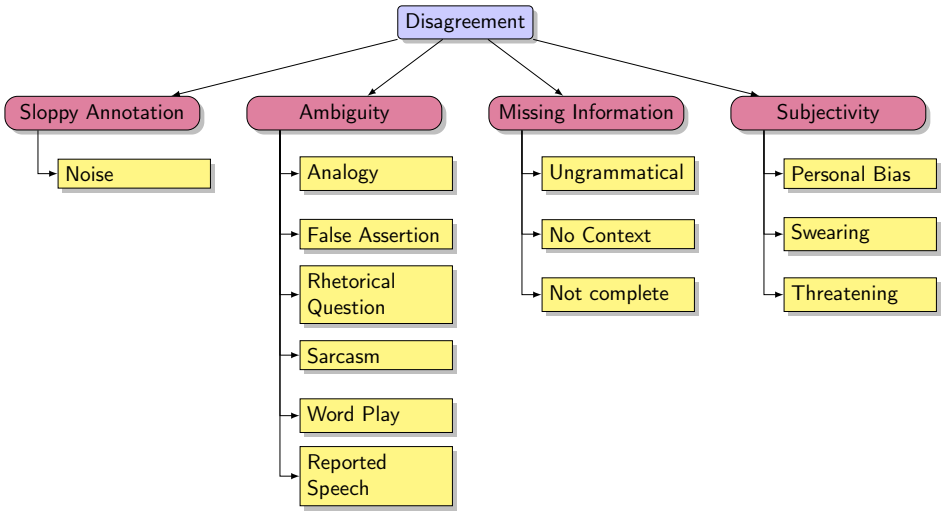

Figure 1: Taxonomy of annotators' disagreement in subjective tasks. Pink boxes represent coarse-grained categories, while yellow boxes correspond to subtypes.

Sarcasm. Sarcasm is characterized by words employed to express the opposite of their literal meaning, mainly used to make fun of a specific topic or person. Sarcasm needs context and other extra-linguistic expressions like pauses and intonation to be understood and it features a disproportion of emphasis regarding the real situation, giving a caustic effect (Gibbs, 1986). However, in writing, these extra-linguistic hints are not possible, thus we must rely on our knowledge of the world or of the addressee to understand sarcasm.

(5) *Who knew a side effect of COVID would be gross incompetence.*

Word Play. Word Play is a figure of speech that involves literary devices to alter some words, with the purpose of giving proof of someone's wit. Among the literary techniques used to convey word play are acronyms, alliterations, i.e. repeating the same sounds in a sentence, and puns, i.e. using words with multiple meanings to obtain a humorous result.

(6) *The only people ripping this country apart are your fellow liberal #DemocRATS and your militant concubines.*

Reported Speech. It is commonly defined as the presentation of discourse that purports to be from a prior occasion, and may originate from another author (Holt, 2009). It can be ambiguous because it can be mistaken as something written by the same author of the text to be annotated.

(7) *White Bystanders With Rifles Stare Down George Floyd Protesters: 'You Ain't Got No Guns'*

3.3 Context sensitivity: Missing Information

In this category we group all cases in which annotators' disagreement may be caused by a lack of information or of context to unambiguously interpret a text (Donaldson and Lepore, 2012). It includes the three following subtypes.

Ungrammatical and Non Standard Language. We assign this label to all the texts that could cause disagreement due to marked uses of the language (e.g. non-standard varieties), use of slang, code-switching or mere typing errors. Intuitively, annotators who do not speak a specific language variety are more likely to misunderstand or misinterpret it (Sap et al., 2019), leading to disagreement. See the example below:

(8) *mane it's hard for some of da blaxk people out dere when dey go into a store they got everybody looking at them "what they doing"*

No Context. This class covers well-known linguistic phenomena which need context to be unequivocally interpreted such as anaphora and deixis (Poesio and Artstein, 2005). We include in this

category also messages containing links, typically used in online communication. Deictic expressions include devices such as demonstratives (*this, that*, etc.), personal pronouns (*I, you, he, she*, etc.), possessives (*his, her, their*, etc.) (Levinson, 2004). As regards anaphoric expressions, instead, the interpretation of the textual content is determined by that of the antecedent (Levinson, 1987), therefore, if there is not enough context, crucial information may be missing. We report below an example containing both a deictic expression (*this*) and an anaphoric pronoun (*them*):

(9) *Dude this guy is serious? And trump retweeted this?????? Please anonymous take them out*

Not Complete. In the last subclass of *Missing Information*, we group all cases that do not convey complete information and do not fall into any other subcategory. A typical example is social media threads written by a user. If just one message is extracted from a thread and annotated, it may be very difficult to understand what its tone and meaning is. As an example, see below:

(10) *Wtaf, a farce in three parts:*

3.4 Bias and Desemanticization: Subjectivity

We group in this category cases of disagreement due to the annotators' identity, beliefs and background, which have been recognised as leading to biased judgments (Sap et al., 2022). This is a major source of disagreement in subjective tasks such as hate speech detection, sentiment analysis or political stance detection. The category includes three subtypes.

Personal Bias. Although it is not easy to identify cases of disagreement due to personal biases, in particular when annotators are unknown, this category can be inferred because it is very likely to occur when the text to be labeled belongs to a divisive topic, such as politics, covid-19, social movements, etc. (Wich et al., 2020). See for example the sentence below:

(11) *#DemocratsAreDestroyingAmerica #BlackLivesMatter is a terrorist organization*

Swearing. Another cause of disagreement related to annotators' subjectivity is how they interpret swear words. Indeed, some annotators perceive specific swear words as offensive and hurtful, while for others they appear to be desemanticized (Ljung, 2011). A typical example in English is the use of the word *bitch*:

(12) *2nd wave about to be a bitch*

Threatening. Disagreement may arise from texts containing linguistically violent expressions or threats and depend on annotators' sensitivity to verbal violence and menaces (Storey, 1995), see for instance, the following text:

(13) *U r going to jail.*

4 Data Annotation and Analysis

We apply the taxonomy illustrated above to the dataset of disagreement in abusive language presented in Leonardelli et al. (2021), which contains more than 10k English tweets labeled as offensive or not offensive by five crowd-workers, and covers three topics: covid-19, US Presidential elections and Black Lives Matter movement. The dataset has been designed so as to include a balanced set of tweets with full agreement (A^{++}), with partial agreement (A^{+} class, 4 matching labels versus 1) and with disagreement (A^{0} class, 2 vs. 3 labels) and has been released divided into a balanced training and test split.

We manually annotate all A^{+} and A^{0} tweets in the test set (1,756 in total), plus a portion of tweets from the training set (809 tweets). A total of 2,574 tweets is annotated, divided in 1,518 for the A^{0} agreement level and 1,056 for the A^{+} agreement level. Annotation was performed by a trained linguist, and during the process the initial taxonomy was adjusted by refining the category definitions or introducing new ones when needed. A second linguist annotated a sample of 200 tweets divided equally into A^{0} and A^{+} following the annotation guidelines (see Appendix B). Cohen's Kappa is 0.591, which corresponds to a moderate agreement. After computing agreement, a discussion and adjudication phase was conducted.

Based on the annotator's feedback, we foresaw the possibility to assign multiple labels to the same tweet when the source of disagreement could refer to two or more categories in the taxonomy, considering however the first label as the most probable interpretation. A multi-category annotation scheme was adopted for the same reason also in Jiang and de Marneffe (2022).

In Table 1 we report the distribution of annotated tweets by category and by agreement level.

In case of multiple labels, we include the first one in the statistics. However, tweets annotated as belonging to more than one category are around 40%, showing that disagreement is often due to a mix of different linguistic phenomena. The category with the most multiple annotations is *Subjectivity*, labeled together with *Missing Information* in most of the cases.

	A^0	A^+
Subjectivity	996	703
Ambiguity	302	201
Missing Information	218	142
Sloppy Annotation	2	10
Total	**1,518**	**1,056**

Table 1: Number of annotated tweets by category and agreement level. A^0 is 3 vs. 2 judgments; A^+ is 4 vs. 1 judgment.

Table 1 shows that the cases leading to disagreement because of annotators' subjectivity cover most of the tweets in the dataset, and that the ranking of the four categories is the same for the two agreement levels. As expected, clear annotation mistakes (i.e. *Sloppy Annotation* category) are more frequent for A^+, but in general they have a minimal impact on the cases of disagreement. This is probably due to the fact that the dataset in Leonardelli et al. (2021) was created adopting a strict quality control protocol aimed at excluding low-quality crowd-workers.

In Figure 2 we report how the subtypes are distributed in the annotated dataset. As shown in previous studies, swear words are often a signal for a hateful attitude, but they are also used in casual contexts with positive social functions (Pamungkas et al., 2020). This double interpretation is likely the main reason why the *Swearing* subtype is very frequent in our dataset of disagreement. Concerning *Personal Bias*, its relevance is probably related to the fact that the annotated tweets deal with controversial topics such as US American elections, Black Lives Matter and covid-19. For instance, few tweets targeting Trump were labeled as not offensive by annotators, who were likely to be Biden supporters, and vice versa.

If we compare the above statistics with the analysis reported in Jiang and de Marneffe (2022) on disagreement in natural language inference (NLI), we observe that the most frequent causes of disagreement are task-specific: in our dataset, they

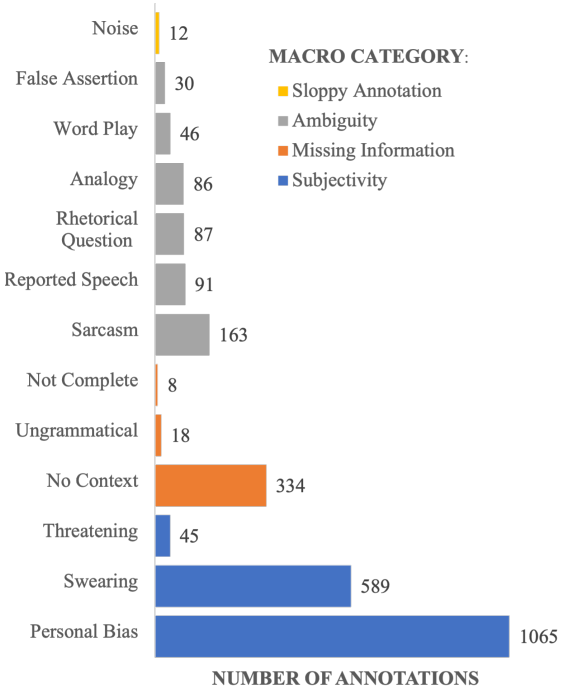

Figure 2: Summary of annotated data by main category and subtype.

are mostly due to annotators' subjectivity and their perception of what is offensive, while in NLI they often stem from the underspecified meaning of lexical items in the sentence pairs to be labeled, or in the probabilistic nature of the inferred content.

5 Experiments

In the next subsections we describe a series of experiments conducted on the annotated dataset, aimed at analysing the relationship between annotators' disagreement and various aspects of automatic offensive language detection. For all the experiments described below, we use the MaChAmp v0.2 toolkit (van der Goot et al., 2021), a classification tool that allows easy implementation of transformers-based classification tasks and supports single-task and multi-task learning. For all the experiments we employ BERT-base uncased (Devlin et al., 2019) (110M parameters) and perform 20 restarts. We keep the default hyperparameter setting of MaChAmp, i.e. max seq length 128, batch size 32, 0.3 dropout, 10 epochs. All the results reported in the following subsections are the average values of 20 runs. All experiments are run on a NVIDIA Quadro RTX 5000 GPU.

Subtype	Main category	Test size	micro F1		
			All	Offensive	Not offensive
Swearing	*Subjectivity*	427	75.74 ± 0.64	89.45	39.83
Rhetorical Question	*Ambiguity*	60	71.75 ± 4.00	65.79	74.51
Not complete	*Missing Info*	7	71.43 ± 12.78	-	-
No context	*Missing Info*	183	71.26 ± 2.35	43.75	77.09
Reported Speech	*Ambiguity*	82	70.37 ± 2.62	60.24	73.85
Threatening	*Subjectivity*	41	68.78 ± 4.12	74.12	65.00
Word Play	*Ambiguity*	33	65.61 ± 3.22	65.31	65.88
Personal Bias	*Subjectivity*	729	64.98 ± 2.05	63.35	65.89
Ungrammatical	*Missing Info*	14	64.29 ± 5.05	53.12	79.17
Sarcasm	*Ambiguity*	97	63.71 ± 3.49	58.75	65.34
Analogy	*Ambiguity*	62	60.56 ± 2.91	70.52	51.82
False Assertion	*Ambiguity*	20	53.00 ± 6.20	37.22	65.91

Table 2: Classification performance of the best system from Leonardelli et al. (2021) for each category and subtype of disagreement. We report the average F1 obtained from 20 restarts and, for the overall results, also the standard deviation.

5.1 Classification performance on disagreement categories

Our first experiment aims at analysing differences in classification performance among disagreement categories and subtypes. To this end, we use the best model for offensive language classification previously described in Leonardelli et al. (2021), which was trained using only tweets with perfect (A^{++}) and high (A^{+}) agreement. We run this model on the tweets in our dataset that belong to the test set of the original work, and calculate separated performance scores for each category and subtype. We report the results in Table 2.

The best performance is obtained on the *Swearing* subtype (75.74 micro-F1). However, the results on the two classes, i.e. Offensive and Not offensive, show that this high F1 mainly depends on the good performance yielded on the offensive class. Indeed, swear words tend to be very predictive of offensive content, and have already been recognised in previous studies as so-called *authentic artifacts*, i.e. highly-discriminating and informative tokens in conveying hatefulness (Ramponi and Tonelli, 2022). On the contrary, swear words in non-offensive tweets are both controversial for human annotators and difficult to detect for classifiers (Pamungkas et al., 2020).

For the most numerous subtype, *Personal Bias*, performance is rather low compared to the other types. However, the classifier yields a comparable performance on the offensive and not offensive class, showing that the two classes are equally challenging when annotators' beliefs and background come into play. In general, classification performance on offensive tweets is lower than on not offensive ones, except for *Swearing*, *Threatening* and *Analogy*. Offensive language detection systems tend to perform better on the not offensive class, because it is usually represented by more examples in the training set. Our experiment confirms this trend with few exceptions.

As a further analysis, in Figure 3 we report the classification results (average of 20 runs) on the different subtypes for A^0 and A^+ tweets. For all the categories except for *Analogy*, the classification performance is better on A^+ cases (low disagreement) compared to A^0 (high disagreement). This is probably because assigning a label to A^0 cases through majority voting is rather arbitrary, leading to cases that a system can poorly classify.

5.2 Training with disagreement

Several works showed that training a classifier using data with a low level of agreement is detrimental to the system performance (Reidsma and op den Akker, 2008; Jamison and Gurevych, 2015). We delve further into this issue by evaluating whether this negative effect depends on the presence of a specific class of disagreement in the training set. To this end, we retrain a classification model for offensive language detection using the original training split used in Leonardelli et al. (2021), and we compare it with the performance obtained including in the same training set only the subset of A^0 tweets belonging to a specific category, i.e. either *Subjectivity* or *Missing Information* or *Ambiguity*. The performance of the models is evaluated on the same three categories in the test set. Results are reported in Table 3.

To reliably compare differences between models' performances, we use Almost Stochastic Order (Dror et al., 2019; Del Barrio et al., 2018) in its

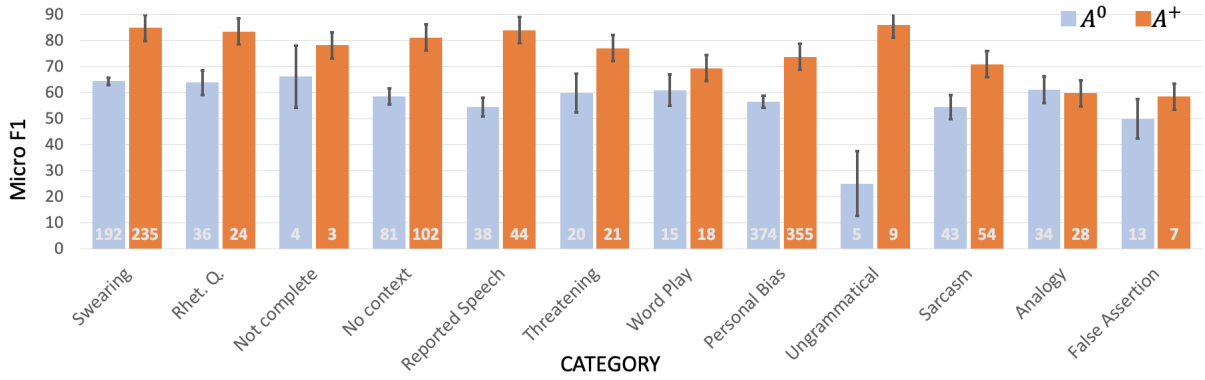

Figure 3: Performance of the classifier for the different agreement levels and categories. Error bars represent standard deviation obtained from 20 restarts.

Training split	Train. size	Testing on *Subj.*	*Missing*	*Amb.*
$A^{++/+}$	1,800	68.95	70.78	65.64
$A^{++/+/0}$	2,700	68.79	69.71	**66.31***
$A^{++/+/0(SUBJ)}$	2,206	68.55	69.29	65.48
$A^{++/+/0(MISS)}$	1,927	68.93	**70.84***	65.79
$A^{++/+/0(AMB)}$	1,923	**69.07**	69.48	64.58

Table 3: Classifier performance (F1) with different versions of the training set, with and without specific classes of disagreement. Statistically significant results (compared to the lowest F1) are marked with (*).

implementation by Ulmer et al. (2022). For each of the three test sets, we compare the models scores across the 20 restarts. For statistical significance a threshold of $\tau = 0.2$ is considered.[1]

As shown in Table 3, while the differences in performance when testing on the *Subjectivity* category are not statistically significant, the best scores obtained on *Ambiguity* and on *Missing Information* show a statistically significant improvement over the lowest F1 and, in the case of *Missing Information*, also over training with $A^{++/+/0(AMB)}$.

For the classification of examples from *Missing Information*, adding only the A^0 examples belonging to the same category of disagreement yields to the best performance among all models. On the contrary, when classifying tweets in the *Ambiguity* class, the best performance is obtained when all A^0 are added to the training set regardless of the tweet category or disagreement level. The lowest performance, instead, is the obtained when adding only A^0 examples from the *Ambiguity* class. This may be due to the fact that this class is the most heterogeneous one, with different subtypes all represented in the data with few examples, covering very different linguistic phenomena. Overall, these results suggest that removing A^0 instances from training is not always the best solution, contrary to what was suggested in previous works (Leonardelli et al., 2021). Instead, distinctions should be made among different types of disagreement when deciding whether to remove training instances or not.

5.3 Multi-task learning with disagreement

Finally, we investigate whether using information about disagreement can improve offensive language detection. We employ a multi-task framework, which has already been used to include disagreement information in classification tasks (Fornaciari et al., 2021; Davani et al., 2022; Ramponi and Leonardelli, 2022). In a multi-task setting, the encoder component is unique and shared between both tasks that during training are jointly fine-tuned. In our case, we consider offensive language detection as the primary task, and disagreement detection as the secondary one, to test whether the latter can provide useful signals to potentially improve the performance on the main task. We test two variants in this respect: *i)* we cast the auxiliary task as a three-way classification aimed at recognising tweets labeled as A^{++}, A^{+} and A^{0}, and *ii)* we implement a more fine-grained version of the previous task, aimed at assigning tweets to one of six classes: A^{++} offensive and not offensive, A^{+} offensive and not offensive and A^{0} offensive and not offensive. All these labels were already provided in the dataset by Leonardelli et al. (2021), so no additional annotation was required. The classifier for offensive language detection is tested separately on the three classes of disagreement *Subjectivity*, *Missing Information* and *Ambiguity*, like in the ex-

[1] Based on Ulmer et al. (2022), this threshold is comparable to a Type I error rate of p-value .05

periments in Section 5.2.

Results are shown in Table 4. The first row presents the single-task setting, i.e. offensive language detection, compared with the multi-task ones, i.e. three-class and six-class classification.

Multitask		**Testing on**		
Task 1	**Task 2**	*Subj.*	*Missing*	*Amb.*
Offensive language	-	68.79	69.61	**66.31**
Offensive language	Agr. level $A^{++,+,0}$	69.15	**69.84**	66.26
Offensive language	Agr. level $N/O^{++,+,0}$	**69.24**	69.34	65.82

Table 4: Classification performance (F1) with multi-task learning. N=Not offensive; O=Offensive.

Although the differences across settings are slight, the best result for the *Subjectivity* category is obtained within the multitask framework with 6-way classification as an auxiliary task. For *Missing Information*, instead, the multitask setting with three-way classification is the best one, while for *Ambiguity* providing auxiliary information on disagreement levels does not seem to yield any improvement. These differences support our intuition that we should distinguish among the different types of disagreement, since different strategies would be necessary to deal with them during classification. However, a statistical analysis similar to the one presented in previous section failed to reveal any significant difference between the models' performances.

6 Conclusions

In this work, we first introduced a two-layered taxonomy for the classification of annotators' disagreement in subjective tasks, consisting in four main categories and a number of subtypes aimed at covering different linguistic phenomena. We then annotated part of an existing dataset developed to study disagreement with the above classes and subtypes. A first analysis shows that the *Subjectivity* class is the prevalent one in the dataset, and that *Personal Bias* and *Swearing* are two major reasons leading to frequent cases of disagreement among annotators for the task of offensive language detection. Secondly, we run several experiments to gain novel insights into disagreement phenomena. In particular, we investigate whether a system for offensive language detection is more prone to wrong classification on specific classes of disagreement. Our results show that the presence of *Sarcasm*, *Analogy* and *False Assertions* negatively affects classifier performance, while *Swearing* obtains the best classification results, despite showing a bias in favour of offensive tweets. Furthermore, cases with high disagreement are generally more difficult to classify than those with mild disagreement for all categories except for *Analogy*. In a second experiment, we show that adding instances of *Missing Information* to the training set, even if they belong to the A^0 class, has a positive effect on the classification of this specific class, while this is not true for *Ambiguity*, probably because it contains more heterogeneous data. Finally, we show that adding disagreement information as an auxiliary task in a multi-task setting, having offensive language detection as the main task, is not generally better, and has different effects on the three above classes.

As regards tweets annotated with multiple labels, we observe that it is rather frequent to find more than one cause of disagreement (around 40% of the tweets in our dataset). We performed a preliminary experiment (not reported in this paper) comparing the performance of the best system configuration on single-label and multi-label examples, and we observed no significant difference. We will further investigate this aspect and compare single- and multi-label items in detail in the future.

In general, we hope that this work can contribute to the ongoing debate on the importance of considering, and not removing, disagreement when creating datasets and when developing classifiers. Furthermore, we advocate for a differentiation of the types of disagreement, showing that their presence in training and test data can have different effects on classification.

Limitations

While the taxonomy of disagreement is designed to be language-independent and to cover the annotation of subjective tasks, we have applied it only to a dataset for offensive language detection in English. Its applicability to other tasks will be investigated in the near future, together with its portability across languages. Also, the small size of the annotated dataset may limit the generalisability of our findings. In particular, the differences in performance across settings in our experiments are not always statistically significant.

Ethics Statement

We do not foresee specific ethical risks related to the current work. On the contrary, the analysis of different types of disagreement is aimed also at making the role of subjective annotations accepted within the NLP research community, making sure that the voices of minorities are included. Indeed, this work contributes to providing methodologies to distinguish subjective annotations from mistakes and poor-quality judgments.

Acknowledgements

Elisa Leonardelli and Sara Tonelli have been funded by the STAND BY ME European project (REC-RDAP-GBV-AG-2020) on "Stop online violence against women and girls by changing attitudes and behaviour of young people through human rights education" (GA 101005641). They have also been supported by the STAND BY ME 2.0 project (CERV-2021-DAPHNE) on "Stop gender-based violence by addressing masculinities and changing behaviour of young people through human rights education" (GA 101049386).

References

Gavin Abercrombie, Valerio Basile, Sara Tonelli, Verena Rieser, and Alexandra Uma, editors. 2022. *Proceedings of the 1st Workshop on Perspectivist Approaches to NLPerspectives@LREC 2022, Marseille, France, 20th June 2022*. European Language Resources Association.

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2019. A new measure of polarization in the annotation of hate speech. In *Proceedings of AI*IA*.

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

V. 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 DP@AI*IA*.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Valerio Basile, Tommaso Caselli, Alexandra Balahur, and Lun-Wei Ku. 2022. Bias, subjectivity and perspectives in natural language processing. *Frontiers in Artificial Intelligence*, 5.

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

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.

Beata Beigman Klebanov, Eyal Beigman, and Daniel Diermeier. 2008. Analyzing disagreements. In *Coling 2008: Proceedings of the workshop on Human Judgements in Computational Linguistics*, pages 2–7, Manchester, UK. Coling 2008 Organizing Committee.

Alessandra Teresa Cignarella, Cristina Bosco, Viviana Patti, and Mirko Lai. 2018. Application and analysis of a multi-layered scheme for irony on the Italian Twitter corpus TWITTIRÒ. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

D.A. Cruse. 2004. *Meaning in Language: An Introduction to Semantics and Pragmatics*. Oxford. Oxford University Press.

B. Dancygier and E. Sweetser. 2014. *Figurative Language*. CUP.

Aida Mostafazadeh Davani, Mark Díaz, and Vinodkumar Prabhakaran. 2022. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

Eustasio Del Barrio, Juan A Cuesta-Albertos, and Carlos Matrán. 2018. An optimal transportation approach for assessing almost stochastic order. In *The Mathematics of the Uncertain*, pages 33–44. Springer.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

T. Donaldson and E. Lepore. 2012. Context-sensitivity. In Gillian Russell and Delia Graff Fara, editors, *Routledge Companion to Philosophy of Language*, pages 116–131. New York, NY, USA: Routledge.

Rotem Dror, Segev Shlomov, and Roi Reichart. 2019. Deep dominance - how to properly compare deep neural models. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 2773–2785. Association for Computational Linguistics.

A. Dumitrache. 2019. *Truth in Disagreement: Crowdsourcing Labeled Data for Natural Language Processing*. Ph.D. thesis, Vrije Universiteit Amsterdam.

Tommaso Fornaciari, Alexandra Uma, Silviu Paun, Barbara Plank, Dirk Hovy, and Massimo Poesio. 2021. Beyond black & white: Leveraging annotator disagreement via soft-label multi-task learning. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2591–2597, Online. Association for Computational Linguistics.

R. W Gibbs. 1986. On the psycholinguistics of sarcasm. *Journal of experimental psychology*, 115(1):3.

E. Holt. 2009. The pragmatics of interaction. *Handbook of Pragmatics Highlights*, 4:190–205.

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.

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.

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating Reasons for Disagreement in Natural Language Inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

Kian Kenyon-Dean, Eisha Ahmed, Scott Fujimoto, Jeremy Georges-Filteau, Christopher Glasz, Barleen Kaur, Auguste Lalande, Shruti Bhanderi, Robert Belfer, Nirmal Kanagasabai, Roman Sarrazingendron, Rohit Verma, and Derek Ruths. 2018. Sentiment analysis: It's complicated! In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1886–1895, New Orleans, Louisiana. Association for Computational Linguistics.

J. Kocoń, A. Figas, M. Gruza, D. Puchalska, T. Kajdanowicz, and P. Kazienko. 2021. Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. *Information Processing & Management*, 58(5):102643.

Sabine Lehmann, Stephan Oepen, Sylvie Regnier-Prost, Klaus Netter, Veronika Lux, Judith Klein, Kirsten Falkedal, Frederik Fouvry, Dominique Estival, Eva Dauphin, Herve Compagnion, Judith Baur, Lorna Balkan, and Doug Arnold. 1996. TSNLP - test suites for natural language processing. In *COLING 1996 Volume 2: The 16th International Conference on Computational Linguistics*.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10528–10539, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

S. C. Levinson. 2004. Deixis. In *The handbook of pragmatics*, pages 97–121. Blackwell.

S.C. Levinson. 1987. Pragmatics and the grammar of anaphora: a partial pragmatic reduction of binding and control phenomena. *Journal of Linguistics*, 23(2):379–434.

M. Ljung. 2011. *Swearing: A Cross-Cultural Linguistic Study*. Palgrave Macmillan.

Héctor Martínez Alonso, Anders Johannsen, Oier Lopez de Lacalle, and Eneko Agirre. 2015. Predicting word sense annotation agreement. In *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics*, pages 89–94, Lisbon, Portugal. Association for Computational Linguistics.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2020. Do you really want to hurt me? predicting abusive swearing in social media. In *The 12th Language Resources and Evaluation Conference*, pages 6237–6246. European Language Resources Association.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Barbara Plank, Dirk Hovy, and Anders Søgaard. 2014. Linguistically debatable or just plain wrong? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 507–511, Baltimore, Maryland. Association for Computational Linguistics.

Massimo Poesio and Ron Artstein. 2005. The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account. In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 76–83, Ann Arbor, Michigan. Association for Computational Linguistics.

Massimo Poesio, Jon Chamberlain, Silviu Paun, Juntao Yu, Alexandra Uma, and Udo Kruschwitz. 2019. A crowdsourced corpus of multiple judgments and disagreement on anaphoric interpretation. 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 1778–1789, Minneapolis, Minnesota. Association for Computational Linguistics.

Alan Ramponi and Elisa Leonardelli. 2022. DH-FBK at SemEval-2022 task 4: Leveraging annotators' disagreement and multiple data views for patronizing language detection. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 324–334, Seattle, United States. Association for Computational Linguistics.

Alan Ramponi and Sara Tonelli. 2022. Features or spurious artifacts? data-centric baselines for fair and robust hate speech detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3027–3040, Seattle, United States. 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.

Paul Rottger, Bertie Vidgen, Dirk Hovy, and Janet Pierrehumbert. 2022. Two contrasting data annotation paradigms for subjective NLP tasks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 175–190, Seattle, United States. Association for Computational Linguistics.

Y. Sang and J. Stanton. 2022. The origin and value of disagreement among data labelers: A case study of individual differences in hate speech annotation. In *Information for a Better World: Shaping the Global Future*, pages 425–444, Cham. Springer International Publishing.

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, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A. Smith. 2022. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5884–5906, Seattle, United States. Association for Computational Linguistics.

Roy Schwartz, Omri Abend, Roi Reichart, and Ari Rappoport. 2011. Neutralizing linguistically problematic annotations in unsupervised dependency parsing evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 663–672, Portland, Oregon, USA. Association for Computational Linguistics.

T. Stivers and N.J. Enfield. 2010. A coding scheme for question–response sequences in conversation. *Journal of Pragmatics*, 42(10):2620–2626.

Kate Storey. 1995. The language of threats. *International Journal of Speech Language and the Law*, 2:74–80.

D. Tuggy. 1993. Ambiguity, polysemy, and vagueness. *Cognitive Linguistics*, 4:273–290.

Dennis Ulmer, Christian Hardmeier, and Jes Frellsen. 2022. deep-significance: Easy and meaningful signifcance testing in the age of neural networks. In *ML Evaluation Standards Workshop at the Tenth International Conference on Learning Representations*.

A. Uma, T. Fornaciari, D. Hovy, S. Paun, B. Plank, and M. Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, Vol. 72.

Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 176–197, Online. Association for Computational Linguistics.

Maximilian Wich, Jan Bauer, and Georg Groh. 2020. Impact of politically biased data on hate speech classification. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 54–64, Online. Association for Computational Linguistics.

A Example Annotations

We report in Table 5 some example tweets with different degrees of annotators' agreement and the corresponding category & subtype. In particular, A^{++} corresponds to total agreement (5/5 crowdsourced annotations with the same label); A^{+} - mild agreement (4 vs 1 labels); A^{0} - disagreement (3 vs 2 labels).

Agreement	Offensiveness	Example	Category	Subtype
A^{++}	N^{++}	Google bans ads on coronavirus conspiracy theory content -url	-	-
A^{+}	N^{+}	@user Freaking love it	Missing Info	No Context
A^{0}	N^{0}	@user How many more George Floyd will die under white domination! Sad!	Ambiguity	Rethorical question
A^{0}	O^{0}	He is another nut case.	Subjectivity	Personal bias
A^{+}	O^{+}	@user You're a bumbling fool #elections2020	Subjectivity	Personal bias
A^{++}	O^{++}	f %* king insanity	-	-

Table 5: Examples of tweets with different degrees of crowd-workers' agreement. N=Not offensive, O=Offensive. $++/+/0$ correspond to high, medium and low agreement respectively. The last two columns include disagreement category and subtype.

B Annotation Guidelines

In order to annotate a text by assigning a category and a subtype of disagreement from the proposed taxonomy, annotators should perform the following steps:

1. Check the text for sources of disagreement: detect the main one and assign it to a category among *Sloppy annotation*, *Ambiguity*, *Missing Information* and *Subjectivity*;
2. Choose a fine-grained class to specify the main source of disagreement
3. Check for a secondary source of disagreement (if any): assign the text to another category (it could be the same as the previous one, with a different subtype)
4. Choose a secondary subtype to specify the secondary source of disagreement

We also report questions to guide the assignment of labels to text instances.

- **Sloppy Annotation** (Label = Sloppy_Annotation)
  - **Noise**: is the text clearly not offensive but marked as such (or vice versa)? (Label = Noise)
- **Ambiguity**: are there multiple interpretations to the text but is it not clear which is the correct one? (Label = Ambiguity)
  - **Analogy**: does the text include a figure of speech that comprehends mechanisms of comparison (included: simile and metaphor) or is the user referring to someone with a periphrasis (e.g., "orange monkey", "bunker boy" for Donald Trump)? (Label = Analogy)
  - **False assertion**: does the user express the opposite of what they think or something wrong with respect to a context? (Label = False_Assertion)
  - **Rhetorical question**: does the text include a question asked in order to make a point rather than to elicit an answer? (Label = Rhetorical_Question)
  - **Sarcasm**: is the text employed to communicate the opposite of its surface meaning in a humorous way or to mock someone/something? (Label = Sarcasm)
  - **Word Play**: does the text include any acronyms, alliterations or puns? (Label = Word_Play)
  - **Reported Speech**: does the text report something someone else stated? For example a newspaper headline? (Label = Reported_Speech)
- **Missing Information**: is the disagreement caused by difficulty of interpretation? (Label = Missing_Info)
  - **Ungrammatical**: does the text include typos or non standard expressions nullifying its comprehension? Do not consider "your/you're", "its/it's" since they are very frequent and do not affect text comprehension (Label = Ungrammatical)
  - **No context**: does the text contain (Label = No_context):
    - Reference to other users?
    - Links?
    - Anaphoric or deictic pronouns without an explicit referent?
    - Demonstrative pronouns?
  - **Not complete**: some parts are missing: is the text not complete? (Label

= Not_Complete)

- **Subjectivity**: does the text contain information that makes annotators' opinions interfere in their judgment? (Label = Subjectivity)
  - **Personal Bias**: does the text contain specific words that can be interpreted in a subjective way by the annotator (for example: "racist", "fascist", "clown", "pathetic", "liar", "pig") or refer to specific, critical opinions (no vax, wearing or not wearing masks)? (Label = Personal_Bias)
  - **Swearing**: does the text include swearing words (for example: "prick", "turd", "crap", "bullshit", "moron", "dumb")? Does it include ableist insults such as "retarded", "psycho" and expressions containing "shit" (for example: "cut the shit", "don't know shit"). Do not consider WTF, SMFH and similar acronyms containing "fuck". (Label = Swearing)
  - **Threatening**: does the text contain linguistic violence (for example: "shut up", "get out", "you are going to prison")? (Label = Threatening)