A Disaggregated Dataset on English Offensiveness Containing Spans

Pia Pachinger[1], Janis Goldzycher[2], Anna M. Planitzer[3], Julia Neidhardt[1], Allan Hanbury[1]
[1]Faculty of Informatics, TU Wien. Correspondence: pia.pachinger@tuwien.ac.at
[2]Institute for Computational Linguistics, University of Zurich
[3]Political Commmunication Research Group, University of Vienna

Abstract

Toxicity labels at sub-document granularity and disaggregated labels lead to more nuanced and personalized toxicity classification and facilitate analysis. We re-annotate a subset of 1983 posts of the *Jigsaw Toxic Comment Classification Challenge* and provide disaggregated toxicity labels and spans that identify inappropriate language and targets of toxic statements.

Manual analysis shows that five annotations per instance effectively capture meaningful disagreement patterns and allow for finer distinctions between genuine disagreement and that arising from annotation error or inconsistency. We find that disagreement often stems from divergent interpretations of edge-case toxicity. This pattern is particularly pronounced in cases of toxic statements involving non-human targets, where disagreement is especially high. Further, when evaluating whether a span consists of inappropriate language, disagreement occurs not only on inherently questionable terms, but also on words that may be inappropriate in specific contexts while remaining acceptable in others. Lastly, we show that transformer-based models effectively learn from aggregated data that reduces false negative classifications by being more sensitive towards minority opinions for posts to be toxic. We publish the new annotations under the CC BY 4.0 license [1].

Content warning: This paper contains examples of offensive language to describe the data.

1 Introduction

The amount of toxic[2] content on the internet is increasing and causes harm. Especially implicit offensiveness still often goes undetected (Zhang et al., 2022). Creating effective automated content moderation systems requires two key elements: nuanced understanding of online norm violations and incorporation of diverse opinions on what content warrants moderation.

The need for more perspectivist offensiveness detection Perceptions of what content is harmful depend on individual, contextual, and geographical factors (Hershcovich et al. 2022; Sandri et al. 2023; Abercrombie et al. 2023 i.a.). Researchers report disagreements in the annotation of toxicity explained by previous annotations by the same annotator (Wich et al., 2020), sociodemographics (e.g. Kocoń et al. 2021; Aroyo et al. 2023) beliefs (Sap et al., 2022), and moral values and geocultural factors (Davani et al., 2023). The human perception of what constitutes harmful language is inherently reflected in the classifiers trained on human labelled data to identify toxic speech. If perceptual differences in what makes a remark toxic are not taken into account when training these systems, what is considered toxic may be disproportionately influenced by the societal majority or, in the worst-case scenario, by an arbitrary group of annotators. Therefore, a one-size-fits-all approach to content moderation is unable to account for the diverse needs of different users (Cresci et al. 2022; Plank 2022; Jhaver et al. 2023 i.a.).

Perspectivism in machine learning (ML) represents a paradigm shift from consensus-based labeling to embracing annotator diversity. This approach collects multiple labels for each data instance and aims to preserve disaggregated annotations along with annotator metadata throughout the entire ML pipeline (Cabitza et al., 2023). This paradigm aligns with descriptive annotation approaches, which embrace annotator subjectivity as meaningful signal rather than noise. Descriptive annotation enables the modeling of diverse beliefs and perspectives, contrasting with prescrip-

[1] https://github.com/pi-pa/disaggregated_offensiveness
https://web.ds-ifs.tuwien.ac.at/toxic_spans

[2] As there are no generally accepted distinctions for *offensiveness* and *toxicity* (Pachinger et al., 2023), we use these terms interchangeably.

tive annotation that enforces uniform interpretation through strict guidelines (Röttger et al., 2022).

The need for offensiveness annotations beyond classes In (semi-) automated content moderation, explainability contributes to a greater understanding and trust of users (Molina and Sundar, 2022) and content moderators (Bunde, 2021). Experienced moderators work more efficiently when provided with structured explanations that pinpoint harmful content and articulate why it violates community standards (Calabrese et al., 2024). In text annotation, a span refers to a contiguous sequence of tokens within a document that is marked and labeled. Rather than annotating entire documents or sentences, spans allow annotators to identify and categorize specific portions of text. Annotators can pinpoint exactly which parts of a text exhibit the phenomenon of interest. Text classification models can learn from the specific linguistic features within marked spans, leading to more accurate predictions about similar text segments. Different spans within the same text can receive different labels, capturing the complexity of real-world documents where multiple phenomena may coexist, facilitating error analysis, model debugging, and model explainability (Lyu et al., 2024).

Main contributions of this paper By re-annotating 1,983 comments from the Jigsaw Toxic Comment Classification dataset, 1,561 of which received annotations from four or five annotators, we classify toxic utterances at the post level and we identify spans comprising the targets of toxic utterances, and spans comprising vulgar expressions while maintaining disaggregated labels from multiple annotators.

Our analysis shows that five annotations per instance effectively capture meaningful disagreement patterns and allow for finer distinctions between genuine disagreement and that arising from annotation error or inconsistency. Disagreements originate from divergent interpretations of edge-case toxicity, differential assessment of specific lexical choices, and varying judgments regarding edge cases in vulnerability of targets. For example, we find substantial disagreement on toxicity classifications involving non-human targets. Further, when evaluating whether a span consists of inappropriate language, disagreement occurs not only on inherently questionable terms, but also on words that may be inappropriate in specific contexts while remaining acceptable in others. Lastly, while annotators generally recognize targets of toxic language, repeated target mentions within comments can pose hurdles for human annotators and for ML approaches to extracting spans comprising targets.

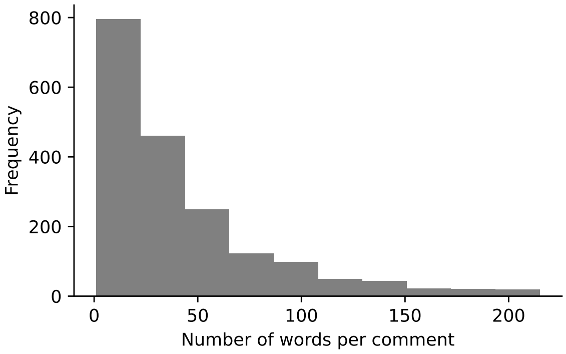

Figure 1: The distribution of the comment lengths of all comments shorter than the 95% percentile of the comments sorted by length.

Despite differences in definitions and annotation approaches, we find broad agreement between the Jigsaw annotations and our own. Additionally, experiments show that transformer-based models effectively learn from aggregated data that reduces false negative classifications by being more sensitive towards minority opinions for posts to be toxic. We release the new annotations under CC BY 4.0 licensing with the underlying comment text being governed by Wikipedia's CC-SA-3.0.

2 Related Work

Recent advances in toxicity annotation include the development of granular labeling frameworks, publication of disaggregated annotation datasets, and diversification of annotator demographics.

Disaggregated annotations on toxicity In recent years, researchers have started to publish disaggregated toxicity annotations. For example, Kumar et al. 2021 release a labeled toxicity dataset that contains 107,620 texts and annotations by 17,280 annotators. Another example is the dataset publised by Kennedy et al. 2020, which contains 50,000 texts and annotations by 11,000 mechanical turkers. See Frenda et al. 2024 for more perspectivist datasets on online toxicity.

Toxicity annotations beyond classes Additionally, there has been a surge in datasets related to offensive text detection with span and free-text annotations. Existing data with annotated spans include spans of the targets of offensive statements

(Calabrese et al., 2022a; Zampieri et al., 2023; Pachinger et al., 2024), the spans contributing to the offensiveness label (Mathew et al., 2021; Pavlopoulos et al., 2021), spans comprising a violation of a moderation policy (Calabrese et al., 2022a), and the spans comprising vulgar language (Pachinger et al., 2024). More recently, free-text annotations related to toxicity labels were released (Sap et al., 2020; Zhang et al., 2022; Zhou et al., 2023). The spans and free text can be used to create inherently faithful explain-then-predict methods for offensive text detection (Kim et al., 2022; Zhang et al., 2022; Zhou et al., 2023). Furthermore, they can be used to create post-hoc explanations (Risch et al., 2020).

Annotator populations in related work Toxicity annotation studies on English toxic content typically rely on annotators from English-speaking countries, particularly the United States. Zhou et al. 2023 engage native English speakers for labeling offensive content, while Sap et al. 2020 recruit annotators exclusively from the U.S. and Canada. Calabrese et al. 2022b similarly restrict their pool to English-speaking countries. While Zhang et al. 2022 broaden their criteria to include anyone with English proficiency, they do not systematically ensure demographic diversity. This geographic concentration raises questions about the generalizability of toxicity judgments, particularly given that English is the dominant lingua franca spoken by a wide variety of people and perceptions of harmful content vary significantly across countries.

3 Data Source

We source the data from the *Toxic Comment Classification Challenge* [3] from Jigsaw. It contains Wikipedia comments which have been labeled by human raters for toxic behavior. Annotated labels in the dataset are *toxic*, *severe toxic*, *obscene*, *threat*, *insult*, *identity hate*. The data is published under the CC0 License, with the underlying comment text being governed by Wikipedia's CC-SA-3.0. As we are interested in nuanced toxicity cases, we select comments labeled as *toxic* without any additional toxic categories or *insult* without any additional toxic categories, excluding comments with more severe or multiple toxic labels. From this pool of data, we randomly source 1983 posts.

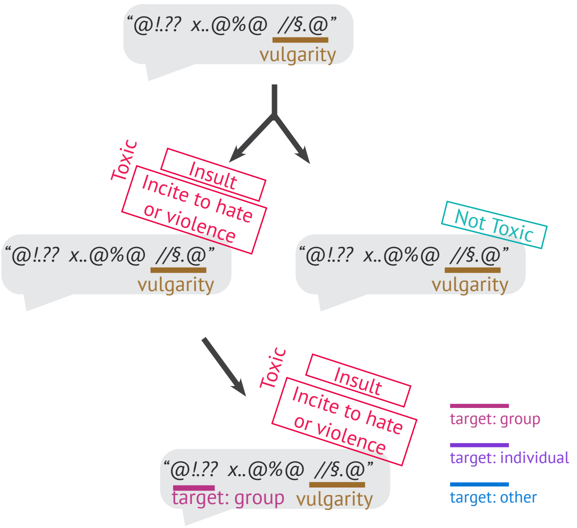

Figure 2: The annotation strategy for this dataset

4 Annotation Schema

We adopt the annotation schema used for the German AustroTox dataset (Pachinger et al., 2024), making the two datasets containing different data sources and cohorts of annotators compatible and allowing for multilingual analyses. Observe the annotation strategy in Figure 2. We classify each comment as insult, incite to hate or violence or not offensive. Since we do not source for incites to hate or violence in the Jigsaw dataset, the amount of posts labelled as *Incite to hate of violence* in our dataset is limited. Therefore, we create an *Offensiveness / Toxicity*[4] class by merging classes *Insult* and *Incite to hate or violence*. For non-offensive and offensive comments, we annotate vulgarities since both, offensive and non-offensive posts can contain vulgarities. For offensive posts, we additionally annotate the targets of the offensive statement and the type of target. If the target is only mentioned via a pronoun, we annotate the pronoun as the target. Adopting a definition of *vulgarity* similar to that employed by Risch et al. (2021), We use the following definitions for classes and spans:

Insult An insult pursues the recognisable goal of disparaging the addressee or the object of reference.

Incite to Hate or Violence An incite to hate or violence against a person or a group of people. It is often hard to draw the line between insults and incites to hate, as insults always somewhat incite

[3] https://www.kaggle.com/c/jigsaw-toxic-comment-classification-challenge

[4] As there are no generally accepted distinctions for *offensiveness* and *toxicity* (Pachinger et al., 2023), we use these terms interchangeably.

hate. For this annotation task, we define insults to be less severe than incites to hate or violence.

Offensive / Toxic An insult or an incite to hate or violence.

Not Offensive / Not Toxic Not an insult nor an incite to hate or violence.

Vulgarity Rude, obscene, foul or boorish language that is inappropriate for civilized discourse.

Target Group The target of an offensive post is a group of persons or an individual insulted based on shared group characteristics.

Target Individual The target of an offensive post is a single person not insulted based on shared group characteristics.

Target Other The target of an offensive post is not a person or a group of people.

We position our work within the descriptive annotation paradigm, recognizing that toxicity perception contains inherent subjective elements. Determining whether someone intends to disparage a target, distinguishing between hate incitement and mere insults, and identifying when comments cross into inciting violence all involve subjective judgment calls. Similarly, the threshold for what constitutes inappropriate passages in civilized conversation varies across readers and contexts. Among our annotation tasks, identifying the target of offensive statements and categorizing the target type represents the most objectively answerable component. Our definition of vulgar language is deliberately expansive, extending beyond conventional sexual, scatological, and religious profanity to include any language inappropriate for civilized discourse. We recognize that determinations of vulgarity are both context-dependent and inherently subjective, as language deemed acceptable in casual forums may prove inappropriate in structured, goal-oriented discussions and edge-case acceptability varies by reader.

5 Annotation Campagin

Annotator cohort We conduct the annotation with master's students in data science. Thirty percent of annotators are registered as female in this course, though this institutional designation may not reflect their actual gender identity. The majority of annotators are between 19 and 26 years old, with all demonstrating at least B2-level English proficiency. Most annotators originate from Eastern Europe.

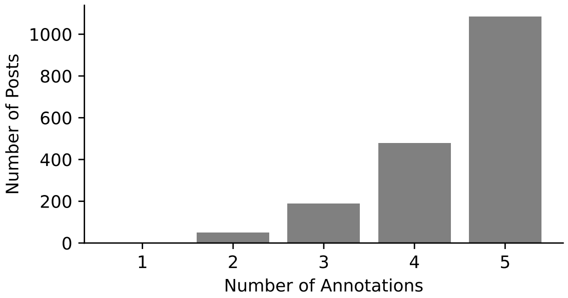

Figure 3: The number of annotators per post.

We acknowledge the importance of representing the perspectives of a population by including diversity in the annotator pool (Clemmensen and Kjærsgaard, 2022). Our annotator pool's homogeneous sociodemographics, constrained by resources and participants in the course, present a limitation in our data. On the other hand, the demographic composition in our dataset contrasts with typical NLP annotation practices, where annotators are predominantly recruited from English-speaking countries. However, English-language online discussions reach global audiences with diverse cultural backgrounds and perspectives on toxicity. We therefore argue that our annotator pool provides valuable demographic diversity to the current landscape of toxicity datasets. Furthermore, opinions on what constitutes toxicity are influenced by a multitude of factors beyond just sociodemographics. While we do not explicitly capture annotator characteristics for individual annotators, these factors are implicitly reflected in the disaggregated annotations (Geva et al., 2019; Wich et al., 2020). Consequently, we view our dataset as a valuable addition to the broader collection of resources, capturing user perspectives in various ways.

Annotation campaign The annotation campaign was reviewed by the ethics committee of our institution. Each annotator annotates about 200 comments, that takes approximately 1.5 to two hours. The dataset contains a higher proportion of offensive comments than the typical distribution in a user forum, but we only source comments with labels *toxic* or *insult* and exclude more severe labels. The annotators are explicitly informed that they have the option to cease annotation if they feel overwhelmed by the task without facing con-

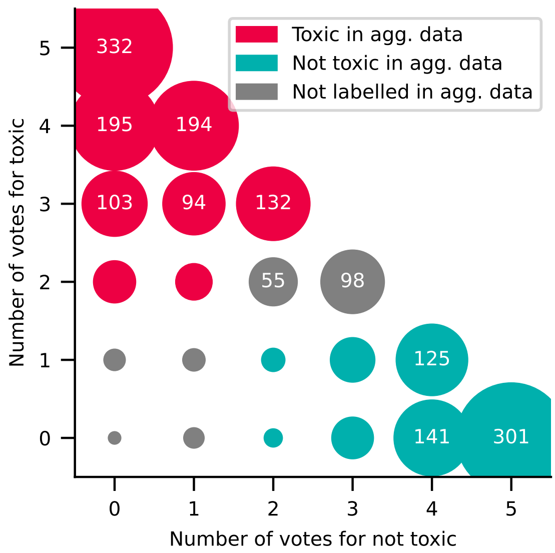

Figure 4: The disagreement in the toxicity annotations. The colors denote the label of the aggregated dataset.

sequences and about the publication of the data and they receive a comprehensive compensation through course credits for their efforts. Figure 3 shows the number of annotators per comment. A vast majority of posts is annotated by five annotators.

6 Disagreements in the Annotations

We calculate inter-annotator agreement using Krippendorff's Alpha across all annotation categories and conduct manual analysis to identify factors correlating with disagreement at both the post and span levels.

Disagreements in offensiveness annotations Figure 4 visualizes the distribution of annotator agreement on offensiveness labels on the post-level. Most posts show complete annotator agreement on whether they are toxic or non-toxic. We report a Krippendorff's alpha of 0.57 for binary offensiveness classification. While this falls below the $\alpha \geq 0.667$ threshold typically recommended for tentative conclusions in prescriptive annotation paradigms (Krippendorff, 2018), it aligns with values reported in comparable toxicity detection studies: Sap et al. (2020) report a Krippendorff's Alpha of 0.51, and Wulczyn et al. (2017) report an Alpha of 0.45.

To understand the origins of disagreements in posts where annotators highly disagree, we manually analyze 60 posts that are annotated as toxic by 3 annotators and as non-toxic by 2 annotators. Table 1 presents all factors related to disagreement we identify through this analysis. We identify subjective elements in 46 of these posts, primarily involving grey-zone or nuanced toxicity that falls into borderline categories. Observe an example where toxicity is subjective and open to interpretation: "*Quiet, you. Whether you are a troll or not is irrelevant - your edits are trolling, are uncivil, and are ridiculous.*"

Additionally, 12 posts contain toxicity directed at non-human targets. We find that toxicity toward non-human entities is typically perceived as less severe and therefore more open to interpretation. The example from above illustrates such a case. Despite viewing our annotation guidelines as descriptive, in some cases, it is possible to definitively say that an utterance is an insult or incitement to hate or violence. We identify 7 such clear-cut cases among the 60 posts we analyze. 5 of the 60 high-disagreement posts we analyze contain quotes with toxic language, such as: "*Vandalism. How's about I stick ""W*nkers Haa HAAa"" in your block log?*". We obfuscate vulgarities and specific targets in this paper.

Subjective whether insult	41
Non-human target	12
Definitely insult	6
Subjective whether incite to h. or v.	5
Toxicity in quote	5
Calls target to leave conversation	5
Definitely not toxic	4
Lack of context	4
Particularly long post	4
Toxic against self	3
Spam (not in our toxicty definition)	2
Definitely incite to hate or violence	1

Table 1: Factors related to disagreements in the offensiveness classifications we identified in 60 comments

Disagreements in vulgarity annotations Figure 5 displays disagreement patterns in span annotations comprising vulgarities. We visualize only spans that at least one annotator marks as vulgar. Most spans perceived as vulgar receive annotations from only one or two annotators. Correspondingly, vulgar span annotations achieve a Krippendorff's Alpha of 0.05, indicating substantial disagreement among annotators.

To understand this low agreement, we manually evaluate 50 spans that receive 3 votes for vulgarity

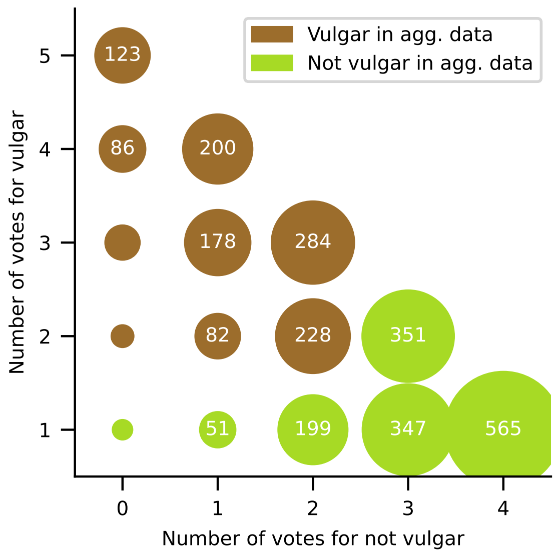

Figure 5: The disagreement on the spans comprising vulgarities. The colors denote the label of the aggregated dataset.

while remaining unannotated by the other 2 annotators. Our analysis reveals that contextual factors critically influence vulgarity perception. Out of these 50 spans, 10 spans are unambiguously vulgar. Take for example the post "*Hey, I said it was ""a"" seat, not ""the"" seat, you dumb motherf#$ker!!*". We classify the span "*motherf#$ker*" as unambigously vulgar. Further, 17 of the 50 spans are subjectively vulgar (where the word's inherent nature is debatable). Recall that our definition of vulgar language is expansive, extending beyond conventional sexual, scatological, and religious profanity to include any language inappropriate for civilized discourse. Taking the previous example, consider the span "*dumb*". While this term carries multiple meanings, within this particular setting it functions as a clear synonym for stupid. The acceptability of such language varies significantly among individuals. We hypothesize that the imagined forum context plays a critical role, where informal forums may accept different language than goal-oriented discussion spaces.

Moreover, 15 spans demonstrate context-dependent vulgarity (where the same word becomes vulgar or benign depending on usage). Take for example the post "*He jailed 50 000 murders, thiefes, rapers, criminals, drug-sellers, prostitutes and many more in only 9 month what you couldn´t do in your 6000 years of history. Stupid losers*" The span "*Stupid*" falls under both categories. Whether it is appropriate for civilized discourse depends on the reader. But, additionally, the word is used in direct speech and against a group that might be vulnerable, which might make it appear more inappropriate to some than in other settings. Lastly, 6 spans are clearly non-vulgar, and one is incomprehensible. This distribution demonstrates that vulgarity annotation involves both lexical ambiguity, whether words are inherently vulgar, and contextual complexity, whether usage renders otherwise benign words inappropriate.

Disagreements in target annotations Figure 6 displays the distribution of annotator votes for spans constituting targets of toxic statements. The visualization includes all target types and only spans that at least one annotator identifies as a target. Most target spans receive annotations from only one annotator. We calculate disagreement based on annotators who label posts as toxic and obtain a Krippendorff's alpha of -0.05 for target annotations, indicating substantial disagreement similar to vulgarity span annotations. Identical pans labeled as different target types are treated as disagreements.

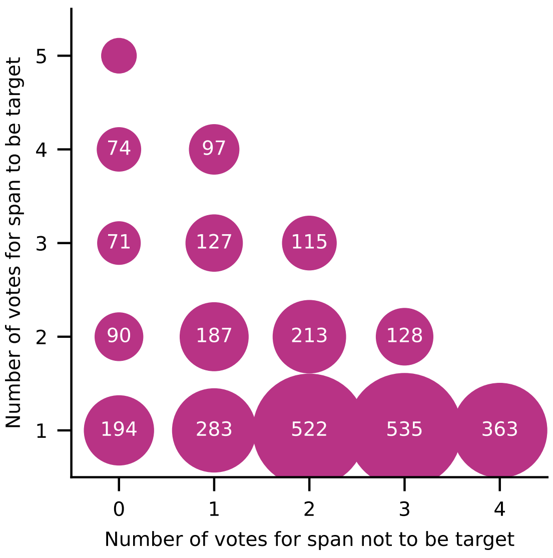

Figure 6: The disagreement on the spans comprising targets

We analyze 30 posts with 5 votes identifying a span as a target. All except one are correctly annotated. In 10 posts, the target appears multiple times, and in 2 cases, other potential targets appear in the data. We further analyze 30 spans with 3 votes for target classification and 2 votes against. All except one span are indeed targets of toxic remarks. In 16 cases, the target appears multiple

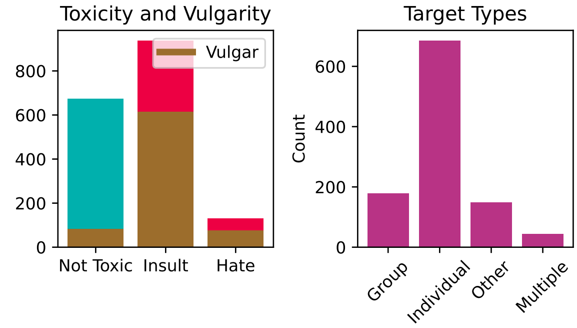

Figure 7: Post-level classes defined by the classes and spans labeled in the aggregated data

times in the post, and in 7 cases, multiple potential targets are mentioned.

In summary, the frequent repetition of targets within posts and target mentions as pronouns creates annotation challenges. This repetition creates inconsistency, some annotators mark the target closest to the most toxic passage, while others mark its first appearance in the post. We advise authors to provide clear instructions for annotating targets of toxic statements, given their highly diverse manifestations.

7 Disagreement Between the Jigsaw Dataset and our Dataset

To enable meaningful comparisons with both the Jigsaw dataset annotations and the German AustroTox (Pachinger, 2024) dataset using a different data source but shared annotation framework, we aggregate the data using an approach that reduces false negative classifications by being more sensitive towards minority opinions for posts to be toxic.

Data aggregation We adopt the same aggregation strategy as the AustroTox dataset for both classes and spans. This approach prioritizes avoiding false negatives for minority perspectives by creating broader decision boundaries. Specifically, we exclude examples with high disagreement that lean slightly toward non-toxic, while labeling examples with disagreement that lean slightly toward toxic as toxic. Figure 4 illustrates how different combinations of toxic versus non-toxic vote counts are labeled for post-level offensiveness classification. We label posts as non-toxic only when they receive at most one toxic vote and at least two non-toxic votes. We discard posts with less than two votes for one class, posts with 2 votes for a post to be toxic and 2 or 3 votes for a post to be not toxic. The remaining posts are labeled as offensive. The aggregated dataset results in a Krippendorff's alpha of 0.64. This value is higher than for the disaggregated data due to the fact that we discard instances with high disagreement and viewer perceptions of toxicity.

Figure 5 shows how we aggregate the spans comprising vulgarities. We label spans as vulgarities if they are annotated by at least two annotators and are not left out by more than two annotators. Spans in the comments comprising the different target types are annotated by majority voting of those who labeled the comment as offensive. If two spans receive a majority for a target span, both are annotated as the respective type of target. We combine the aggregated post-level classifications (*offensive* and *not offensive*) with span annotations to create fine-grained categories. This allows us to identify which types of toxicity are most prevalent in the dataset according to broad annotator consensus. Figure 7 shows the distribution of these categories and the frequency of different target types appearing in toxic utterances.

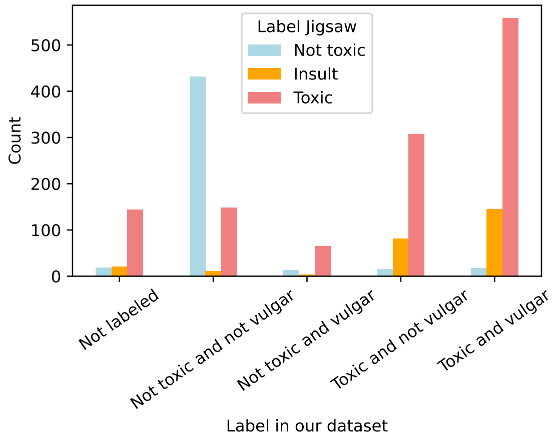

Figure 8: Labels in our vs. labels in the Jigsaw dataset

Disagreement between the Jigsaw dataset and our dataset Figure 8 compares the label distributions between the Jigsaw dataset and our aggregated dataset. The majority of posts classified as offensive in our dataset correspond to the labels *toxic* or *insult* in the Jigsaw dataset. While the definition of toxicity differs between the two datasets, this broad alignment suggests general agreement among majority opinions regarding what constitutes offensive content. In the area of toxic comment classification and hate speech detection, the labels *profane* and *vulgar* are often used similarly. Comments labeled as *not toxic and vulgar* in our dataset are categorized as *toxic* in the Jigsaw dataset, yet they do not receive the label *profane* in the Jigsaw dataset, suggesting that we use a broader definition for vulgarity than was used for the original data. The most significant divergence involves 155 posts that our dataset labels as *not toxic* and *not vulgar* but that Jigsaw categorizes as *toxic*.

We manually review 50 of the 155 posts that our dataset classified as non-toxic and non-vulgar but that the Jigsaw dataset labeled as toxic. We find 14 comments to be genuinely non-toxic, while 12 fall into a subjective gray area where toxicity judgments could reasonably vary. An additional 8 comments appear toxic under broader definitions than ours, for instance, posts containing accusations of lying or spam that we would not classify as toxic, and one comment is toxic according to our definition. Further, 8 comments contain vulgar language, with 5 additional posts falling into a subjective category for vulgarity. Five comments use slang or specialized abbreviations that could lead to different interpretations across annotators. Lastly, 4 posts lack sufficient context for reliable assessment.

In summary, our analysis reveals three primary sources of disagreement between the datasets. Subjective interpretation challenges are the most prevalent issue, affecting 26 of the 50 comments. These include posts lacking sufficient context, containing specific language or slang, or falling into gray areas for toxicity or vulgarity assessment where annotators can reasonably disagree. Definitional differences explain 8 cases where comments appear toxic under Jigsaw's broader criteria but not ours. Potential labeling inconsistencies exist as well, with 8 comments appearing vulgar despite being labeled non-vulgar in our dataset, and 14 comments seeming non-toxic despite Jigsaw's toxic classification.

8 Neural Experiments

We conduct experiments on the aggregated data in order to show that the labels provide learnable signals. We conduct experiments on binary offensiveness classification, token classification of vulgar passages, and passages constituting the different types of targets.

Models We fine-tune and evaluate encoder-based models and we evaluate the few-shot performance of decoder-based models in a 10-fold-cross validation setting. We fine-tune encoder-based models on all three tasks independently. This means that the target detection task inherently includes offensiveness classification, as we only annotate targets of offensive statements. We choose ELECTRA Large[5] (Clark et al., 2020) and Roberta Large[6] (Liu et al., 2019) for our experiments, as they exhibit good performance at the SemEval-2023 task 10: explainable detection of online sexism (Kirk et al., 2023). Additionally, we assess the in-context learning performance of the following large language models: GPT 3.5[7] (*gpt-3.5-turbo-1106*) (Ouyang et al., 2022), GPT 4 [8] (*gpt-4-1106-preview*) (et al., 2024), and Mistral [9] (Jiang et al., 2023). We use the same prompts as Pachinger et al. 2024. They contain an offensiveness definition, the post to be classified and for the five-shot scenario, randomly sampled annotated example posts. Due to limited performance, we define the token with the higher logit as the Llama3 and Mistral's prediction. We tokenize the spans generated by the generative models with the Roberta tokenizer. We compute the Micro F1 by adding up the values of the confusion matrix for the three target classes using Nakayama's (2018) framework.

Results Table 2 presents the evaluation results. Several important limitations should be noted, particularly for the in-context learning experiments. We did not fine-tune the decoder-based models or perform prompt optimization, meaning the decoder-based model results represent a lower bound of achievable performance rather than optimal outcomes. Nevertheless, these results demonstrate that the models can achieve reasonable performance on several tasks, which serves our primary objective. The models perform better on our dataset than on AustroTox, which results in Binary and F1 score of 0.76 for offensiveness classification and 0.71 for vulgarity token-classification, and a Micro F1 score of 0.24 for target classification (Pachinger et al., 2024). We attribute this to the general prevalence of English in NLP and to the distinct data sources. Further, the fine-tuned smaller language models

[5] https://huggingface.co/google/electra-large-discriminator

[6] https://huggingface.co/FacebookAI/roberta-large

[7] https://platform.openai.com/docs/models/gpt-3-5

[8] https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo

[9] https://huggingface.co/mistralai/Mistral-7B-v0.1

			Offensive		**Vulgarity**		**Target**	
			Post-level, 2 cls		Token-level, 2 cls		Token-level, 4 cls	
		Params	Binary	Macro	Binary	Macro	Micro	Macro
Electra	Large	335M	**.88** ± 04	.79 ± 15	.64 ± 24	**.87** ± 07	.08 ± 12	.35 ± 16
Roberta			**.90** ± 02	**.86** ± 03	**.77** ± 03	**.89** ± 02	**.27** ± 03	**.59** ± 04
Mistral	0-Shot	7.24B	.48 ± 05	.55 ± 04	-	-	-	-
	5-Shot		.77 ± 03	.73 ± 03	-	-	-	-
Llama3	0-Shot	8B	.78 ± 03	.75 ± 04	-	-	-	-
	5-Shot		.82 ± 02	.75 ± 03	-	-	-	-
GPT 3.5	0-Shot	-	**.89** ± 02	**.85** ± 02	.46 ± 04	.72 ± 02	.16 ± 02	.50 ± 02
	5-Shot		**.89** ± 02	**.85** ± 03	.47 ± 02	.73 ± 01	.18 ± 03	.52 ± 03
GPT 4	0-Shot	-	.87 ± 03	**.84** ± 03	.41 ± 06	.70 ± 03	.15 ± 02	.49 ± 02
	5-Shot		**.89** ± 02	**.86** ± 02	.43 ± 04	.71 ± 02	.18 ± 02	.52 ± 02

Table 2: Mean F_1 scores and standard deviations of ten-fold cross-validation on the different tasks. Cls stands for the number of classes for the respective task. The Micro F1 scores were computed leaving out the negative class since the negative class is highly prevalent. Values in bold are statistically insignificantly different.

perform better in all tasks on our data. However, fine-tuning the decoder-based models would likely improve their performance significantly. These results suggest that fine-tuning yields better outcomes in this setting, particularly for detecting vulgar content.

Inline with the results of the experiments on the AustroTox dataset, we find that especially the vulgar token detection task profits from fine-tuning. None of the models achieve good performance on target token detection. However, several factors explain these poor results. First, this is a challenging four-class classification task where the evaluation using the Micro-F1 score focuses only on target classes, excluding the predominant non-target class. This evaluation approach, combined with the sparse distribution of target spans in the data, inherently produces lower scores compared to the other two tasks. The high level of human annotator disagreement provides additional insight into these performance issues. Targets frequently appear multiple times within posts and are often referenced only through pronouns, creating ambiguity. Given that human annotators struggled with the task, the poor model performance becomes more understandable.

9 Conclusion

We re-annotate posts from the Jigsaw Toxic Comment Classification Challenge, providing disaggregated toxicity labels and spans that identify inappropriate language and targets. This sub-document granularity enables more nuanced and personalized toxicity classification. Manual analysis demonstrates that five annotations per instance effectively distinguish meaningful disagreement from annotation inconsistencies. We find high levels of disagreement on borderline toxicity cases, particularly for toxic statements targeting non-human entities. Additionally, when annotating spans comprising inappropriate language, disagreement occurs both on inherently questionable terms and on context-sensitive words that may be acceptable in some settings but inappropriate in others. Finally, experiments show that transformer-based models effectively learn from aggregated data that reduces false negative classifications by being more sensitive towards minority opinions for posts to be toxic.

Acknowledgements

At the time of conducting the main parts of this project, Pia and Anna were funded by the Vienna Science and Technology Fund (WWTF) [10.47379/ICT20015]. Janis is funded by the University of Zurich Research Priority Program project *Digital Religion(s)*. We thank the participants in the annotation campaign. Their dedication and effort have been invaluable to the success of this project. Lastly, the financial support by the Christian Doppler Research Association is gratefully acknowledged.

References

Gavin Abercrombie, Dirk Hovy, and Vinodkumar Prabhakaran. 2023. Temporal and second language influence on intra-annotator agreement and stability in hate speech labelling. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 96–103, Toronto, Canada. Association for Computational Linguistics.

Lora Aroyo, Alex Taylor, Mark Diaz, Christopher Homan, Alicia Parrish, Gregory Serapio-García, Vinodkumar Prabhakaran, and Ding Wang. 2023. Dices dataset: Diversity in conversational ai evaluation for safety. *Advances in Neural Information Processing Systems*, 36:53330–53342.

Enrico Bunde. 2021. Ai-assisted and explainable hate speech detection for social media moderators–a design science approach.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 6860–6868.

Agostina Calabrese, Leonardo Neves, Neil Shah, Maarten Bos, Björn Ross, Mirella Lapata, and Francesco Barbieri. 2024. Explainability and hate speech: Structured explanations make social media moderators faster. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 398–408.

Agostina Calabrese, Björn Ross, and Mirella Lapata. 2022a. Explainable abuse detection as intent classification and slot filling. *Transactions of the Association for Computational Linguistics*, 10:1440–1454.

Agostina Calabrese, Björn Ross, and Mirella Lapata. 2022b. Explainable abuse detection as intent classification and slot filling. *Transactions of the Association for Computational Linguistics*, 10:1440–1454.

Kevin Clark, Minh-Thang Luong, Quoc V Le, and Christopher D Manning. 2020. Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555*.

Line H Clemmensen and Rune D Kjærsgaard. 2022. Data representativity for machine learning and ai systems. *arXiv preprint arXiv:2203.04706*.

Stefano Cresci, Amaury Trujillo, and Tiziano Fagni. 2022. Personalized interventions for online moderation. In *Proceedings of the 33rd ACM Conference on Hypertext and Social Media*, pages 248–251.

Aida Mostafazadeh Davani, M. C. D'iaz, Dylan K. Baker, and Vinodkumar Prabhakaran. 2023. Disentangling perceptions of offensiveness: Cultural and moral correlates. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*.

OpenAI et al. 2024. Gpt-4 technical report.

Simona Frenda, Gavin Abercrombie, Valerio Basile, Alessandro Pedrani, Raffaella Panizzon, Alessandra Teresa Cignarella, Cristina Marco, and Davide Bernardi. 2024. Perspectivist approaches to natural language processing: a survey. *Language Resources and Evaluation*, pages 1–28.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1161–1166.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Shagun Jhaver, Alice Qian Zhang, Quan Ze Chen, Nikhila Natarajan, Ruotong Wang, and Amy X Zhang. 2023. Personalizing content moderation on social media: User perspectives on moderation choices, interface design, and labor. *Proceedings of the ACM on Human-Computer Interaction*, 7(CSCW2):1–33.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Chris J Kennedy, Geoff Bacon, Alexander Sahn, and Claudia von Vacano. 2020. Constructing interval variables via faceted rasch measurement and multitask deep learning: a hate speech application. *arXiv preprint arXiv:2009.10277*.

Jiyun Kim, Byounghan Lee, and Kyung-Ah Sohn. 2022. Why is it hate speech? masked rationale prediction for explainable hate speech detection. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6644–6655, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Hannah Kirk, Wenjie Yin, Bertie Vidgen, and Paul Röttger. 2023. SemEval-2023 task 10: Explainable detection of online sexism. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2193–2210, Toronto, Canada. Association for Computational Linguistics.

Jan Kocoń, Alicja Figas, Marcin Gruza, Daria Puchalska, Tomasz Kajdanowicz, and Przemysław

Kazienko. 2021. Offensive, aggressive, and hate speech analysis: From data-centric to human-centered approach. *Information Processing & Management*, 58(5):102643.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing toxic content classification for a diversity of perspectives. In *Seventeenth Symposium on Usable Privacy and Security (SOUPS 2021)*, pages 299–318.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards faithful model explanation in nlp: A survey. *Computational Linguistics*, pages 1–70.

Binny Mathew, Punyajoy Saha, Seid Muhie Yimam, Chris Biemann, Pawan Goyal, and Animesh Mukherjee. 2021. Hatexplain: A benchmark dataset for explainable hate speech detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 14867–14875.

Maria D Molina and S Shyam Sundar. 2022. When ai moderates online content: effects of human collaboration and interactive transparency on user trust. *Journal of Computer-Mediated Communication*, 27(4):zmac010.

Hiroki Nakayama. 2018. seqeval: A python framework for sequence labeling evaluation. Software available from https://github.com/chakki-works/seqeval.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback.

Pia Pachinger. 2024. Austrotox: A dataset for target-based austrian german offensive language detection. *Comited to ACL*.

Pia Pachinger, Janis Goldzycher, Anna Planitzer, Wojciech Kusa, Allan Hanbury, and Julia Neidhardt. 2024. Austrotox: A dataset for target-based austrian german offensive language detection. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11990–12001.

Pia Pachinger, Allan Hanbury, Julia Neidhardt, and Anna Planitzer. 2023. Toward disambiguating the definitions of abusive, offensive, toxic, and uncivil comments. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 107–113, Dubrovnik, Croatia. Association for Computational Linguistics.

John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, and Ion Androutsopoulos. 2021. SemEval-2021 task 5: Toxic spans detection. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 59–69, Online. Association for Computational Linguistics.

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.

Julian Risch, Robin Ruff, and Ralf Krestel. 2020. Offensive language detection explained. In *Proceedings of the second workshop on trolling, aggression and cyberbullying*, pages 137–143.

Julian Risch, Anke Stoll, Lena Wilms, and Michael Wiegand. 2021. Overview of the germeval 2021 shared task on the identification of toxic, engaging, and fact-claiming comments. In *Proceedings of the GermEval 2021 Shared Task on the Identification of Toxic, Engaging, and Fact-Claiming Comments*, pages 1–12.

Paul Röttger, 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.

Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. Why don‘t you do it right? analysing annotators’ disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, Dubrovnik, Croatia. 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*, pages 5477–5490, Online. 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.

Maximilian Wich, Hala Al Kuwatly, and Georg Groh. 2020. Investigating annotator bias with a graph-based approach. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 191–199, Online. Association for Computational Linguistics.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399.

Marcos Zampieri, Skye Morgan, Kai North, Tharindu Ranasinghe, Austin Simmmons, Paridhi Khandelwal, Sara Rosenthal, and Preslav Nakov. 2023. Target-based offensive language identification. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 762–770, Toronto, Canada. Association for Computational Linguistics.

Qiang Zhang, Jason Naradowsky, and Yusuke Miyao. 2022. Rethinking offensive text detection as a multi-hop reasoning problem. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3888–3905, Dublin, Ireland. Association for Computational Linguistics.

Xuhui Zhou, Hao Zhu, Akhila Yerukola, Thomas Davidson, Jena D Hwang, Swabha Swayamdipta, and Maarten Sap. 2023. Cobra frames: Contextual reasoning about effects and harms of offensive statements. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6294–6315.

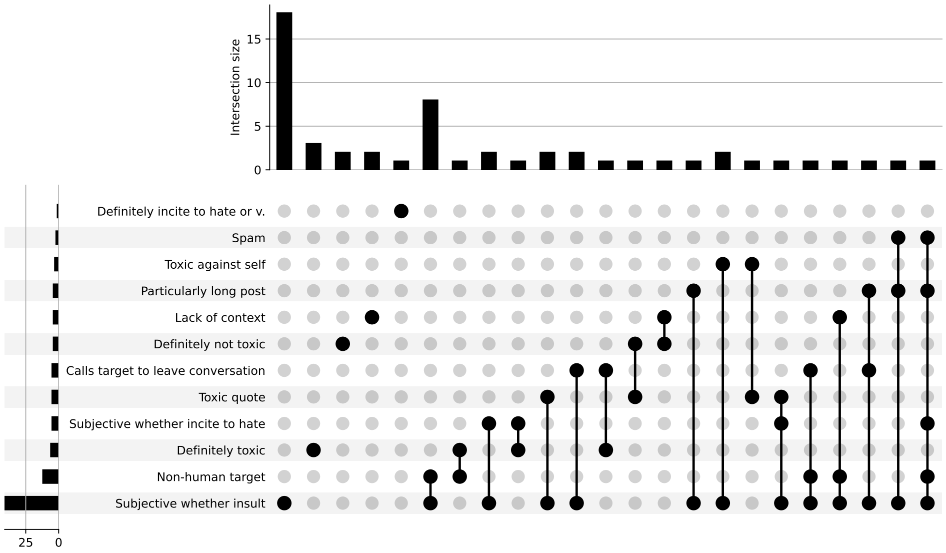

Figure 9: Factors related to disagreement between annotators in offensiveness labels in 60 posts with 3 annotators saying that the post is toxic and 2 annotators saying that it is not toxic.

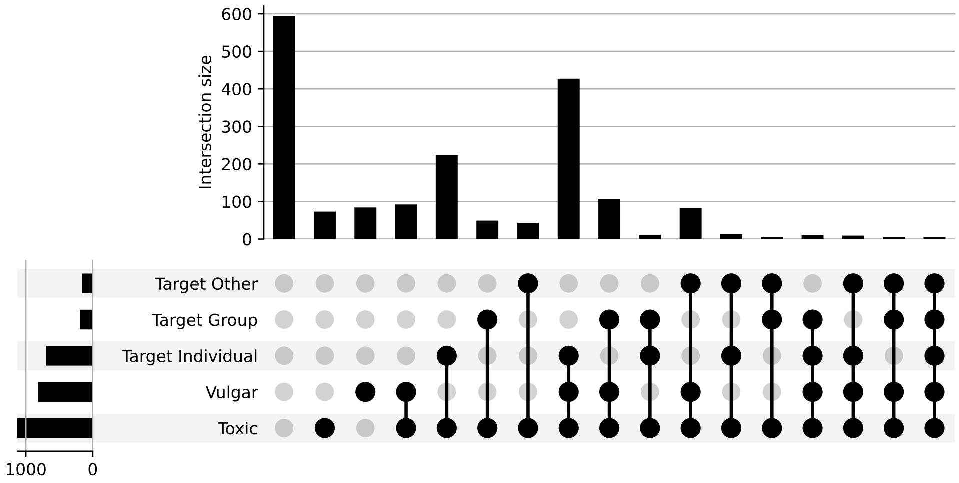

Figure 10: The fine grained class labels of the aggregated data and their co-appearance.

~~~
You receive a user comment. Your task is to answer
the following questions about the comment:

1. Is the comment toxic? (1 = toxic, 0 = non-toxic)
Definition: An offensive or toxic comment includes derogatory remarks towards
individuals, groups of people, or other entities. It may also incite hatred or
violence against individual persons or groups of people.

2. Who or what is the target of toxicity?
Mark at least one of the following targets of toxicity if the comment is toxic:
”Target_Group”, ”Target_Individual”, or ”Target_Other”.

3. Mark vulgarities with ”Vulgarity”. Vulgarities can occur in toxic and
non-toxic comments.

Respond in JSON format with the following fields:
```json
{
    ”Label”: <0 or 1>,
    ”Tags”: [
        {
            ”Tag”: <”Target_Group”, ”Target_Individual”, ”Target_Other”,
                    or ”Vulgarity”>,
            ”Token”: <Span of the target or the vulgarity>
        },
        ...
    ]
}
```
~~~

Figure 11: The multitask system prompt we use for the neural experiments.