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, Vienna University of Technology
[2]Institute for Computational Linguistics, University of Zurich
[3]Political Commmunication Research Group, University of Vienna

Abstract

More detailed annotation schemes and disaggregated labels lead to more nuanced toxicity classification. We publish new annotations of the *Jigsaw Toxic Comment Classification Challenge* under the CC BY-SA 4.0 license [1]. The annotations contain disaggregated toxicity labels and spans. We further publish an aggregated version of the annotations. Benchmarking the aggregated data shows that the labels provide learnable patterns.

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). It is important to create (semi-) automated content moderation systems that include a more nuanced understanding of online norm violations.

In the case of automated content moderation, explainability contributes to a greater understanding and trust of users (Molina and Sundar, 2022) and content moderators (Bunde, 2021). Annotated spans in text allow for evaluating whether certain elements in a text can be detected and foster model interpretability. And while annotated spans can lead to more differentiated classifications, they additionally improve model explainability (Lyu et al., 2024).

Further, 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.). Therefore, a one-size-fits-all approach to content moderation is unable to account for the diverse needs of different users (Plank, 2022; Sap et al., 2022; Jhaver et al., 2023) and perspectivist data and models on online toxicity need to be developed.

Main Contributions

1. We publish annotations of by in most cases 5 annotators per post for 1983 posts of the *Jigsaw Toxic Comment Classification Challenge* under the CC BY-SA 4.0 license
2. The annotations contain disaggregated toxicity labels fostering research on strongly perspectivist approaches to toxicity classification
3. Further, the annotations contain disaggregated annotated spans. To the best of our knowledge, this is the first dataset annotated for offensiveness classification containing disaggregated annotated spans
4. We additionally publish aggregated annotations, benchmarking the aggregated data yields a Binary F1 score of 0.9 in offensiveness classification and 0.7 in vulgar tokens detection in the data, showing that the annotations provide strong signals

2 Related Work and Background

There exists a wide corpus of work studying why individuals rate the toxicity of utterances differently. These differences are called annotator bias and can have various reasons. Researchers report annotator bias on 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).

Furthermore, in recent years, researchers have started to publish the disaggregated toxicity annotations. For example, Kumar et al. 2021 publish a

[1]`https://anonymous.4open.science/r/perspectivist_toxicity_data-02B7/README.md`

[2]As there are no generally accepted distinctions for *offensiveness* and *toxicity* (Fortuna et al., 2020), we use these terms interchangeably.

Annotation	Type	Authors
Call for violence, Explicit / Implicit, Target group	Class	Kennedy et al. 2018
Targeted / Not targeted	Class	Zampieri et al. 2019
Target groups	Text	Sap et al. 2020, Zhou et al. 2023
	Span	Zampieri et al. 2023
Annotators' feelings, Discriminated attribute	Class	Ousidhoum et al. 2019
Implied Statement, Stereotype, Power dynamic	Text	Sap et al. 2020
Inferiority Language, White Grievance	Class	ElSherief et al. 2021
Offensive argument	Span	Mathew et al.; Pavlopoulos et al.
	Text	Demus et al. 2022
Violation of policy	Span	Calabrese et al. 2022
Criminal relevance	Class	Demus et al. 2022
Reader attribute, Chain of reasoning Corresponding explicit / non-offensive statement	Text	Zhang et al. 2022
Situation, Speaker, Listener, Intent, Impact, Reactions	Text	Zhou et al. 2023
Individual, Other, or Group Target; Vulgar	Span	Pachinger et al. 2024

Table 1: Non exhaustive list of annotations in publicly available datasets for improving offensive text detection performance and making offensive text detection more nuanced and explainable. Span is the same as rationale. Text denotes free-text.

dataset that contains 107,620 texts and annotations by 17,280 annotators. It is available on request. Another example is the dataset publised by Kennedy et al. 2020, which contains 50,000 texts and annotations by 11,000 mechanical turkers. It is openly accessible. See Frenda et al. 2024 for more perspectivist datasets on online toxicity.

Additionally, in recent years, there has been a surge in datasets related to offensive text detection with span and free-text annotations which can be used to evaluate the faithfulness of language models (Lyu et al., 2024). Existing data with annotated spans include spans of the targets of offensive statements (Zampieri et al., 2023; Pachinger et al., 2024), the spans contributing to the offensiveness label (Mathew et al., 2021; Pavlopoulos et al., 2021), and the spans comprising a violation of a moderation policy (Calabrese et al., 2022), 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). A more detailed list of fine-grained information annotated in offensive utterances can be observed in Table 1.

3 Dataset Creation

3.1 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. 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 cases of toxicity, we only source comments with labels *toxic* or *insult* and exclude more severe labels.

Observe some examples of utterances labeled as *toxic* in the dataset for the Toxic Comment Classification Challenge:

> "*Calling me a joke is not a personal attack? Hypocrite!*"

> "*Hey, I said it was A seat, not THE seat, you dumb motherf#@ker!! Learn how to read English. Soon I shall be an administrator and have you purged from this noble experiment.*"

> "*Good morning. Why do you sit at a computer waiting to delete other peoples' additions? The topic I uploaded was of a fictional organisation whose*

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

comedy nature DOES appeal to those people blessed with a pulse. Before any of us here could editlink or expand our article...you deleted it.seems to me you are a lonely, Spurs-supporting, bird loving (and to what extent?), non-humourous, southern fairy cake...who probably smells. And whose big dumb face is as dumb as a butt. We sound childish yes, however, we cannot beleive that there are people as pathetic as you lording yourself around the internet. It was people like you, sir, who were responsible for carrying every major dictator and despot of the last century to power on a wave of apathy, ill-humour and rubber-desk-johnny procedure-bound mentral dross. You are, by any reasonable and objective measure, worse than Stalin. You are the reason this world is full of crashing bores!do not delete our article again...Or we will 180 the internet."

Observe some examples of utterances labeled as *insult* in the dataset for the Toxic Comment Classification Challenge:

"*This IP is from a school and as so is used by everyone, even the idiots who want to vandalize wikipedia without thought to everyone else; so all apologies from me as one of those who use the IP*"

"*LOL, congrats on your google skills in digging comments I made on a soccer website. Grow up and crawl back in your black van.*"

We observe that the comments are hard to understand and subsequently to classify. Further, they considerably vary in length.

3.2 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 1. We classify each comment as insult, incite to hate or violence or not offensive. Subsequently we merge classes *Insult* and *Incite to hate or violence* into an *Offensiveness* class. 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.

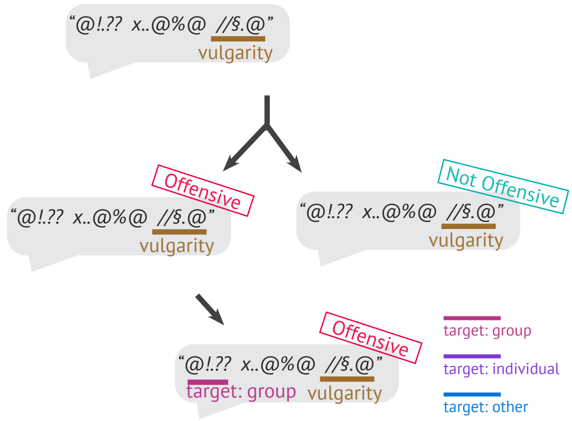

Figure 1: The annotation strategy for this dataset

We use the following definitions for classes and spans:

Offensive An offensive comment includes disparaging statements towards persons, groups of persons or other entities or incites to hate or violence against a person or a group of people.

Not Offensive A non-offensive comment does not include disparaging statements or incites to hate or violence.

Vulgarity 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.

3.3 Annotation Campaign

We conduct the annotation with master's students. 30% of the annotators are registered as female at our institution, that does not necessarily reflect the gender they most identify with. The majority of the annotators are between 19 and 26 years old. All annotators have a level of English of at least B2. The majority of annotators origins from Eastern Europe.

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 consequences. The annotators are informed about the publication of the data and they receive a comprehensive compensation through course credits for their efforts. 1750 posts are annotated by five annotators, 189 are annotated by four annotators, and the remaining 44 posts are annotated by three annotators or less. Figure 2 shows the different votes for posts to be toxic or not.

Observe an example in the resulting disaggregated dataset:


```
{
    "Index":"7af3262004ea8400",
  "Comment":"lol nuffin much bighead , ma whole
    fam jus went to dinner w\/o me =[ but
   ummm yeah lol yooh join gaiaonline yet ?",
    "Annotators_not_toxic":[
        33,
        13,
        30
    ],
    "Annotators_toxic":[
        36,
        47
    ],
    "Annotators_insult":[
        36
    ],
    "Annotators_hate":[
        47
    ],
    "Tags":[
        {
            "Tag":"Vulgarity",
            "Token":"gaiaonline",
            "Annotators":[
                47
            ]
        },
        {
            "Tag":"Vulgarity",
            "Token":"bighead",
            "Annotators":[
                13,
                47
            ]
        }
    ],
    "Jigsaw_toxic":1,
        "Jigsaw_severe_toxic":0,
        "Jigsaw_obscene":0,
        "Jigsaw_threat":0,
        "Jigsaw_insult":0,
        "Jigsaw_identity_hate":0
    },
```


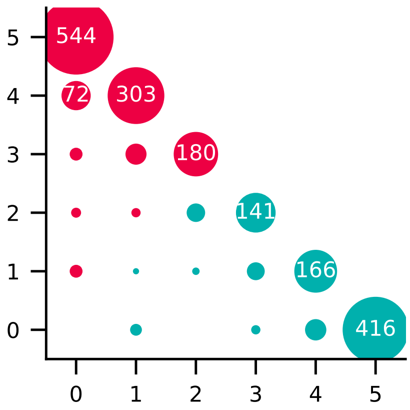

Figure 2: Votes for a post to be toxic (y-axis) versus votes for a post not to be toxic (x-axis). The colors denote the label of the aggregated dataset.

	Target	Not vulgar	Vulgar
All posts		913	886
Not off.		567	79
Offensive	Group	46	120
	Individual	191	472
	Other	35	91
	Multiple	18	29
	None	56	95

Table 2: The size of the aggregated dataset. *Off* stands for offensive.

4 Aggregating the Dataset

In order to ensure compatibility with the AustroTox dataset, we use the same aggregation strategy for classes and spans. I.e., we aggregate the posts by majority vote but discard the 141 posts labeled as toxic by two annotators and as non-toxic by three annotators in order to enlarge the decision boundary (observe the 141 posts in Figure 2). We keep spans comprising vulgarities if they are annotated by at least two annotators and at least n - 2 annotators, where n is the total number of annotators for that post. Spans in the comments comprising the different target types are annotated by major-

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

Table 3: 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.

ity voting of those who labelled the comment as offensive. Table 2 contains the size of the resulting dataset. We report a Krippendorff's Alpha of 0.62 on the binary offensiveness classification.

Observe and example in the resulting disaggregated dataset:

```
{
    "Index":"5755717a4353c136",
  "Comment":"\"\n\n Personal Attack \nYou made
    a personal attack with your comment. Why
    don't you learn to eat a decroted
    piece of crap? Oh wait, you've mastered
    that already!\"",
    "Label":1,
    "Annotators_not_toxic":[
        50
    ],
    "Annotators_toxic":[
        48,
        49,
        11,
        5
    ],
    "manually_cleaned":0,
    "vulgar":1,
    "target_group":0,
    "target_individual":1,
    "target_other":0,
    "Label_fine":"Target_Individual_Vulgar",
    "Tags":[
        {
            "Tag":"Target_Individual",
            "Token":"You"
        },
        {
            "Tag":"Vulgarity",
            "Token":"crap",
            "Votes":4
        }
    ]
}
```

5 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. We fine-tune and evaluate smaller language models and we evaluate the few-shot performance of large language models in a 10-fold-cross validation setting.

We fine-tune smaller language 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[4] (Clark et al., 2020) and Roberta Large[5] (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).

Further we prompt the following large language models for our experiments: GPT 3.5[6] (*gpt-3.5-turbo-1106*) (Ouyang et al., 2022), GPT 4 [7] (*gpt-4-1106-preview*) (et al., 2024), and Mistral [8] (Jiang

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

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

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

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

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

et al., 2023).

We use the same prompt as Pachinger et al. 2024. Observe the prompt in Appendix 6. The prompts 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.

Observe the evaluation outcomes in Table 3. The models perform better on our dataset than on the German data, 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 perform better in all tasks on our data. Nonetheless, we did not fine-tune the larger language models. Inline with Pachinger et al. 2024, we find that especially the vulgar token detection task profits from fine-tuning. None of the models detects tokens comprising targets sufficiently well, further analysis of the annotations is needed in order to understand where the complexity of this task lies.

6 Conclusion and Potential Further Use of the Data

We publish new annotations for the *Jigsaw Toxic Comment Classification Challenge* containing disaggregated labels and spans under the CC-BY-SA 4.0 license[9].

The new annotations allow for fine-grained classification based on the annotated classes and spans (e.g. *offensive* and *targeted group*, *non-offensive* and *vulgarity*). This leads to more inherent explainability and facilitates error analysis. Further, they allow for research on perspectivist modelling approaches such as for example replicating each statements with disagreeing annotations in a number proportional to the disagreements (Cabitza et al., 2023), different language models for two groups of annotators bi-partitioned by polarization (Akhtar et al., 2019, 2020), soft-labels (Fornaciari et al., 2021), or adding categorical annotator information to the input (Vallecillo-Rodríguez et al., 2023).

[9] https://anonymous.4open.science/r/perspectivist_toxicity_data-02B7/README.md

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.

Sohail Akhtar, Valerio Basile, and Viviana Patti. 2019. A new measure of polarization in the annotation of hate speech. In *AI* IA 2019–Advances in Artificial Intelligence: XVIIIth International Conference of the Italian Association for Artificial Intelligence, Rende, Italy, November 19–22, 2019, Proceedings 18*, pages 588–603. Springer.

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

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, Björn Ross, and Mirella Lapata. 2022. 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*.

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*.

Christoph Demus, Jonas Pitz, Mina Schütz, Nadine Probol, Melanie Siegel, and Dirk Labudde. 2022.

Detox: A comprehensive dataset for German offensive language and conversation analysis. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 143–153, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

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.

Paula Fortuna, Juan Soler, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the 12th language resources and evaluation conference*, pages 6786–6794.

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.

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*.

Brendan Kennedy, Mohammad Atari, Aida Mostafazadeh Davani, Leigh Yeh, Ali Omrani, Yehsong Kim, Kris Coombs, Shreya Havaldar, Gwenyth Portillo-Wightman, Elaine Gonzalez, et al. 2018. The gab hate corpus: A collection of 27k posts annotated for hate speech. *PsyArXiv. July*, 18.

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.

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.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. 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 4675–4684, Hong Kong, China. Association for Computational Linguistics.

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, 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.

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.

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.

María Estrella Vallecillo-Rodríguez, FMP del Arco, Luis Alfonso Ureña-López, María Teresa Martín-Valdivia, and Arturo Montejo-Ráez. 2023. Integrating annotator information in transformer fine-tuning for sexism detection. *Working Notes of CLEF*.

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.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of NAACL-HLT*, pages 1415–1420.

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. *arXiv preprint arXiv:2306.01985*.

Multitask System Prompt

```
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 3: The multitask system prompt