

Detecting Online Hate Speech: A Comparison of Annotation Approaches

Natalia Umansky[†], Mael Kubli[‡], Karsten Donnay[§],
Fabrizio Gilardi[¶], Dominik Hangartner^{||}, Ana Kotarcic^{*,*},
Selina Kurer^{††} and Philip Grech^{‡‡}

August 16, 2023

Abstract

Efforts to study and curb online hate speech depend on the capacity to detect it at scale. A critical step in state-of-the-art classification pipelines remains the production of a large number of high-quality annotations to train classifiers. While machine-learning methods assume the availability of adequate labeled data, this step is usually poorly documented. Consequently, there is little guidance available to construct a training set effectively and efficiently. This paper conducts a comprehensive comparison of various annotation approaches and their performance using a unique corpus of online comments developed for detecting and countering online hate speech in Switzerland involving diverse groups of coders: research assistants, activists, crowd workers, citizen scientists, and generative AI models like ChatGPT. Due to the subjectivity of the task, intercoder reliability is typically low, as different coders may assign inconsistent labels to the same comments. To overcome this challenge, we propose a threefold comparison framework. Instead of relying on a single annotation per comment, we use three annotations and apply the majority vote method, yielding more robust labels that substantially improve accuracy and reliability, regardless of the type of annotators used. The results illustrate the challenge of constructing a high-quality training set and provide guidance for various kinds of difficult classification tasks, including (but not limited to) online hate speech.

Keywords— Quantitative Text Analysis, Computational Methods, Hate Speech

[†]Postdoctoral Research Fellow, University of Zurich

[‡]PhD Researcher, University of Zurich

[§]Assistant Professor of Political Behavior and Digital Media, University of Zurich

[¶]Professor of Policy Analysis, University of Zurich

^{||}Professor of Public Policy, ETH Zurich

^{*,*}Postdoctoral Research Fellow, University of Zurich

^{††}PhD Researcher, ETH Zurich

^{‡‡}Executive Director, Immigration Policy Lab, ETH Zurich

1 Introduction

Detecting and addressing online hate speech is a pressing concern, necessitating the ability to identify such content on a large scale (Parker & Ruths, 2023; Kotarcic, Hangartner, Gilardi, Kurer, & Donnay, 2022). Hate speech¹, according to the United Nations, is defined as ‘any kind of communication in speech, writing or behavior, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, color, descent, gender or other identity factor’ (*United Nations Strategy and Plan of Action on Hate Speech*, 2019). This definition serves as a guideline for the creation of classification pipelines aimed at detecting and stopping such conduct. A crucial aspect of these state-of-the-art pipelines, both for hate speech detection and other tasks, is the generation of a substantial number of high-quality annotations to train classifiers (Barberá, Boydston, Linn, McMahon, & Nagler, 2021; Benoit, Conway, Lauderdale, Laver, & Mikhaylov, 2016; Boukes, Van de Velde, Araujo, & Vliegthart, 2020; Van Atteveldt, Van der Velden, & Boukes, 2021). However, the process of producing these annotations needs to be more adequately documented, resulting in a need for more guidance on how to construct an effective and efficient training set (Barberá et al., 2021).

In this paper, we address this gap by comparing different annotation approaches and assessing their performance using a unique corpus of online comments developed within a project focused on detecting and countering online hate and toxic speech in Switzerland. We begin by acknowledging that intercoder reliability is typically poor, meaning that different coders may assign inconsistent labels to the same comments (Grimmer & Stewart, 2013; Scharkow & Bachl, 2021). To address this challenge, we propose the adoption of a threefold comparison framework. We involve various groups of coders, including domain experts, research assistants, activists, crowd workers, citizen scientists, and even generative AI models like ChatGPT. Instead of relying on a single annotation for each comment, we use three annotations and employ the majority vote method. This means

¹Appendix A for detailed conceptualization

that if one coder performs reliably and the other two perform at random, on average, their annotations will be anti-correlated, and the majority vote will still yield a more robust label for the comment.

Despite low intercoder agreement, we observe that in the aggregate, the majority of labels are either identical or nearly identical. Our analysis demonstrates that even with a low intercoder agreement, using multiple annotators and applying the majority vote can substantially improve accuracy and reliability, regardless of the type of annotators we employ. Furthermore, our results emphasize that it is crucial to validate the accuracy of annotations and not solely rely on intercoder agreement. The majority vote approach, or majorization as we refer to it, makes sense in this context because it helps mitigate the lack of a single trusted annotator, especially for subjective tasks like hate speech detection. This diversity enriches the training set and enhances reliability, mitigating potential biases, subjective judgments, and errors (Van Atteveldt et al., 2021). By using majorization as a form of regularization, we can make more confident inferences from the training set.

The findings of this study shed light on the complexities involved in creating a high quality training set, particularly for challenging classification tasks such as identifying online hate speech (Grimmer & Stewart, 2013; Scharkow & Bachl, 2021; Benoit et al., 2016; Boumans & Trilling, 2016). Moreover, they provide valuable guidance for future annotation efforts and recommendations for enhancing the efficiency and effectiveness of training set construction. Even with the advent of powerful Large Language Model models (LLMs), manual annotations will remain essential, and the majority vote method can further improve their accuracy (Gilardi, Alizadeh, & Kubli, 2023a). It is worth noting that different coders may exhibit varying degrees of consistency (intercoder agreement) in their annotations. This diversity underscores the importance of using multiple coders and establishing a benchmark for comparison. Looking ahead, future research should explore the optimal number of annotators required for tasks involving multiple classes. As we delve into the pursuit of quality annotations, we recognize that we need more knowledge

about the manual annotations used to train and test classifiers. Even with highly trained research assistants, intercoder reliability can remain a challenge. Therefore, the majority vote approach emerges as a promising alternative.

Our study has broader implications for the field of natural language processing and computational social science (Abadi et al., 2016; Barberá et al., 2021). It highlights the importance of moving beyond mere labels and focusing on quality labels in training sets. By adopting multiple coders and the majority vote method, we enhance the accuracy and reliability of annotations for challenging classification tasks like online hate speech detection. This approach provides valuable insights into constructing effective training sets and offers robust solutions for addressing other complex classification tasks in online communication and social media analysis, including (but not limited to) online hate speech.

While our study makes significant strides in addressing the challenge of intercoder reliability by proposing adopting a threefold comparison framework, certain limitations should be acknowledged. One such limitation is the need for more exploration into determining the optimal number of coders required for tasks involving multiple classes, such as hate speech detection. The focus of this research is on leveraging multiple annotators and employing the majority vote method to enhance the robustness of annotations. While this approach has demonstrated promising results in mitigating the subjectivity of the task and increasing the reliability of the training set, it needs to directly address the question of how many coders are sufficient for achieving a desired level of annotation quality.

2 Data and Methods

2.1 Annotators and Annotator Characteristics

In this study, a total of 207 annotators participated across six groups. Among them, three annotators were experts, three research assistants (RA Group), 17 were crowd workers

from Appen (Appen Group), 53 were crowd workers from Prolific (Prolific Group), 24 were citizens participating in Citizen Science (Citizen Group), and finally, 107 were volunteer community members of a partner NGO (NGO Group). Additionally, the entire set of comments was also coded by ChatGPT three times using the standard temperature of the model. For more details on the annotation procedure used with chatGPT, see [B](#). An increasing body of evidence suggests that ChatGPT and other large language models may perform well for a variety of annotation tasks ([Gilardi, Alizadeh, & Kubli, 2023b](#)).

2.2 Annotator Scheme and Instructions

To ensure comparability between the three groups of annotators, all coders received the exact same instructions. Annotators were first instructed to determine whether a comment or tweet contained hate or toxic speech. They were asked to indicate the targeted groups if hate speech was present. In cases where no specific target group could be identified, annotators were instructed to label the comment as toxic speech. It is important to note that while a comment or tweet may target multiple groups, toxic speech is exclusive, meaning that once a comment or tweet is deemed toxic, it cannot be assigned to any other category. In this paper, we do not differentiate between toxic speech and hate speech but aggregate both as hate speech. This is not as precise as differentiating between them, but in a broad sense, toxic speech is also defined as hate speech. See [Appendix A](#) for a detailed definition of hate speech. Annotators were provided with the title and the entire article text for additional context, as hate speech is a profoundly contextual phenomenon. However, the comment history per article could not be provided to the annotators.

2.3 Data and Data Sampling

The dataset initially consisted of 500 unique comments obtained from a Swiss national media outlet, including published, moderated, and deleted comments which served as the gold standard for the construction of the training set used to build a hate speech classifier

for Swiss content (Kotarcic et al., 2022). The classifier was built for the stop hate speech project led by the Public Discourse Foundation and implemented in collaboration with the Digital Democracy Lab (UZH), the Public Policy Group, and the Immigration Policy Lab (ETH). The project combined natural language processing and machine learning with civil society engagement to counter online hate speech to improve the quality of public discourse and minimize offline consequences of hostile online behavior. Due to platform settings, six comments could not be consistently annotated across all groups, resulting in a total of 494 unique comments for analysis. While the RA and experts set had only three annotations per comment, some comments on Appen and the NGO platforms were annotated more than three times. Three annotations per comment were randomly sampled in such cases while preserving the per-comment distribution of positive and negative labels. If annotators did not follow the annotation instructions and marked a comment as both hate speech and toxic, the hate speech label was retained if a target group was identified. Except for the experts, we determined the final annotation per comment by majority rule. The experts joined together after coding each comment alone to address any discrepancies in their assessments, engaging in thorough discussions to reach a consensus and finalize their conclusions instead of a majority vote. The experts classified 238 of the comments as hate speech and 256 as non-hate speech.

2.4 Evaluation Metrics

Quantitative evaluations are conducted to assess and compare the qualities of individual annotation sets. We employ the joint probability of agreement to measure the intercoder agreement to determine the overall coding quality of different groups (Bayerl & Paul, 2011). This measure enables the calculation of inter-rater reliability among multiple annotators, both overall and for the actual labels. In the context of this study, the measure aligns well with the emphasis on consensus and decision-making, offering an interpretable and direct understanding of the degree of agreement among coders. We use this straightforward approach to measure the intercoder agreement using the following

formula (Cohen, 1960; Fleiss, 1971):

$$Agreement = \frac{\sum_i^n \delta_i}{n}$$

Where δ is an indicator function that equals 1 when all annotators have the same value for one comment i and 0 otherwise, the summation symbol \sum indicates that we are summing over all n comments, which we then divide through the number of instances n .

Moreover, this study conducted a comparative analysis amongst the distinct annotator groups, with their accuracy measured against a benchmark group comprising experts in the field. Notably, the benchmark group consisted of three experts, all of whom are co-authors of this paper. The experts are the gold standard due to their extensive experience and expertise. In the annotation process, the experts independently coded each row to ensure individual evaluations. Subsequently, the experts collaborated to address any discrepancies in their assessments, engaging in thorough discussions to reach a consensus and finalize their conclusions. The use of inter-coder agreement aligns with this collaborative approach, emphasizing the frequency of agreement over individual variations and offering a suitable measure for comparative analysis. The harmonized judgments attained through this collaborative approach were then employed as the benchmark data for the study. The majority voting system, embedded within this approach, facilitates the resolution of discrepancies and is congruent with the direct interpretation provided by the inter-coder agreement measure. This comparison allowed us to gauge the inter-coder agreement and quality of the other annotator groups in instances where the experts identified a comment as hate speech and vice versa.

Finally, we employed a correlation matrix analysis to assess the degree of correlation among the decisions made by the different annotator groups. The correlation matrix analysis was employed to assess the degree of correlation between the decisions made by the different annotator groups. Specifically, it aimed to quantify the level of agreement or disagreement between each pair of annotator groups in their assignment of labels to the comments.

For each comment in the dataset, the correlation matrix was constructed by comparing the labels assigned by one annotator group to the labels given by another. This process was repeated for all possible pairs of annotator groups, generating a matrix where each cell represents the correlation between two groups.

3 Results

To evaluate the performance and reliability of different annotation groups in detecting and classifying online hate speech, we first assessed the intercoder agreement within each group. Intercoder agreement measures the level of consistency or agreement between multiple coders when assigning labels to the same data set (Barberá et al., 2021; Van Atteveldt et al., 2021). As expected for a subjective task like hate speech detection, we observed low intercoder agreement within all groups except for the experts (Table 1). This finding aligns with previous research that highlights the inherent difficulty in achieving high consensus among annotators for such tasks (Akhtar, Basile, & Patti, 2020; Kralj Novak et al., 2022). Nevertheless, this variation in annotations across different coders underscores the need for a robust approach to construct reliable training sets.

Platform	Agreement	Krippendorff's Alpha
Experts	0.822	0.761
Research Assistants	0.632	0.506
NGO	0.66	0.458
Prolific	0.549	0.391
Appen	0.31	0.08
Citizen Science	0.508	0.335
Chat GPT	0.783	0.682
Combined ICR	0.289	0.399

Table 1: Intercoder agreement

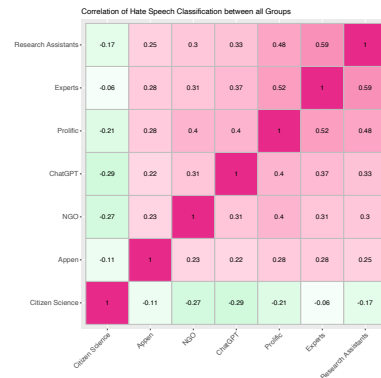


Figure 1: Two-by-two between group agreement

To mitigate the impact of low intercoder agreement, we employed a majority vote method, referred to as majorization. Instead of relying on a single coder’s annotation for

each comment, we used three annotations and applied the majority vote to determine the final label. By doing so, we enhanced the accuracy and reliability of the annotations. The intercoder agreement between the groups also revealed an interesting pattern: groups that performed better in intercoder agreement with their own members generally exhibited higher correlations with other groups (Figure 1). Conversely, groups that struggled with intercoder agreement within their ranks also showed lower correlations with other groups. This indicates that groups that were more consistent internally tended to align better with other groups, leading to increased reliability in the majority vote approach.

Even for the research assistants who have the highest performance apart from the expert group, we found that in most cases, the final decision came down to a majority vote. When looking at each annotator’s (RA) performance separately, we find that the proportions for their individual agreement with the final majority vote were 0.681, 0.582, and 0.736 for the annotator, respectively. A chi-square test on these proportions indicated a significant difference (χ -squared = 9.956, $df = 2$, p -value = 0.006888), implying that the annotators’ performances were not entirely equivalent in achieving majority agreement. This means that while one coder performs reliably, the other two perform almost at random. Therefore, on average, their annotations are anti-correlated, and the majority vote will still yield a more robust label for the comment. In other words, the majorization approach effectively helped mitigate the impact of low intercoder agreement, leading to more consistent and robust annotations.

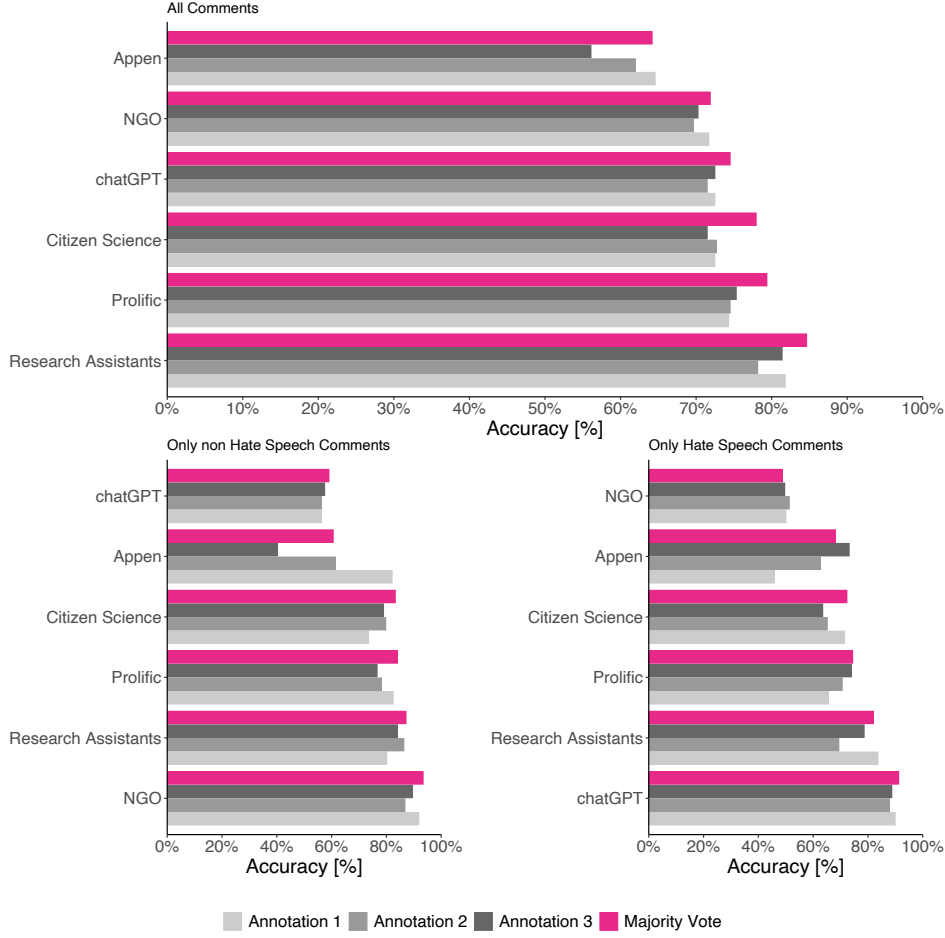


Figure 2: Accuracy of all five groups and ChatGPT compared with the gold standard produced by experts in the field.

Next, we evaluated the accuracy of annotations across different groups compared to the gold standard established by the expert group (Figure 2). Despite the RA group outperforming others with 84.6 %, some notable patterns emerged. Notably, Prolific (79.4 %), Citizen Science (77.9 %), and ChatGPT (74.5 %) demonstrated overall solid accuracy, closely followed by the NGO (71.9 %). On the other hand, Appen performed relatively poorly, with an overall accuracy of just 64.2 %.

The overall accuracy only describes part of the picture, and a more comprehensive examination of the accuracy was undertaken by considering both the majority votes and single annotation accuracies. As expected, the accuracy of single annotations was consistently lower than that of the majority vote for all groups. This trend was observed even for research assistants and ChatGPT, where each annotation is done by one individual,

as well as for other groups where multiple annotators were involved in each annotation round. The findings reveal that a majority vote approach positively impacts accuracy, bolstering the robustness of the annotation process.

Furthermore, when examining the accuracy of hate speech and non-hate speech instances separately, we observed that different groups exhibited distinct performances for each class. Citizen Science, Appen, and Prolific performed consistently in both categories. The NGO, while achieving the highest accuracy for non-hate speech cases, struggled to classify hate speech comments accurately with only 48.7 % accuracy. On the other hand, ChatGPT, despite excelling in identifying hate speech, showed the lowest accuracy in classifying non-hateful content with 59.0 %. While in most cases, researchers will settle for reporting only the overall accuracy of their annotations, disentangling the performance by class is of utmost importance. In this case, for example, the sample of online comments to be annotated is balanced (relatively similar number of hate speech and non-hate speech comments). However, the reported performance disparity for different classes leads to an unbalanced training set, significantly impacting classifier performance and evaluation metrics. These results are very promising and indicate that thanks to the majorization approach, similar levels of accuracy can be achieved by less trained and inexpensive groups, making the annotation process more accessible.

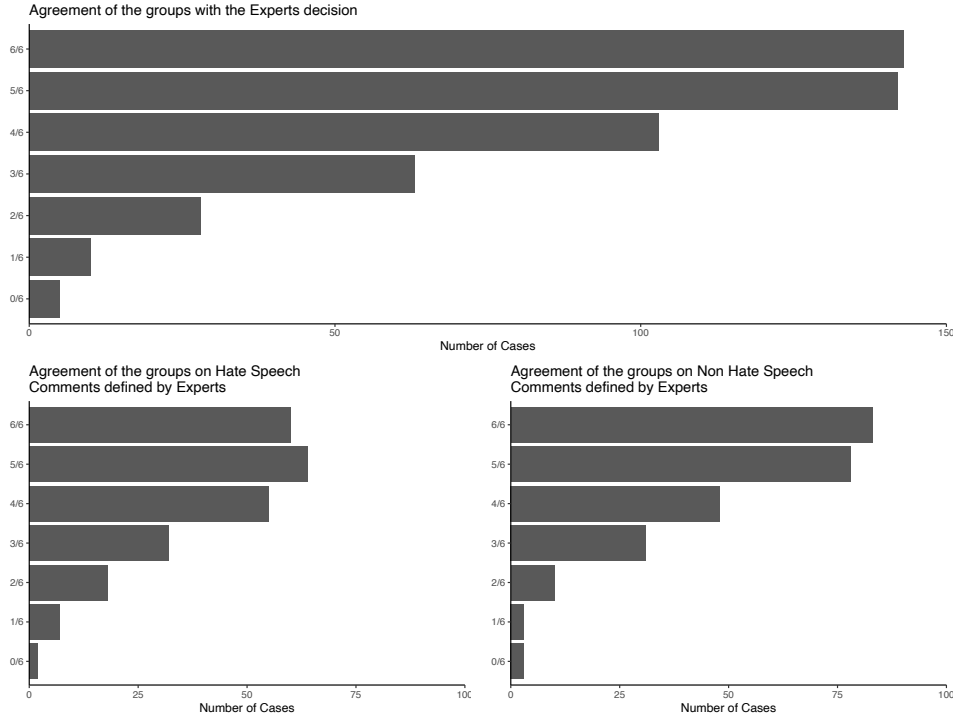


Figure 3: Agreement count of the five groups and ChatGPT compared to the gold standard of experts in the field

To validate the accuracy of annotations and assess their agreement with expert judgments, we compared the annotations of different groups with those of the experts. Overall, approximately 80% of the comments received the same label from all groups, indicating a relatively high level of agreement with expert judgments (Figure 3). This is a promising result obtained from our majorization approach. However, the remaining 20 % of cases show that Appen induced the biggest number of discrepancies compared to other groups. This suggests that, despite the majority vote approach, annotations from Appen were more prone to diverging from expert judgments, potentially due to the nature of the crowd-sourcing process or other factors.

In conclusion, our results indicate that employing the majority vote method with multiple annotators significantly improves the accuracy and reliability of annotations for hate speech detection. While the research assistants demonstrated the highest performance, other groups, particularly Citizen Science and Prolific, were not far behind despite being more cost-effective and receiving less training. While the intercoder agreement within groups was low, as expected for subjective tasks such as hate speech detection, the ma-

majority vote approach and inter-group agreement helped overcome this challenge and create more robust annotations. These findings provide valuable guidance for constructing high-quality training sets, particularly for complex classification tasks like identifying online hate speech.

4 Conclusion & Avenues for Future Research

Our study sheds light on the challenges and strategies involved in constructing a high-quality training set for detecting and classifying online hate speech. The subjective nature of hate speech classification led to a low intercoder agreement within groups, as indicated by the alpha coefficient. However, we found that the majority vote approach also referred to as majorization, proved to be an effective solution in dealing with this challenge.

Our research compared various groups of coders, including research assistants, activists, crowd workers, citizen scientists, and generative AI models like ChatGPT. Despite differences in expertise and training, we observed that the annotations produced by these diverse groups were not too far off from each other in the aggregate. The majority vote approach played a crucial role in achieving higher levels of agreement across all groups. When employing this method, we found that approximately 80 % of the comments received consistent labels from the various annotators, aligning with the experts' judgments. This demonstrates the robustness and reliability of the majority vote approach in harmonizing annotations and mitigating the subjectivity inherent in hate speech classification.

We also recognized that hate speech classification remained challenging even for the group of experts, and reaching an agreement for all 500 entries required considerable time and effort. This underscores the need to pay meticulous attention to producing a large number of high-quality annotations when training classifiers for state-of-the-art classification pipelines.

Moving forward, the findings of our study have significant implications for the development of hate speech detection systems and the creation of effective training sets. As

we navigate the complexities of identifying and addressing online hate speech, it becomes evident that leveraging multiple annotators and employing the majority vote approach enhances the accuracy and reliability of the annotations, irrespective of the annotators' background or expertise. However, future research should focus on the disparities that emerge when comparing the annotation accuracy of the two classes (hate speech and non-hate speech), even when employing the majorization approach, and explore the annotation process for tasks that are not binary (see Appendix C).

In conclusion, our research advocates for a comprehensive documentation and guidance framework in generating annotations for challenging classification tasks like online hate speech detection. By incorporating these insights into future annotation efforts, we can improve training set construction's overall effectiveness and efficiency. Ultimately, this will contribute to the development of more accurate classifiers that can enhance the reliability and effectiveness of online hate speech detection systems, contributing to the development of safer and more inclusive online environments.

References

- Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., & Zheng, X. (2016). Tensorflow: A system for large-scale machine learning. In *12th usenix symposium on operating systems design and implementation (osdi 16)* (pp. 265–283).
- Akhtar, S., Basile, V., & Patti, V. (2020). Modeling annotator perspective and polarized opinions to improve hate speech detection. In *Proceedings of the aaai conference on human computation and crowdsourcing* (Vol. 8, pp. 151–154).
- Barberá, P., Boydstun, A. E., Linn, S., McMahon, R., & Nagler, J. (2021). Automated text classification of news articles: A practical guide. *Political Analysis*, 29(1), 19–42.
- Bayerl, P. S., & Paul, K. I. (2011). What Determines Inter-Coder Agreement in Manual Annotations? A Meta-Analytic Investigation. *Computational Linguistics*, 37(4), 699–725.
- Benoit, K., Conway, D., Lauderdale, B. E., Laver, M., & Mikhaylov, S. (2016). Crowdsourced text analysis: Reproducible and agile production of political data. *American Political Science Review*, 110(2), 278–295.
- Boukes, M., Van de Velde, B., Araujo, T., & Vliegenthart, R. (2020). What’s the tone? easy doesn’t do it: Analyzing performance and agreement between off-the-shelf sentiment analysis tools. *Communication Methods & Measures*, 14(2), 83–104.
- Boumans, J. W., & Trilling, D. (2016). Taking stock of the toolkit: An overview of relevant automated content analysis approaches and techniques for digital journalism scholars. *Digital Journalism*, 4(1), 8–23.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1), 37–46.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5), 378.
- Gilardi, F., Alizadeh, M., & Kubli, M. (2023a). Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056*.

- Gilardi, F., Alizadeh, M., & Kubli, M. (2023b). *ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks*.
- Grimmer, J., & Stewart, B. M. (2013). Text as data: The promise and pitfalls of automatic content analysis methods for political texts. *Political Analysis*, 21(3), 267–297.
- Kotarcic, A., Hangartner, D., Gilardi, F., Kurer, S., & Donnay, K. (2022). Human-in-the-loop hate speech classification in a multilingual context. In *Proceedings of the 2022 conference on empirical methods in natural language processing*.
- Kralj Novak, P., Scantamburlo, T., Pelicon, A., Cinelli, M., Mozetič, I., & Zollo, F. (2022). Handling disagreement in hate speech modelling. In *Information processing and management of uncertainty in knowledge-based systems: 19th international conference, ipmu 2022, milan, italy, july 11–15, 2022, proceedings, part ii* (pp. 681–695).
- Parker, S., & Ruths, D. (2023). Is hate speech detection the solution the world wants? *Proceedings of the National Academy of Sciences*, 120(10), e2209384120.
- Scharkow, M., & Bachl, M. (2021). Computer-assisted content analysis of social media data. In P. Vorderer, P. H. Rössler, C. A. Klimmt, & S. Böcking (Eds.), *The oxford handbook of media psychology* (pp. 455–475). Oxford University Press.
- United nations strategy and plan of action on hate speech*. (2019). https://www.un.org/en/genocideprevention/documents/advising-and-mobilizing/Action_plan_on_hate_speech_EN.pdf. United Nations. (Accessed on 23.02.2022)
- Van Atteveldt, W., Van der Velden, M. A., & Boukes, M. (2021). The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures*, 15(2), 121–140.

A Hate Speech Definition

For this study, we chose to combine top-down and bottom-up approaches in defining hate speech, drawing from existing definitions as used by (Kotarcic et al., 2022). At the core of their definition lies the UN statement that characterizes hate speech as "any form of communication, whether in speech, writing, or behavior, that attacks or employs pejorative or discriminatory language targeting an individual or a group based on their religion, ethnicity, nationality, race, color, descent, gender, or other identity-related factors" (*United Nations Strategy and Plan of Action on Hate Speech*, 2019). Furthermore, we identified additional characteristics and targeted groups frequently observed in online discourse within Switzerland. As a result, the expanded definition of hate speech encompasses attacks and insults directed towards sex, age, gender, religion, nationality/skin color/origin, as well as mental and physical impairments, which is building upon the UN definition. Additionally, they incorporated social status (such as income, education, and job), political orientation, appearance, and other factors (e.g., Covid-19, cyberbullying) into their operational definition, which we use here again.

In practice, it is not straightforward to delimit toxic speech, i.e., insults and derogatory use of language, from statements targeting a specific person or group based on their (inherent) characteristics. Many toxic claims linguistically closely resemble hate speech statements. Therefore, we included toxic speech in our analysis as hate speech for several reasons, and this approach can be characterized as a broad definition of hate speech. While hate speech, as traditionally defined, targets individuals or groups based on their inherent characteristics, toxic speech often involves insults and derogatory language without explicitly referencing these specific characteristics. However, toxic speech can perpetuate harmful stereotypes, foster a hostile environment, and incite discrimination or violence against specific individuals or groups. Furthermore, this improves the resulting figures to explain the need to document more adequately show the inherent differences between annotation groups and their merits.

B Annotation Process with chatGPT

Our methodology incorporates the usage of the OpenAI-developed ChatGPT (API model 3.5 Turbo) to annotate the designated comments. We employ a predefined codebook as the core reference guide or prompt to guide the ChatGPT’s processing and responses. In essence, the codebook comprises the requisite categories or themes which the ChatGPT was expected to assign to each comment.

The model individually processed each comment and then categorized it based on the themes and the conceptual rules defined in the codebook. To ensure the same number of codings for each comment as with all six other groups, we let chatGPT code the 500 comments three times independently, resulting in a total of 1,500 coding operations.

To perform this, each comment was fed as input to the model along with the codebook. In this context, the codebook served as a priming sequence to direct the model’s attention towards the desired facets of the comment text, essentially, the attributes we wished to encode. With the themes and the rules from the codebook in mind, ChatGPT-3.5 Turbo generated an output representing the annotated comment.

In configuring the model, we utilized the standard temperature setting for ChatGPT. The temperature, essentially a measure of the randomness of the model’s output, is set at a moderate level to balance consistency and diversity in the generated responses. Lower values near 0 produce more focused and deterministic outputs, while higher values closer to 1 introduce more randomness. This standard-setting helped ensure the annotation process was adaptable yet consistent across the wide range of comments.

By utilizing the ChatGPT-3.5 Turbo with the standard temperature setting, we ensured a balance between coherent and deterministic outputs and the introduction of creative variability. This resulted in the model generating reasonable, informative, and varied codings, ensuring the robustness and richness of the coded dataset.

C Multi-class Results using the narrow definition of Hate Speech

In this study, we employ a binary classification approach in our primary analysis to gain valuable insights into the gap between different annotation approaches and to assess their performance using a unique corpus of online comments developed for detecting and countering online hate speech in Switzerland. Our broad definition of hate speech encompasses toxic comments within its scope. By adopting the binary classification, we can effectively compare and contrast various annotation strategies, understanding their strengths and weaknesses in classifying hate speech versus non-hate speech instances. The corpus of online comments used in this research project is specifically curated for studying hate speech and related phenomena in the Swiss context. The binary classification serves as a basis for comprehending the challenges in creating an efficient and effective training set.

However, for a more detailed and nuanced perspective, we present the results in the appendix using a multiclass labeling approach, which includes three distinct classes. This fine-grained approach allows us to explore and differentiate between hate speech, toxic speech, and non-hate speech instances. By incorporating this multiclass labeling, we provide a more comprehensive understanding of hate speech classification and its variations within our dataset.

Binary classification is instrumental in revealing the disparities between different annotation methods. At the same time, the multiclass approach in the appendix enriches our understanding of hate speech by providing a more detailed and differentiated measurement encompassing three distinct classes. Together, these analyses contribute to our efforts in comprehensively studying hate speech and improving its classification, detection, and mitigation.

C.1 Multi-class Results

We evaluate the performance and reliability of different annotation groups in detecting and classifying online hate speech and toxic messages by assessing the intercoder agreement within each annotation group plus chatGPT. Unlike for the binary case, we now observe slightly lower intercoder agreement over all groups (see: Table 2). This was to be expected since adding more levels generally decreases performance (Akhtar et al., 2020; Kralj Novak et al., 2022). The variation in annotations across different coders underscores the need for a robust approach to construct reliable training data again.

Platform	Agreement	Krippendorff's Alpha
Experts	0.762	0.728
Research Assistants	0.538	0.435
NGO	0.617	0.433
Prolific	0.549	0.391
Appen	0.269	0.129
Citizen Science	0.455	0.308
Chat GPT	0.698	0.632
Combined ICR	0.211	0.310

Table 2: Intercoder agreement of different groups using the distinction between hate speech and toxic speech

As with the binary classification scheme, we use majorization. Now let us look at the accuracy between groups more closely than our expert group. We again see similar patterns in Figure 4. When we look at the accuracy of the five groups plus chatGPT for comments labeled by the experts as hate speech, toxic and non-hate speech, we again see that different groups have different strengths when labeling hate speech, toxic speech, and normal speech. Hence, understanding the strengths and weaknesses of these groups in hate speech labeling becomes crucial for a nuanced assessment of the classification task. The interesting finding here is that when we start differentiating between toxic- and hate speech, we see that the overall accuracy for non-hate speech does not change much. Still,

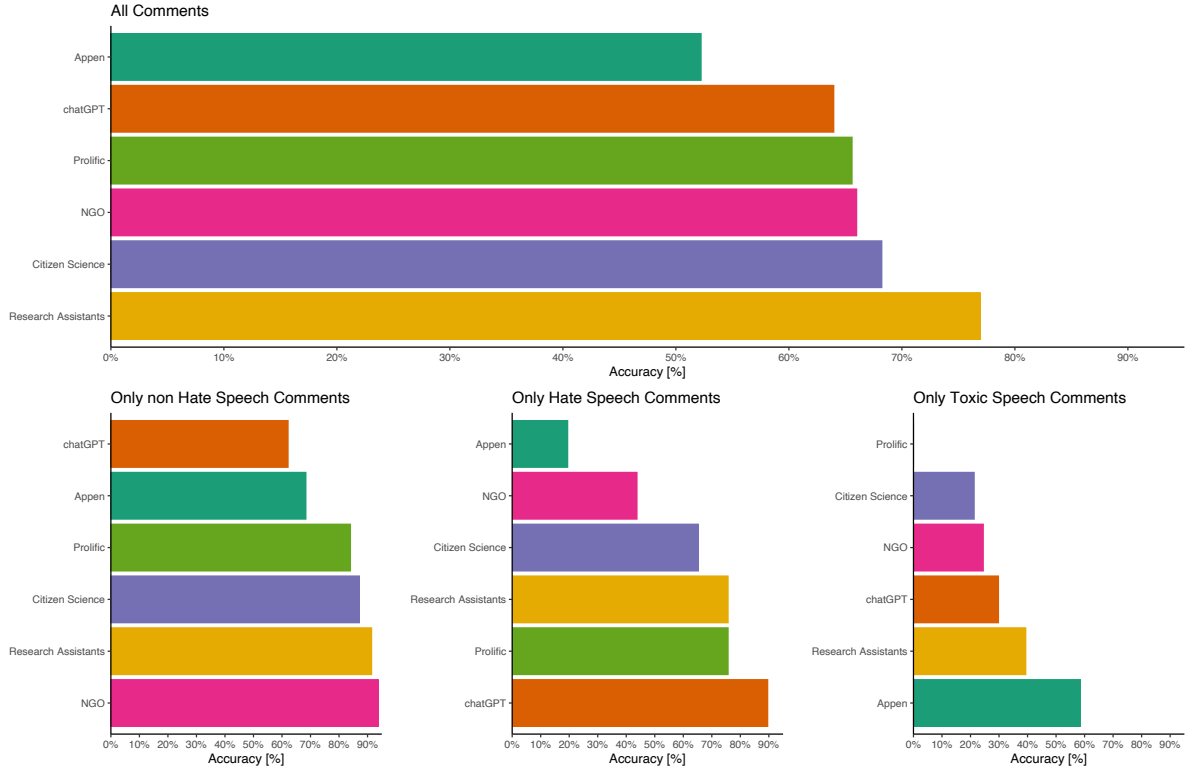


Figure 4: Accuracy of all five groups and ChatGPT compared with the gold standard produced by experts in the field.

we see some big changes concerning labeling hate speech now that we exclude toxic labels. Appen’s crowd workers seemed to have large troubles with actual hate speech while doing great with toxic speech. Apart from that, we also see that the prolific crowd completely omitted to label any comment as toxic, which indicates once more how subjectively difficult this task is. This again highlights that disentangling the performance by class is of utmost importance.

The results for the agreement of the different groups against the gold standard group reveal nothing new to the results of the binary results except for the way lower agreement between the groups for hate speech and toxic speech (see: Figure 5). This finding underpins the difficulty of building reliable and valid training data in a multi-class setting. It impressively shows that with many classes building training data comparable to the experts in the field is getting very challenging.

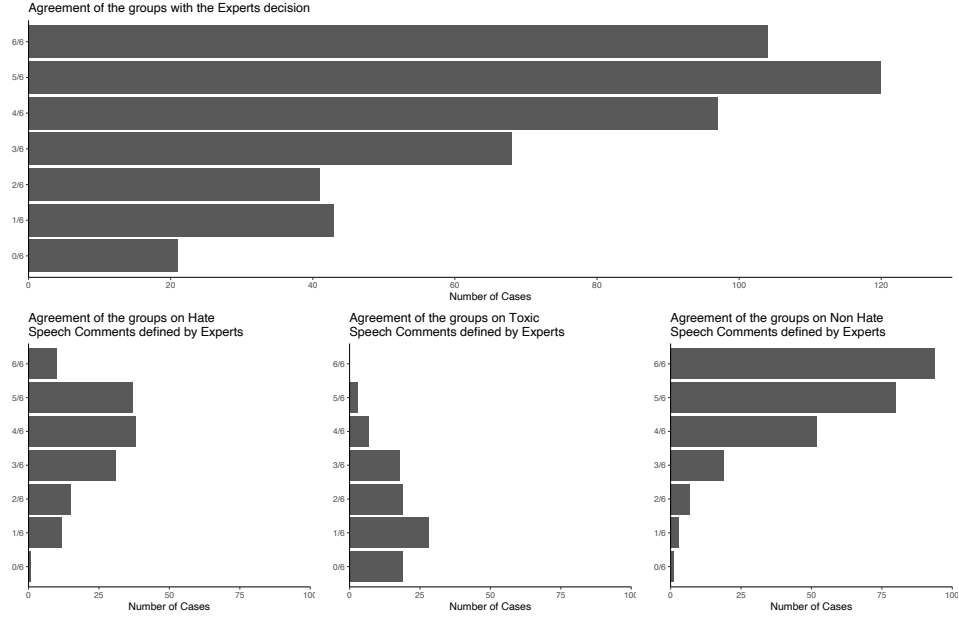


Figure 5: Agreement count of the five groups and ChatGPT compared to the gold standard of experts in the field

In conclusion, these results again indicate that employing the majority vote method with multiple annotators significantly improves the accuracy and reliability of annotations for hate speech detection. Moreover, it shows that more research is needed when it comes to multi-class labeling of data, especially in the area of annotator group size.