Content Moderation in Online Platforms: A Study of Annotation Methods for Inappropriate Language

Baran Barbarestani, Isa Maks, Piek Vossen
Vrije Universiteit Amsterdam
{b.barbarestani, isa.maks, p.t.j.m.vossen}@vu.nl

Abstract

Detecting inappropriate language in online platforms is vital for maintaining a safe and respectful digital environment, especially in the context of hate speech prevention. However, defining what constitutes inappropriate language can be highly subjective and context-dependent, varying from person to person. This study presents the outcomes of a comprehensive examination of the subjectivity involved in assessing inappropriateness within conversational contexts. Different annotation methods, including expert annotation, crowd annotation, ChatGPT-generated annotation, and lexicon-based annotation, were applied to English Reddit conversations. The analysis revealed a high level of agreement across these annotation methods, with most disagreements arising from subjective interpretations of inappropriate language. This emphasizes the importance of implementing content moderation systems that not only recognize inappropriate content but also understand and adapt to diverse user perspectives and contexts. The study contributes to the evolving field of hate speech annotation by providing a detailed analysis of annotation differences in relation to the subjective task of judging inappropriate words in conversations.

Keywords: Online Content Moderation, Subjectivity in Annotation, Inappropriate Language

1. Introduction

In the digital age, online communication has become an integral part of human interaction. As individuals engage in discussions and share opinions across various platforms, the issue of inappropriate content emerges as a significant concern. Addressing the challenge of identifying and annotating inappropriate content, regardless of the question whether this is hate speech or not, is crucial for maintaining a safe and respectful online environment. But also for the purpose of detecting explicit and implicit hate speech, inappropriate language detection can play a role in online (platform) conversations. Within the context of a conversation, interlocutors can start generalizing and targeting a group at some stage of the conversation and start using inappropriate language at another point. We, therefore, are studying both inappropriate and targeting language within the context of complete online conversations. In this work, we are reporting on detecting inappropriate language within conversational context regardless of whether specific groups of people are being targeted. In future work, we will also report on targeting language in conversation and how both targeting and inappropriate language evolve during interactions.

Inappropriate content encompasses text that is considered offensive, harmful, or objectionable based on social, cultural, or ethical standards (Yenala et al., 2018). These standards are expected to vary from community to community, which makes annotation of the data subjective: people will experience inappropriate language differently. Annotating social media data and training models is, therefore, not just a matter of wrong or right but also of taste and standards. In this paper, we describe a curated data set of English Reddit conversations that are likely to contain inappropriate language. We applied a series of different annotation methods to these data to analyze the subjectivity of the annotations: 1) expert annotation, 2) crowd annotation, 3) prompted ChatGPT annotation and 4) lookup using lexicons including toxic words. We observe that the agreement across the annotations is very high for all types of annotations while an error analysis shows that, besides differences in span annotations, most disagreements are subjective. This suggests that models should be value-aware but also be able to differentiate between interlocutors to judge conversations as inappropriate given the context.

Our contributions are: 1- We curated an English Reddit data set with discussion threads having a high probability of toxic language, which can be used to study conversational contexts. 2- We applied four different annotation methods to this data set to mark inappropriate words in the comments. 3- We analyzed the agreement across the annotations using different methods and pplied an error analysis to the disagreements. 4- We report on the subjectivity of the annotations.

Our data, code, and guidelines are available on our Github repository [1].

[1] https://github.com/cltl/InappropriateLanguageDetection

2. Related Work

2.1. Annotation Methodologies and Taxonomies

Hate speech has been subject to diverse annotation methodologies. Vidgen and Derczynski (2020) analyzed expert annotation approaches. The study emphasized the need for well-defined tasks, carefully selected language(s) for annotation, and a clear taxonomy of abuse categories, showcasing the importance of engaging relevant social scientific theory. Furthermore, the paper highlighted the significance of annotator expertise and diversity, urging the selection of annotators based on skill sets, experiences, and demographic backgrounds. The study also sheds light on the scarcity of systematic information about annotators in existing data sets, underlining the necessity for detailed demographic information and guidelines. While it underscores the value of annotation guidelines and iteratively developed data sets, the study acknowledges the challenges tied to nuanced aspects of abusive language, such as irony and intent. The approach outlined in (Babakov et al., 2021) leverages a large-scale crowdsourcing study to annotate sensitive topics and appropriateness in Russian-language texts. They present a process involving manual labeling, automated classification, and the identification of inherent keywords associated with sensitive topics. Despite successfully collecting a substantial data set, the paper acknowledges several shortcomings, including challenges in ensuring accurate manual labeling and potential biases in crowdsourced annotations due to topic complexity.

2.2. Challenges with Respect to Disagreements among Annotators

(Davani et al., 2023) investigate how normative social stereotypes can influence the annotation process and subsequently impact hate speech classifiers. The research demonstrates the necessity of understanding annotators' biases and the incorporation of social scientific theories to improve hate speech annotation. It introduces the concept of annotation biases related to social stereotypes, emphasizing that a diversified pool of annotators can help reduce these biases. As researchers continue to refine hate speech annotation methods, they provide a valuable perspective on the challenges and opportunities in this evolving field. Nonetheless, the paper does not extensively address the specific methods or guidelines that could effectively minimize the influence of social stereotypes during the annotation process. While the study identifies the issue and highlights the value of recognizing disagreements among annotators, it falls short in providing concrete recommendations or countermeasures to mitigate these biases. In addition, (Sang and Stanton, 2022) try to understand the origin and significance of disagreements among data labelers, offering a case study on individual differences in hate speech annotation.

2.3. Analysis and Impact of Context

Previous research by (Qiu et al., 2023) has acknowledged the challenge of detecting and moderating Not Safe for Work (NSFW) content within open-domain dialogue systems but often lagged in detecting NSFW language, especially within dialogues. Notably, The paper introduces CENSORCHAT, a data set for NSFW dialogue detection, leveraging knowledge distillation with GPT-4 and ChatGPT. Nevertheless, it presented limitations, including a reliance on predefined prompts for annotations, potential biases, and limited coverage of NSFW contexts. (Ljubešić et al., 2022) analyze the significance of context in hate speech annotation. While (Ljubešić et al., 2022) extensively discuss the impact of context on annotation quality, they do not delve deeply into the potential biases introduced by annotators, which can affect the study's outcomes. In addition, (Zhang et al., 2018) introduce the phenomenon of conversational derailment, where civil discussions take a negative turn with one participant attacking another. The study constructs a labeled data set for personal attacks through an annotation procedure involving manual inspection and crowdsourced filtering. However, the process of annotating conversations for personal attacks is subjective and prone to biases, which the study does not fully address.

These studies collectively highlight the complexities and challenges associated with annotating and detecting abusive language in online discourse. They emphasize the importance of well-defined tasks and clear taxonomies of abuse categories. Moreover, they underscore the significance of annotator expertise, diversity, and demographic information, as well as the need for nuanced understanding and context in annotation guidelines. However, many studies fall short in addressing biases and discrepancies inherent in the annotation process. Our work contributes to the above by analyzing differences across annotators and annotation methods in more detail in relation to a highly subjective task to judge whether words in conversations are inappropriate.

3. Data Set

3.1. Data Description

The dataset utilized in this study comprises English conversation threads sourced from various subreddits on Reddit, where the comments within these threads have been banned. A total of 28 subreddits were included in the data set. The data set comprises 67,677 submissions and 1,168,546 comments. The combined number of tokens in the data set is 4,017,460.

The selection approach to collect data was inspired by (Vidgen et al., 2021). Since we want to study the impact of conversational context on interpretation, it is essential to capture the structure and dynamics of the conversation threads. We processed the data to reconstruct separate conversation threads from the branching comments in each conversation. In this approach, the first comment of a conversation thread became the start of a new node in the original conversation, ensuring that conversation threads (also known as subthreads) did not overlap with each other as the comments in each subthread are unique.

After constructing the branching subthreads in the data set, we selected subthreads using the following criteria:

Total Number of Comments: Subthreads were filtered to have a minimum of 3 and a maximum of 17 comments. This range was chosen to strike a balance between having enough data for meaningful analysis and avoiding excessively long conversations that might introduce outliers or complicate the analysis.

Number of Tokens per Subthread, Including Punctuation: After observing the distribution of the number of tokens across the subthreads, we selected a token count range from 51 to 1,276 tokens. This range was chosen based on the observation that the majority of subthreads contained at least 51 comments.

Maximum Number of Tokens per Comment: The maximum token count was set to 38 tokens across all the comments within the subthread.

Toxicity Level: Subthreads were selected based on their proportion of toxic words out of all the tokens in each subthread using three lexicons: Wiegand (Wiegand et al., 2018), Hurtlex (Bassignana et al., 2018), and a lexicon created by (Schouten et al., 2023) with the methodology presented in (Zhu et al., 2021). We categorized the subthreads into 10 bins with the highest toxicity and based on their normalized toxicity scores ranging from 0.08 to 0.2.

The majority of the comments and subthreads in Reddit do not contain toxic words. A random selection of subthreads is, therefore, very likely to contain no inappropriate words. Therefore, we selected 400 subthreads from the higher toxicity bins and an additional 98 subthreads with a toxicity score of 0. The final statistics for both toxic and non-toxic subthreads can be found in Table 1.

Statistic	Toxic	Non-Toxic
# of tokens	23,393	4,984
# of comments	1,778	367
# of subthreads	400	98
Avg. # comments x sub	4	8
Max. # comments x sub	15	9
Min. # comments x sub	3	3
Avg. # tokens x comment	13	13
Max. # tokens x comment	35	31
Min. # tokens x comment	1	1

Table 1: Selected Subthreads Statistics

4. Annotation Task Design

The annotation task focuses on identifying and classifying instances of inappropriate language within the context of comments.

4.1. Definitions

We define two key terms: context, which refers to the previous comment(s), and explicitly inappropriate language, illustrated in the next example.

Title: The Wall Is Hitting Much Sooner?

Context: Yeah man these gym thots I see all the time might not even be 35, but they look like they are in their 40's! Wrinkles, tattoos, fucking disgusting.

User ID: Infinitewisdom1984

Comment 2: Eww! I forgot about all the fucking middle-aged crossfitters too. Cringiest shit on earth.

Explicitly inappropriate language: This applies to sentences that contain specific words generally recognized as inappropriate. Examples of explicitly inappropriate language include slurs, swear words, profanity, and other terms with inherently offensive or derogatory meanings. For instance, in the sentence, "She is not being a bitch. She is just less likely to put up with your shit," the words "bitch" and "shit" are explicitly inappropriate due to their generally inappropriate meanings.

4.2. Annotation Instructions

The annotators were provided with basic instructions. We explained explicitly inappropriate language as comprising swear words, slurs, and any other kind of profanity, such as f*ck, sh*t, b*tch, n*gger, etc. The annotators were instructed to mark all inappropriate words and also to indicate if a comment contained no explicitly inappropriate words at all. We designed the task through the LingoTURK platform (Pusse et al., 2016) and used the Prolific platform (Palan and Schitter, 2018) for annotator recruitment.

5. Expert Annotations

To have an independent evaluation of the crowd-annotation, we decided first to apply expert annotation to a subset of the data. Out of the initial pool of 498 subthreads, 39 were selected as the gold set and annotated by 3 expert annotators, i.e. the authors of this study. The selected subthreads contain a total of 209 comments and 2491 tokens. Two annotators followed the instructions of the crowd strictly by annotating inappropriate words regardless of the context, whereas one annotator applied the instructions loosely by considering the context to decide whether the inappropriate words were intended to offend somebody. A summary of average Cohen's Kappa values at the token level across different annotator pairs and all gold data can be seen in Table 2. Overall, annotators demonstrate moderate to high agreement, with the highest agreement for annotations between A1 and A2, both of whom followed the strict interpretation. Annotator A3, following the loose interpretation, has clearly the lowest agreement with both of the others.

Annotators	Kappa	Lenient (%)	Exact (%)
A1 vs. A2	0.805	83.25	76.25
A1 vs. A3	0.587	77.0	46.0
A2 vs. A3	0.573	77.0	45.0

Table 2: Inter-Annotator Agreement and Inappropriate Span Agreement among Experts

To explore the sources of disagreements among the annotators we calculated lenient and exact agreement scores following the approach outlined by (Somasundaran et al., 2008). Specifically, the "Exact" span agreement score assesses agreement when two text spans match precisely. On the other hand, the "Lenient" span agreement score considers an overlap relation between the two annotators' retrieved spans as a hit. If strict and lenient scores are close (as for A1 and A2, see Table 2) span differences are not an important source of disagreement. If the differences are bigger (as for A3 vs. A1 and A2, respectively) span differences are an issue.

We prioritize token-level evaluation to analyze short spans (mostly 1 or 2 tokens) for a more detailed examination of inappropriate language in online discussions. This approach is chosen over character-level evaluation as our analysis focuses on short phrases and individual words rather than individual characters.

To further compare with other annotation approaches, we adjudicated the expert annotations by following the strict interpretation and majority vote, which we label as AdjExpert annotation from here onwards.

6. Crowd Annotations

Crowd annotations were conducted for all 498 subthreads by five annotators. The selection of annotators followed the approach outlined by (Barbarestani et al., 2022), involving a pre-screening and post-screening selection process to ensure the quality and reliability of annotations. Inter-annotator agreement scores are obtained from all 498 subthreads (see Table 3) The average percentage agreement score offers a measure of consensus among annotators, with a higher agreement score on all tokens (0.98) compared to tokens labeled as inappropriate (0.90), indicating that tokens labeled as inappropriate may present more difficulty in accurate assessment. Additionally, the average Cohen's Kappa demonstrates substantial agreement (0.76) among pairs of crowd annotators. We also adjudicated the crowd annotations by taking the majority vote. From now onwards, the adjudicated set will be called 'AdjCrowd'.

Metric	Value
Avg. Cohen's Kappa on all tokens	0.76
Avg. % agreement on inappropriate tokens	0.90
Avg. % agreement on all tokens	0.98

Table 3: Inter-Annotator Agreement among Crowd Annotators

7. ChatGPT-Generated Annotations

In our annotation process, we utilized ChatGPT to generate annotations for inappropriate tokens in 498 subthreads, considering the entire conversation context. Instructions and examples provided to ChatGPT were consistent with those for human annotators. Despite clear instructions, ChatGPT occasionally generated tokens from the broader context rather than the target comment, necessitating re-generation for accuracy. We maintained consistency by specifying the correct order of columns in the input CSV file and accommodating all response formats from ChatGPT. Sporadic service unavailability was encountered due to internal errors, addressed by implementing a retry mechanism with a 5-second pause to ensure continuity in the annotation process. This comprehensive approach ensured a consistent and rigorous assessment of inappropriate tokens in the given data set. An example prompt for ChatGPT is provided in Figure 1:

8. Lexicon-Based Annotations

The process of lexicon-based annotations, involved the integration of the three distinct lexicons dis-

You are a language model that helps annotate inappropriate words and harmful messages in comments. Do not generate words that are not present in the comment. Given a CSV file, First, you should read the following columns per row: "title text," "comment," "context," "user ID," and "post text." "Context" refers to the previous comments. You should read each "title text," "post text," and "context." Then, you should read "comment."

Instructions

You should first read the previous comments. Then, read the current comment carefully and identify the inappropriate words in the comment. Inappropriate words are swear words, slurs, and any other kind of profanity, such as fuck, shit, bitch, nigger, etc. Such words are inappropriate in themselves such that one does not need to read them in a sentence to understand that they are inappropriate. Below are some examples:

Figure 1: ChatGPT Prompt for Annotating Inappropriate Words in Comments

cussed in 3.1. To enhance the comprehensiveness of our annotations, we constructed a combined lexicon by uniting toxic words from these three lexicons. This combined lexicon, comprising 3451 tokens, served as a comprehensive reference for identifying inappropriate language in the data set. Among these tokens, 54 were found to be shared among the three lexicons. To generate annotations for individual tokens within comments, we utilized this combined lexicon. If a token was found within the lexicon, we labeled it as "inappropriate." Conversely, tokens not present in the combined lexicon were labeled as "not inappropriate." Examples of the shared tokens among the three lexicons are the following: fucking, fucks, asshole, fat, gay

9. Inter-Annotator Agreement Across Four Methods

9.1. Annotation Approach Comparison and Analysis

Here, we provide insights into the annotation results, shedding light on both the quantity and average span length of inappropriate tokens for different annotation methods in both gold and non-gold sets. In our study, we use the term "annotation" and not "classification" to encompass a broad range of methods (including manual methods) as our intention is to capture the process of labeling inappropriate words within the context of online discussions.

Method	Inappr. Tokens	Avg. span length
AdjCrowd	130	1.08
AdjExpert	192	1.25
Expert (A1)	167	1.21
Expert (A2)	201	1.24
Expert (A3)	310	2.06
ChatGPT	146	1.29
Lexicon	297	1.19
AdjCrowd	1408	1.1
ChatGPT	1332	1.26
Lexicon	3056	1.19

Table 4: Inappropriate Token Annotations (Upper Part: Gold, Lower Part: Non-Gold)

Table 4 (column Inappr. Tokens - upper part) displays the number of inappropriate tokens in the gold set for the four annotation methods. The counts range from 130 tokens annotated by the crowd to 310 tokens annotated by expert annotator A3. The expert annotators seem to identify a larger number of inappropriate tokens compared to the crowd. ChatGPT and the lexicon-based approach identified 146 and 297 inappropriate tokens, respectively.

Table 4 (column Inappr. Tokens - lower part) presents the number of inappropriate tokens across non-gold data for all annotation methods, except for the experts, as the expert set does not include annotations of the non-gold set. The counts range from 1408 tokens annotated by the crowd to 3056 tokens annotated using the lexicon-based approach. Interestingly, while the lexicon-based approach identified a substantial number of inappropriate tokens, it also marked a significant number of tokens not marked as inappropriate in the AdjEpert set, indicating its tendency to over-flag tokens as inappropriate. This suggests that the lexicon-based approach may lack nuanced understanding and context. Many of the tokens mentioned in the lexicon are not not toxic at all or are not toxic in particular contexts.

Table 4 (column Avg. span length) demonstrates the average span lengths of inappropriate tokens for the four annotation methods. Interestingly, almost all annotation methods appear to annotate on average short spans (ranging from 1.08 to 1.29) with the exception of Expert (A3) who annotates spans with an average length of 2 tokens (2.06). We already saw in section 5 that this annotator adopted a loose interpretation of the guidelines as compared to the other expert annotators. Here is an example, where all annotators agree on the token "shit" while A3 has also annotated "comments" as part of the larger span:

Example 1.

your rotten brain and shit comments belong with the other addicts

9.2. Token-Level Agreement

To assess the consistency and potential subjectivity of the different annotations, we conducted a cross-annotation comparison on the gold set. For the experts, we utilized the AdjtExpert data, and for the crowd, we used the AdjCrowd set.

Pair	Kappa (Token)	% Agreement (Comment)	% Agreement (Subthread)
AdjCrowd-AdjExpert	79.5%	92.08%	98.08%
ChatGPT-AdjExpert	68.3%	87.33%	100.00%
Lexicon-AdjExpert	62.3%	84.44%	97.92%
AdjCrowd-ChatGPT	63.2%	88.12%	98.08%
AdjCrowd-Lexicon	54.3%	79.69%	95.99%
ChatGPT-Lexicon	50.3%	78.11%	97.92%

Table 5: Comparison of Annotation Agreements at Different Levels

Cohen's Kappa values for the comparison of the four approaches are presented in Table 5: AdjCrowd annotations, AdjExpert annotations, responses generated by ChatGPT, and Lexicon-based annotations. Notably, we observed the highest Cohen's Kappa between AdjCrowd and AdjExpert (79.5%), suggesting a reliable alignment of judgments. Similarly, moderate to substantial agreement was observed in the comparisons between AdjCrowd vs. ChatGPT (63.2%), AdjCrowd vs. Lexicon (54.3%), AdjExpert vs. ChatGPT (68.3%), and AdjExpert vs. Lexicon (62.3%). While ChatGPT demonstrates superior performance compared to the lexicon-based approach, the crowd still outperforms ChatGPT.

.

9.3. Comment-Level Agreement

Furthermore, we compared annotations at the comment level, defining a comment as inappropriate if it contained at least one token marked as such. The findings, summarized in Table 5, reveal varying levels of agreement among annotators. The highest percentage agreement, at 92.08%, is observed between domain AdjExpert and AdjCrowd, indicating strong alignment of opinions. Comparatively, agreement between experts and ChatGPT is slightly lower at 87.33%, suggesting less alignment between ChatGPT's annotations and expert judgments. Additionally, agreement between ChatGPT and the crowd is 88.12%, with slightly less alignment compared to the expert-crowd agreement. The alignment between lexicon-based and ChatGPT annotations is 78.11%, while the agreement between lexicon-based and crowd annotations is 79.69%. Furthermore, the agreement between lexicon-based annotations and experts is 84.44%, suggesting less alignment compared to the crowd.

9.4. Subthread-Level Agreement

We also conducted a comparison of annotations at the subthread level, considering a subthread as inappropriate if it contained at least one inappropriate comment. The results can be seen in Table 5. The table summarizes the overall percentage agreements between annotations provided by the experts, crowd, ChatGPT, and lexicon-based approach at the subthread level. The values range from 95.99% to 100.00% across different pairs, demonstrating a high level of agreement. The "ChatGPT-AdjExpert" pair consistently achieved 100.00% agreement across all gold data, indicating a high level of agreement between expert annotations and ChatGPT's generated annotations. Additionally, lexicon-based annotations show a high level of agreement with expert and crowd annotations, further validating their reliability.

10. Error Analysis

We conducted an error analysis to identify the sources of discrepancies observed across expert, crowd, and ChatGPT annotations. This analysis was done only on the gold set. The tokens on which there is disagreement are underlined. Table 6 presents a breakdown of the sources of disagreements after assessing each case individually for each set of annotations explained in previous sections. We extracted distinct disagreement cases across annotations, the numbers of which vary. Regarding disagreements among the experts, we isolated instances where one annotator diverged from the consensus of the other two. However, for disagreements among the crowd, we identified cases where two annotators dissented from the collective judgment of the remaining three, which yielded a percentage agreement of 0.6, signifying a significant level of discord among annotators.

Source	Experts	Crowd	ChatGPT vs. AdjExpert	AdjExpert vs. AdjCrowd
Subj. interpretation	92 (41.25%)	24 (82.76%)	69 (69%)	51 (82.26%)
Span difference	97 (43.5%)	0 (0%)	6 (6%)	5 (8.1%)
Difficult language	15 (6.73%)	2 (6.9%)	5 (5%)	3 (4.84%)
Annotation error	7 (3.14%)	0 (0%)	-	0 (0%)
Target group	12 (5.38%)	3 (10.34%)	8 (8%)	3 (4.84%)
Lack of consist.	-	-	12 (12%)	-
Total	223	29	100	62

Table 6: Comparison of Sources of Disagreements

10.1. Expert Annotation

Subjective interpretation: Instances where annotators had differing interpretations based on subjectivity and personal judgment

Example 2.

> Furries should be in the same mental institutions as trannies. What in the fuck happened to this country.

The term "Furries" refers to individuals interested in anthropomorphic animal characters, with its appropriateness subject to context and annotator perspective. In the provided context, it is used derogatorily, equating "Furries" with mental illness, which can be offensive. Opinions vary; some view it as innocuous in certain contexts, while others find it offensive. "Fuck," a vulgar term expressing strong emotions, varies in appropriateness based on context and community norms. In this context, it conveys frustration about the country's state. While deemed inappropriate in formal settings, it is more accepted in casual or online discourse. Opinions on its appropriateness also differ among annotators.

Span difference: Disagreements arising from varying opinions regarding the inappropriate text spans for annotation

Example 3.

> Post the picture of Donald Jr with his kids. Jesus christ hes an ugly son of a bitch - that's the cringe.

The comment features the phrase "ugly son of a bitch," where the term "bitch" is often considered inappropriate for its derogatory nature. However, annotators may differ in their interpretation of the span of inappropriate language. Some may annotate only the word "bitch" as inappropriate, while others may deem the entire phrase "an ugly son of a bitch" inappropriate due to its derogatory connotation.

Difficult/ ambiguous/ complex language: Cases involving complex, ambiguous, or challenging language, leading to differing annotations

Example 4.

> go back to your fucking estro weed subs my dude. your rotten brain and shit comments belong with the other addicts.

The use of "estro," "weed," and "subs" in the provided comment presents challenges for annotators due to their slang or abbreviated nature and the lack of clear context. "Estro" is an informal abbreviation for estrogen, but its specific meaning might not be immediately clear to all readers, leading to ambiguity. "Weed," typically understood as marijuana or cannabis, lacks context in this instance, causing uncertainty about its intended reference. Similarly, "subs," likely short for "subreddits," could be interpreted in various ways without explicit clarification, contributing to uncertainty among annotators. These factors make interpreting these terms difficult and contribute to ambiguity in the comment.

Target group annotation: Disagreements related to associating inappropriate language with specific target groups and annotating the associated target group as well as or instead of the inappropriate token

Example 5.

> They awoke the sleeping neck-bearded giant by trying to fuck with his video games. Now the angry neck-beard giant has found a new game - fucking up the SJW/Marxist/Globalist establishment.

The comment targets the group "SJW/Marxist/Globalist" negatively, implying opposition or attack against them. "SJW" refers to social justice warriors, often used derogatorily for those advocating progressive causes. "Marxist" and "Globalist" are also used pejoratively. The comment portrays these groups as being challenged by a metaphorical "neck-bearded giant" and suggests aggressive retaliation against them, conveying a hostile attitude. Some annotators have also highlighted these target groups as well as the inappropriate language associated with them.

Annotation error: Discrepancies arising from errors made during the annotation process were identified and addressed through thorough discussion among expert annotators. Each disagreement case was individually examined and deliberated upon to recognize and acknowledge any errors that may have occurred during the annotation process.

The analysis showed a considerable number of span differences, but also a high count of subjective interpretation as the main sources of disagreements. We can clearly see the impact of a more loose contextual interpretation versus a more strict interpretation that ignores the context.

10.2. Crowd Annotation Error Analysis

Similar to the expert annotation error analysis, we conducted an error analysis for the crowd annotations. As can be seen in Table 6, most cases of disagreement were related to subjective interpretation. There were no disagreements in span differences, and only a few cases related to difficult/ambiguous language.

10.2.1. AdjExpert vs. AdjCrowd

We identified discrepancies between the final labels in the AdjExpert and AdjCrowd sets. This comparison yields valuable insights into the nature of disagreements, which can be observed in Table 6. Notably, AdjExpert marked a significantly higher number of tokens as inappropriate compared to AdjCrowd, indicating differing standards and subjective interpretations between the two groups. In all instances, it was noted that the crowd did not flag the tokens as inappropriate, indicating a trend toward stricter criteria among expert annotators who demonstrate greater sensitivity to such content.

10.3. ChatGPT Annotation Error Analysis

We performed an error analysis on ChatGPT annotations by comparing them to the AdjExpert data, as can be seen in Table 6. It is important to note that ChatGPT may interpret language with bias, sentiment, or viewpoint, which probably differ from human experts' consensus opinion. Here is an example of a case on which ChatGPT disagreed with AdjExpert:

Example 6.

Lol freedom fighter. You're a redneck faggot bro foh

This discrepancy is due to subjective interpretation because the appropriateness of the term "redneck" can vary depending on context and individual perspectives. In certain contexts, "redneck" may be used as a neutral or even affectionate term to describe someone from a rural or working-class background. However, in the provided example, the term is used alongside "faggot," which is a derogatory and offensive slur targeting individuals based on their sexual orientation. While in the AdjExpert set the term "redneck" was considered to be inappropriate in this context due to its derogatory connotation when paired with "faggot," ChatGPT may have failed to recognize the offensiveness of the term "redneck" in this specific context.

A particularly noteworthy disagreement category added to our analysis of this set of annotations is the 'lack of consistency in word forms' category. This inconsistency includes variations in word forms, such as singular, plural, conjugated, and other linguistic transformations. For instance, consider the sentence below:

Example 7.

They awoke the sleeping neck-bearded giant by trying to fuck with his video games. Now the angry neck-beard giant has found a new game - fucking up the SJW/Marxist/Globalist establishment.

In this example, ChatGPT identifies "fuck" as inappropriate but fails to flag "fucking," which is another form of the same word. In some cases, ChatGPT even failed to recognize the same repeated word in the same sentence as inappropriate. Since it was challenging to determine whether a response from ChatGPT was genuinely an error, we excluded the "annotation error" category. Unlike human annotators, we cannot discuss each case individually with ChatGPT to conclude whether it was really an error made by ChatGPT or not. For target group annotation discrepancies, We examined all cases of disagreement, where either ChatGPT or AdjExpert annotated a target group.

Overall, in analyzing the data presented in Table 6 across different methods, several key insights emerge. The prevalence of subjective interpretation and span differences underscores the significance of interpretive flexibility in content moderation, with different annotators holding varying perspectives.

11. Conclusion

This study examined various methods for annotating inappropriate language in online discussions, including expert, crowd, ChatGPT-generated, and lexicon-based annotations. It identified sources of disagreement among annotation sets, such as subjective interpretation, span differences, and language difficulty. Each annotation method exhibits strengths suitable for different content moderation contexts: crowd annotations for scalability and diverse perspectives, ChatGPT-generated annotations for real-time moderation, lexicon-based annotations for customizable filters, and expert annotations for high-stakes content or legal compliance. It is important to note that the inconsistencies between ChatGPT and the crowd suggest a need for further investigation in future studies. Emphasizing adaptable content moderation approaches, the study lays groundwork for exploring implicit hate speech and advocates for nuanced understanding within broader contexts. By analyzing inter-annotator agreement and addressing subjective disagreements among human annotators, the research aims to maintain variation and mitigate errors through revised task instructions. It refrains from directly adjudicating subjective disagreements and offers flexibility upon data release, allowing researchers to combine annotations or designate specific annotations as gold references. Future plans could involve exploring a hybrid annotation pipeline integrating expert, crowd, and ChatGPT-generated annotations to enhance subjective variation, evaluated through empirical studies.

12. Acknowledgements

This research was supported by Huawei Finland through the DreamsLab project. All content represented the opinions of the authors, which were not necessarily shared or endorsed by their respective employers and/ or sponsors.

13. Bibliographical References

Nikolay Babakov, Varvara Logacheva, Olga Kozlova, Nikita Semenov, and Alexander Panchenko. 2021. Detecting inappropriate messages on sensitive topics that could harm a company's reputation. *arXiv preprint arXiv:2103.05345*.

Baran Barbarestani, Isa Maks, and Piek Vossen. 2022. Annotating targets of toxic language at the span level. In *Proceedings of the Third Workshop on Threat, Aggression and Cyberbullying (TRAC 2022)*, pages 43–51.

Elisa Bassignana, Valerio Basile, Viviana Patti, et al. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *CEUR Workshop proceedings*, volume 2253, pages 1–6. CEUR-WS.

Aida Mostafazadeh Davani, Mohammad Atari, Brendan Kennedy, and Morteza Dehghani. 2023. Hate speech classifiers learn normative social stereotypes. *Transactions of the Association for Computational Linguistics*, 11:300–319.

Nikola Ljubešić, Igor Mozetič, and Petra Kralj Novak. 2022. Quantifying the impact of context on the quality of manual hate speech annotation. *Natural Language Engineering*, pages 1–14.

Stefan Palan and Christian Schitter. 2018. Prolific. ac—a subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, 17:22–27.

Florian Pusse, Asad Sayeed, and Vera Demberg. 2016. Lingoturk: managing crowdsourced tasks for psycholinguistics. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 57–61.

Huachuan Qiu, Shuai Zhang, Hongliang He, Anqi Li, and Zhenzhong Lan. 2023. Facilitating nsfw text detection in open-domain dialogue systems via knowledge distillation. *arXiv preprint arXiv:2309.09749*.

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, Marseille, France. European Language Resources Association (ELRA).

Yisi Sang and Jeffrey Stanton. 2022. The origin and value of disagreement among data labelers: A case study of individual differences in hate speech annotation. In *International Conference on Information*, pages 425–444. Springer.

Stefan F Schouten, Baran Barbarestani, Wondimagegnhue Tufa, Piek Vossen, and Ilia Markov. 2023. Cross-domain toxic spans detection. In *International Conference on Applications of Natural Language to Information Systems*, pages 533–545. Springer.

Swapna Somasundaran, Josef Ruppenhofer, and Janyce Wiebe. 2008. Discourse level opinion relations: An annotation study. In *Proceedings of the 9th SIGdial Workshop on Discourse and Dialogue*, pages 129–137.

Bertie Vidgen and Leon Derczynski. 2020. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *Plos one*, 15(12):e0243300.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021. Introducing cad: the contextual abuse dataset.

Michael Wiegand, Josef Ruppenhofer, Anna Schmidt, and Clayton Greenberg. 2018. Inducing a lexicon of abusive words–a feature-based approach. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1046–1056.

Harish Yenala, Ashish Jhanwar, Manoj K Chinnakotla, and Jay Goyal. 2018. Deep learning for detecting inappropriate content in text. *International Journal of Data Science and Analytics*, 6:273–286.

Justine Zhang, Jonathan P Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Nithum Thain, and Dario Taraborelli. 2018. Conversations gone awry: Detecting early signs of conversational failure. *arXiv preprint arXiv:1805.05345*.

Qinglin Zhu, Zijie Lin, Yice Zhang, Jingyi Sun, Xiang Li, Qihui Lin, Yixue Dang, and Ruifeng Xu. 2021. Hitsz-hlt at semeval-2021 task 5: Ensemble sequence labeling and span boundary detection for toxic span detection. In *Proceedings of the 15th international workshop on semantic evaluation (SemEval-2021)*, pages 521–526.