Unraveling the Complexities of Offensive Language: A Detailed Analytical Framework for Understanding Offensive Communication Dynamics

Anonymous ACL submission

Abstract

Offensive online content can marginalize and cause harm to groups and individuals. To prevent harm while ensuring speech rights, fair and accurate detection is required. However, current models and data struggle to distinguish offensive language from acceptable non-toxic linguistic variations related to culture or subjective interpretation. This study presents a comprehensive toxicity assessment with two annotated datasets focusing on nuances of human interpretation with objective evaluation. The significant improvement in inter-annotator agreement suggests uncontrollable subjectivity and research biases can arise without structured guidelines. Additionally, we explore the effectiveness of in-context learning with few-shot examples to improve toxicity detection from large language models (LLMs), GPTs specifically, finding that explicit assessment criteria significantly improve agreement between automated and human evaluations of offensive content. The feasibility of criteria-based auto-annotations is evidenced by the better performance of smaller models fine-tuned on 10 times less auto-annotated data with multi-language variations. The findings demonstrate notable efficiency in combining contextual understanding of LLMs with criterion-guided learning.

Content Warning: This article only analyzes offensive language for academic purposes. Discretion advised.

1 Introduction

In the digital age, the anonymity of the Internet and the lack of direct interaction have led to increased offensive and hateful speech (D. Citron and Helen L. Norton, 2011; Mondal et al., 2017). This variation in perception and regulation of offensive speech across different regions, from free speech protection in the US to legal restrictions in Europe (Kocoń et al., 2021), highlights the subjectivity involved and the need for effective detection and analysis methods.

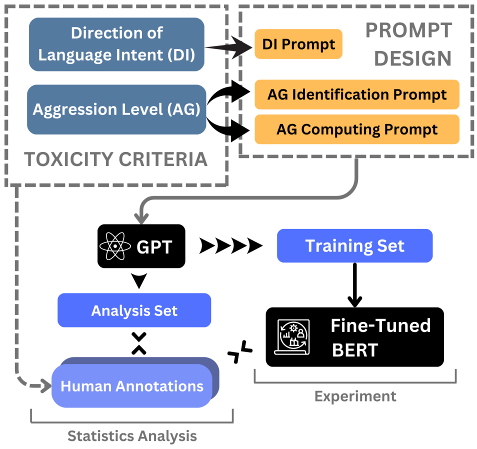

Figure 1: Research Framework

Current datasets typically employ multifaceted methodologies for content categorization, taking into account not just the presence of offensive language but also its context, target, and the intent behind it (Zampieri et al., 2019; Basile et al., 2019; Mollas et al., 2020). Dataset annotations commonly highlight the significance of context in interpreting offensive content. The concept of hate speech often overlaps with offensive language in the construction of corpora and in offensive language detection tasks. Prominent datasets such as Hate Speech and Offensive Language (Davidson et al., 2017), ETHOS (Mollas et al., 2020), HatEval at SemEval-2019 Task 5 (Basile et al., 2019), and HateXplain (Mathew et al., 2021) focus either on hate speech or offensive language, or on the interplay between the two. These datasets adopt varied approaches in handling the relationship between offensive language and hate speech. For instance, datasets like HateXplain, Hate Speech and Offensive Language, and two datasets at SemEval—HatEval and Identification Dataset (OLID) at SemEval-2020 Task 12 (Zampieri et al., 2019)—treat offensive language and hate speech as distinct entities. In contrast, other research integrates the two under the broader term 'abusive language', suggesting commonalities between hate speech and offensive language (Calabrese et al., 2021). The varied usage of terminology in the field has led to some degree of academic ambiguity (Fortuna et al., 2020).

In this work, we focus on the critical distinction between objective aspects, where consensus is achievable, and subjective elements, which are often the subject of debate, in assessing offensiveness and toxicity. We challenge the polarized views that either consider toxicity as entirely subjective or entirely uniform. Our approach argues against relying solely on ambiguous definitions or exhaustive lists for evaluations. By implementing concrete criteria, we address vulnerabilities in the annotation process, such as personal biases and preferences, enhancing the accuracy and reliability of the assessments. The methodology and results of our approach are depicted in Figure 1.

We make the following contributions:

1. We contributed two datasets, one annotated with predefined criteria and the other without, to illustrate the impact of these criteria on annotation.

2. We ensured that our criteria are transparent and replicable, facilitating their application by humans and Large Language Models (LLMs).

3. The results demonstrate the improvement in the agreement and consistency of GPT annotations guided by our criteria.

4. By processing data with GPTs prompted by the proposed criteria, we have successfully fine-tuned smaller models with significantly smaller and diverse annotated datasets to produce better concordance.

2 Related Works

2.1 Lexical Bias

Despite the influence of individual preferences and the potential for over-judgment, lexical bias is a common learning bias shown in many current datasets. This issue of non-offensive yet aggressive language mislabeled as offensive is also called unintended bias (Dixon et al., 2018) or, more specifically, lexical bias (Garg et al., 2023) or linguistic bias (Fan et al., 2019). For instance, (1) and (2) are identified as offensive based on the emotional emphasis `FUCK` in (1), racial terms `nigga` and slang `bitch` in (2):

> (1) And apparently I'm committed to going to a new level since I used the key. Well FUCK. Curiosity killed the Cat(hy) (Barbieri et al., 2020)

> (2) I ain't never seen a bitch so obsessed with they nigga😂" I'm obsessed with mine &#128529 (Davidson et al., 2017)

However, it is unnecessary that the appearance of these terms inherently conveys offensiveness or an intent to harm. Emotional emphasis sounds aggression, but there is no intention to offend others. Racial expressions in the African American Language (AAL) also pose challenges to simplistic judgments that rely solely on the presence of aggressive language (Deas et al., 2023). The lexical form of racial terms, such as n-words, is not intrinsically derogatory. Whether these terms are slurs depends on their perlocutionary effect, which considers the context and circumstances of their usage and reception (Allan, 2015; Rahman, 2012). `nigga` is employed in a romantic context (Garcia et al., 2003; Smitherman; Rahman, 2012), and `bitch` is not used in a gender-offensive manner.

2.2 Analysis and evaluation

Analyzing and annotating subjective content involves several inherent challenges, primarily due to the variability and complexity of human perception and expression (Reidsma and op den Akker, 2008; Hayat et al., 2022). A significant issue in this process is the potential inadequacy of individual annotations, which may result in an unrepresentative sample of viewpoints (Burmania et al., 2015; Leonardelli et al., 2021; Chen and Joo, 2021). Additionally, contextual misinterpretation poses a major problem – a lack of or misrepresentation of context can lead to inaccurate labeling. The influence of the social environment on annotators' decisions cannot be understated, often affecting their judgments subtly (Joseph et al., 2017; Haliburton et al., 2023).

The task of detecting offensiveness is particularly challenging, requiring a balance between subjective interpretation and the need to avoid overt subjectivity. Given the range of valid interpretations, the human annotation should also represent this feature. However, most offensive datasets are constructed based on one single subjective annotation, neglecting other potential interpretations (de Gibert et al., 2018; Basile et al., 2019; Zampieri et al., 2019). Highly unified annotations will neglect the language variations as well as embedded understanding divergence. Highlighting the complexities and challenges in annotating subjective content, we consider the agreement as an additional evaluation approach that does not assume the comparison item is the sole standard rather than solely depending on accuracy measures. Annotations but rather treats it as one possible reference point.

3 Annotation Methodology

The methodology for evaluating linguistic offensiveness consists of two sections: defining the core concepts and proposing criteria corresponding to the definitions. Regarding the overall annotating process, we adopt the tweet-centric annotation approach, focusing solely on individual tweets and contextual thread information. While more practical, enabling streamlined annotator workflows and clear evaluation units, it limits human annotators to evaluating tweet content without considering preceding/subsequent conversational exchanges that provide context. However, this study does not employ a majority ruling to determine singular “correct” annotations per tweet, which risks overlooking nuance. Instead, an inter-annotator agreement is considered when evaluating annotation reliability. This allows more nuanced and reliable assessment, recognizing language’s complexity and the value of diverse perspectives.

3.1 Defining Offensive Language

Some previous studies have also equated toxic speech with hate speech when examining different facets of this language use (Koratana and Hu, 2019; Moon et al., 2020). Toxic language represents another term associated with an offensive language capable of inflicting harm through various mechanisms (Buell, 1998). However, as it lacks an intrinsic association with emotions of anger per se, herein, we treat it as a semantically broader, more neutral substitute nomenclature for offensive language. Hate speech, on the other hand, is more informal, angrier, and often explicitly attacks the target (Elsherief et al., 2018), which could only be one kind of toxic language but is not equivalent to toxic language. Treating toxicity and hatred separately avoids potential confusion arising from treating them as interchangeable concepts while maintaining conceptual alignment with the larger literature on technology-mediated linguistic aggression and harm.

`Offensiveness` and **`Toxicity`** emphasize different aspects of language used to harm people (Kocoń et al., 2021), but these two terms do not distinguish from each other as offensive language and hate speech do. **`Offensiveness`** or **`Toxicity`** in language can be characterized by its capacity to evoke negative or adverse reactions, distinguishing it from the mere use of swear words (Legroski, 2018). This concept is intrinsically tied to notions of linguistic politeness and social decorum (Archard, 2014), where the primary concern is the intention to denigrate or demean, rather than the actual harm inflicted (Archard, 2008). In essence, offensiveness often hinges on the speaker’s intention to belittle or insult, and this intentionality is a crucial aspect in understanding and identifying offensive content. However, the term "aggressiveness" in sociological and psychological studies also has positive connotations (Hawley and Vaughn, 2003). Aggressiveness is a vital component of dominating behavior (Kacelnik and Norris, 1998), but dominating behaviors are not equivalent to behaviors that affect others negatively, which differs from toxic behaviors. When it co-occurs with outward language intention, the language can trigger antisocial or harmful outcomes and, therefore, is offensive and toxic (Stokes and Cox, 1970). Aggressiveness or Aggression alone does not constitute toxicity. Aggressive language components may contribute to offensive speech, but only when coupled with explicit intents to cause harm or distress to a target. Identifying the language used explicitly toward others will prevent annotating bias while retaining some space for different interpretations. In short, offensive language requires both aggressive elements as well as clear directional intent toward a target.

3.2 Criteria for Toxicity

Adapted from definition, two indicators are assessed by both human annotators and included in auto-annotation:

`Direction of Intent (DI)` indicates whether the language is directed internally (denoted 0) or externally (denoted 1). Since a tweet may contain multiple sentences with shifting targets, the annotated focus or intent could vary. Therefore, keeping

Level	Item	Category	Example
Lexical	Aggressive NP/DP[a]	*Aggressive Item*	Steretyped NP/DP (nigga, chingchong, etc), bitch, shit, dumbass, etc
Lexical	Aggressive VP[b]	*Aggressive Item*	fuck, hate, etc
Lexical	Aggressive AdjP[c]	*Aggressive Item*	retarded, psycho, stupid, etc
Lexical	Aggressive AdvP[d]	*Aggression Catalyzer*	fucking, etc
Syntactic	Strong Expression	*Aggression Catalyzer*	should, must, definitely, etc
Syntactic	Rhetorical Question	*Aggression Catalyzer*	Doesn't everyone feel the same? etc
Syntactic	Imperative	*Aggression Catalyzer*	Shut the door, etc
Discourse	Ironic Expression	*Aggression Catalyzer*	Clear as mud, etc
Discourse	False Construct	*Aggressive Item* or *Aggression Catalyzer*	Those are people who only believe in flat earth, etc
Discourse	Controversial Content	*Aggressive Item*	Inappropriate Content (adult, religious, etc), jeering at others' mistakes or misfortunes, etc

[a] NP stands for noun phrase, and DP for determiner phrase.
[b] VP stands for verb phrase.
[c] AdjP stands for adjective phrase.
[d] AdvP stands for adverbial phrase.

Table 1: Relative Aggression Computing Reference

such disagreement in annotations is necessary.

Aggression (AG) is annotated by categorizing negative, rude, or hostile attitudes as mild (0.1-1 point) or intense (>1) based on a reference table 1 of weighted linguistic characteristics such as slurs or vulgarities. The first thing to notice is that the classification of different types of aggression is not absolute or fixed. What constitutes a specific category of aggression could shift over time as cultural norms and language usage evolve. Additionally, it can sometimes be difficult to precisely categorize certain expressions of aggression due to variations in language, influences from popular culture, and other contextual factors. The following criteria only try to grasp a more objective overview of aggression, which does not rule out all subjectivity. In calculating the relative aggression score for each piece, we count each unique linguistic item only once. Putting values on categories assesses the functional diversity of different language components, providing a more precise evaluation of the aggression level. The cumulative aggression scores are computed from various distinct aggressive lexical items, syntactic structures, and discourse strategies. However, in certain instances, merely adding more terms from a single category can decrease the perceived aggression. This is because excessive repetition of similar aggressive language might come across as impotent rage, reducing the overall impact of the aggression expressed. The specific target(s) of each aggressive expression are also extracted as full noun phrases. The reference table provides a framework for categorizing and quantifying linguistic aggression across multiple levels of language. Four main levels are identified: lexical, syntactic, and discourse. Within each level, linguistic items are classified as aggressive items (AI) that independently convey aggression (1 point), or aggression catalyzers (ACs) which intensify aggression but are not inherently aggressive (0.5 points). AIs include slurs, vulgarities, and controversial content. ACs include emphatic language, rhetorical questions, imperatives, and ironic expressions. To compute an overall aggression score, AIs are weighted 1 point, and ACs 0.5 points. However, the false construct is a special case. A false construct is a systematic error or preexisting belief that leads to flawed evaluations or unfair treatment of individuals or groups. If it is paired with ACs, it becomes AIs worth 0.5 points, as they form an aggression base. This multi-layered approach allows for a nuanced analysis of how various linguistic devices work together to convey varying degrees of aggression. The table provides a few examples for each category.

3.3 Auto-annotation

Leveraging in-context learning is a promising approach to mitigate various learning biases while ensuring low-cost and highly generalizable processing. In-context learning is a paradigm where a language model learns a downstream task by being

Comparison	CK	AC1	Agreement (Agr.) %
Without Criteria			
1T & 2T	0.5172	0.5094	76.50
With Criteria			
1AG_C & 2AG_C	0.8422	0.8419	90.75
1DI_C & 2DI_C	0.5913	0.5908	91.50
1T_C & 2T_C	0.7487	0.7486	92.50

Table 2: Inter-Annotator Agreement for Annotations With and Without Guidelines

Comparison	CK	AC1	Agr. %
1T & Davidson et al., 2017	-0.0475	-0.2552	51.25
2T & Davidson et al., 2017	-0.0566	-0.1742	62.25
1T_C & Davidson et al., 2017	-0.0884	-0.1237	75.00
2T_C & Davidson et al., 2017	-0.0405	-0.0698	77.00

Table 3: Inter-annotator Reliability Evaluation on annotations with and without criteria and original annotation.

conditioned on restricted prompts, thereby enhancing flexibility (Hao et al., 2022). This learning method involves the model improving at a specific task after being provided with a selection of relevant examples or demonstrations (Lampinen et al., 2022; Margatina et al., 2023; Coda-Forno et al., 2023). The model uses the context from a single prompt or interaction to discern the expectations for that particular instance (Han et al., 2023). Similarly, few-shot learning enables large language models (LLMs) to rapidly adapt to tasks for which they were not explicitly trained (Gao et al., 2020; Perez et al., 2021; Mahabadi et al., 2022). By analyzing a limited set of examples, the model can deduce the desired output format and content for new tasks, contrasting with traditional machine learning methods that typically require extensive training data (Wertheimer and Hariharan, 2019).

This study utilizes GPT-3.5 and GPT-4, known for their proficiency and accessibility in in-context and few-shot learning. GPT-3.5's extensive architecture allows it to grasp and generate contextually relevant responses with limited input (Yang et al., 2021). GPT-4 further enhances this capability due to its even more extensive training and sophisticated design (OpenAI, 2023). We accessed both models via APIs to use small amounts of task-specific data to adapt to this task. Unlabeled data were processed with carefully constructed prompts to generate annotations consistent with pre-established formats. These prompts were designed for two components: direction of intent and level of aggression. The direction of intent prompt used general descriptive instructions, while the aggression level prompt combined descriptive instructions with few-shot examples sourced from 'AI' and 'AC' categories to demonstrate specific scenarios. Given the subjective nature of aggression, including some examples in the latter prompt was crucial for ensuring some uniformity in annotations. Additionally, the challenge of neurotoxic degeneration is tackled by employing a method similar to Instruction Augmentation (INST) Prabhumoye et al., 2023. We divided the aggression level prompt into two sections: one for language use assessment and another for aggression scoring. This division adheres to INST principles, enhancing the clarity and precision of instructional prompts, thereby improving the performance and dependability of language models in complex tasks.

4 Statistics Analysis on 400 Pieces

4.1 Inter-annotator Reliability and Agreement

For manual annotation and statistic analysis, the dataset was randomly extracted from the Offensive and Hate speech dataset (Davidson et al., 2017), comprising 400 tweets. It is characterized by dense occurrences of various categories of offensive language and includes instances of non-standard English, providing a comprehensive sample for analysis. Two separate annotation processes were conducted with and without predefined criteria. Two annotators with distinct backgrounds - one a marketing graduate student without linguistics training, the other a linguistics graduate student - were se-

GPT4	CK	AC1	Agr. %	GPT3.5	CK	AC1	Agr. %
Without Criteria							
1T	0.2030	0.0685	62.75	1T	0.3149	0.2532	**67.50**
2T	0.2819	0.2190	73.75	2T	0.3534	0.3331	**74.50**
With Criteria							
1DI_C	0.3376	0.3361	87.00	1DI_C	0.1999	0.1799	**87.75**
2DI_C	0.5647	0.5646	**92.25**	2DI_C	0.2820	0.2704	90.25
1AG_C	0.3460	0.3016	**62.5**	1AG_C	0.2813	0.2605	59.25
2AG_C	0.3849	0.3565	**66.5**	2AG_C	0.2700	0.2588	60.0
1T_C	0.5299	0.5282	**87.00**	1T_C	0.4013	0.3887	85.5
2T_C	0.6103	0.6094	**89.50**	2T_C	0.4015	0.3910	86.0

Table 4: Agreement percentages between GPT predictions and human annotations.

lected to illustrate how academic foundations can influence judgments. The marketing student had no formal linguistics knowledge, while the linguistics student possessed a comprehensive understanding of language. Both were asked to evaluate offensiveness, assuming an intuitive understanding of offensive language. In contrast, the annotators with criteria were linguistics graduate students trained on established guidelines. They first annotated intention direction and aggression level, then rated offensiveness based on those indicators. Annotation without criteria took under 5 hours; with criteria, over 10. The increased duration resulted from precisely evaluating relevant language per outlined criteria and calculating aggression scores, necessitating more detailed analysis.

Annotation distribution is displayed in Appendix B, and confusion matrices for annotator agreements are depicted in Appendix A. For a comprehensive evaluation of annotator consistency, we calculated Cohen's Kappa (CK) (McHugh, 2012) and Gwet's AC1 (AC1)(Cicchetti, 1976), as detailed in Table 2. Initially, we assessed the inter-annotator reliability for both our annotations without criteria and those from Davidson et al., 2017, displayed in Table 3. Gwet's AC1 can help avoid the paradoxical behavior and biased estimates associated with Cohen's Kappa, especially in situations of high agreement and prevalence (Zec et al., 2017).

According to Table 2[1], it is evident that incorporating specific criteria in the annotation process significantly enhances the consistency and agreement between raters. This conclusion is supported by the observed values in Cohen's Kappa and Gwet's AC1 metrics and the Agreement Percentages. Cohen's Kappa and Gwet's AC1 values that adjust for chance agreement indicate a moderate agreement without criteria. However, these values markedly increased when criteria were applied as the first and last pairs approached near-perfect agreement levels, underscoring the critical role of well-defined criteria in enhancing the reliability and validity of qualitative assessments. Unlike our annotations, the comparison with the original annotations presents contrasting results in Table 3. Cohen's Kappa and Gwet's AC1 values are negative across all comparisons, suggesting a level of disagreement more pronounced than random chance. This starkly contrasts the earlier findings where criteria application resulted in near-perfect agreement levels in certain pairs. Although the Agreement Percentages showed some level of surface agreement, they do not align with the deeper discordance indicated by the antagonistic Cohen's Kappa and Gwet's AC1 values. This discrepancy underscores the complexities in achieving inter-rater reliability and emphasizes the need for a thorough review of annotation guidelines and processes to understand and rectify the underlying causes of such significant misalignments.

4.2 Agreement between Human Annotations and GPT Annotations

As Cohen's Kappa and Gwet's AC1 were originally created to assess inter-rater reliability between human annotators, directly applying them to evaluate agreement between machine and human annotations may not be entirely apt (Popović and Belz,

[1] 1T - Toxicity, no guidelines, marketing student; 2T - Toxicity, no guidelines, linguistics student; 1AG_C - Aggression, with guidelines, Annotator 1; 2AG_C - Aggression, with guidelines, Annotator 2; 1DI_C - Intent direction, with guidelines, Annotator 1; 2DI_C - Intent direction, with guidelines, Annotator 2; 1T_C - Toxicity, with guidelines, Annotator 1; 2T_C - Toxicity, with guidelines, Annotator 2

Model (Fine-Tuning Data)	DI (Acc.)	AG (Acc.)	T (Acc.)
RoBERTa-base (Davidson et al., 2017)	-	-	0.937
DeBERTa-base (Davidson et al., 2017)	-	-	0.943
RoBERTa-base (G3P)	0.908	0.749	0.920
DeBERTa-base (G3P)	0.918	0.723	0.922
RoBERTa-base (G4P)	0.944	0.821	0.890
DeBERTa-base (G4P)	0.938	0.856	0.863

Table 5: Accuracy Metrices for BERT models Fine-tuned on Davidson et al., 2017 baseline and GPT-annotated Datasets

Model (Fine-Tuning Data)					**1T**	**2T**
RoBERTa-base (Davidson et al., 2017)					54.00	66.50
DeBERTa-base (Davidson et al., 2017)					50.70	62.75
	1DI_C	**2DI_C**	**1AG_C**	**2AG_C**	**1T_C**	**2T_C**
RoBERTa-base (Davidson et al., 2017)	-	-	-	-	81.25	82.25
DeBERTa-base (Davidson et al., 2017)	-	-	-	-	78.00	79.00
RoBERTa-base (G3P)	87.50	90.25	**61.00**	**62.50**	84.50	86.00
DeBERTa-base (G3P)	89.50	86.25	57.50	60.25	83.25	85.25
RoBERTa-base (G4P)	89.25	**91.00**	51.75	56.75	85.50	**86.50**
DeBERTa-base (G4P)	**89.75**	90.50	52.50	57.25	**85.75**	86.25

Table 6: Agreement (%) Performance of BERT models fine-tuned on Davidson et al., 2017 baseline and GPT-annotated data

2021). While primarily intended for only human judgment scenarios, we include evaluations using these metrics when comparing GPT model predictions and human labels since dedicated methods for assessing machine-human agreement have yet to be established. We analyzed concordance between human annotations and those generated by Generative Pre-trained Transformer models, namely GPT-4 (OpenAI, 2023) and GPT-3.5 (OpenAI, 2022), across two annotation categories.

The trinary evaluations in Table 4 demonstrate reasonable consistency and agreement between human annotations and those from GPT-3.5 and GPT-4. Without criteria, GPT-3.5 agreement was slightly higher than GPT-4. Refining the prompts enabled more effective synergy between automated analysis and human oversight. Using specific criteria significantly improved alignment with human judgment for both models. Under criteria-based scenarios, GPT-4 annotations showed comparable agreement and consistent inter-rater reliability. The inter-annotator reliability statistics show that GPT annotations have even higher agreement and consistency than the original human annotations. Overall, establishing criteria enhanced model concurrence with human annotators, with GPT-4 consistently demonstrating higher agreement and suggesting aptitude for criteria-based analysis. The notable improvement in agreement when using explicit criteria motivates fine-tuning smaller models with these guided GPT annotations. Our next exploration will assess whether annotations from prompted GPTs enhance performance beyond unrefined prompts. We will use GPTs with meticulous prompts to automatically annotate text, then train and evaluate other models on these datasets. By comparing agreement for models with and without criteria-based fine-tuning, we can evaluate this approach's efficacy.

5 Experiment on Fine-tuning Small Models

Two baselines were fine-tuned on RoBERTa-base (Liu et al., 2019) and DeBERTa-base (He et al., 2021) with 2,4384 pieces tweets from Hate Speech and Offensive Language dataset (Davidson et al., 2017), excluding 400 pieces used in manual annotation. Experiment data consists of 295 Reddit posts in AAL, 341 tweets from OLID (Zampieri et al., 2019), 311 tweets from the offensive and hate speech dataset (Davidson et al., 2017), and 1000 tweets from Hateval (Basile et al., 2019), 1942 pieces in total for GPT auto-annoations. Data samples are randomly selected. This approach mitigates biases from linguistic patterns in any dataset. Including diverse social media (e.g., Reddit, Twitter) facilitates robust exposure to vernacular language and dialects, which also increases the challenge of matching human annotations on the 400-piece compared to baselines fine-tuned on the same dataset, excluding the 400 pieces. RoBERTa-base and DeBERTA-base were fine-tuned using a batch size of 8 for training and 16 for evaluation with the default learning rate. Models were trained for 3 epochs with 10% of data reserved for testing.

5.1 Result Analysis and Discussion

As shown in Table 5, when fine-tuned on different datasets, DeBERTa-base slightly outperforms RoBERTa-base on the Hate speech and Offensive language dataset, but RoBERTa-base achieves higher accuracies in specific categories like Language Intent and Aggression when trained on GPT-annotated datasets (G3P[2] and G4P[3]).

Table 6 shows that fine-tuned models align well with human annotations in identifying language intent but struggle with aggression categorization. When fine-tuned on a baseline dataset, BERT models moderately agree with human toxicity annotations (78-79%), similar to the 76.5% agreement rate without criteria. Notably, criteria-based auto-annotations improve model performance, with higher agreement rates (85.75%, 86.50%) using the G4P dataset. DeBERTa-base consistently outperforms RoBERTa-base, indicating better complex language understanding. This analysis emphasizes the importance of high-quality annotations and the benefits of GPT-based annotations for language model training. Despite improvements, fine-tuned BERT models still lag behind human annotators (92.50%) and GPT-4 (85.75%, 86.50%) in agreement rates, possibly due to small dataset sizes. The performance of models fine-tuned with G3P and G4P are similar. In comparison with baselines, these results indicate that GPT-annotated training data better aligns models with human judgment and shows stability across language variances and genres. Further research into context-specific tuning and criteria design is needed for detailed analysis and improved data annotation.

[2]Annotated data by GPT-3.5 with prompt

[3]Annotated data by GPT-4 with prompt

6 Conclusion

This work provides insights to advance the understanding of offensive language detection and analysis. Initially, we emphasize the importance of defining explicit criteria for constructing datasets on toxicity and offensiveness. This methodology effectively manages subjectivity, thereby reducing the risks of over-generalization and personal bias in dataset compilation. Secondly, our findings reveal the enhanced efficacy of large language models, specifically GPT-3.5 and GPT-4, when employing in-context learning supplemented with few-shot examples. We observed a substantial improvement in the agreement rates between GPT-generated assessments and human evaluations when explicit criteria were utilized. This underscores the significance of criterion-based instruction in enhancing model accuracy. Finally, we investigated the potential benefits of fine-tuning smaller models, RoBERTa-base and DeBERTa-base, with datasets auto-annotated by GPTs under explicit criteria. This strategy resulted in higher agreement rates compared to models trained on datasets without such criteria, demonstrating the effectiveness of integrating advanced LLMs with criterion-guided auto-annotation. These findings hold substantial importance for improving toxic content moderation systems, thereby contributing towards fostering a more responsible and respectful digital communication environment.

Limitations

We identified some limitations that are important for guiding future research. The scope of human annotation within our dataset could be expanded. First of all, we conduct human annotation on a dense toxic corpus; if the corpus switches to a more controversial one, the agreement would to expected to be lower. So, the human agreement in this research is only a reference, not a solid upper bound. Although we relied on a significant amount of human input, the complexities and nuances of offensive language suggest that a broader and more diverse set of human annotations could enhance the model's understanding and accuracy. Another limitation lies in the size of our auto-annotated dataset, which comprises less than 2,000 entries. While this dataset has been critical for training and evaluating our models, its relatively limited size may not fully capture the extensive range of linguistic variations in offensive language. Expanding the dataset could offer a more comprehensive perspective, potentially leading to more accurate and generalizable outcomes. Additionally, there is room for improvement in the performance of smaller models on the auto-annotated dataset, even though it surpasses that of GPT-4 with criteria. Exploring different configurations, experimenting with various model architectures, and further tuning could enhance performance.

References

Keith Allan. 2015. When is a slur not a slur? the use of nigger in 'pulp fiction'. *Language Sciences*, 52:187–199.

David Archard. 2008. Disgust, offensiveness and the law. *Journal of Applied Philosophy*, 25(4):314–321.

David Archard. 2014. Insults, free speech and offensiveness. *Journal of Applied Philosophy*, 31(2):127–141.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa-Anke, and Leonardo Neves. 2020. TweetEval:Unified Benchmark and Comparative Evaluation for Tweet Classification. In *Proceedings of Findings of EMNLP*.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 54–63.

Lawrence Buell. 1998. Toxic discourse. *Critical Inquiry*, 24:639 – 665.

Alec Burmania, Srinivas Parthasarathy, and Carlos Busso. 2015. Increasing the reliability of crowdsourcing evaluations using online quality assessment. *IEEE Transactions on Affective Computing*, 7(4):374–388.

Agostina Calabrese, Michele Bevilacqua, Björn Ross, Rocco Tripodi, and Roberto Navigli. 2021. Aaa: Fair evaluation for abuse detection systems wanted. In *Proceedings of the 13th ACM Web Science Conference 2021*, pages 243–252.

Yunliang Chen and Jungseock Joo. 2021. Understanding and mitigating annotation bias in facial expression recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 14980–14991.

Domenic V Cicchetti. 1976. Assessing inter-rater reliability for rating scales: resolving some basic issues. *The British Journal of Psychiatry*, 129(5):452–456.

Julian Coda-Forno, Marcel Binz, Zeynep Akata, Matthew Botvinick, Jane X. Wang, and Eric Schulz. 2023. Meta-in-context learning in large language models.

D. Citron and Helen L. Norton. 2011. Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *Proceedings of the international AAAI conference on web and social media*, volume 11, pages 512–515.

Ona de Gibert, Naiara Perez, Aitor García-Pablos, and Montse Cuadros. 2018. Hate Speech Dataset from a White Supremacy Forum. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 11–20, Brussels, Belgium. Association for Computational Linguistics.

Nicholas Deas, Jessi Grieser, Shana Kleiner, Desmond Patton, Elsbeth Turcan, and Kathleen McKeown. 2023. Evaluation of african american language bias in natural language generation.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Mai Elsherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth M. Belding-Royer. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *International Conference on Web and Social Media*.

Lisa Fan, Marshall White, Eva Sharma, Ruisi Su, Prafulla Kumar Choubey, Ruihong Huang, and Lu Wang. 2019. In plain sight: Media bias through the lens of factual reporting. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. 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.

Tianyu Gao, Adam Fisch, and Danqi Chen. 2020. Making pre-trained language models better few-shot learners. *arXiv preprint arXiv:2012.15723*.

A Garcia et al. 2003. Nigger: The strange career of a troublsome word. *Society*, 40(5):93.

Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. 2023. Handling bias in toxic speech detection: A survey.

Luke Haliburton, Sinksar Ghebremedhin, Robin Welsch, Albrecht Schmidt, and Sven Mayer. 2023. Investigating labeler bias in face annotation for machine learning. *arXiv preprint arXiv:2301.09902*.

Xiaochuang Han, Daniel Simig, Todor Mihaylov, Yulia Tsvetkov, Asli Celikyilmaz, and Tianlu Wang. 2023. Understanding in-context learning via supportive pretraining data. *arXiv preprint arXiv:2306.15091*.

Yaru Hao, Yutao Sun, Li Dong, Zhixiong Han, Yuxian Gu, and Furu Wei. 2022. Structured prompting: Scaling in-context learning to 1, 000 examples. *ArXiv*, abs/2212.06713.

Patricia H. Hawley and Brian E. Vaughn. 2003. Aggression and adaptive functioning: The bright side to bad behavior. *Merrill-Palmer Quarterly*, 49:239 – 242.

Hassan Hayat, Carles Ventura, and Agata Lapedriza. 2022. Modeling Subjective Affect Annotations with Multi-Task Learning. *Sensors*, 22(14):5245.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2021. Deberta: Decoding-enhanced bert with disentangled attention. In *International Conference on Learning Representations*.

Kenneth Joseph, Lisa Friedland, William Hobbs, Oren Tsur, and David Lazer. 2017. Constance: Modeling annotation contexts to improve stance classification. *arXiv preprint arXiv:1708.06309*.

Alejandro Kacelnik and Sasha Norris. 1998. Primacy of organising effects of testosterone. *Behavioral and Brain Sciences*, 21:365 – 365.

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.

Animesh Koratana and Kevin Hu. 2019. Toxic speech detection.

Andrew K Lampinen, Ishita Dasgupta, Stephanie CY Chan, Kory Matthewson, Michael Henry Tessler, Antonia Creswell, James L McClelland, Jane X Wang, and Felix Hill. 2022. Can language models learn from explanations in context? *arXiv preprint arXiv:2204.02329*.

Marina Chiara Legroski. 2018. Offensiveness scale: how offensive is this expression? *Estudos Linguísticos (São Paulo. 1978)*, 47(1):169–180.

Elisa Leonardelli, Stefano Menini, Alessio Palmero Aprosio, Marco Guerini, and Sara Tonelli. 2021. Agreeing to disagree: Annotating offensive language datasets with annotators' disagreement. *arXiv preprint arXiv:2109.13563*.

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. *CoRR*, abs/1907.11692.

Rabeeh Karimi Mahabadi, Luke Zettlemoyer, James Henderson, Marzieh Saeidi, Lambert Mathias, Veselin Stoyanov, and Majid Yazdani. 2022. Perfect: Prompt-free and efficient few-shot learning with language models. *arXiv preprint arXiv:2204.01172*.

Katerina Margatina, Timo Schick, Nikolaos Aletras, and Jane Dwivedi-Yu. 2023. Active learning principles for in-context learning with large language models. *arXiv preprint arXiv:2305.14264*.

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.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Ioannis Mollas, Zoe Chrysopoulou, Stamatis Karlos, and Grigorios Tsoumakas. 2020. Ethos: an online hate speech detection dataset. *arXiv preprint arXiv:2006.08328*.

Mainack Mondal, Leandro Araújo Silva, and Fabrício Benevenuto. 2017. A Measurement Study of Hate Speech in Social Media. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*. ACM.

Jihyung Moon, Won Ik Cho, and Junbum Lee. 2020. Beep! korean corpus of online news comments for toxic speech detection. *ArXiv*, abs/2005.12503.

OpenAI. 2022. Gpt-3.5: Language models are few-shot learners. `https://openai.com/blog/gpt-3-5-update/`. Accessed: [Insert current date here].

OpenAI. 2023. Gpt-4 technical report.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. *Advances in neural information processing systems*, 34:11054–11070.

Maja Popović and Anya Belz. 2021. A reproduction study of an annotation-based human evaluation of mt outputs. Association for Computational Linguistics (ACL).

Shrimai Prabhumoye, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2023. Adding instructions during pretraining: Effective way of controlling toxicity in language models. In *Conference of the European Chapter of the Association for Computational Linguistics*.

Jacquelyn Rahman. 2012. The n word: Its history and use in the african american community. *Journal of English Linguistics*, 40(2):137–171.

Dennis Reidsma and Rieks op den Akker. 2008. Exploiting 'subjective' annotations. In *Proceedings of the Workshop on Human Judgements in Computational Linguistics - HumanJudge '08*. Association for Computational Linguistics.

Geneva Smitherman. Black talk: Words and phrases from the hood to the amen corner (paperback)-common.

Allen W Stokes and Lois M Cox. 1970. Aggressive man and aggressive beast. *BioScience*, 20(20):1092–1095.

Davis Wertheimer and Bharath Hariharan. 2019. Few-shot learning with localization in realistic settings. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6558–6567.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2021. An empirical study of gpt-3 for few-shot knowledge-based vqa. *ArXiv*, abs/2109.05014.

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. *arXiv preprint arXiv:1902.09666*.

Slavica Zec, Nicola Soriani, Rosanna Comoretto, and Ileana Baldi. 2017. Suppl-1, m5: high agreement and high prevalence: the paradox of cohen's kappa. *The open nursing journal*, 11:211.

A Appendix

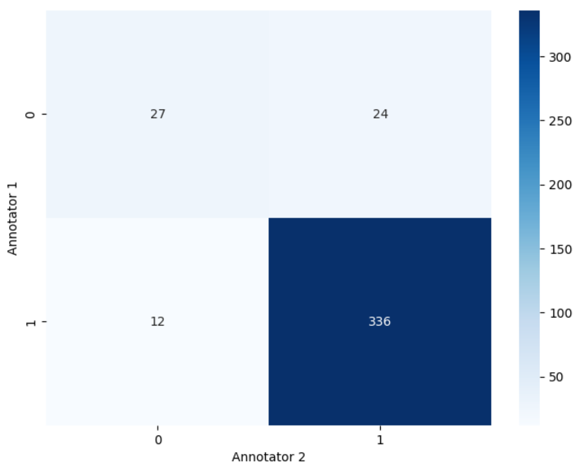

Figure 2: Confusion Matrix on Direction Intent Annotation

B Appendix

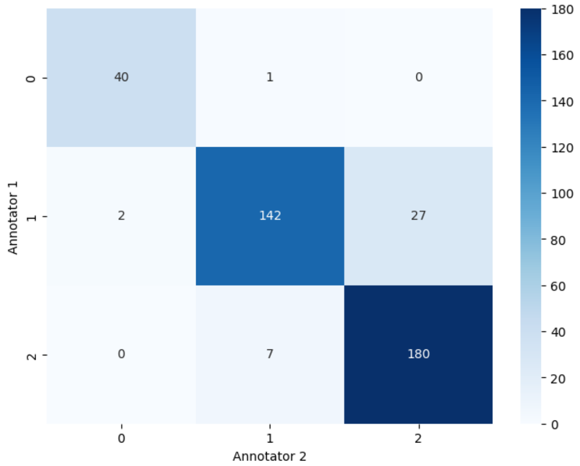

Figure 3: Confusion Matrix on Aggression Annotation

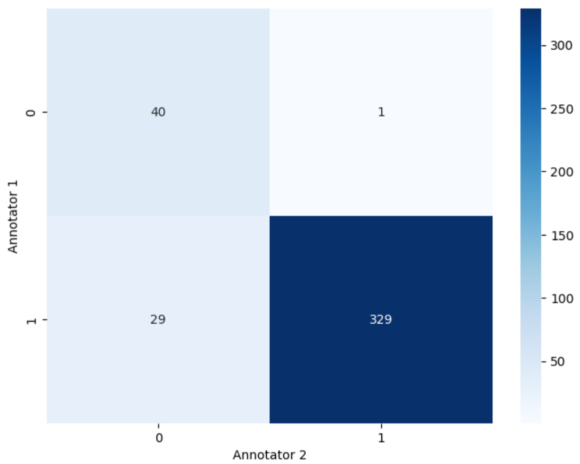

Figure 4: Confusion Matrix on Toxicity Annotation with Criteria

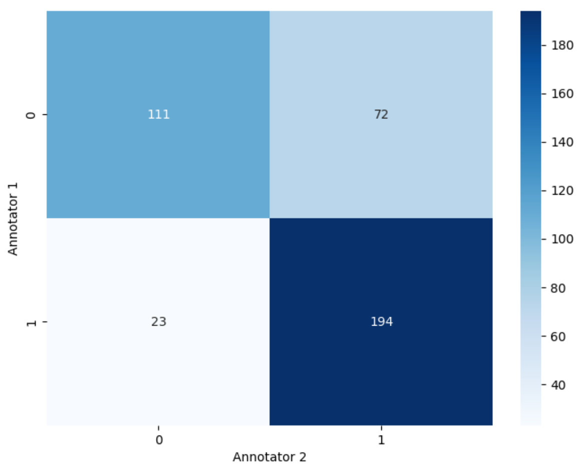

Figure 5: Confusion Matrix on Toxicity Annotation without Criteria

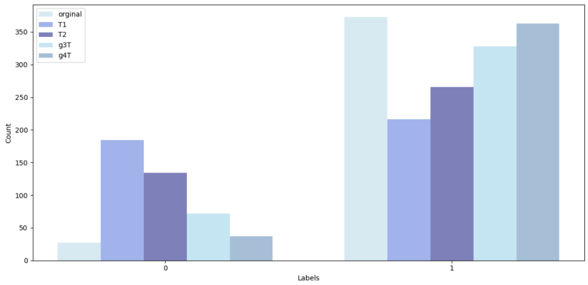

Figure 6: Distribution of Toxicity Annotation without Criteria

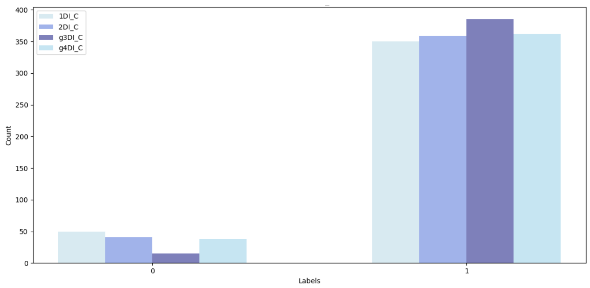

Figure 7: Distribution of Direction of Language Intent Annotation with Criteria

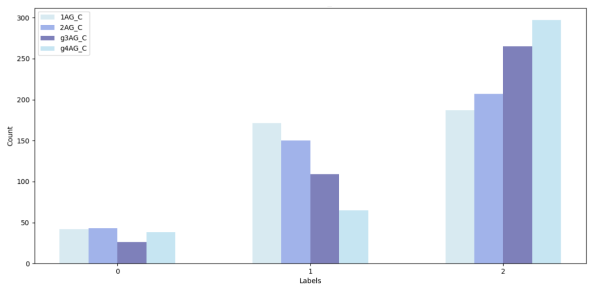

Figure 8: Distribution of Aggressive Level Annotation with Criteria

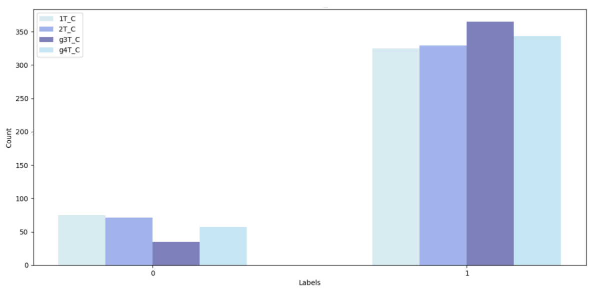

Figure 9: Distribution of Toxicity Annotation with Criteria