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 language 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 substantial increase in inter-annotator agreement indicates the effectiveness of structured guidelines at controlling subjective variability and strengthening result consistency. 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 automatic 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 in-context learning with limited data size and heterogeneous language types.

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

1 Introduction

In the digital age, the anonymity of the Internet and the lack of direct interaction have led to increased offensive language (Mondal et al., 2017). In order to properly offer people the option to avoid potentially offensive language while also protecting minoritized language varieties from being misidentified, accurate detection that can identify languages despite changes over time is required. 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 underlying intent (Zampieri et al., 2019; Basile et al., 2019; Mollas et al., 2020). Abusive, toxic, or offensive language and hate speech were often directly identified based on finite lists of phrases (Davidson et al., 2017), annotators' interpretation of the textual content (de Gibert et al., 2018; Founta et al., 2018; Sap et al., 2019), or a combination of both (Vargas et al., 2021; Basile et al., 2019). This brings up the first issue of an unclear research subject, described as inconsistency in terminology and categorization (Fortuna et al., 2020). The terminology used in this issue is complex, with substantial overlap between related concepts like toxic language, offensive language, and hate speech. However, these concepts are not completely interchangeable. Additionally, inconsistencies in terminology alone do not fully address underlying annotation and learning biases.

Biases in annotation refer to the systematic tendency of human annotators that leads to errors or skewed labels in the training data used for machine learning models (Davani et al., 2023). The most common approach for mitigating annotator bias is diversifying annotation teams and increasing annotation on each raw piece (Davani et al., 2023; Sap et al., 2019; Geva et al., 2019). However, questions remained regarding how diverse the annotator team should be and how many annotators were required to eliminate bias efficiently. While diversification and scale help address bias, the root issue often lies in subtle differences in interpretations addressing complex socio-cultural dynamics that are especially vulnerable (Al Kuwatly et al., 2020; Kuwatly et al., 2020). Therefore, rather than treating annotator disagreement as mere "noise" or using majority vote labels that cover up disagreement, inevitable disagreements should be modeled rather than directly discarded in the aggregation process (Davani et al., 2023, 2021) Classifying language as simply "toxic" or "non-toxic" risks introducing biases against minority groups (Sap et al., 2019), which can perpetuate biases in model learning. Specifically, a key issue is lexical bias caused by high-frequency terms in the training data (Tan and Celis, 2019). While adding more diverse data could reduce this frequency bias, doing so without care could undermine meaningful information from actual high-frequency usage. Furthermore, more data may simply be unavailable, especially for endangered or minority languages (Liu et al., 2022). If humans can identify less common non-toxic uses of potentially toxic terms, yet models cannot, it suggests key features of toxicity may not be overtly demonstrated. Particularly in low-resource situations, the core issue is how to effectively represent the underlying linguistic features that indicate toxicity.

As depicted in Figure 1, our research comprises three components: criteria proposal, small-scale statistical analysis, and experiments. To evaluate the quality of human annotations based on new criteria, we analyzed inter-rater reliability among three sets of annotations: 1) original annotations following general definitions and finite word list, 2) non-criteria-based annotations on the 400-piece set that follows the descriptive data annotation paradigm, and 3) annotations on the 400-piece set based on the new criteria following the prescriptive data annotation paradigm. To simulate scenarios with limited human resources, we use LLMs as stand-ins for professional, well-trained human annotators to label the 400-piece descriptively and prescriptively. Our experiments aim to demonstrate the performance of smaller models fine-tuned on the prescriptive annotations on the 1942-piece set to simulate restricted data resources - both small in size and containing a hybrid mix of language types and genres. Finally, we compare against models fine-tuned on original annotations.

The major contributions and findings are:

1. This research proposes detailed annotation criteria to enable consistent offensive language data labeling, particularly when using LLMs as annotators with limited human resources.

2. This research contributes two newly annotated offensive language detection datasets created based on the proposed standardized annotation criteria.

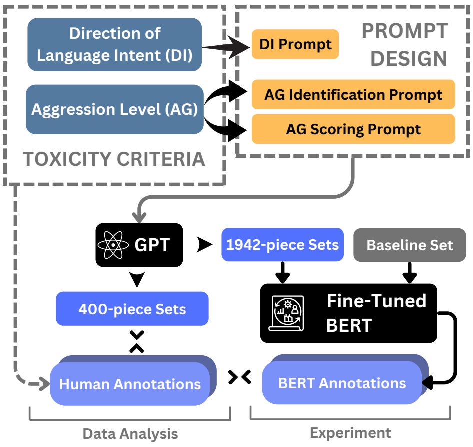

Figure 1: Research Framework: This research establishes standardized criteria for toxic language annotation and analyzes inter-annotator reliability. Experiments on BERT models across language types tend to demonstrate the broader applicability of the proposed annotation criteria, even with limited resources.

3. Proposed criteria yield higher inter-annotator agreement for prescriptive annotations and improved quality for criteria-guided LLM annotations over originals.

4. Smaller models fine-tuned on a criteria-guided LLM multi-source dataset outperform models trained on a single far larger original annotated dataset.

2 Related Works

The issue of non-offensive language being 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 example, both (1) and (2) were identified as offensive:

> (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&#128514;" I'm obsessed with mine &#128529 (Davidson et al., 2017)

In (1), F**K is used as emotional emphasis. Similarly, slang does not always induce toxicity as presented in (2); race-related term n***a is a neutral word often found in African American English (AAE) and gender-related b***h. The three terms are not definitely appropriate on all occasions, but whether they actually mean harm to others depends on their perlocutionary effect, considering the context and circumstances of their usage and reception (Allan, 2015; Rahman, 2012). Contextual swearing and Minority language pose the major challenges to simplistic judgments that rely solely on the presence of a list of phrasal units and general definitions (Pamungkas et al., 2023; Deas et al., 2023). Simple reminders of exceptions and rare cases that should be taken into annotating consideration will not work out, as unrestricted context interpretation based on individual assumptions will inevitably introduce biases (Rast, 2009). Making educative annotation decisions regarding context needs to follow prior defined instruction (Giunchiglia et al., 2017; Röttger et al., 2021). The descriptive data annotation paradigm has the advantage of embracing subjectivity to gain insights into diverse viewpoints but faces challenges in eliciting, representing, and modeling those viewpoints effectively (Röttger et al., 2021; Alexeeva et al., 2023). Whereas the descriptive data annotation paradigm embraces subjectivity to provide insights into diverse perspectives, effectively eliciting, capturing, and modeling those varied viewpoints poses significant challenges (Röttger et al., 2021; Ruggeri et al., 2023).

Regarding relevant concepts, some previous studies have equated toxic speech with hate speech when examining different facets of this language use (Koratana and Hu, 2019; Moon et al., 2020), which are confusing and less productive for data annotation (Fortuna et al., 2020). Toxic language refers to harm-inflicting expressions (Buell, 1998). Since it generally lacks an inherent link to directed anger(Gelber, 2019), toxic language is often used as a semantically broader, more neutral term substituting for offensive language(Radfar et al., 2020; Baheti et al., 2021). However, hate speech tends to be more emotional and directly aggressive towards targets (Gelber, 2019; Elsherief et al., 2018), constituting one manifestation of toxic language rather than being equivalent to it (Fortuna et al., 2020). Treating toxicity and hatred separately avoids potential confusion arising from treating them as interchangeable concepts.

As the focus of this study, **Offensiveness** and **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). A related term, "aggressiveness," that is usually viewed as harboring hostile intentions has positive connotations in sociological and psychological studies (Hawley and Vaughn, 2003). While aggressiveness is fundamental to dominating behavior (Kacelnik and Norris, 1998), such behavior differs from outward toxicity that adversely impacts others. 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). Aggressive components may contribute to offensive speech, but only when coupled with explicit intents to cause harm or distress to a target. **In short, toxic offensive language is the language that shows explicit aggression towards others.** Carefully analyzing the language used toward others, decomposing it into basic elements, can help annotators focus on fundamental components that make speech truly offensive or toxic, rather than mislabeling those emotional but nonharmful minority language usages. This approach aims to direct human judgment in annotation onto the most relevant textual features to avoid biases and improve agreement by not erroneously marking provocative but ultimately inoffensive speech as inappropriate.

3 Methodology

Adapted from the definition, the direction of language intent (DI) and the presence of aggression (AG) are two components that need to be assessed to determine toxicity. To support future studies on relevant issues such as hate speech, aggression was evaluated regarding how intense it is. DI has two labels: 1 for explicitly targeting other people and 0 for other cases. AG has three labels: 0 for non-aggressive, 1 for mildly aggressive, and 2 for intensely aggressive. A piece of data will be categorized as **toxic or offensive if and only if it is labeled as 1 for Directed Insults (DI) and also labeled as either 1 or 2 for Aggression (AG).**

3.1 Annotation Criteria

Direction of Intent (DI) indicates whether the language is directed externally (label 1) or not (label 0). DI utilizes a binary labeling system to cate-

Level	Item	Category	Example
Lexical	Aggressive NP/DP[a]	*Aggressive Item*	Stereotyped 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 Scoring Reference: Assigns numerical values for aggressive speech: 1 point for Aggressive Items (overtly toxic statements) and 0.5 points for Aggression Catalyzers (toxicity booster). The false construct will be an exception.

gorize whether statements explicitly target other individuals. Text segments receive a label of 1 if they directly refer to or address a specific person or group using second-person pronouns, proper nouns, or clear contextual references that signal an interpersonal attack or criticism. Alternatively, text segments receive a label of 0 if the statements implicate others more implicitly, as is commonly the case with ironic expressions, or focus primarily on the speaker themselves rather than targeting external subjects directly. This simplified dichotomization aims to delineate clear instances of directive aggressive speech from more ambiguous cases as a first step. Since a tweet may contain multiple sentences with shifting targets, the annotated focus or intent could vary. Therefore, keeping such disagreement in annotations is necessary.

Aggression (AG) is annotated by categorizing negative, rude, or hostile attitudes into three levels: non-aggression (label 0) assigned an aggression score of 0, mild aggression (label 1) assigned an aggression score of 1, and intense aggression (label 2) assigned an aggression score interval $(1, \infty)$. Table 1 provides a relative score reference for categorizing and quantifying linguistic aggression across multiple language levels. Three main levels are identified: lexical, syntactic, and discourse. Within each level, linguistic items are classified as aggressive items (AI) that independently convey aggression or aggression catalyzers (ACs), which intensify aggression but are not inherently aggressive. To compute an overall aggression score, AIs are weighted 1 point, and ACs 0.5 point. AIs include slurs, vulgarities, and inflammatory content ACs include emphatic language, rhetorical questions, imperatives, and ironic expressions. However, the false construct is a special case that appears as a systematic error or preexisting belief that leads to flawed evaluations or unfair treatment of individuals or groups. If a false construct is paired with AC(s), it becomes an AI to form an aggression base but is still worth 0.5 points. In calculating the relative aggression score for each piece, we count each unique linguistic item only once. For example, if only two biased noun phrases (AI) are used in one piece, the score should be one, not two. Table 1 provides a few examples for each item. The pieces that receive 0 will be non-toxic. The pieces with scores equal to 1 will be mild aggression, and pieces with scores larger than 1 will be intense aggression.

Please note that the aggressive expression classifications are not fixed. What constitutes a specific category of aggression could shift over time as cultural norms and language use 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 intend to rule out all subjectivity. Putting values on categories assesses the functional diversity of different language components, providing a more precise evaluation of the aggression level. 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.

3.2 Human Annotation

Two separate annotation processes were conducted, one with predefined criteria and one without. For the non-criteria-based human annotation, two annotators were given the question prompt, "Is the tweet toxic or offensive? If toxic or offensive, label 1; if it is not, label 0." allow unrestricted subjectivity , following the descriptive data annotation paradigm. To examine the reliability of the original annotation, two annotators with academic backgrounds were chosen to resemble the diverse and unspecified backgrounds of CrowdFlower(CF) workers who were randomly employed and coded for Davidson et al., 2017. The first annotator was a graduate marketing student familiar with internet culture but with no formal linguistic knowledge. The second was a graduate linguistics student with sufficient linguistic knowledge and socio-linguistic practices. Choosing annotators this way allowed evaluation of the reliability between the original and the descriptive data annotation under similar annotation conditions. The annotation with criteria was conducted by two linguistics graduate students who were trained with prescriptive instructions as presented in Appendix A . Please find more information about annotators and more details about the annotation process in Appendix B.

3.3 LLM Annotation

Leveraging in-context learning is a promising approach to mitigate various learning biases while ensuring low-cost and highly generalizable processing (Lampinen et al., 2022; Margatina et al., 2023; Coda-Forno et al., 2023). Few-shot learning enables language models to rapidly adapt to new downstream tasks by analyzing a small set of relevant examples or interactions to discern expected outputs without extensive retraining (Gao et al., 2020; Perez et al., 2021; Mahabadi et al., 2022).

This study uses GPT-3.5-turbo and GPT-4 to generate prototypical responses with proposed criteria prompts. 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 instruction to adapt to this task. Unlabeled data were processed with carefully constructed prompts to generate annotations consistent with pre-established formats. For descriptive LLM annotation, the question prompt used for human annotation was directly entered. For criteria-based LLM annotation, prompts were designed separately for the direction of intent, aggression recognition, and aggression scoring. The direction of intent prompt used general prescriptive instructions, while the aggression level prompt combined prescriptive 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 assessing language use and another for aggression scoring. This division adheres to INST principles, enhancing the clarity and precision of instructional prompts for saving effects in cleaning the outcomes.

4 Data Analysis

We randomly collected 400 tweets from the Offensive and Hate Speech dataset of the Davidson 2017 dataset (Davidson et al., 2017). This dataset contains a high frequency of various types of offensive language and non-mainstream English. We chose this dataset because its dense toxic content and casual language use make it relatively straightforward for both human annotators and language models to process. The prevalence of clear toxic content reduces potential confusion and ambiguity that could skew the analysis.

4.1 Inter-annotator Reliability and Agreement

Confusion matrices for all annotations are listed in Appendix C, and the distributions are displayed in Appendix D. For a comprehensive evaluation

Pair	CK	AC1	Agr.%
Without Criteria			
1T & 2T	0.5172	0.5094	76.50
With and Without Criteria			
1T & 1T_C	0.3000	0.2406	66.75
2T & 1T_C	0.3889	0.3718	75.75
1T & 2T_C	0.2883	0.2229	66.25
2T & 2T_C	0.3966	0.3769	76.25
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: 1T - descriptive toxicity, marketing student; 2T - descriptive toxicity, linguistics student; 1AG_C - prescriptive aggression, Annotator 1; 2AG_C - prescriptive aggression, Annotator 2; 1DI_C - prescriptive intent direction, Annotator 1; 2DI_C - prescriptive intent direction, Annotator 2; 1T_C - prescriptive toxicity, Annotator 1; 2T_C - prescriptive toxicity, Annotator 2

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, incorporating specific criteria in the annotation process significantly enhances consistency and agreement between raters. This conclusion is supported by the larger positive values of trinary metrics for with-criteria pairs compared to without-criteria pairs and with-without-criteria pairs. Cohen's Kappa and Gwet's AC1 values, which adjust for chance agreement, indicate only moderate agreement without criteria. However, these values markedly increased when criteria were applied, as the first and last pairs approached near-perfect agreement levels. This underscores the critical role of well-defined criteria in enhancing reliability and validity of qualitative assessments. Interestingly, the reliability evaluations for with-without-criteria pairs are even lower than without-criteria pairs, suggesting the annotation logics for the two annotation types are completely different.

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 also indicates underlying distinctions in how the annotations were carried out, and the fact that the majority vote labels they used for the final label were not from the same annotator could be a reason why reliability tests exhibit so much difference. These statistics starkly contrast the earlier findings where criteria application resulted in a near-perfect agreement for certain pairs. Although the agreement percentages showed some surface agreement, they do not align with the deeper discordance indicated by the negative Cohen's Kappa and Gwet's AC1 values. As a result, prescriptive data annotations (1T_C, 2T_C) show higher reliability compared to descriptive data annotations (1T, 2T). Prescriptive data annotation paradigms are more appropriate for this task. This discrepancy highlights the complexities in achieving inter-rater reliability and the need to thoroughly review annotation guidelines and processes to understand and rectify the 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, 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 GPT 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 prompted criteria, GPT-3.5 slightly outperforms GPT-4 in both agreement and reliability, but refining the prompts enabled more effective and reliable synergy between automated toxicity analysis and human-like interpretation. Using the proposed criteria significantly improved the alignment with human judgment for both models, especially for GPT-4 annotations. Inter-rater reliability Under criteria-based scenarios, GPT-4 annotations

Pair	**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.

Pair	**CK**	**AC1**	**Agr. %**	**Pair**	**CK**	**AC1**	**Agr. %**
Without Criteria							
1T & G4T	0.2030	0.0685	62.75	1T & G3T	0.3149	0.2532	**67.50**
2T & G4T	0.2819	0.2190	73.75	2T & G3T	0.3534	0.3331	**74.50**
With Criteria							
1DI_C & G4DI_C	0.3376	0.3361	87.00	1DI_C & G3DI_C	0.1999	0.1799	**87.75**
2DI_C & G4DI_C	0.5647	0.5646	**92.25**	2DI_C & G3DI_C	0.2820	0.2704	90.25
1AG_C & G4AG_C	0.3460	0.3016	**62.5**	1AG_C & G3AG_C	0.2813	0.2605	59.25
2AG_C & G4AG_C	0.3849	0.3565	**66.5**	2AG_C & G3AG_C	0.2700	0.2588	60.0
1T_C & G4T_C	0.5299	0.5282	**87.00**	1T_C & G3T_C	0.4013	0.3887	85.5
2T_C & G4T_C	0.6103	0.6094	**89.50**	2T_C & G3T_C	0.4015	0.3910	86.0

Table 4: Agreement percentages between GPT predictions and human annotations: G4T - descriptive toxicity, GPT-4; G3T - descriptive toxicity, GPT-3.5-turbo; G4DI_C - prescriptive intent direction, GPT-4; G4AG_C - prescriptive aggression, GPT-4; G4T_C - prescriptive toxicity, GPT-4; G3DI_C - prescriptive intent direction, GPT-3-turbo; G3AG_C - prescriptive aggression, GPT-3.5-turbo; G3T_C - prescriptive toxicity, GPT-3.5-turbo

showed comparable agreement and consistent inter-rater reliability. The reliability statistics show that GPT annotations have even higher agreement and consistency than the original human annotations and without-criteria human annotations following the descriptive paradigm. The established criteria improved accuracy. Additionally, GPT-4 outperformed GPT-3.5 on this task. This suggests an aptitude for criteria-based analysis. After implementing the proposed criteria, these notable improvements demonstrate that prescriptive data annotation instructions can help researchers overcome the lack of human annotator resources.

5 Experiments

Two baselines were fine-tuned: RoBERTa-base (Liu et al., 2019) and DeBERTa-base (He et al., 2021). 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 (Vaswani et al., 2023). Models were trained for 3 epochs with 10% of data reserved for testing. Baseline models are fine-tuned on 2,4384 pieces of tweets from the Davidson 2017 dataset (Davidson et al., 2017), excluding 400 pieces used in statistic analysis. 1942-piece dataset consists of 295 Reddit posts in African American English (Deas et al., 2023), 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) for prescriptive LLM annotations. Mixing different datasets will mitigate extrusive language features, and diverse social media (e.g., Reddit, Twitter) facilitates robust exposure to diverse language and dialects. According to previous studies and empirical observation which suggest larger datasets, particularly those with language types similar to the target application, tend to lead to higher performance in language models (Sahlgren and Lenci, 2016; Linjordet and Balog, 2019; Kaplan et al., 2020), the Davidson 2017 dataset would likely enable superior performance compared to the smaller, more complex 1942-piece dataset due to its size and domain relevance advantages.

5.1 Result Analysis and Discussion

As shown in Table 5, when fine-tuned on different datasets, DeBERTa-base slightly outperforms RoBERTa-base on baseline dataset, but RoBERTa-base achieves higher accuracy in prescriptive Aggression and prescriptive toxicity when trained on GPT-annotated datasets (G3P[1] and G4P[2]).

Agreements in Table 6 indicate that fine-tuned models align well with human annotations in identifying language intent but struggle more with aggression classifications. When fine-tuned on a baseline dataset, BERT models moderately agree with human toxicity annotations (78-82.25%), which is close to the 76.5% agreement rate between human

[1] 1942-piece set annotated by GPT-3.5-turbo prescriptively

[2] 1942-piece set annotated by GPT-4 prescriptively

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

annotators 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. This result contradicts the previous hypothesis that a baseline dataset with a much larger size and more uniform language patterns would help small models outperform GPT annotations; instead, it strongly suggests the advantages of prescriptively annotating data for language toxicity task.

6 Conclusion

This work provides important insights to advance the understanding of offensive language for analysis and detection. To better evaluate toxicity, we independently assessed language intent and aggression level. This bifurcated approach addresses the core of toxicity while allowing more controlled annotation to mitigate biases. This prescriptive approach effectively manages unwanted interpretations, thereby reducing risks associated with certain lexicons and personal bias in dataset compilation. Secondly, our findings reveal the enhanced efficacy of LLMs when employing in-context learning supplemented with few-shot examples. We observed substantial improvement in agreement rates between GPT-generated and human annotations when explicit criteria were used. Not only did human annotations in this study demonstrate higher reliability and agreement, but the prescriptive LLM annotations also showed higher reliability and agreement than the original annotation. Finally, we investigated the potential benefits of the proposed annotation paradigm in assisting BERT models to adapt with limited data size and complex language patterns. Even under restricted conditions, automatically generated annotations following the proposed instruction enabled BERT models to outperform baselines. These findings hold importance for maximizing data utilization efficiency and preparing toxic content moderation systems to adapt to language patterns with limited resources. This contributes towards fostering a more responsible and respectful digital communication environment.

Limitations

We identified some limitations that are important for guiding future research. While prescriptive annotation paradigms may better identify uniform patterns, they risk overlooking meaningful interpretations not yet identified by linguists and social scientists. The proposed criteria account for variations in English, but their practical application relies heavily on annotators' language knowledge. The dynamic nature of internet language poses additional challenges for human coders to accurately comprehend tweets, as no annotators can fully grasp the breadth of English online language, let alone code-switching usages by multilingual users. On the other hand, annotators lacking contextual understanding of in-group language may erroneously analyze utterances meant to promote within-community comprehensibility, a limitation challenging to resolve through improved annotation design. In contrast, LLMs demonstrate an advantage in aggregating insights from considerably larger data sources. Therefore, determining approaches for incorporating LLMs in detection alongside human rationale remains an important direction for further research.

Furthermore, the scope of human annotation within our dataset could be expanded. Human annotation of a dense toxicity corpus reveals high agreement; however, corpora containing more implicit cultural-related expressions would likely yield lower agreement. 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. Additionally, there is room for improvement in the performance of smaller models on the automatically generated dataset. Open-source LLMs could be possible substitutes. Exploring different configurations, experimenting with various model architectures, and further tuning could enhance performance.

References

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Proceedings of the fourth workshop on online abuse and harms*, pages 184–190.

Maria Alexeeva, Caroline Hyland, Keith Alcock, Allegra Argent Beal Cohen, Hubert Kanyamahanga, Isaac Kobby Anni, and Mihai Surdeanu. 2023. Annotating and training for population subjective views. In *Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*.

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.

Ashutosh Baheti, Maarten Sap, Alan Ritter, and Mark O. Riedl. 2021. Just say no: Analyzing the stance of neural dialogue generation in offensive contexts. *ArXiv*, abs/2108.11830.

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.

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.

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.

Aida Mostafazadeh Davani, M. C. D'iaz, and Vinodkumar Prabhakaran. 2021. Dealing with disagreements: Looking beyond the majority vote in subjective annotations. *Transactions of the Association for Computational Linguistics*, 10:92–110.

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.

Antigoni Founta, Constantinos Djouvas, Despoina Chatzakou, Ilias Leontiadis, Jeremy Blackburn, Gianluca Stringhini, Athena Vakali, Michael Sirivianos, and Nicolas Kourtellis. 2018. Large scale crowdsourcing and characterization of twitter abusive behavior. In *Proceedings of the international AAAI conference on web and social media*, volume 12.

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

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

Katharine Gelber. 2019. Terrorist-extremist speech and hate speech: Understanding the similarities and differences. *Ethical Theory and Moral Practice*, pages 1–16.

Mor Geva, Yoav Goldberg, and Jonathan Berant. 2019. Are we modeling the task or the annotator? an investigation of annotator bias in natural language understanding datasets. *ArXiv*, abs/1908.07898.

Fausto Giunchiglia, Enrico Bignotti, and Mattia Zeni. 2017. Personal context modelling and annotation. *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 117–122.

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

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

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

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models.

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

Hala Al Kuwatly, Maximilian Wich, and Georg Groh. 2020. Identifying and measuring annotator bias based on annotators' demographic characteristics. In *Workshop on Abusive Language Online*.

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.

Trond Linjordet and Krisztian Balog. 2019. Impact of training dataset size on neural answer selection models. In *Advances in Information Retrieval: 41st European Conference on IR Research, ECIR 2019, Cologne, Germany, April 14–18, 2019, Proceedings, Part I 41*, pages 828–835. Springer.

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.

Zoey Liu, Crystal Richardson, Richard J. Hatcher, and Emily Prudhommeaux. 2022. Not always about you: Prioritizing community needs when developing endangered language technology. In *Annual Meeting of the Association for Computational Linguistics*.

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*.

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.

Endang Wahyu Pamungkas, Valerio Basile, and Viviana Patti. 2023. Investigating the role of swear words in abusive language detection tasks. *Language Resources and Evaluation*, 57(1):155–188.

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*.

Bahar Radfar, K. Shivaram, and Aron Culotta. 2020. Characterizing variation in toxic language by social context. In *International Conference on Web and Social Media*.

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

Erich H. Rast. 2009. Context and interpretation.

Paul Röttger, Bertie Vidgen, Dirk Hovy, and Janet B Pierrehumbert. 2021. Two contrasting data annotation paradigms for subjective nlp tasks. *arXiv preprint arXiv:2112.07475*.

Federico Ruggeri, Francesco Antici, Andrea Galassi, Katerina Korre, Arianna Muti, and Alberto Barrón-Cedeño. 2023. On the definition of prescriptive annotation guidelines for language-agnostic subjectivity detection. In *Text2Story@ECIR*.

Magnus Sahlgren and Alessandro Lenci. 2016. The effects of data size and frequency range on distributional semantic models. *arXiv preprint arXiv:1609.08293*.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2019. Social bias frames: Reasoning about social and power implications of language. *arXiv preprint arXiv:1911.03891*.

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

Yi Chern Tan and Elisa Celis. 2019. Assessing social and intersectional biases in contextualized word representations. *ArXiv*, abs/1911.01485.

Francielle Vargas, Fabiana Rodrigues de Góes, Isabelle Carvalho, Fabrício Benevenuto, and Thiago Alexandre Salgueiro Pardo. 2021. Contextual-lexicon approach for abusive language detection. *arXiv preprint arXiv:2104.12265*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

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 Annotator Codebook

A.1 General Definitions

A list of short-cut definitions is presented in Table 7. Please see the methodology for further validations.

A.2 Annotation Instruction for two Indicators

Aggression will be assessed regarding every distinct negative, rude, or hostile attitude. Please see Table 1 and general description below for more information about specific language use. Computation logic: If the score is less or equal to 1, the aggression level will be 1. If the score exceeds 1, the aggression level will be 2. Otherwise, the aggression level will be 0.

Term	Definition
Aggression/Aggressiveness	Aggression in this context indicates hostile or rude attitudes, whether it involves readiness or not.
Aggressive	Being aggressive means showing hostile or rude attitudes, whether it involves readiness or not.
Offensiveness	General rudeness in a way that causes somebody to feel upset or annoyed because it shows a lack of respect.
Offensive	Being rude in a way that causes somebody to feel upset or annoyed because it shows a lack of respect.
External	Towards other people or parties.
Internal	Towards the self.
Construct	The mind-dependent object, namely ideas, perspectives, etc.
Inappropriate Language	Language uses that could have negative and unwanted impacts on people.
Biased Language	Biased Language contains obviously wrong or counterfactual expressions that target an individual or a group not limited to humans.
Offensive Language	Offensive Language shows intended aggressiveness toward others.
Hate Speech	Hate Speech is an offensive language of intense external aggressive intention with explicit targets rooted in explicit or implicit false construct.

Table 7: Definitions of Terms

- Level refers to the general linguistic category of each item.
- Item name includes the names of aggression-related items.
- Category refers to the category that indicates how the item is related to aggression.
  - Aggressive items / AI (1 point): are aggressive by themselves.
  - Aggression catalyzers / AC (.5 point): are unaggressive themselves and function to boost the aggressive level.
  - Expressions from the same item category only count once; for example, if there are two different aggressive noun phrases, the score will be one rather than two.
  - Override Rule: The overall relative aggression score will be 0 if there is no aggressive item.
  - SPECIAL CASE: False constructs are non-aggressive. But when people pair false constructs with other aggressive catalyzers, they become aggressive items (but with .5 point) and should be seen as aggression bases. For example, how come your people really believe in flat earth?
- Example contains examples of each item.

Direction of Language Intent (External or Non-external) evaluates Whether the language targets other(s) explicitly. The direction is decided regarding the direction of aggression, which means even statements about speakers' selves could contain aggression against others.

B Extra Information about Human Annotation based on Surveys

Specialties

- Annotator 1 without criteria: Internet Marketing & Data Analytics
- Annotator 2 without criteria: Corpus Linguistics & Syntax
- Annotator 1 with criteria: Sematics Analysis & Syntax & Corpus Linguistics
- Annotator 2 with criteria: Socio-linguistics & Language Acquisition

Aside from mainstream English, are you familiar with any regional dialects, sociolects, or linguistic styles more common in minority communities and groups?

- Annotator 1 without criteria: Yes
- Annotator 2 without criteria: Yes
- Annotator 1 with criteria: Yes
- Annotator 2 with criteria: Yes

Approximately how many hours did it take you to complete all the annotations assigned to you?

- Annotator 1 without criteria: 4
- Annotator 2 without criteria: 4.5
- Annotator 1 with criteria: 5 (criteria-based training) + 7 (annotation)
- Annotator 2 with criteria: 5 (criteria-based training) + 8 (annotation)

How confident are you in the accuracy of the annotations you completed? (1-5)

- Annotator 1 without criteria: 2. No so confident, many African American English I found hard to understand accurately
- Annotator 2 without criteria: 3. I am confident about my annotations identifying explicit toxic expressions and hate speech, but less confident in others.
- Annotator 1 with criteria: 4.5. I'm pretty confident, though I'm not an African American English native speaker. I studied AAE corpus before, so I consider myself familiar with AAE. About that DI, sometimes I think it could go either way cause their tweets ain't just one sentence. For AG, the score generally matches what I think about aggression. All in all, this dataset is easier than the one with political stuff. I don't know too much about politics.
- Annotator 2 with criteria: 4. Yes, I think AAE is not really an issue. The AG scoring guide helps break things down to the word level. Basically, it doesn't really matter if the phrases are used differently or not; as long as they are seen as aggressive by some people, they'll be taken as aggressive. But it really takes a lot of time and effort just to highlight each aggressive item and categorize the aggression. DI seemed pretty straightforward to me at first, but after our group discussion, I realized there could also be other interpretations.

Looking back at your annotations after a month has passed, how did you feel about the quality and accuracy of the work you originally completed?

- Annotator 1 without criteria: Still confused about many tweets.
- Annotator 2 without criteria: There could be different interpretations. It's really about the larger context.
- Annotator 1 with criteria: Not really much in terms of toxicity. DI's still kinda confusing in a couple of cases.
- Annotator 2 with criteria: Basically the same as when I finished it up

C Confusion Matrices

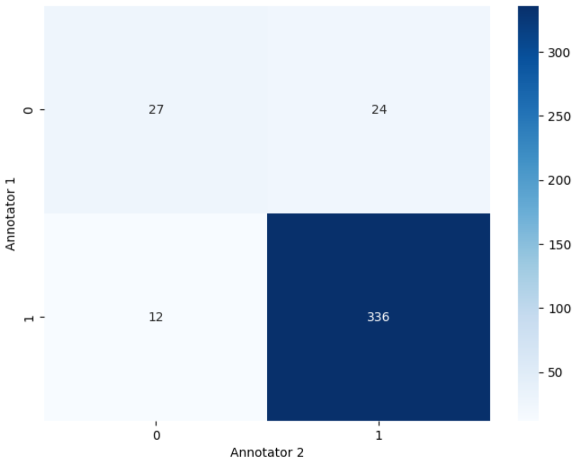

Figure 2: Confusion Matrix on Direction Intent Annotation

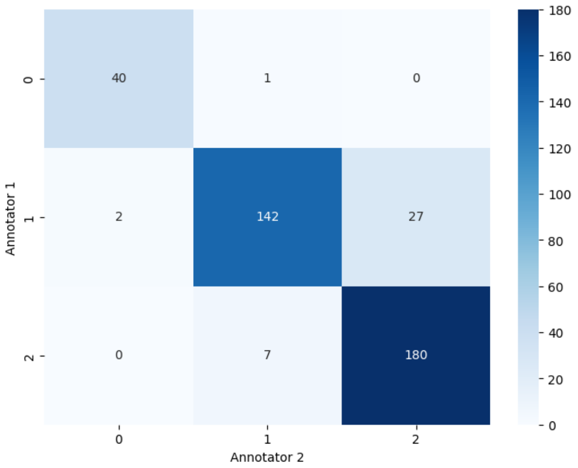

Figure 3: Confusion Matrix on Aggression Annotation

D Annotation Distribution

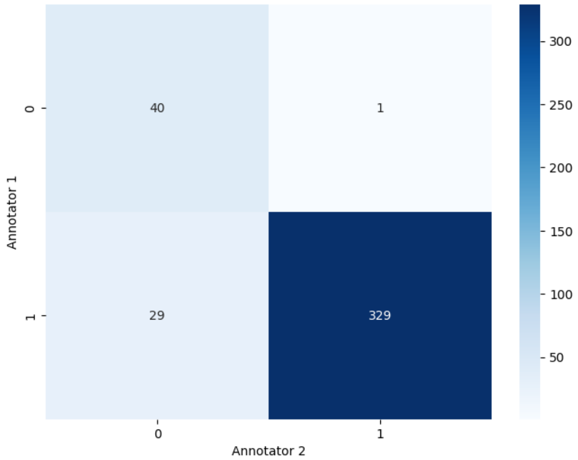

Figure 4: Confusion Matrix on Toxicity Annotation with Criteria

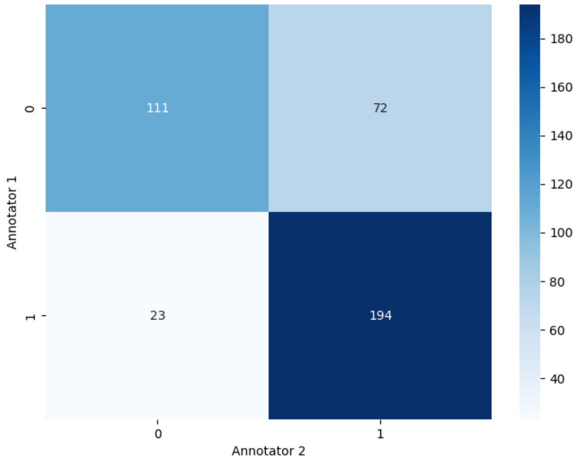

Figure 5: Confusion Matrix on Toxicity Annotation without Criteria

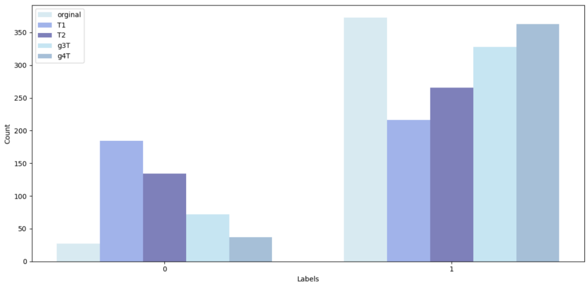

Figure 6: Distribution of Toxicity Annotation without Criteria

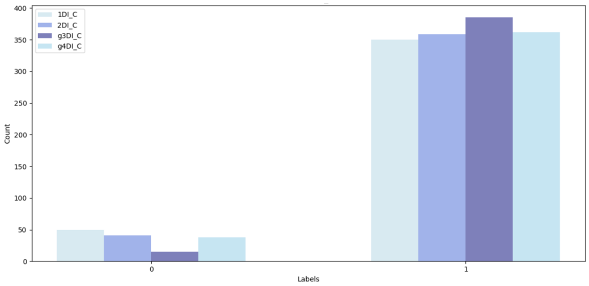

Figure 7: Distribution of Direction of Language Intent Annotation with Criteria

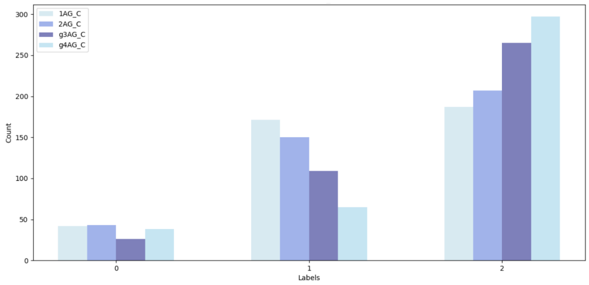

Figure 8: Distribution of Aggressive Level Annotation with Criteria

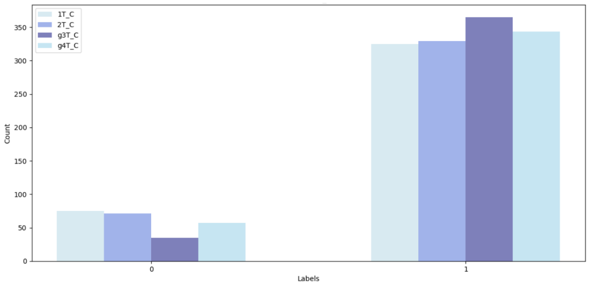

Figure 9: Distribution of Toxicity Annotation with Criteria