Contextualizing Toxicity: An Annotation Framework for Unveiling Pragmatics in Conversations of Online Discussion Forums

Yingxue Fu, Anaïs Ollagnier
Universite Cote d'Azur, CNRS, Inria, I3S
Sophia Antipolis, France
fuyingxue321@gmail.com

Abstract

The role of context has attracted increasing attention in research on toxicity detection. Interpreting toxic language remains a complex and multifaceted challenge, shaped by numerous linguistic, contextual, and social factors. However, current approaches often define "context" narrowly, focusing primarily on surface lexical cues such as hate lexicons, profanity markers, or sentiment polarity. These features, while useful, are insufficient to capture the interactional dynamics, user behaviors, and intentionality that shape such phenomena. To address this gap, this paper introduces a novel and systematic annotation framework, grounded in Speech Act Theory (Austin, 1962), aimed at deciphering the illocutionary and perlocutionary dimensions of conversation, which are unexplored in existing studies. We apply this framework to a new dataset of complete Reddit conversation threads, sampled to include discussions that turn toxic (124 conversations, 1990 messages). We evaluate the performance of GPT models (GPT-3, GPT-4, and GPT-5) on this challenging annotation task, providing insights into how large language models capture pragmatic and contextual dimensions of online toxicity.

Keywords: pragmatics, toxicity, annotation, Reddit conversations

1. Introduction

Online discussion platforms such as Reddit enable people to communicate beyond geographical and temporal boundaries. While these forums provide valuable spaces for practical advice and community support, they can also be misused. Toxic or hateful messages can spread rapidly, and their consequences may escalate beyond control—driving participants to leave a discussion or, in more severe cases, causing psychological or even physical harm to individuals or groups. Effective moderation is therefore essential to maintaining a healthy and supportive online environment. Given the diversity of topics and the vast number of messages, automated systems are increasingly used to assist human moderators (Haythornthwaite, 2023).

Since 2016, a growing number of resources and benchmark corpora have been developed to support research on toxicity detection. Most studies have focused on identifying toxic or harmful content—including offensive, abusive, or hateful speech—primarily collected from social media platforms such as Twitter, Facebook, Gab, and Reddit (Fortuna et al., 2020). Several surveys have provided structured overviews of this growing field by cataloging and characterizing available datasets (Poletto et al., 2020; Alkomah and Ma, 2022; Yu et al., 2024).

Despite this progress, a recurrent limitation concerns the definition and conceptualization of the task, which vary considerably across studies. These definitions often overlap with one another (Waseem et al., 2017), and the boundaries remain ill-defined (Hada et al., 2021), leading to inconsistent annotation guidelines and complicating efforts to build large-scale, comparable language resources and develop robust automatic systems.

A further issue relates to how toxicity is annotated. In most datasets, messages are labeled in isolation from their surrounding context (Pavlopoulos et al., 2020), disregarding the conversational flow in which they appear. As pointed out by Vidgen et al. (2021), this is a well-established shortcoming of abusive content datasets, where content preceding or following a toxic message is often ignored. Building on this observation, recent studies have begun to incorporate thread-level information when assigning labels of offensiveness (Vidgen et al., 2021; Hada et al., 2021). However, contextual features are still generally treated as supplementary rather than central to annotation, and only a few frameworks explicitly encode such information—for instance, Ollagnier (2024), where intentions and roles are annotated to capture the pragmatic dimensions of conversational toxicity (Ma et al., 2025).

Recent studies have increasingly explored the pragmatic dimensions of offensive language, focusing on the implied assumptions and biases underlying toxic discourse (Sap et al., 2020; Zhou et al., 2023; Muti et al., 2024; ElSherief et al., 2021). These studies frame toxicity as both a linguistic and social behavior, where interpretation depends on pragmatic inferences. However, existing datasets capture only a limited scope of pragmatic information (Ma et al., 2025), typically restricted to stereotypes or biases linked to protected characteristics (Meta, 2025). While suitable for hate

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 3961–3974
11–16 May 2026.

speech targeting social groups, such frameworks overlook conversational toxicity, where harmful intent emerges through interactional dynamics rather than explicit bias.

In this study, we try to abstract away the myriad forms of offensive language, following Pavlopoulos et al. (2020) and Wulczyn et al. (2017), and use "toxicity" as umbrella terms to denote "a rude, disrespectful, or unreasonable comment that is likely to make you leave a discussion". We propose a generalizable annotation framework that captures the pragmatic dimensions of toxic conversations beyond the literal text. Our scheme annotates pragmatic inferences, their illocutionary speech act types (e.g., directives, expressives), and whether they stem from an implicature or a presupposition. It also models perlocutionary effects by analyzing replies to assess whether the speaker's intended meaning is understood. We apply this framework to a new dataset of Reddit conversations that turn toxic, comprising 31 manually annotated and 93 GPT-4/5–annotated threads. We further evaluate human and model reliability (GPT-3, GPT-4, GPT-5), demonstrating the framework's scalability and applicability. The dataset is publicly released[1].

2. Related Work

Toxicity detection has been widely studied through diverse taxonomies, classifying messages as general profanity or targeted insults (Zampieri et al., 2019), focusing on specific targets (Muti et al., 2024; Sanguinetti et al., 2018), or analyzing the intentions behind aggressive language (Sanguinetti et al., 2018; Ollagnier, 2024; Vidgen et al., 2021). However, across these approaches, several studies have questioned the practice of annotating toxic messages in isolation and have emphasized the importance of incorporating contextual information to better capture the interactional and pragmatic nature of toxicity (Schmidt and Wiegand, 2017; Zampieri et al., 2019; Pavlopoulos et al., 2020; Wang et al., 2025; Zhou et al., 2023).

Yet, in most existing work, context is narrowly defined—typically limited to a parent–reply message pair (Pavlopoulos et al., 2020) or the title of a news article (Gao and Huang, 2017). Even when considered, contextual cues are often treated as supplementary features rather than integral to annotation. Only a few studies (Qian et al., 2019; Ollagnier et al., 2025; Zhang et al., 2018; Sap et al., 2020; ElSherief et al., 2021; Zhou et al., 2023) explicitly encode contextual or pragmatic information—for instance, intentions and situational context in Ollagnier (2024). Table 1 summarizes context-aware datasets for toxicity detection, showing the types of contextual information used to annotate offensiveness and the degree to which pragmatic dimensions are represented.

To address these limitations, it is crucial to draw on insights from pragmatics, the study of language use in context. Pragmatic phenomena—such as implicature, presupposition, speech acts, and intention recognition—help explain how meaning is conveyed and interpreted beyond literal content (Ma et al., 2025). Recent research has leveraged these concepts to improve interpretability in language tasks. For instance, Kim et al. (2023) and Kim et al. (2021) show how modeling presuppositions and implicatures enhances question answering, while Srikanth et al. (2024) demonstrate that false assumptions are often encoded as implicatures. Beyond QA, Corvi et al. (2025) apply Speech Act Theory (Austin, 1962) to conceptualize representational harms—such as stereotyping and demeaning—as perlocutionary effects of illocutionary acts.

Similarly, in toxicity detection, integrating pragmatic information (e.g., underlying biases, intentions, and social cues) can improve the ability of automatic systems to detect implicit or subtle toxicity (Ocampo et al., 2023a). Moreover, pragmatic modeling offers a lens for identifying potential harms (Sap et al., 2020) and promoting transparency and accountability in content moderation (Yang et al., 2023).

3. Theoretical Framework

According to the Stanford Encyclopedia of Philosophy[2], pragmatics can be divided into *near-side* and *far-side* pragmatics. The former concerns the determination of what is said (e.g., deictic expressions and ambiguity resolution), while the latter examines what is done beyond saying—namely, the *speech acts* performed and the *implicatures* conveyed.

Toxic utterances often operate through intentions and implied meanings (Wang et al., 2025). Consequently, our framework centers on **speech acts**, to model the communicative functions of toxic messages, and on **implicatures**, to capture their implicit, context-dependent meanings.

3.1. Three Dimensions of A Speech Act

A speech act can be analyzed through a three-dimensional framework (Austin, 1962): the *locutionary act*, corresponding to the production of an utterance according to linguistic conventions; the *illocutionary act*, representing the performance of an action *in* saying something (e.g., requesting, questioning, warning); and the *perlocutionary act*,

[1] https://github.com/yingxueF/pragmaticsAggressionRedditConversations/

[2] https://plato.stanford.edu/entries/pragmatics/ , accessed on Oct 2, 2025

Work	Source	Size	Context	Pragmatics	Topic
Gao and Huang (2017)	Fox News user comments	1528 comments from 10 complete discussion threads	all the comments in the same thread, the news article the comment is written for, and user name	no	hate speech
Pavlopoulos et al. (2020)	Wikipedia Talk Pages	250 comments (duly labeled w/o context)+10k comments	parent post in the thread and discussion title	no	toxicity
Contextual Abuse Dataset (Vidgen et al., 2021)	Reddit	25k comments	the conversation thread that a message is part of (Every annotation has a label for whether contextual information was needed to make the annotation.)	rationales for judgment: For each entry, the part of the text which contains the abuse is highlighted.	abusive language
Ruddit (Hada et al., 2021)	Reddit	6k comments	Reddit thread that a comment is part of	no	offensive language
Qian et al. (2019)	Reddit and Gab	5k (Reddit) + 12k (Gab)	conversations + title and post(Reddit)	intervention responses	hate speech
CyberAgressionAdo-Large (Ollagnier et al., 2025)	French dataset of aggressive conversations collected in role-playing scenarios	5,789 messages	full conversations	role and intention	cyberbullying
Wikipedia Abusive Conversations (Cécillon et al., 2020)	Wikipedia talk pages	193k conversations and 383k messages annotated as being abusive or not	reconstructed conversations	no	abusive language
WhatsApp Dataset (Sprugnoli et al., 2018)	Italian chats	14,600 tokens divided in 10 chats	full conversations	role, cyberbullying type, presence of sarcasm, and whether the expression containing insults is really offensive	cyberbullying
Zhang et al. (2018)	Wikipedia talk pages	1,270 paired awry-turning and on-track conversations	reconstructed conversations	politeness strategies	antisocial behavior
Social Bias Inference Corpus (Sap et al., 2020)	Reddit, Twitter and hate sites including Gab, Stormfront and banned Reddits	44,671 posts	no	social biases, stereotypes and whether the intention is to offend	offensive language
ElSherief et al. (2021)	Twitter	4,153 implicit hate (19,112 total)	no	target demographic group and its associated implied statement	hate speech
Zhou et al. (2023)	offensive statements from Toxigen (Hartvigsen et al., 2022)	around 33k offensive statements paired with machine-generated contexts and free-text explanations	social and situational context for an offensive statement	situation, identities of speaker and listener, communicative intent behind a statement, target group, sociocultural power differential, implied biases or stereotypes, listener's emotional and cognitive reactions, and offensiveness	offensive language

Table 1: Context-aware datasets for detecting or countering toxicity.

referring to the effect the utterance has on the listener's thoughts, feelings, or actions. For example, in *Could you close the window?*, the locutionary act is producing a grammatical English sentence, the illocutionary act is making a request, and the perlocutionary act is the listener responding by closing the window—or choosing not to.

The locutionary act has been extensively studied in toxicity research, primarily through analyses of syntactic patterns and lexical indicators of abusive language (Bassignana et al., 2018; Stamou et al., 2022). However, much less attention has been given to the illocutionary and perlocutionary dimensions, which are essential for understanding how toxicity emerges and evolves through interaction. Focusing on these dimensions allows us to move beyond surface linguistic features toward capturing the communicative functions and social effects of utterances within conversation threads. This framework thus enables the annotation of fine-grained pragmatic information that reflects the dynamic and contextual nature of online exchanges. Importantly, it also accounts for non-aggressive utterances, which, as shown by Ollagnier (2024), can contribute to the escalation or mitigation of toxic interactions. Yet, such messages are rarely considered in existing datasets. To address this gap, our framework integrates pragmatic information across all messages within a conversation, providing a more comprehensive view of the interactional dynamics of toxicity.

3.2. Five Types of Illocutionary Acts

A notable extension of Austin's framework is the taxonomy of illocutionary acts proposed by Searle (1975). Table 2 summarizes the five types: *representatives*, which commit the speaker to the truth of a proposition; *expressives*, which convey psychological or emotional states; *directives*, which aim to prompt the hearer to act; *commissives*, which bind the speaker to a future action; and *declarations*, which bring about a new state of affairs through the act of saying itself, provided the speaker has the authority. Each type can be identified through typical performative verbs and, together, they form an exclusive and exhaustive classification (Birner, 2025).

Type	Explanation
representatives	The purpose is to commit the speaker to the truth of the expressed proposition. Typical verbs include *think*, *believe*, *assert*, *claim*, *conclude*, and *deny*.
expressives	The illocutionary point is to express the psychological state specified in the sincerity condition about a state of affairs specified in the propositional content. Typical verbs include *thank*, *congratulate*, *apologize*, *condole*, *deplore* and *welcome*.
directives	This illocutionary act is an attempt by the speaker to get the hearer to do something. Typical verbs include *order*, *command*, *request*, *ask*, *question*, *beg*, *plead*, *advise*, and *permit*.
commissives	This illocutionary act commits the speaker to some future course of action. Typical verbs include *promise* and *vow*.
declarations	The illocutionary act denotes cases where one brings a state of affairs into existence by declaring it to exist, such as *I resign*, *You're fired*, *I appoint you chairman*, and *I declare the opening of the ceremony*.

Table 2: Five types of illocutionary acts.

Previous research has shown the close relationship between emotion and toxic language (Schäfer and Kistner, 2023; Mohammad, 2018). Since expressives are likely to contain emotional expressions, they might be closely associated with toxicity. Moreover, utterances that focus on the truth of a proposition might trigger aggressive behavior when the interlocutors disagree with each other. Directives could signal a command or suggestion from a speaker, and commissives could also lead to or be triggered by disagreements or threats. Therefore, the types of speech acts represent a potentially useful dimension to study toxicity.

3.3. Presuppositions and Implicatures

Grice (1975) distinguishes between literal and non-literal aspects of meaning, and develops a framework for analyzing the non-literal aspects. The assumption is that rational conversation participants follow some general cooperative principles to make communication efficient: *Make your conversational contribution such as is required, at the stage at which it occurs, by the accepted purpose or direction of the talk exchange in which you are engaged* (Grice, 1975, p.45). Four maxims are proposed, and if they are followed, it will yield results consistent with the cooperative principle. We do not elaborate on the maxims here since they are used as supplementary materials for annotators to determine whether an utterance involves pragmatic inferences. We leave it to future work to investigate toxicity through the lens of these maxims.

Following Srikanth et al. (2024), pragmatic inferences refer to assumptions derivable from a message, including presuppositions and implicatures. Implicatures have been studied extensively by Grice (1975), focusing on the non-literal aspects of utterances. Presuppositions are implicit assumptions of utterances that interlocutors take for granted (Kim et al., 2021). These implicit assumptions must be true in order for a message to make sense. For instance, *The King of France is bald* presumes that there is a king in France. As such, presuppositions are stronger assumptions than implicatures, for they require that both interlocutors have this belief/knowledge/emotional reaction. Presuppositions are generally triggered by lexicons or some syntactic structures, such as definite articles and possessives (Kim et al., 2021). However, Srikanth et al. (2024) show that domain or world knowledge is often needed to infer presuppositions in real world data.

As claimed by Srikanth et al. (2024), presuppositions are based on mutual acknowledgment of facts: when one makes a presupposition, the presumed content is believed to be shared by the interlocutor. In contrast, implicatures are a softer way to express uncertainty. Therefore, disagreements over presuppositions may prompt clarification from the speaker side or doubt about the speaker, whereas implicatures might lead to disputes over the implied content.

3.4. Capturing Perlocutionary Act

An important dimension of a speech act is its effect on the interlocutor. Typically, among several possible pragmatic inferences, one is more saliently or more readily derived because it offers greater relevance (Wilson and Sperber, 2004). On social media platforms, interlocutors often come from diverse educational backgrounds and hold differing beliefs, making it not uncommon for a speaker's utterance to be interpreted differently from its intended meaning. The hearer may focus on a less salient inference or construe the most salient inference in an unexpected way. Such violations of expectation can prompt clarification of the original message or trigger aggressive responses. Therefore, this dimension is incorporated into our annotation scheme.

3.5. Taxonomy

Table 3 shows the taxonomy. Pragmatic inferences are expressed in free text, and the type of illocutionary act for each inference is categorical. More than one inference is possible, since pragmatic inferences are defined to include presuppositions and implicatures, and as shown by Srikanth et al. (2024), multiple inferences are common in real world data. Among these, the most salient inference is selected. If the reply message indicates agreement with this inference, the label "yes" is used to represent the relationship under the category "as intended". Otherwise, the label "no" is applied. This design models the dynamics of a conversation thread, similar to the study by Procter et al. (2019), where replies to tweets are annotated based on whether they agree with the original message. The row "PRE/IMP" denotes whether the **most salient inference** belongs to a presupposition or an implicature. When no pragmatic inferences are identified, the annotator can use "literal" as the inference label and only needs to annotate the aggressiveness information.

Tag	Explanation	Categories
inferences	content of pragmatic inferences	free-text
	illocutionary act types	representatives, expressives, directives, commissives, declarations
most salient inference	the most salient inference chosen from row 1	free text
as intended	whether the reply message agrees with the most salient inference (perlocutionary act)	yes/no
PRE/IMP	whether a message belongs to presuppositions/implicatures	presupposition (PRE), implicature (IMP)
aggressive	toxicity of the message	yes/no

Table 3: The annotation scheme.

We follow a broad definition of toxicity, encompassing directed abuse towards (a) conversation participant(s) or a third party of the conversation, and generalized abuse defined in Waseem et al. (2017).

4. Data Collection

Data Source We use complete conversation threads from Reddit in our study. However, the method is generalizable to other online discussion platforms. Reddit is a large, publicly available social media platform, which is organized based on topics, known as *subreddits*. A user starts a discussion about a question by making a *post* and optionally, a description of the post in a subreddit, and other users can make *comments* to it. Thus, users can participate in threaded, asynchronous discussions, making it a large source of conversational data. A main thread, triggered by a post, often branches into multiple subthreads, which are localized interactions between users on a topic. We use "messages" in the following to refer to posts and comments collectively.

Data Selection Following previous work using Reddit data (Hada et al., 2021), we extract data from the Pushshift repository (Baumgartner et al., 2020)[3]. Unlike existing studies, however, we focus on sampling full conversation threads containing toxicity, which remains under-explored.

We start with community-based sampling (Qian et al., 2019; Vidgen et al., 2021), which means that subreddits that are likely to contain toxic language are selected. Clues for choosing such subreddits include subreddits chosen by Hada et al. (2021) and subreddits containing *LGBT*, *lesbian*, *immigrants*, *black*, *Asian*, *Muslim*, *Jew*, *Trump* and so on in their names. Subreddits with titles in other languages than English are filtered out. We also refer to recommendations of Reddit users[4]. Then we perform a preliminary filtering on extracted subreddits: we only choose subthreads that are public, with more than 10 comments, and contain mainly texts, rather than images. We define a conversation as an interaction involving *at least two* human speakers who alternate turns, resulting in nonlinear and intertwined discourse. Based on this definition, we remove threads that do not meet the following conditions: a) at least two participants; b) participants must send at least two/three messages in a thread; and c) a minimum of three distinct turns of speech is required. In addition, we remove threads with more than 20% of non-English messages using a language detection tool[5], and filter threads containing more than 20% of non-textual data, such as urls. To reduce the workload of annotation, we only keep threads with posts containing [5, 120] tokens and fewer than 30 messages.

Sampling Method After completing the initial curation pass, we specifically focus on identifying threads containing toxic language. To do this, first, we use a lexicon of toxic words—Hurtlex (Stamou et al., 2022) to detect threads containing occurrences of these lexicon items. From this information, we compute indicators that measure the in-

[3] https://www.reddit.com/r/pushshift/comments/1itme1k/separate_dump_files_for_the_top_40k_subreddits/, containing top 40k subreddits (ranked based on number of posts), from 2005-06 to 2024-12.

[4] https://www.reddit.com/r/NoStupidQuestions/comments/1co8ish/what_are_the_most_toxic_subreddits/

[5] https://pypi.org/project/fasttext-langdetect/

tensity and concentration of toxic terms within a thread. At the message level, intensity is computed as the ratio of toxic tokens over the total number of tokens of a message, and the intensity score of the whole thread is computed by averaging intensity scores of all the messages. Higher values of intensity indicate a greater overall presence of toxicity in the conversation. Concentration assesses the density of toxic terms within specific conversation windows. We define it as the density of toxic terms within specific conversation windows (i.e., subsets of consecutive messages within a thread). This helps capture localized bursts of toxicity rather than its overall distribution. Inspired by Ollagnier (2024), we use a window size of two in our implementation. Similarly, the concentration score of the whole thread is computed by averaging concentration scores of all the windows.

Following Hada et al. (2021), we incorporate valence and arousal scores when identifying targeted threads to increase the likelihood of capturing implicit forms of toxicity[6]. Valence measures the positive–negative dimension of emotion, while arousal captures the active–passive dimension. We use the NRC VAD lexicon developed by Mohammad (2018), which provides real-valued scores between 0 and 1 for tokens along these two dimensions.

Because intensity and concentration scores measure different properties of messages within a conversation, we compute thread-level valence and arousal scores using the same method as for intensity and concentration scores of toxic terms. The thresholds for valence and arousal are set as in Hada et al. (2021): 0.25 for valence and 0.75 for arousal.

The thread-level intensity index is obtained by summing the intensity scores of toxic terms with the valence and arousal scores. Similarly, the thread-level concentration index is computed by summing the corresponding concentration scores. Finally, threads are ranked based on the sum of the thread-level intensity and concentration indices. We leave it to future work to explore weighted combinations of these indicators and compare their effectiveness.

In our pilot experiments testing the effectiveness of this method, we selected the top 50 threads from a total of 196 threads after the data selection procedure. Among these, 10 threads were found to contain toxicity upon manual verification. Additionally, we examined the remaining 146 threads and identified 14 threads containing toxicity. For comparison, a random baseline on the same data yielded a precision of 12.24% and a recall of 25.50%, whereas our method achieved a precision of 20.00% and a recall of 41.67%.

We then adjusted the top N values until the corpus reached a reasonably large size and performed a manual examination to ensure that the threads contained toxicity. In the end, 124 threads were selected.

[6]These are primary dimensions of emotion.

5. Data Annotation

Agreement Measure It is challenging to measure consistency in annotating pragmatic inferences. In some cases, there may be no pragmatic inference (labeled as "literal"). The number of inferences can vary across instances, and the inferences in the two annotation sets are presented in random order by the annotators. Moreover, the illocutionary act types are specific to individual inferences, so their agreement cannot be measured independently. We adopt the following procedure: 1) If both annotations are "literal" in pragmatic inferences, then this counts as a total match[7]; 2) If one annotation is "literal", while the other is not, this is considered a total disagreement; 3) The cosine similarity between sentence embeddings (Reimers and Gurevych, 2019) of two sets of annotations is computed, which suggests that each annotation from an annotator (set A with m entries) will be compared with all the inference annotations from the other annotator (set B with n entries), forming an $m \times n$ matrix. For each item i of m, check if its highest cosine similarity score with items in n exceeds a threshold t (we set it to 0.40 based on empirical results). If the maximum similarity $\geq t$, count it as a match (semantic match). Then, precision is computed as the number of matches divided by m. The same computation is performed over n, and the result is considered as recall. The canonical F1 is used.

As is clear, precision and recall can be used exchangeably here. However, the distinction between precision and recall is necessary in some cases. For example, if set A contains human/reference annotations, set B contains GPT annotations, and we focus on the performance of GPT, the precision defined above would be better changed to recall because it measures the fraction of all human annotations that are matched, and the recall above would be precision, as it means the fraction of GPT annotations that are true matches. For computing the agreement on illocutionary act types, we first compute the number of entries from reference annotations that are matched by entries in the other set of annotations in cosine similarity (semantic match) and at the same time, in illocutionary act types (type match). The agreement is calculated as the number of entries with both a semantic match and a type match divided by the number of entries with a semantic match. By definition, the most salient inference is selected from the annotated inferences,

[7]This label suggests that no pragmatic inference is involved.

which are presented in random order. We choose to use semantic match to compute agreement on the selected most salient inference. In future work, the annotation procedure can be refined to better facilitate the evaluation of this property. Cohen's Kappa (Cohen, 1960) is used to measure agreement on the remaining tags.

Semi-automated Annotation Workflow Manual pragmatic annotation is resource-intensive and requires trained annotators well versed in linguistics. Large language models (LLMs), such as GPT, have demonstrated strong performance in complex semantic and pragmatic reasoning (Sravanthi et al., 2024). We therefore explore their use as semi-automatic annotators to accelerate corpus construction while maintaining quality. One of the authors, who has a background in computational linguistics, performed the annotation on 31 threads (25% of the whole corpus) first, and three LLMs (GPT-3, GPT-4, and GPT-5) independently annotated the same subset. Inter-annotator agreement (Inter-AA, also referred to as IAA) was then computed between the human and each model's annotations to assess reliability. Based on the IAA results, GPT-4 and GPT-5 were selected to annotate the remaining corpus. The prompt template is shown in Appendix A (section 9.1) and model parameter settings are shown in Appendix B (section 9.2).

Intra-Annotator Consistency As pragmatic inference is a subjective and elusive phenomenon and free-text annotation is used, we assessed intra-annotator consistency (Intra-AA) by having the same annotator re-annotate a random subset of 10 conversations three weeks after the initial annotation. This procedure allows us to estimate the stability of annotation decisions over time and to ensure that the annotation scheme yields consistent judgments by the same individual. According to the annotator's account, the time gap between the two stages was sufficient to minimize the influence of the first round of annotation on the second round.

Analysis Corpus statistics are shown in Table 4 and the results of IAA and Intra-AA are presented in Table 5. IAA is computed between human annotation and the GPT model's annotation. We recommend using the human annotated part (31 threads) as test set and the rest as training set[8], as shown in Table 4.

It is noticeable from the Intra-AA scores that the same annotator would use different texts to express pragmatic inferences and the perception regarding the most salient inference changes over time. As is clear from Table 4, the messages are long and vary considerably in length. When a message is long, its content is likely to be multifaceted, and multiple pragmatic inferences are possible. When the reply message is also long and rich in content, it is challenging to maintain high consistency in identifying the most salient inference. Among GPT models[9], GPT-4's annotations of pragmatic inferences and the corresponding illocutionary act types are more similar to those produced by humans. It can be observed from Table 4 that the number of tokens per inference generated by GPT4 is closer to those annotated by humans, while the inferences generated by GPT5 are much shorter. GPT5 and GPT3 are close to each other in annotating pragmatic inferences and selecting the most salient inference, with GPT5 outperforming GPT3 in annotating illocutionary act types. Moreover, GPT5 outperforms the others in annotating perlocutionary acts (as intended) by a large margin, indicating a stronger

Properties	Test (Human)	Training (GPT4)	Training (GPT5)
avg. number of messages/thread	15 $\pm$ 3.22	14.84 $\pm$ 3.14	—
avg. number of tokens/message	42.50 $\pm$ 42.70	42.26 $\pm$ 43.32	—
avg. number of unique users/thread	3.94 $\pm$ 1.03	3.97 $\pm$ 1.03	—
avg. number of inferences/message	1.94 $\pm$ 0.32	2.02 $\pm$ 0.22	3.11 $\pm$ 0.10
avg. number of tokens per inference/message	14.16 $\pm$ 1.82	15.17 $\pm$ 1.85	10.78 $\pm$ 1.21
ratio of aggressive labels	0.44	0.50	0.48

Table 4: Statistics of the annotated corpus. $\pm$ indicates standard deviation. The entries of GPT5 in dashes are the same as GPT4.

Tag	Intra-AA	IAA(GPT3)	IAA(GPT4)	IAA(GPT5)
inferences (semantic)	0.85	0.56	0.64	0.55
illocutionary act types	0.69	0.38	0.50	0.48
most salient inference (semantic)	0.55	0.40	0.43	0.40
as intended (κ)	0.74	0.01	0.24	0.66
PRE/IMP (κ)	0.53	0.77	0.35	0.49
aggressive (κ)	0.70	0.25	0.56	0.60

Table 5: Annotators' agreement. The word "semantic" denotes semantic match, and κ refers to Cohen's Kappa.

[8]GPT-4 and GPT-5 annotations are produced on the same text, and the data can be used jointly or selected based on their respective strengths.

[9]GPT3 = gpt-3.5-turbo, GPT4 = gpt-4.1-2025-04-14, and GPT5= gpt-5

capability of understanding content concordance between messages[10]. As our target label is aggressiveness, an expert human annotator worked on annotating this label on 9 full threads. After moderate adjunction, the κ between human annotators on this label is 0.60, which suggests that GPT5's annotation on this label is close to human level. Therefore, GPT5's annotation on the target label can be extracted and GPT4's annotation of pragmatic inferences may be used in training automatic systems.

Statistical Correlation Features such as illocutionary act types, as intended, and PRE/IMP are potentially useful for predicting aggressiveness. As a first step, we use mutual information (MI) (Shannon, 1948) to measure the correlation strengths between different pragmatic features and the target label. To better control the variables, features represented by free texts are not considered. In addition to pragmatic features, we also incorporate thread-level features for reference, including message_author, parent_username (user name of the parent message), subthread_depth (measured as the distance from a leaf node to the message within the subthread), and turn_index (the distance from the root to the message in the subthread, i.e., the reverse of subthread_depth). To better interpret the results, we randomly shuffle the target labels and compute the correlation strengths[11]. As can be seen from Table 6, the correlation strengths drop considerably when the target labels are shuffled. Thread-level features including message_author and parent_username are the strongest indicators of aggressiveness. However, these features are tied to specific users. The feature *expressiveness* is also one of the strongest indicators of aggressiveness. The feature *inference type* in the table means the illocutionary act type of the most salient inference. It is known from Table 5 that intra-AA for selecting the most salient inference is low, and consequently, the corresponding inference-type annotations are noisy. While selecting threads with shorter messages could potentially improve the results, we argue that this would misrepresent the characteristics of real data. The other features show only weak correlations with the target label. A complicating factor is imbalanced distribution of feature classes. For example, in the human-annotated portion, *representatives* constitute 63.04%, followed by *expressives* at 27.11%. *Directives* and *commissives* are rare, comprising 8.69% and 1.16%, respectively, and no instances of *declarations* are present. In the *PRE/IMP* dimension, *PRE* represents only 2.92% of the instances. These are the features that show the weakest correlations.

Annotated Labels	Labels Randomly Shuffled
message_author 0.22 ± 0.15	parent_username 0.14 ± 0.08
parent_username 0.21 ± 0.12	inference_type 0.13 ± 0.13
expressives 0.20 ± 0.22	message_author 0.12 ± 0.08
inference_type 0.14 ± 0.13	as_intended 0.08 ± 0.06
as_intended 0.09 ± 0.06	expressives 0.08 ± 0.14
representatives 0.06 ± 0.10	representatives 0.07 ± 0.16
subthread_depth 0.05 ± 0.11	commissives 0.05 ± 0.09
turn_index 0.04 ± 0.06	subthread_depth 0.04 ± 0.06
declarations 0.03 ± 0.06	turn_index 0.04 ± 0.07
commissives 0.03 ± 0.06	declarations 0.04 ± 0.06
PRE/IMP 0.03 ± 0.05	PRE/IMP 0.03 ± 0.05
directives 0.03 ± 0.05	directives 0.02 ± 0.07

Table 6: Correlation strengths of pragmatic features and thread features, computed on the human annotated part.

To further investigate this question, we run a Random Forest model (Breiman, 2001) (details are shown in Appendix C, section 9.3) to predict aggressiveness using the above features. The ensemble-based structure of a Random Forest model provides a convenient means of quantifying how much each feature contributes to predictive performance. We use the implementation from Scikit-learn (Pedregosa et al., 2011). The annotations of GPT4 and GPT5 are used jointly as the training set. Note that we only use the features above.

The test set contains 259 non-aggressive and 214 aggressive instances. Proportional random guessing yields an overall F1 score of 0.45. With a random seed of 42, we achieve a macro-averaged F1 score of 0.52. From Table 7, it is clear that a computational model tends to rely on thread-level features. Similar to the observations by Pavlopoulos et al. (2020), pragmatic features do not improve the performance on toxicity detection. This may suggest that pragmatic features encode higher-level discourse and interactional cues rather than surface-level indicators of toxicity. Consequently, such features may be better suited for tasks that focus on detecting early conversational derailment or shifts in interactional stance before explicit toxicity arises.

Features	Importance
message_author	0.25
parent_username	0.23
turn_index	0.14
subthread_depth	0.12
representatives	0.06
as_intended	0.06
expressives	0.04
inference_type	0.04
directives	0.03
commissives	0.02
PRE/IMP	0.01
declarations	0.00

Table 7: Importance scores of features.

[10]Since auto-regressive GPT models cannot see future messages, we expect their predictions to differ from human judgments, which are informed by full context.

[11]MI is computed at the thread level and averaged over the dataset.

6. Conclusion

We propose a generalizable framework for annotating pragmatics in datasets containing toxicity, enabling a deeply contextualized approach to studying toxicity in online discussion forums. The framework is theoretically grounded in Speech Act Theory, delving into the illocutionary and perlocutionary dimensions of language, which remains underexplored in existing studies. For the implementation of this new framework, we demonstrate a workflow from data selection and sampling to corpus annotation and evaluation. Owing to the elusive nature of pragmatic inference, we test intra-annotator consistency and find that it is possible to achieve reasonable consistency for the same annotator. We also test GPT models' capability in this task in order to achieve scalability and cost-effectiveness. In future work, we plan to explore using the corpus for detecting warning signs of whether a conversation will derail (Zhang et al., 2018).

7. Ethical Considerations and Limitations

Ethical Considerations The dataset contains Reddit discussion threads. It is intended to support a contextualized study of pragmatics of toxic behaviors in online forums. In the final release, usernames will be replaced with their cryptographic hash digests in the conversation, and only comment IDs will be shared in place of the original messages. This research has been reviewed by our institutional ethics committee.

Limitations We present our theoretical framework and describe the complete workflow for its implementation. At the same time, we acknowledge that additional human annotations are valuable for checking **inter-annotator agreement**. However, due to limited resources, this step was not performed. As stated in *Semi-automated Annotation Workflow*, manual annotation of pragmatic phenomena is resource-intensive and requires annotators to have considerable expertise in linguistics. If large language models (LLMs) can be used effectively, we consider this a positive indicator of the scalability and cost-effectiveness of the proposed approach. In future work, the annotated pragmatic information could support multiple applications, such as predicting from the onset of a conversation whether it will derail, or improving transparency in toxicity detection and mitigation.

Thanks to the suggestion made by a reviewer, we highlight the **complexity and cognitive demand** of annotation guidelines here. Based on our annotation experience, when a message is long, it poses significant challenges on identifying the most salient inference and its perlocutionary act, i.e., whether the reply message agrees with it or not, and whether it is an implicature or a presupposition. A remedy is to segment long messages into smaller argumentative units first. However, this assumes the conversation is argumentative in style—a condition that may not always hold. This reveals a deeper issue familiar to toxicity detection research: **data bias**. Specifically, datasets tend to overrepresent conflict-heavy conversations, and the distinctive features of platforms like Reddit further limit the generalizability of methods across domains.

8. Bibliographical References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Fatimah Alkomah and Xiaogang Ma. 2022. A literature review of textual hate speech detection methods and datasets. *Inf.*, 13(6):273.

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

John Langshaw Austin. 1962. *How to do things with words*. Oxford University Press.

Elisa Bassignana, Valerio Basile, and Viviana Patti. 2018. Hurtlex: A multilingual lexicon of words to hurt. In *Proceedings of the Fifth Italian Conference on Computational Linguistics (CLiC-it 2018), Torino, Italy, December 10-12, 2018*, volume 2253 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. The pushshift reddit dataset. In *Proceedings of the international AAAI conference on web and social media*, volume 14, pages 830–839.

Betty J. Birner. 2025. Pragmatics. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University. Accessed: 2 October 2025.

Leo Breiman. 2001. Random forests. *Machine learning*, 45(1):5–32.

Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. 2019. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *Proceedings of the*

2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), pages 1664–1674.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Noé Cécillon, Vincent Labatut, Richard Dufour, and Georges Linarès. 2020. WAC: A corpus of Wikipedia conversations for online abuse detection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 1382–1390, Marseille, France. European Language Resources Association.

Bharathi Raja Chakravarthi, Prasanna Kumaresan, Ruba Priyadharshini, Paul Buitelaar, Asha Hegde, Hosahalli Shashirekha, Saranya Rajiakodi, Miguel Ángel García, Salud María Jiménez-Zafra, José García-Díaz, Rafael Valencia-García, Kishore Ponnusamy, Poorvi Shetty, and Daniel García-Baena. 2024. Overview of third shared task on homophobia and transphobia detection in social media comments. In *Proceedings of the Fourth Workshop on Language Technology for Equality, Diversity, Inclusion*, pages 124–132, St. Julian's, Malta. Association for Computational Linguistics.

Lu Cheng, Ahmadreza Mosallanezhad, Yasin N Silva, Deborah L Hall, and Huan Liu. 2021. Mitigating bias in session-based cyberbullying detection: A non-compromising approach. In *The Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, volume 1.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Emily Corvi, Hannah Washington, Stefanie Reed, Chad Atalla, Alexandra Chouldechova, P. Alex Dow, Jean Garcia-Gathright, Nicholas J Pangakis, Emily Sheng, Dan Vann, Matthew Vogel, and Hanna Wallach. 2025. Taxonomizing representational harms using speech act theory. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 3907–3932, Vienna, Austria. Association for Computational Linguistics.

Jonathan Culpeper. 1996. Towards an anatomy of impoliteness. *Journal of pragmatics*, 25(3):349–367.

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.

Karthik Dinakar, Birago Jones, Catherine Havasi, Henry Lieberman, and Rosalind Picard. 2012. Common sense reasoning for detection, prevention, and mitigation of cyberbullying. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 2(3):1–30.

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

Paula Fortuna, Juan Soler Company, and Leo Wanner. 2020. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. pages 6786–6794.

Lei Gao and Ruihong Huang. 2017. Detecting online hate speech using context aware models. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 260–266, Varna, Bulgaria. INCOMA Ltd.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Rishav Hada, Sohi Sudhir, Pushkar Mishra, Helen Yannakoudakis, Saif Mohammad, and Ekaterina Shutova. 2021. Ruddit: Norms of offensiveness for english reddit comments. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2700–2717.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3309–3326, Dublin, Ireland. Association for Computational Linguistics.

Caroline Haythornthwaite. 2023. Moderation, networks, and anti-social behavior online. *Social Media + Society*, 9(3):20563051231196874.

Yiping Jin, Leo Wanner, and Aneesh Moideen Koya. 2025. What the#?*!: Disentangling hate across target identities. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 199–221.

David Jurgens, Libby Hemphill, and Eshwar Chandrasekharan. 2019. A just and comprehensive strategy for using nlp to address online abuse. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3658–3666.

Najoung Kim, Phu Mon Htut, Samuel R. Bowman, and Jackson Petty. 2023. (QA)^2: Question answering with questionable assumptions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8466–8487, Toronto, Canada. Association for Computational Linguistics.

Najoung Kim, Ellie Pavlick, Burcu Karagol Ayan, and Deepak Ramachandran. 2021. Which linguist invented the lightbulb? presupposition verification for question-answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3932–3945, Online. Association for Computational Linguistics.

Brigitte Krenn, Johann Petrak, Marina Kubina, and Christian Burger. 2024. GERMS-AT: A sexism/misogyny dataset of forum comments from an Austrian online newspaper. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 7728–7739, Torino, Italia. ELRA and ICCL.

Prasanna Kumar Kumaresan, Rahul Ponnusamy, Dhruv Sharma, Paul Buitelaar, and Bharathi Raja Chakravarthi. 2024. Dataset for identification of homophobia and transphobia for Telugu, Kannada, and Gujarati. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4404–4411, Torino, Italia. ELRA and ICCL.

Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. 2025. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8679–8696, Vienna, Austria. Association for Computational Linguistics.

Meta. 2025. Hateful conduct. https://transparency.meta.com/policies/community-standards/hateful-conduct/. Accessed: October 1, 2025.

Saif Mohammad. 2018. Obtaining reliable human ratings of valence, arousal, and dominance for 20,000 English words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 174–184, Melbourne, Australia. Association for Computational Linguistics.

Arianna Muti, Federico Ruggeri, Khalid Al Khatib, Alberto Barrón-Cedeño, and Tommaso Caselli. 2024. Language is scary when over-analyzed: Unpacking implied misogynistic reasoning with argumentation theory-driven prompts. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21091–21107, Miami, Florida, USA. Association for Computational Linguistics.

Nicolás Benjamín Ocampo, Elena Cabrio, and Serena Villata. 2023a. Unmasking the hidden meaning: Bridging implicit and explicit hate speech embedding representations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6626–6637.

Nicolás Benjamín Ocampo, Ekaterina Sviridova, Elena Cabrio, and Serena Villata. 2023b. An in-depth analysis of implicit and subtle hate speech messages. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1997–2013, Dubrovnik, Croatia. Association for Computational Linguistics.

Anais Ollagnier. 2024. CyberAgressionAdo-v2: Leveraging pragmatic-level information to decipher online hate in French multiparty chats. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4287–4298, Torino, Italia. ELRA and ICCL.

Anaïs Ollagnier, Elena Cabrio, Serena Villata, and Valerio Basile. 2024. Cyberagressionado-large: French multiparty chat dataset to address online hate. *Revue TAL: traitement automatique des langues*, 65(3):21–44.

Anaïs Ollagnier, Elena Cabrio, Serena Villata, and Valerio Basile. 2025. CyberAgressionAdo-large: French multiparty chat dataset to address online hate. *Traitement Automatique des Langues*, 65(3):21–44.

Anaïs Ollagnier, Elena Cabrio, Serena Villata, and Catherine Blaya. 2022. CyberAgressionAdo-v1: a dataset of annotated online aggressions in French collected through a role-playing game. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 867–875, Marseille, France. European Language Resources Association.

OpenAI. 2025. GPT-5 system card. Technical report, OpenAI. Accessed: October 1, 2025.

John Pavlopoulos, Jeffrey Sorensen, Lucas Dixon, Nithum Thain, and Ion Androutsopoulos. 2020. Toxicity detection: Does context really matter? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4296–4305.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2020. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477 – 523.

R Procter, H Webb, M Jirotka, P Burnap, W Housley, A Edwards, and M Williams. 2019. A study of cyber hate on twitter with implications for social media governance strategies. In *Conference for Truth and Trust Online 2019*. Conference for Truth and Trust Online.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4755–4764.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Manuela Sanguinetti, Fabio Poletto, Cristina Bosco, Viviana Patti, and Marco Stranisci. 2018. An Italian Twitter corpus of hate speech against immigrants. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Johannes Schäfer and Elina Kistner. 2023. HS-EMO: Analyzing emotions in hate speech. In *Proceedings of the 19th Conference on Natural Language Processing (KONVENS 2023)*, pages 165–173, Ingolstadt, Germany. Association for Computational Lingustics.

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10.

John R. Searle. 1975. A taxonomy of illocutionary acts. In K. Gunderson, editor, *Language, Mind and Knowledge*, pages 344–369. University of Minnesota Press.

Claude E. Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Peter K Smith, Jess Mahdavi, Manuel Carvalho, Sonja Fisher, Shanette Russell, and Neil Tippett. 2008. Cyberbullying: Its nature and impact in secondary school pupils. *Journal of child psychology and psychiatry*, 49(4):376–385.

Rachele Sprugnoli, Stefano Menini, Sara Tonelli, Filippo Oncini, and Enrico Piras. 2018. Creating a WhatsApp dataset to study pre-teen cyberbullying. In *Proceedings of the 2nd Workshop on Abusive Language Online (ALW2)*, pages 51–59, Brussels, Belgium. Association for Computational Linguistics.

Settaluri Sravanthi, Meet Doshi, Pavan Tankala, Rudra Murthy, Raj Dabre, and Pushpak Bhattacharyya. 2024. Pub: A pragmatics understanding benchmark for assessing llms' pragmatics capabilities. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 12075–12097.

Neha Srikanth, Rupak Sarkar, Heran Mane, Elizabeth Aparicio, Quynh Nguyen, Rachel Rudinger, and Jordan Boyd-Graber. 2024. Pregnant questions: The importance of pragmatic awareness in maternal health question answering. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7253–7268, Mexico City, Mexico. Association for Computational Linguistics.

Vivian Stamou, Iakovi Alexiou, Antigone Klimi, Eleftheria Molou, Alexandra Saivanidou, and Stella Markantonatou. 2022. Cleansing & expanding the HURTLEX(el) with a multidimensional categorization of offensive words. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 102–108, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Zeerak Talat and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on twitter. In *Proceedings of the NAACL student research workshop*, pages 88–93.

Wondimagegnhue Tsegaye Tufa, Ilia Markov, and Piek TJM Vossen. 2024. The constant in hate: Toxicity in reddit across topics and languages. In *Proceedings of the Fourth Workshop on Threat, Aggression & Cyberbullying@ LREC-COLING-2024*, pages 1–11.

Daniel Vanderveken. 1985. *Foundations of illocutionary logic*. Cambridge University Press.

Bertie Vidgen, Dong Nguyen, Helen Margetts, Patricia Rossini, and Rebekah Tromble. 2021. Introducing CAD: the contextual abuse dataset. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2289–2303, Online. Association for Computational Linguistics.

Xinyu Wang, Sai Koneru, Pranav Narayanan Venkit, Brett Frischmann, and Sarah Rajtmajer. 2025. The unappreciated role of intent in algorithmic moderation of abusive content on social media. *Harvard Kennedy School Misinformation Review*.

Zeerak Waseem, Thomas Davidson, Dana Warmsley, and Ingmar Weber. 2017. Understanding abuse: A typology of abusive language detection subtasks. In *Proceedings of the First Workshop on Abusive Language Online*, pages 78–84, Vancouver, BC, Canada. Association for Computational Linguistics.

Jiaxin Wen, Pei Ke, Hao Sun, Zhexin Zhang, Chengfei Li, Jinfeng Bai, and Minlie Huang. 2023. Unveiling the implicit toxicity in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1322–1338, Singapore. Association for Computational Linguistics.

Deirdre Wilson and Dan Sperber. 2004. Relevance theory. In Laurence R. Horn and Gregory Ward, editors, *The Handbook of Pragmatics*, pages 607–632. Blackwell Publishing, Oxford.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399.

Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se-Young Yun. 2023. HARE: Explainable hate speech detection with step-by-step reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5490–5505, Singapore. Association for Computational Linguistics.

Michael Yoder, Lynnette Ng, David West Brown, and Kathleen Carley. 2022. How hate speech varies by target identity: A computational analysis. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 27–39, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Zehui Yu, Indira Sen, Dennis Assenmacher, Mattia Samory, Leon Fröhling, Christina Dahn, Debora Nozza, and Claudia Wagner. 2024. The unseen targets of hate - A systematic review of hateful communication datasets. *CoRR*, abs/2405.08562.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420.

Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1350–1361, Melbourne, Australia. Association for Computational Linguistics.

Xuhui Zhou, Hao Zhu, Akhila Yerukola, Thomas Davidson, Jena D. Hwang, Swabha Swayamdipta, and Maarten Sap. 2023. COBRA frames: Contextual reasoning about effects and harms of offensive statements. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6294–6315, Toronto, Canada. Association for Computational Linguistics.

9. Appendices

9.1. Appendix A. Prompt Template for Annotating Pragmatic Inferences

Table 8 shows the prompt template for pragmatic annotation using LLMs.

9.2. Appendix B. Model Parameters

max_completion_tokens =4096
temperature=0
n=1

9.3. Appendix C. Details of Computational Models

9.3.1. Proportional random guessing

Total instances: 214 + 259 = 473
aggressive probability (p): $\frac{214}{473}$ = 0.452
non-aggressive probability: 1- p = 0.548

Table 9 shows the expected values for predicting the aggressive class. Following this, we compute:
precision: $\frac{96.7}{96.7+117.0}$ = 0.453
recall: $\frac{96.7}{96.7+117.3}$ = 0.452

F1 2 $\times$ $\frac{precision \times recall}{precision+recall}$ = 2 $\times$ $\frac{0.453 \times 0.452}{0.453+0.452}$ = 0.452

9.3.2. Random Forest Model

Parameter settings of the Random Forest model:
random seed = 42
n_estimators=200
max_depth=None
class_weight= "balanced"

You are a language expert.
Let's analyse step by step.
Given a thread of messages from an online forum, generate 2-3 or more, if possible, most important pragmatic inferences for EACH message.
The inferences can be sentiments, such as emotional state or stance towards some issues, and hidden knowledge from the message, including stereotypes and social constructs, pre-assumptions held by the author of the message, and implications derivable from the message. Try to avoid trivial inferences.
Be CONCISE. For example, given a sentence "What kind of music should I play to my baby in the womb?", you need to generate inferences like the following:
1. Babies can hear sound in the womb.
2. Babies can differentiate different sounds.
3. Hearing music positively influences fetal development.
4. Certain kinds of music are more beneficial to babies in the womb than the others.
The following inferences are also possible. However, these are trivial inferences that do not contribute to conveying the main point.
There is a baby in the interlocutor's womb.
Different kinds of music are available.
Music is something that can be played.
b) When giving an inference, you need to determine the "type" of the inference from one of the following:
1. representatives: statements that tell how things are, focusing on the truth of a proposition.
2. directives: statements attempting to persuade the hearer to do something, such as ordering, advising, and warning the hearer.
3. commissives: statements that commit the speaker to a course of action, such as actions of vowing, promising, offering to do something, or negatively, refusing or preventing someone from doing something by threatening.
4. expressives: statements expressing a psychological state, such as apologizing, congratulating, condoling.
5. declarations: statements effecting an immediate change in the institutional state of affairs (firing from employment, christening). e.g."I declare you husband and wife", and "You are fired!"
c) If no pragmatic inferences are considered important for the conversation but only literal interpretation of the messages is used by the conversation participants, simply output "literal" and "none" in "type".
d) Give your confidence score [0, 100] for each of the generated INFERENCES.
e) Determine the most salient inference for a message BASED ON its reply message, which has the message id in REPLY_To, and copy its type to an entry "inference_type".
f) Determine whether the most salient inference is agreed in the reply message. Put "yes" or "no" or "uncertain" (when no messages has the message id in "Reply_To") in an entry "as_intended".
g) Determine wehther the most salient inference belongs to a "presupposition" (PRE, something taken to be true by the speaker, without proving) or "implicature" (IMP, something the speaker tries to convey with the message).
h) Determine whether the message is aggressive (using abusive language) (OAG), covertly aggressive (not using abusive language) (CAG), or not aggressive (NAG).
Your output should be in a valid json format:
{ "15zppta": { "Pragmatic_Inferences": { "1": { "content": "...", "type": "...", "confidence": 0.90 }, "2": { "content": "...", "type": "...", "confidence": 0.85 }, "3": { "content": "...", "type": "...", "confidence": 0.80 } }, "as_intended": "yes", "aggressive": "CAG", "most_salient_inference": "1", "inference_type": "...", "PRE/IMP": "IMP" }, "jxiaamu": { "Pragmatic_Inferences": { "1": { "content": "...", "type": "...", "confidence": 0.92 }, "2": { "content": "...", "type": "...", "confidence": 0.88 } }, "as_intended": "yes", "aggressive": "CAG", "most_salient_inference": "1", "inference_type": "...", "PRE/IMP": "IMP" } } ...
Input:
"Message_ID": "kq2jlzg",
"Message_Content":"...",
"Reply_To": "kq2hrm8"
Your output: ?

Table 8: Prompt template for pragmatic annotation.

	Predict AG	Predict NAG
True AG	214 $\times$ p = 96.7 (TP)	214 $\times$ (1-p) = 117.3 (FN)
True NAG	259 $\times$ p = 117.0 (FP)	259 $\times$ (1-p) = 141.9 (TN)

Table 9: Expected values for the aggressive class.