From Intrinsic Toxicity to Reception-Based Toxicity: A Contextual Framework for Prediction and Evaluation

Anonymous ACL submission

Abstract

Most toxicity detection models treat toxicity as an intrinsic property of text, overlooking the role of context in shaping its impact. In this position paper, drawing on insights from psychology, neuroscience, and computational social science, we reconceptualise toxicity as a socially emergent signal of stress. We formalise this perspective in the **Contextual Stress Framework (CSF)**, which defines toxicity as a stress-inducing norm violation within a given context and introduces an additional dimension for toxicity detection. As one possible realisation of CSF, we introduce **PONOS (Proportion Of Negative Observed Sentiments)**, a metric that quantifies toxicity through collective social reception rather than lexical features. We validate this approach on a novel dataset, demonstrating improved contextual sensitivity and adaptability when used alongside existing models.

1 Introduction

Warning: *This paper contains examples of language that may be perceived as offensive or toxic.*

When asked to define obscenity, Judge Potter Stewart famously said: "I know it when I see it". To this day, we have not come far beyond this intuition.

Natural language processing (NLP) has made remarkable technical progress in detecting "toxic" language. Yet, despite ever-larger models and rising benchmark scores, a foundational question remains unresolved: **what exactly is toxicity?**

An influential work by Dixon et al. (2018) defines toxic speech as "rude, disrespectful, or unreasonable language likely to make someone leave a discussion". However, terms such as "rude" and "disrespectful" are inherently subjective, varying considerably across cultures, individuals, and situational contexts. Crucially, such definitions offer no quantitative basis for measurement, leading to labelling inconsistencies that undermine model performance (Garg et al., 2023).

Despite its vagueness and limitations, this definition remains widely used in both research and industry.

This ambiguity prevents consistent classification even for seemingly simple phrases - for instance, "I don't like you". Some may interpret it as toxic, while others may not. Asking annotators with differing perspectives to label toxicity therefore leads to disagreement, introducing noise that undermines model accuracy (Liu et al., 2021; Goyal et al., 2022; Sap et al., 2021).

At the level of academia as a whole, the lack of a standardised definition has led researchers to employ different terms for similar phenomena (Vidgen, 2019), assigning labels such as "toxic", "hateful", "offensive", or "abusive" to effectively the same data (Fortuna, 2020; Madukwe et al., 2020). In this paper, we use *toxic speech* as an umbrella term.

Another major issue in toxicity detection research is the field's overreliance on benchmark-driven evaluation. Improvements in benchmark scores are frequently treated as evidence of real-world progress, despite the absence of ground-truth metrics linking these scores to reductions in social harm, such as exclusion, stress, or violence.

Most toxicity detection systems implicitly assume that toxicity is an intrinsic property of language - something that can be detected from surface features, isolated utterances, or predefined keyword lists. This assumption neglects the crucial role of context. Without context, toxicity detection systems are prone to racial, sexual, political, religious, and geographical biases (Garg et al., 2023; Sap et al., 2019). These systems often degrade into simple profanity filters (Chen, 2022), lacking the nuance necessary for real-world use (Xenos et al., 2021; Pavlopoulos et al., 2020; Menini et al., 2021) and becoming susceptible to numerous adversarial attacks (Berezin et al., 2025, 2024, 2023).

1.1 Our Contribution: A New Frame, A Measurable Signal

To address this, we argue for a shift in emphasis: from viewing toxicity as purely intrinsic to the text towards modelling it as an emergent signal of perceived social stress.

Drawing on interdisciplinary insights, we introduce the **Contextual Stress Framework (CSF)**, which frames toxicity as a stress-inducing norm violation *within a given context*. CSF adds an additional dimension to existing text-intrinsic approaches by capturing how toxicity is perceived and socially received.

Within CSF, we introduce **PONOS** (Proportion Of Negative Observed Sentiments) as one possible realisation of the framework: a metric that quantifies toxicity via the collective emotional reception a message provokes. Rather than determining what a message intrinsically *is*, PONOS captures how it is received in context.

2 Related Work and Limitations of Existing Definitions

Contemporary research demonstrates that words can inflict measurable psychological and physiological harm (Struiksma et al., 2022b; Sapolsky, 2004).

Toxic communication has been linked to increased anxiety, chronic stress, radicalisation, and, at its extreme, self-harm and suicide (van Geel et al., 2014; Kiritchenko et al., 2021; Cervone et al., 2021). At the societal level, hate speech exacerbates discrimination, fosters violence, and erodes the conditions for open public discourse (Vidgen and Derczynski, 2021).

Despite the urgency of addressing these harms, the field lacks a coherent operational definition of toxicity. Definitions across academic, legal, and industrial contexts remain fragmented, often vague, inconsistent, and mutually incompatible (Mnassri et al., 2024).

In an effort to systematise existing approaches, Herz and Molnár (2012) categorise definitions of harmful speech along four dimensions:

1. **Harm-based**: speech that causes psychological or social harm;
2. **Content-based**: speech conveying hateful ideas;
3. **Lexical-based**: speech characterised by the use of derogatory words;
4. **Dignity-based**: speech undermining the inherent dignity of target groups.

Each category, however, exhibits critical limitations.

Harm-based definitions rely on an inherently underspecified notion of harm, encompassing both genuinely toxic speech and socially necessary but distressing acts (e.g. delivering bad news).

Content-based definitions risk overreach, potentially classifying legitimate political disagreement or critique as hate speech (Parekh, 2012; Hare and Weinstein, 2009; Anderson and Barnes, 2023).

Lexical-based approaches overlook toxic messages expressed through ostensibly neutral language (e.g. "Women, back to the kitchen!").

Dignity-based models risk conflating legitimate criticism with toxicity, thereby suppressing vital social and political discourse (Anderson and Barnes, 2023).

Subsequent work by Khurana et al. (2022) identifies five critical dimensions: target group, dominance, perpetrator identity, type of negative reference, and type of consequences. Yet, it finds no universal definition applicable across datasets or social contexts. Even broader formulations, such as Fiser et al.'s notion of "socially unacceptable discourse" (Fišer et al., 2017), fail to resolve the underlying contextual instability.

As Brown (2017) observes, harmful speech is not merely contested but systematically ambiguous, acquiring different meanings depending on the cultural, political, and historical frames in which it is situated.

3 An Interdisciplinary Basis for CSF

Across psychology, neuroscience, and sociolinguistics, verbal harm is understood as fundamentally context-dependent: the same utterance can be harmless or harmful depending on who interprets it, under what expectations, and within which normative setting. A common thread is an *appraisal* process: communicative acts are evaluated relative to local moral and interactional norms, and when an act is perceived as threatening, disrespectful, or boundary-violating, it can trigger affective responses and stress-related dysregulation in recipients and bystanders (Lazarus, 1991; Eisenberger et al., 2003; Lieberman and Eisenberger, 2009).

At the **individual level**, research on stress and emotion emphasises that harmful impact depends on interpretation rather than surface form alone.

Predictive and context-sensitive accounts of cognition model perception as expectation-driven; violations of anticipated social or emotional states can elicit stress responses even when intent is ambiguous (Friston, 2010; Barrett and Simmons, 2015; Kleckner et al., 2017). This link between social threat and physiological regulation is reflected in established correlates such as heart-rate variability (Shaffer et al., 2014).

At the **interactional level**, harm is jointly constructed and becomes legible through sequences of action and reaction: norms are maintained and renegotiated through observable responses such as condemnation, withdrawal, or escalation. Work on norm enforcement shows that deviance is regulated through collective emotional and behavioural reactions (Bicchieri, 2005; Cialdini et al., 1991; van Kleef and Fischer, 2015).

At the **cultural level**, communities differ in how they sanction profanity, sarcasm, reclaimed slurs, or in-group teasing; what appears "toxic" out of context may be benign in-group signalling, and vice versa. Pragmatic accounts emphasise that meaning is inferred from shared background and interactional framing rather than encoded in words alone (Clark, 1996; Sperber and Wilson, 1986/1995). Politeness and impoliteness research likewise treat offence as an emergent property of interaction (Brown and Levinson, 1987; Culpeper, 1996), while language ideology work highlights how judgements of "toxic" versus "neutral" language are shaped by power and dominant cultural narratives (Woolard and Schieffelin, 1994; Irvine, 2001). Finally, meaning and offensiveness shift across time and across languages, undermining any fixed universal standard (Traugott and Dasher, 2001; Dewaele, 2004).

Taken together, these literatures motivate CSF's core composition: toxicity arises when an utterance is perceived to violate local norms and this perception induces stress or disruption in an audience (additional discussion in Appendix A).

4 Definition of Toxicity

Building on the interdisciplinary insights outlined above, we model verbal harm as comprising two components: (i) a perceived violation of social or moral norms in a given interactional context; and (ii) an ensuing stress response in those who interpret the act.

Consider a speaker who writes a profane, intentionally offensive message in Chinese to a recipient who does not speak the language. Although the intent is hostile and the string contains profanity, no harm is experienced if the recipient cannot interpret it. There is no perceived norm violation, no affective response, and no stress. Under our framework, such a message registers as low in toxicity in this interaction - not because it is morally defensible, but because it fails to produce a measurable signal of harm in the audience that actually receives it. The same message would become toxic if it reached an informed audience.

Conversely, imagine a speaker makes a flirtatious joke believing it to be benign or even complimentary. If the recipient experiences it as inappropriate or violating - due to context, power dynamics, or personal boundaries, stress and social disruption may occur. In many professional and legal contexts, this is recognised as harm: offence is taken, not given. Under our model, such an interaction can be toxic even in the absence of harmful intent.

These examples motivate a crucial point: toxicity is defined not by intent or language alone, but by its reception - by the stress it induces and the norms it is perceived to violate.
Based on this synthesis, we propose:

Definition 1. ***Toxicity*** *— the characteristic of causing stress via a perceived contradiction of accepted morality and norms of interaction, within the context in which it is interpreted.*

Consequently:

Definition 2. ***Toxic speech*** *— speech that induces stress via a perceived contradiction of accepted morality and norms of communication, within the context in which it is interpreted.*

These definitions deliberately foreground context, perception, and reception. We distinguish between toxicity and toxic speech to account for the broader scope of harmful communicative acts, including visual, multimodal, or behavioural signals that may not be strictly linguistic.

Operational implication (two-axis view). The two components of the definition above yield two complementary measurement routes. Text-only toxicity detectors (e.g. moderation models trained on annotator judgements) primarily operationalise perceived norm violation from the content itself under assumed or averaged contexts. In contrast, reception-based signals operationalise experienced stress as it manifests in interaction (e.g. in replies),

capturing context that is not present in the message text alone. Our framework therefore treats intrinsic scores and reception signals as distinct axes rather than interchangeable "toxicity scores": an utterance may be norm-violating without eliciting strong negative reception (high violation, low stress), or elicit negative reception without being flagged by text-only detectors (low violation, high stress).

While our framework treats stress as a measurable proxy for harm, we do not claim that all stress-inducing communication is inherently toxic. Delivering bad news or asserting uncomfortable truths may elicit stress without violating social or moral norms. Toxicity, as defined here, arises from the intersection of perceived norm violation *and* stress response, not from stress in isolation. This distinction is crucial to avoid conflating necessary discomfort with communicative harm.

This reconceptualisation challenges the assumption that toxicity can be located in isolated words or surface features. Instead, it positions toxicity as a socially emergent signal - a disruption of shared expectations that manifests as emotional or relational stress. Any detection system intended to mitigate real-world harm must therefore adopt an interactional, context-sensitive perspective.

We do not deny the relevance of speaker intent, but argue that intent alone is insufficient and should instead be treated as one component of context. Here, *context* refers to the full set of conditions under which a communicative act is interpreted, including linguistic, situational, cognitive, cultural, and relational dimensions. A more detailed discussion of contextual factors is provided in Appendix B.

4.1 Towards Measurement: Stress as a Measurable Signal

If toxicity detection aims to mitigate harm, then the relevant notion of harm must be grounded in a signal that is (i) theoretically motivated and (ii) measurable at scale. Across disciplines, **stress** provides such an anchor: it refers to a family of physiological and behavioural responses that arise when an individual appraises a situation as threatening, boundary-violating, or socially unsafe. This makes stress a principled correlate of communicative harm in cases where harm is symbolic, ambiguous, or socially mediated.

Unlike global labels such as "offensive" or "rude", stress has a well-characterised biological basis and a long measurement tradition in psychology and neuroscience (McEwen, 1998; Thayer et al., 2012).

Stress can be estimated through multiple classes of indicators, each trading off fidelity against scalability. **Direct physiological measures** (e.g. cortisol levels) are biologically faithful but require invasive sampling and are impractical outside controlled settings. **Non-invasive biosignals** such as heart-rate variability, galvanic skin response, respiration, pupil dilation, and facial or vocal affect provide scalable approximations and correlate with social-threat responses (Sharma and Gedeon, 2012; Giannakakis et al., 2019), but still depend on specialised hardware and stable observation conditions.

For NLP systems deployed in online environments, the most practical evidence is therefore **behavioural**: stress and disruption often surface in interaction through reaction patterns, shifts in tone, escalation, withdrawal, or affective language in replies. Prior work shows that collective response dynamics can serve as indirect indicators of perceived norm violation and emotional disruption in online discussion (Jafari et al., 2023; Aleksandric et al., 2024). Importantly, behavioural proxies do not measure internal physiology directly; they capture how social tension manifests publicly in interaction - precisely the level at which toxicity becomes consequential for communities and moderation decisions.

5 PONOS: A Metric for Reception-Based Toxicity

Building on the stress-based account of toxicity developed above, we now introduce a concrete metric for capturing reception-based toxicity: how a communicative act is received by its audience in context.

Our definition implies a causal pathway: a post is toxic when it is *perceived* to contradict local moral or interactional norms, and this perceived violation induces a stress response in the audience.

In online interaction, stress and social disruption often surface as expressions of disapproval, outrage, distress, or condemnation in replies. Accepting that toxicity is reflected in the stress it evokes, we treat patterns of community response as a proxy signal for reception-based toxicity: when a post elicits a disproportionate share of negative reactions, this provides evidence that it violated local

norms *and* induced tension in its audience.

To formalise this intuition, we introduce **PONOS** (**P**roportion **O**f **N**egative **O**bserved **S**entiment; from Greek πόνος, "pain") as one instantiation of the Contextual Stress Framework. PONOS quantifies toxicity not through lexical content or opaque model scores, but through the emotional reception of a message within a specific community and context.

5.1 Defining PONOS

Let a post x receive a set of replies $R(x) = \{r_1, r_2, \ldots, r_n\}$. Each reply $r_i \in R(x)$ is assigned a sentiment label $s(r_i) \in \{-1, 0, +1\}$, where -1 denotes negative sentiment, 0 neutral sentiment, and $+1$ positive sentiment.

We define **PONOS** as the proportion of negative replies:

$$\text{PONOS}(x) = \frac{1}{|R(x)|} \sum_{r_i \in R(x)} \mathbb{I}[s(r_i) = -1], \quad (1)$$

where $\mathbb{I}[\cdot]$ denotes the indicator function.

Intuitively, PONOS(x) estimates the share of observed replies that express negative affect. Higher values indicate stronger collective expressions of disapproval or distress, consistent with reception-based harm in context.

PONOS is (i) interpretable as a proportion of negative replies, (ii) context-sensitive via local audience norms, and (iii) statistically grounded and comparable, as an estimate of a negative-reception rate whose uncertainty decreases with $|R(x)|$ (Guo et al., 2017).

PONOS is a flexible framework that admits multiple extensions to better reflect community dynamics, reply timing, and user context. In Appendix C we describe its variants, such as PONOS-Net, PONOS-Weighted, and PONOS-Early.

6 Empirical Evaluation

This section provides empirical support for CSF: toxicity is not a fixed property of text in isolation, but a contextual social signal that is realised in interaction. Our definition decomposes toxicity into two components: (i) *perceived norm violation* (contradiction of accepted morality and norms of interaction) and (ii) the *stress or disruption* this perceived violation induces in an audience. Accordingly, we evaluate PONOS as a reception-based measure of the second component (stress or disruption manifested behaviourally), and we treat widely used text-only toxicity instruments as reference measures that primarily operationalise the first component (perceived norm violation) under assumed or averaged contexts.

We therefore design the evaluation as validity evidence for a reception-based axis, rather than as a leaderboard. Concretely, we test (i) whether the reception signal can be instantiated reliably in-domain, (ii) whether it is estimable without access to future replies, and (iii) whether it provides information that is distinct from intrinsic (text-only) instruments. We operationalise these three goals as the following **evaluation questions:**

- **Q1 (Operationalisation / instrument selection):** Which sentiment model yields an in-domain sentiment function $s(\cdot)$ that best matches human judgements of reply sentiment in this community?
- **Q2 (Estimability at inference time):** Can measured PONOS be predicted from the post text alone, and does explicitly modelling likely reactions improve estimation?
- **Q3 (Construct divergence / complementarity):** How does reception-based measurement (PONOS) relate to intrinsic-toxicity instruments (text-only toxicity detectors)? Where do they agree, and where do they systematically diverge? Does PONOS provide a complementary axis aligned with CSF, or is it reducible to existing intrinsic scores?

6.1 Dataset Creation

We curate a dataset from the Reddit community *r/BlackPeopleTwitter*. We choose this community for two reasons: (i) its scale (5.8M members) and (ii) explicit encouragement of African-American Vernacular English (AAVE) and community-specific slang use. This setting is useful for studying context sensitivity, since dialectal and in-group expressions are frequently misinterpreted by generic toxicity systems trained on broad web data, effectively penalising speakers for communicating in their own dialect.

In total, we collected 401,931 posts containing 15,922,851 comments.

6.1.1 Human Annotation for Selecting $s(\cdot)$

To instantiate the sentiment function $s(\cdot)$ in Eq. 1, we construct a small, high-quality labelled set of

posts and replies, since replies constitute the reception signal from which PONOS is computed.

We sample 56 user messages, each with at least four replies. If a message had too many replies, we retained only the top seven by score, resulting in a total of 273 replies.

Annotators and quality control. We recruit 11 annotators via Prolific platform, all self-identifying as Black/African American U.S. nationals and regular internet users (7 men, 4 women; ages 21–60). We include attention checks and examine consistency. Four annotators show strong engagement and reliable performance; we retain their annotations for analysis.

Agreement and aggregation. Inter-annotator agreement is Fleiss' $\kappa = 0.48$ with 67.8% pairwise agreement, consistent with moderate agreement for subjective judgement tasks. Final labels are obtained by majority vote; in rare 2-2 ties, an additional label is provided by a researcher following the same guidelines.

6.2 Experimental Setup

All experiments were run on NVIDIA H100 GPUs, totalling 883 compute hours. Model training and evaluation were performed using the February 2025 release of the `unsloth` library (unslothai, 2025). Metrics were computed using scikit-learn v1.5.0 (Pedregosa et al., 2011).

Context provided to models. For reply sentiment classification, we provide as much conversational context as feasible: the original comment, the reply text, the title of the post, and (when applicable) the set of replies under classification. We found that providing this context improved model's performance. Prompts used with LLMs are provided in Appendix E.

6.3 Step 1 (Q1): In-Domain Instrument Selection for $s(\cdot)$

We evaluate multiple candidate sentiment models on the labelled reply set in order to select $s(\cdot)$ for Eq. 1. This step is best viewed as *instrument selection*: we choose the model that most reliably reproduces in-domain human judgements of negative reception, rather than optimising for generic sentiment benchmarks.

Table 1 reports performance for detecting the negative sentiment class on our labelled data.

Model	Prec.	Rec.	F1
RoBERTa (2023)	0.63	0.22	0.32
ModernBERT (2024)	0.71	0.21	0.32
Twitter-RoBERTa (2020)	0.58	0.64	0.61
GPT-4o (2024a)	0.67	**0.82**	0.74
LLaMA 3 70B	**0.81**	0.70	0.75
Qwen 3 70B	0.76	0.80	**0.78**

Table 1: Precision, recall, and F1 scores for the negative sentiment class across candidate models on our labelled reply set.

Qwen 3 (Yang et al., 2025) achieved the strongest overall performance and is used as $s(\cdot)$ in all subsequent experiments. During error analysis, we found that most misclassifications occur in examples with low inter-annotator agreement. These instances often involved ambiguous language that led both assessors and models to varied interpretations.

6.4 Step 2 (Q2): Estimating PONOS from Post Text

With $s(\cdot)$ fixed, we compute PONOS from observed replies and then ask whether it can be accurately predicted at inference time (i.e. before replies are available).

Approach A: Direct PONOS Estimation We train regression models to predict the measured PONOS value from the post text alone. This estimates reception without explicit reaction modelling and serves as a lightweight baseline suitable for low-latency settings.

Approach B: PONOS via Modelled Reactions We also evaluate a reaction-modelling pipeline: a language model generates plausible replies to a post, and we apply $s(\cdot)$ to these generated replies to estimate PONOS. This directly instantiates CSF's hypothesis that toxicity is reflected in likely community responses.

To separate the contributions of domain adaptation and reaction modelling, we evaluate two variants using LLaMA 3 70B (Llama Team, 2024): (i) reply prediction (RP) with PONOS inference via $s(\cdot)$, and (ii) domain-tuned reply prediction (DTRP).

6.4.1 Domain-Specific Pretraining Details

For DTRP, we pretrained LLaMA 3 70B on *r/BlackPeopleTwitter* data to better capture community style and norms. We used the `unsloth` framework

to load and fine-tune the base model in 4-bit precision. The model was adapted using LoRA with rank 64 for efficient parameter updates. Gradient checkpointing and `rslora` were employed to further reduce memory overhead. All training and inference were conducted using bfloat16 precision and FlashAttention 2 to maximize memory efficiency.

Each training example consists of a Reddit comment and a prompt requesting five likely replies in the tone and style of *r/BlackPeopleTwitter*, separated by `<|reply|>`. The prompt template appears in Appendix E.

6.4.2 Step 3 (Q3): Comparing PONOS to Text-only Toxicity Scorers

We characterise how PONOS reception-based signal relates to intrinsic-toxicity instruments. To study divergence, we include intrinsic toxicity systems as reference instruments: the OpenAI Moderation API (`omni-moderation-latest`) (2024b) and the Google Perspective API (2024).

These systems estimate toxicity as a property of text and are therefore aligned with the perceived norm-violation component of our definition. In contrast, PONOS measures the reception or stress-disruption component via replies. We therefore focus on the relationship between these signals and on systematic regions of disagreement.

6.5 Results: Estimability and Construct Divergence

Estimating continuous PONOS. Table 2 reports mean squared error (MSE) and mean absolute error (MAE) for continuous PONOS estimation. Two patterns emerge. First, strong text-only estimators (e.g. fine-tuned SentenceBERT) achieve competitive error, suggesting that some correlates of reception are recoverable from post text alone. Second, reaction modelling benefits substantially from domain adaptation: reply prediction without adaptation (RP) yields high error, while the domain-tuned reply prediction (DTRP) model attains the lowest error (MSE 0.042; MAE 0.152).

Together, these results support the claim that approximating community response can improve reception prediction when the generator is calibrated to community norms.

Construct relationship analysis: PONOS vs. intrinsic instruments. We quantify the association between reception-based measurement and intrinsic-toxicity scoring using large-scale comparisons ($n = 88{,}717$) between measured PONOS and two widely deployed toxicity detection instruments: the OpenAI Moderation API and the Google Perspective API. For OpenAI, we summarise each post by the maximum category score ("max moderation score") and additionally report correlations for individual categories. For the Perspective API, we use the continuous toxicity score.[1]

Model	MSE	MAE
Direct Prediction		
TF–IDF + Ridge	0.046	0.173
SentenceBERT FT	0.043	0.166
Reply Prediction		
LLaMA 3 8B RP	0.278	0.402
LLaMA 3 70B RP	0.138	0.304
GPT-4o RP	0.214	0.308
LLaMA 3 70B DTRP	**0.042**	**0.152**

Table 2: Mean squared error (MSE) and mean absolute error (MAE) for PONOS prediction. FT = fine-tuned model.

Both instruments show positive but modest association with reception. For OpenAI, PONOS correlates with the max moderation score at $\rho = 0.203$ (95% CI [0.197, 0.210]) and $r = 0.195$ (95% CI [0.189, 0.201]); among categories, `harassment` shows the strongest monotonic association ($\rho = 0.207$). For the Perspective toxicity score, the association is $\rho = 0.176$ (95% CI [0.170, 0.183]) and $r = 0.173$ (95% CI [0.167, 0.179]). Together, these results support the construct claim: intrinsic-toxicity scoring and reception-based PONOS are related but non-equivalent signals.

Finally, to test whether these patterns depend on a particular instrument, we measure inter-tool agreement. OpenAI max moderation score and Perspective toxicity correlate strongly ($\rho = 0.866$, 95% CI [0.864, 0.868]), indicating that intrinsic instruments form a coherent axis that is weakly aligned with the reception-based PONOS axis.

Quadrant breakdown. We discretise both axes into high versus low using quantile thresholds and count posts in each quadrant. Using top-25% thresholds on PONOS and max moderation score (Figure 1), we find that off-diagonal disagreement is substantial. For the OpenAI

[1] In all cases, we report Spearman's ρ (monotonic association) and Pearson's r (linear association), with 95% confidence intervals via Fisher's z approximation.

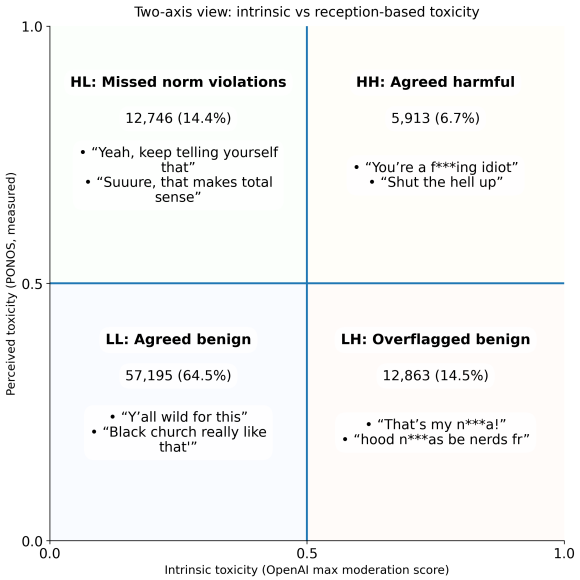

Figure 1: Two-axis view of intrinsic toxicity (OpenAI max moderation score; x-axis) versus reception-based toxicity (PONOS; y-axis). Each quadrant reports prevalence (count and percentage; $n = 88{,}717$) and includes illustrative masked snippets.

max moderation score, 14,513 posts (16.4%) are high-PONOS/low-moderation and 14,158 (16.0%) are low-PONOS/high-moderation (32.3% off-diagonal total). The Perspective API shows the same pattern: 15,174 posts (17.1%) are high-PONOS/low-intrinsic and 14,887 (16.8%) are low-PONOS/high-intrinsic (33.9% off-diagonal).

This supports the view that reception and intrinsic toxicity are related but non-equivalent axes.

To interpret the disagreement regions, we analyse two extreme divergence sets constructed from posts where reception and intrinsic instruments disagree most strongly. We emphasise that these are not universal "model failures"; rather, they are diagnostic cases that make construct mismatch concrete.

Reception-high / instrument-low: missed norm violations. These posts elicit consistently hostile, aggressive, or sarcastic replies (maximal negative reception) yet receive low intrinsic scores. On average, the OpenAI max moderation score is 0.090 and the Perspective API toxicity score is 0.155.

These cases often involve sarcasm, pragmatic antagonism, or contextually coded norm violations that are legible to the community but lack strong lexical markers.

Reception-low / instrument-high: overflagging benign dialect and in-group language. Conversely, these posts receive no negative reception but are flagged as highly toxic by intrinsic scorers. The OpenAI max moderation score averages 0.757 and the Perspective API toxicity score averages 0.674.

The content is dominated by dialectal expressions, humour, and identity-referential language that is benign in-community but resembles outgroup toxic markers in generic training data.

Takeaway. Across both extremes, divergence is systematic: lexically "clean" antagonism can provoke strong negative reception while remaining low-scoring intrinsically, whereas culturally situated profanity and in-group language can be intrinsically flagged while eliciting no negative reception. This provides concrete validity evidence for CSF's two-axis view: reception-based measurement is complementary to intrinsic toxicity, and helps avoid both missed harms and false alarms in dialect- and context-rich communities.

7 Conclusion

We reframed toxicity as a contextual, interactional phenomenon that arises when a communicative act is perceived to violate local norms and thereby induces stress or disruption in its audience.

We proposed the Contextual Stress Framework (CSF) and operationalised a reception-based signal, PONOS, as the proportion of negative reactions to an utterance. Empirically, we provided validity evidence for this reception axis: we selected an in-domain sentiment instrument for $s(\cdot)$, we showed that PONOS is estimable at inference time (with domain-tuned reaction modelling performing best), and demonstrated systematic construct divergence from intrinsic toxicity instruments. In particular, large and interpretable off-diagonal regions show that reception-based measurement is complementary to intrinsic toxicity, and can help avoid both missed harms and false alarms in dialect- and context-rich communities.

Taken together, these findings support CSF's central claim: intrinsic-toxicity scores and reception-based measurement are related but non-equivalent. They form complementary axes for analysing harm in context.

Future work can extend this approach by developing richer reception signals beyond reply polarity, validating CSF across platforms and cultures, incorporating multimodal context, and studying interventions that use both axes to mitigate harm.

8 Limitations

We acknowledge several important limitations of our work:

- PONOS is not a universal or intrinsic measure of toxicity. It is a context-dependent, reception-driven proxy for perceived social stress. Rather than determining what a message *is*, PONOS captures how it is *received* by a particular audience in a particular setting.
- As such, PONOS is best understood as a complementary signal that augments text-intrinsic toxicity models. It is most informative in settings where audience response is observable and sufficiently large. In Appendix C, we describe extensions such as PONOS-Weighted, PONOS-Net, and PONOS-Early, which account for user influence, network structure, and early-stage reception.
- **Stress is not synonymous with toxicity.** Not all norm violations are harmful (e.g. satire, protest), and not all harm produces measurable stress. Our metric captures perceived disruption, not ethical correctness. It defines toxicity at the intersection of norm violation *and* stress response—not stress alone—to avoid conflating necessary discomfort with communicative harm; however, this separation may not always be obvious.
- **Sentiment classification is an imperfect proxy.** PONOS relies on sentiment classifiers that are known to carry cultural, demographic, and linguistic biases. Their subjectivity introduces noise, especially in edge cases such as sarcasm or coded language.
- **Community-level perception $\neq$ moral truth.** PONOS reflects what provokes disapproval in a particular group, not a universal judgement. What is toxic in one community may be normative in another. This is a strength—but also a limitation.
- **Susceptibility to manipulation.** Like all crowd-based signals, PONOS can be gamed through brigading, reply injection, or sockpuppetry. Mitigating these risks may require extensions such as reply weighting (e.g. verified users, community trust scores, vulnerability-aware weighting) or time-window filters.
- **Blind spots.** PONOS cannot capture silent harm—for example, when content is ignored, downvoted, or drives users away without leaving replies. It may miss damage caused by exclusion or indirect effects.
- **Not a moral compass.** PONOS is not a substitute for ethical judgement. It models perception, not morality, and should be used alongside normative frameworks rather than in place of them.

9 Ethical Considerations

Any metric used in content moderation, filtering, or prioritisation carries normative weight. While we position CSF and PONOS as descriptive tools—reflecting community reactions rather than imposing moral standards—we recognise their potential for misuse or overreach.

Moderation decisions based solely on PONOS scores risk reinforcing existing biases or silencing dissent. High PONOS values should not automatically lead to suppression; rather, they should signal the need for human oversight, contextual analysis, and, where appropriate, dialogic intervention.

We explicitly do not claim that CSF and PONOS reflect objective truth. They reflect perceived social friction—a volatile signal that can be shaped by group norms, visibility effects, and adversarial influence.

Instead, we call for responsible deployment:

- Use CSF and PONOS transparently and explainably.
- Pair them with community input, human review, and ethical guidance.
- Monitor for drift, manipulation, and misalignment with social values.
- Consider their use not only for suppression, but also for norm diagnosis, echo-chamber detection, and tracking semantic change.

Ultimately, we advocate for a shift in perspective: from minimising toxicity to mapping social tension. This enables more pluralistic, adaptive, and humane approaches to language technology—aligned not with abstract ideals, but with the lived experience of language in context.

9.1 Data Collection Ethical Considerations

We complied with Reddit's terms of service and did not collect sensitive personal information. Participants were informed about the task and that the data would be used for research. Annotators were compensated at $12/hour in accordance with Prolific's recommended payment guidelines. Full instructions are provided in Appendix D.

References

Ana Aleksandric, Sayak Saha Roy, Hanani Pankaj, Gabriela Mustata Wilson, and Shirin Nilizadeh. 2024. Users' behavioral and emotional response to toxicity in twitter conversations. *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1):29–42.

Keith Allan and Kate Burridge. 2006. *Forbidden Words: Taboo and the Censoring of Language*. Cambridge University Press, Cambridge, UK.

Luvell Anderson and Michael Barnes. 2023. Hate Speech. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Fall 2023 edition. Metaphysics Research Lab, Stanford University, Stanford.

Francesco Barbieri, Jose Camacho-Collados, Luis Espinosa Anke, and Leonardo Neves. 2020. TweetEval: Unified benchmark and comparative evaluation for tweet classification. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1644–1650, Online. Association for Computational Linguistics.

Lisa Feldman Barrett and W. Kyle Simmons. 2015. Interoceptive predictions in the brain. *Nature Reviews Neuroscience*, 16(7):419–429.

Sergey Berezin, Reza Farahbakhsh, and Noel Crespi. 2023. *No offence, Bert - I insult only humans! Multilingual sentence-level attack on toxicity detection networks*. Association for Computational Linguistics, Singapore.

Sergey Berezin, Reza Farahbakhsh, and Noel Crespi. 2024. Read over the lines: Attacking llms and toxicity detection systems with ascii art to mask profanity. *Preprint*, arXiv:2409.18708.

Sergey Berezin, Reza Farahbakhsh, and Noel Crespi. 2025. The tip of the iceberg: Revealing a hidden class of task-in-prompt adversarial attacks on llms. *Preprint*, arXiv:2501.18626.

Cristina Bicchieri. 2005. *The Grammar of Society: The Nature and Dynamics of Social Norms*. Cambridge University Press.

Alexander Brown. 2017. What is hate speech? part 2: Family resemblances. *Law and Philosophy*, 36:1–53.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Studies in Interactional Sociolinguistics. Cambridge University Press.

Carmen Cervone, Martha Augoustinos, and Anne Maass. 2021. The language of derogation and hate: Functions, consequences, and reappropriation. *Journal of Language and Social Psychology*, 40:80–101.

Edwin Chen. 2022. Holy $#!t: Are popular toxicity models simply profanity detectors? https://www.surgehq.ai/blog/are-popular-toxicity-models-simply-profanity-detectors. Accessed: 2024-02-05.

Robert B. Cialdini, Carl A. Kallgren, and Raymond R. Reno. 1991. A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. volume 24 of *Advances in Experimental Social Psychology*, pages 201–234. Academic Press.

Cirimus. 2024. modernbert-large-go-emotions. `https://huggingface.co/cirimus/modernbert-large-go-emotions`. Accessed: 2025-05-19.

Herbert H. Clark. 1996. *Using Language*. Cambridge University Press, Cambridge, UK.

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

Jean-Marc Dewaele. 2004. The emotional force of swearwords and taboo words in the speech of multilinguals. *Journal of Multilingual and Multicultural Development*, 25(2-3):204–222.

Naiyi Ding, Ying Liu, Weiping Wu, Guoqing Xu, and Xiaolin Zhou. 2016. Emotion modulates early auditory response: Erp evidence from mandarin chinese emotional words. *Neuropsychologia*, 89:326–334.

Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. 2018. Measuring and mitigating unintended bias in text classification. pages 67–73.

Naomi I. Eisenberger, Matthew D. Lieberman, and Kipling D. Williams. 2003. Does rejection hurt? an fmri study of social exclusion. *Science*, 302(5643):290–292.

Jon Elster. 2007. *Explaining Social Behavior: More Nuts and Bolts for the Social Sciences*. Cambridge University Press, Cambridge, UK.

Darja Fišer, Tomaž Erjavec, and Nikola Ljubešić. 2017. Legal framework, dataset and annotation schema for socially unacceptable online discourse practices in Slovene. In *Proceedings of the First Workshop on Abusive Language Online*, pages 46–51, Vancouver, BC, Canada. Association for Computational Linguistics.

Paula et al. Fortuna. 2020. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, Marseille, France. ELRA.

Karl Friston. 2010. The free-energy principle: a unified brain theory? *Nature Reviews Neuroscience*, 11(2):127–138.

Tanmay Garg, Sarah Masud, Tharun Suresh, and Tanmoy Chakraborty. 2023. Handling bias in toxic speech detection: A survey. *ACM Comput. Surv.*, 55(13s).

Giorgos Giannakakis, Dimitris Grigoriadis, Katerina Giannakaki, Olympia Simantiraki, Alexandros Roniotis, and Manolis Tsiknakis. 2019. Review on psychological stress detection using biosignals. *IEEE Transactions on Affective Computing*, PP:1–1.

Nitesh Goyal, Ian D. Kivlichan, Rachel Rosen, and Lucy Vasserman. 2022. Is your toxicity my toxicity? exploring the impact of rater identity on toxicity annotation. *Proc. ACM Hum.-Comput. Interact.*, 6(CSCW2).

Jesse Graham, Jonathan Haidt, Sena Koleva, Matt Motyl, Ravi Iyer, Sean P. Wojcik, and Peter H. Ditto. 2013. Chapter two - moral foundations theory: The pragmatic validity of moral pluralism. volume 47 of *Advances in Experimental Social Psychology*, pages 55–130. Academic Press.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning - Volume 70*, ICML'17, page 1321–1330. JMLR.org.

Ivan Hare and James Weinstein. 2009. *Extreme Speech and Democracy*. Oxford University Press.

Michael Herz and Peter Molnár, editors. 2012. *The Content and Context of Hate Speech: Rethinking Regulation and Responses*. Cambridge University Press.

Judith T Irvine. 2001. *" Style" as distinctiveness: the culture and ideology of linguistic differentiation*. na.

Amir Reza Jafari, Guanlin Li, Praboda Rajapaksha, Reza Farahbakhsh, and Noel Crespi. 2023. Fine-grained emotions influence on implicit hate speech detection. *IEEE Access*, 11:105330–105343.

Jigsaw. 2024. Perspective api. `https://www.perspectiveapi.com/`. Accessed: 2025-05-11.

Urja Khurana, Ivar Vermeulen, Eric Nalisnick, Marloes Van Noorloos, and Antske Fokkens. 2022. Hate speech criteria: A modular approach to task-specific hate speech definitions. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 176–191, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Svetlana Kiritchenko, Isar Nejadgholi, and Kathleen Fraser. 2021. Confronting abusive language online: A survey from the ethical and human rights perspective. *Journal of Artificial Intelligence Research*, 71:431–478.

Ian R. Kleckner, Jiahe Zhang, Alexandra Touroutoglou, Lorena Chanes, Chenjie Xia, W. Kyle Simmons, Karen S. Quigley, Bradford C. Dickerson, and Lisa Feldman Barrett. 2017. Evidence for a large-scale brain system supporting allostasis and interoception in humans. *Nature Human Behaviour*, 1:0069.

Richard S. Lazarus. 1991. *Emotion and Adaptation*. Theories of Emotion. Oxford University Press, New York.

Pei Liang, Xiaoxia Wang, Ying Jin, Lili Li, Lei Cui, and Hui Zhang. 2018. Contextual valence and arousal modulate the neural processing of emotional words: Erp evidence from a priming paradigm. *Neuropsychologia*, 117:71–81.

Matthew D. Lieberman and Naomi I. Eisenberger. 2009. Neuroscience. pains and pleasures of social life. *Science*, 323(5916):890–891.

Haochen Liu, Wei Jin, Hamid Karimi, Zitao Liu, and Jiliang Tang. 2021. The authors matter: Understanding and mitigating implicit bias in deep text classification. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 74–85, Online. Association for Computational Linguistics.

AI @ Meta Llama Team. 2024. The llama 3 herd of models. A detailed contributor list can be found in the appendix of this paper.

Sam Lowe. 2023. roberta-base-go_emotions. `https://huggingface.co/SamLowe/roberta-base-go_emotions`. Accessed: 2025-05-20.

Kosisochukwu Madukwe, Xiaoying Gao, and Bing Xue. 2020. In data we trust: A critical analysis of hate speech detection datasets. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 150–161, Online. Association for Computational Linguistics.

Bruce S. McEwen. 1998. Protective and damaging effects of stress mediators. *New England Journal of Medicine*, 338(3):171–179.

Stefano Menini, Alessio Palmero Aprosio, and Sara Tonelli. 2021. Abuse is contextual, what about nlp? the role of context in abusive language annotation and detection. *arXiv preprint arXiv:2103.14916*.

Khouloud Mnassri, Reza Farahbakhsh, Razieh Chalehchaleh, Praboda Rajapaksha, Amir Reza Jafari, Guanlin Li, and Noel Crespi. 2024. A survey on multi-lingual offensive language detection. *PeerJ Computer Science*, 10:e1934. Accessed: 2025-05-11.

OpenAI. 2024a. Gpt-4o system card. https://cdn.openai.com/gpt-4o-system-card.pdf. Accessed: 2025-05-11.

OpenAI. 2024b. Moderation. https://platform.openai.com/docs/guides/moderation. Accessed: 2025-05-11.

Bhikhu Parekh. 2012. *Is There a Case for Banning Hate Speech?* Cambridge University Press, Stanford.

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, Online. Association for Computational Linguistics.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Maarten Sap, Swabha Swayamdipta, Laura Vianna, Xuhui Zhou, Yejin Choi, and Noah A Smith. 2021. Annotators with attitudes: How annotator beliefs and identities bias toxic language detection. *arXiv preprint arXiv:2111.07997*.

Robert M Sapolsky. 2004. *Why zebras don't get ulcers: The acclaimed guide to stress, stress-related diseases, and coping*.

Erkin Sarı, Emine Yücel, and Mehmet Fatih Bükün. 2024. The role of emotions in collective responses to in-group norm violations: The case of university's sensitivity to the natural environment norm. *Current Psychology*, 43:27187–27206.

Fred Shaffer, Rollin McCraty, and Christopher L. Zerr. 2014. A healthy heart is not a metronome: an integrative review of the heart's anatomy and heart rate variability. *Frontiers in Psychology*, 5:1040.

Nandita Sharma and Tom Gedeon. 2012. Objective measures, sensors and computational techniques for stress recognition and classification: A survey. *Computer methods and programs in biomedicine*, 108(3):1287–1301.

George M. Slavich and Steve W. Cole. 2013. The emerging field of human social genomics. *Clinical Psychological Science*, 1(3):331–348.

Dan Sperber and Deirdre Wilson. 1986/1995. *Relevance: Communication and Cognition*. Blackwell, Oxford.

Marijn E. Struiksma, Hannah N. M. De Mulder, and Jos J. A. Van Berkum. 2022a. Do people get used to insulting language? *Frontiers in Communication*, Volume 7 - 2022.

Marijn E Struiksma, Hannah NM De Mulder, and Jos JA Van Berkum. 2022b. Do people get used to insulting language? *Frontiers in Communication*, 7:910023.

Julian F. Thayer, Fredrik Åhs, Mats Fredrikson, John J. Sollers, and Tor D. Wager. 2012. A meta-analysis of heart rate variability and neuroimaging studies: implications for heart rate variability as a marker of stress and health. *Neuroscience & Biobehavioral Reviews*, 36(2):747–756.

Elizabeth Closs Traugott and Richard B. Dasher. 2001. *Regularity in Semantic Change*. Cambridge Studies in Linguistics. Cambridge University Press.

unslothai. 2025. unslothai: Thai natural language processing library. Accessed: 2025-05-16.

Mitch van Geel, Paul Vedder, and Jenny Tanilon. 2014. Relationship Between Peer Victimization, Cyberbullying, and Suicide in Children and Adolescents: A Meta-analysis. *JAMA Pediatrics*, 168(5):435–442.

Gerben A. van Kleef. 2024. Bottom-up influences on social norms: How observers' responses to transgressions drive norm maintenance versus change. *Current Opinion in Psychology*, 60:101919.

Gerben A. van Kleef and Agneta H. Fischer. 2015. Emotions and social norms: Insights from social psychology. In Yuri L. Hanin, editor, *Emotion and Decision Making Explained*, pages 59–75. Routledge.

Bertie Vidgen and Leon Derczynski. 2021. Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):1–32.

Bertie et al. Vidgen. 2019. Challenges and frontiers in abusive content detection. In *Proceedings of the Third Workshop on Abusive Language Online*, Florence, Italy. Association for Computational Linguistics.

Kathryn A. Woolard and Bambi B. Schieffelin. 1994. Language ideology. *Annual Review of Anthropology*, 23:55–82.

Alexandros Xenos, John Pavlopoulos, and Ion Androutsopoulos. 2021. Context sensitivity estimation in toxicity detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 140–145, Online. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, et al. 2025. Qwen3 technical report. https://arxiv.org/abs/2505.09388. Accessed: 10 May 2025.

A Detailed Basis for CSF

To move beyond the limitations of current toxicity detection approaches, we synthesise insights from multiple disciplines to develop a framework for understanding verbal harm.

A.1 Personal level

The experience of verbal harm is fundamentally affective, rooted in how individuals perceive threats to personal and social stability. Research in psychology and neuroscience consistently shows that such harm does not arise from the intrinsic properties of language alone, but from context-dependent appraisals, emotional expectations, and violations of predictive frameworks.

In psychology, appraisal theory (Lazarus, 1991) posits that emotional and stress responses arise when individuals evaluate events as threatening to their goals, values, or social standing. From this perspective, communication becomes harmful not solely because of its content, but because it is perceived as violating significant personal concerns.

Research on social pain further demonstrates that symbolic threats can activate biological stress systems in much the same way as physical injury (Eisenberger et al., 2003; Lieberman and Eisenberger, 2009).

Neuroscientific findings indicate that the brain processes language in an anticipatory, predictive, and context-sensitive manner. Predictive coding models (Friston, 2010; Kleckner et al., 2017; Barrett and Simmons, 2015) suggest that the brain continuously generates expectations about emotional and social states. When communication violates these expectations, it triggers a stress response.

Physiological indicators of stress, such as heart rate variability, are influenced by exposure to social threat (Shaffer et al., 2014). Event-related potential studies (Ding et al., 2016; Liang et al., 2018) show that negative emotional language disrupts neural integration, while chronic exposure to social stressors is associated with pro-inflammatory genetic activity (Slavich and Cole, 2013).

Together, this body of research demonstrates that responses to harmful or toxic speech are not merely psychological but also physiological.

A.2 Group level

Beyond direct victimisation, witnessing harm inflicted on others can also elicit stress responses, signalling potential threats within the social environment (Struiksma et al., 2022a).

Sociology and social psychology demonstrate that responses to verbal behaviour are governed by group norms, emotional expectations, and mechanisms of norm enforcement (Elster, 2007). Research by Bicchieri (2005) and Cialdini et al. (1991) shows that social norms are maintained through collective emotional and behavioural reactions to deviance. Importantly, norm violations are not evaluated in isolation, but interpreted through the emotional dynamics of the group, which shape perceptions of harm.

The concepts of emotional collectives (van Kleef and Fischer, 2015) and emotional norm-shaping processes (van Kleef, 2024; Sarı et al., 2024) further emphasise how shared emotional responses reinforce or reshape social expectations in real time.

Moral Foundations Theory similarly highlights that emotional reactions to norm violations differ across groups and cultures, leading to divergent judgements about what constitutes harmful or toxic speech (Graham et al., 2013).

From this perspective, toxicity functions as a social signal: a disruption in the moral and emotional fabric of the group.

A.3 Cultural and Linguistic Level

Linguistic and sociolinguistic research demonstrates that meaning is co-constructed in interaction and shaped by contextual framing. Harmful speech cannot be reduced to individual words or expressions divorced from their social use.

Pragmatics (Clark, 1996; Sperber and Wilson, 1986/1995) emphasises that meaning is not fully encoded in words, but inferred through shared background knowledge and conversational context. An identical phrase may be benign in one setting and deeply offensive in another, depending on tone, power relations, and prior discourse.

Politeness and impoliteness theory (Brown and Levinson, 1987; Culpeper, 1996) similarly show that perceived rudeness, offence, or aggression are emergent properties of interaction rather than intrinsic features of utterances.

Language ideologies (Woolard and Schieffelin, 1994; Irvine, 2001) further demonstrate that judge-

ments about what counts as "toxic" or "neutral" language are shaped by power relations and dominant cultural narratives.

Historical linguistics and semantics highlight the instability of meaning across time and context (Allan and Burridge, 2006; Traugott and Dasher, 2001). Words such as "nice" (originally meaning "foolish") or "fag" (benign in British English but a slur in American usage) illustrate how emotional and moral associations shift across cultural and historical settings, undermining any fixed, universal standard of toxicity.

Finally, **multilingualism and language variation** introduce further complexity. Emotional resonance and perceived offensiveness vary across languages due to differing socialisation patterns (Dewaele, 2004). What is taboo or offensive in one language may carry no such connotation in another.

Taken together, these perspectives argue strongly against treating harmful language as a static or text-intrinsic phenomenon. Toxicity cannot be reliably detected by identifying “bad words” in isolation, but only by modelling how language interacts with contextual, social, and cultural expectations.

B Details on Understanding of Context

Context structures interpretation. It governs what is considered appropriate, offensive, or harmful within a given setting. To support operational clarity, we distinguish four key dimensions:

Verbal context refers to the surrounding linguistic material — what is said before, after, or alongside the utterance. It determines how meaning is disambiguated, particularly in cases of sarcasm, reappropriation, or coded language.

Situational context includes the physical, social, and institutional setting of communication: who is speaking, to whom, in what place, and under what conditions. Normative expectations can vary dramatically between settings — for example, language deemed acceptable at a music festival may be inappropriate in a courtroom or classroom. This dimension also encompasses digital and symbolic environments such as online platforms, workplace hierarchies, or subcultural spaces.

Cognitive context refers to the mental and emotional states of the communicative participants — including intentions, assumptions, prior knowledge, and emotional stance. Many speech acts, such as jokes, flirtation, or irony, can only be correctly interpreted when cognitive context is inferred or shared.

Cultural context involves the broader moral frameworks, communicative conventions, and ideological norms that vary across communities. A behaviour considered respectful in one culture may be viewed as inappropriate or even threatening in another. For example, direct individualistic expression is normative in many Western contexts, but may be perceived as disruptive or disrespectful in collectivist settings.

Given this multidimensional variability, any model of toxicity detection must be contextually grounded and culturally adaptive. Static or universalist rule sets will inevitably misclassify utterances as they fail to consider how meaning emerges within interaction.

C PONOS Variants

C.1 Extensions and Variants

PONOS is a flexible framework that admits multiple extensions to better reflect community dynamics, reply timing, and user context.

PONOS-Net. We define the **net sentiment** variant as:

$$\text{PONOS}_{\text{net}}(x) = \frac{1}{|R(x)|} \sum_{r_i \in R(x)} s(r_i) \quad (2)$$

This version captures the overall affective balance of replies, distinguishing polarised threads (with both positive and negative responses) from uniformly negative ones.

PONOS-Weighted. To account for user-level features, such as author credibility, social influence, or vulnerability - we define a weighted version:

$$\text{PONOS}_{\text{w}}(x) = \frac{1}{\sum_i w(r_i)} \sum_{r_i \in R(x)} w(r_i) \cdot \mathbb{I}[s(r_i) = -1] \quad (3)$$

Here, $w(r_i)$ is a weight assigned to each reply, reflecting community-specific salience (e.g., upvotes, karma, user type).

PONOS-Early. To mitigate feedback loop effects (e.g., reply priming), we define an early-window variant that considers only the first $k \ll n$ replies:

$$\text{PONOS}_{\text{early}}(x) = \frac{1}{k} \sum_{i=1}^{k} \mathbb{I}[s(r_i) = -1] \quad (4)$$

This variant reflects the immediate, unprimed perception of a post, prior to potential community escalation or reinforcement.

C.2 Implementation Considerations

The sentiment function $s(\cdot)$ can be realised through multiple approaches:

- Manual annotation by culturally aligned raters;
- Off-the-shelf sentiment classifiers (e.g., RoBERTa, BERTweet, GPT-based);
- Community-specific fine-tuned models.

We emphasize that the interpretability and validity of PONOS depend on the alignment between the sentiment model and the communicative norms of the target community. In our experiments, we rely on human-annotated sentiment labels sourced from subreddit-native annotators to ensure cultural fidelity.

D Assessors' Instructions

```
Dear Participant,

Thank you for taking the time to
    participate in our survey. Your
    input is invaluable to us. In this
    survey, you will be asked to
    classify reactions expressed in
    replies to comments from discussions
     on the Reddit community.

Task:
- Given a comment and its replies from a
     discussion on the Reddit community
    r/BlackPeopleTwitter, classify the
    reaction expressed in each reply as
    one of the following:
  - "approval (comment)" - Agreement/
      support of the comment itself.
  - "approval (subject)" - Agreement/
      support of the subject mentioned
      in the comment.
  - "neutral" - No strong reaction for
      or against the comment or subject.
  - "condemnation (comment)" -
      Disapproval of the comment itself.
  - "condemnation (subject)" -
      Disapproval of the subject
      mentioned in the comment.
- Be sure to identify whether approval/
    condemnation is directed at the
    comment or the subject.

Examples:

Topic: "New office policy - no phones
    allowed"
Comment: "The new policy is a disaster,
    it's making everything worse."

1. Reply: "I totally agree, it's a
    complete mess."
   - Classification: approval (comment)

2. Reply: "I think we should give it
    some time to see the effects."
   - Classification: neutral

3. Reply: "This policy is the worst
    decision ever made by this
    administration."
   - Classification: condemnation (
      subject)

4. Reply: "I don't have a strong opinion
     on this."
   - Classification: neutral

5. Reply: "LOL, shut up Karen, nobody
    cares what you think."
   - Classification: condemnation (
      comment)

6. Reply: "I kinda like the new policy."
   - Classification: approval (subject)

By following these instructions, you
    will help us understand the
    community dynamics and how different
     reactions are expressed within the
    r/BlackPeopleTwitter community.
Your participation is crucial for the
    success of this research. Thank you!
```

E Prompts

Sentiment classification prompt:

```
You are a sentiment analysis assistant
    trained to understand African
    American Vernacular English (AAVE)
    and online Black community discourse
    .

Your task is to classify the emotional
    tone of REPLY message as Positive,
    Neutral, or Negative.
Your analysis supports a larger
    framework for detecting *perceived
    toxicity* - understanding how it
    might cause stress or social tension
     within the community.

Use all available context:
- Topic: the post that started the
    thread
- Comment: the message being replied to
- Reply: the specific message to analyze

Consider slang, tone, cultural style,
    and conversational intent.

- Positive: Agreement, Pride, joy,
    affirmation, playfulness, flirting,
    or supportive sarcasm. Being angry
    TOGETHER against something also goes
     here.
- Neutral: Factual comments,
    observations, dry or ambiguous tone,
```

```
     rhetorical questions.
- Negative: Disagreement, confrontation,
     arguing, complaints, passive
    aggression, warnings, critique, or
    social "shade."

Always output one of: 'Positive', '
    Neutral', or 'Negative' without
    explanation. Be sure to not miss any
     'Negative'.
Format your answer as "ANSWER:"

FEW-SHOT EXAMPLES:

    Topic: Trump won
    Comment:" I think this is ok as long
         as we are not getting democrat
        president
    Reply: Nigga this is a HILARIOUS
        take
    ANSWER: Negative

    Topic: New record lol
    Comment: That clip had me dying, fr.
    Reply: Ayo I been laughing for 10
        minutes, why nobody brings up
        that he claimed to be THE MASTER
    ANSWER: Positive

    Topic: Gotta do whacha gotta do
    Comment: Once I was texting with
        girl for 6 months to get a
        picture from her
    Reply: Bro, I've ended up writing
        her real letters. desparate shit
    ANSWER: Neutral

    Topic: Economy is fucked
    Comment: I finna strat looting just
        to get food on the table
    Reply: they are making us do it
    ANSWER: Positive
```

Reaction prediction prompt:

```
        {"role": "system", "content": "
            You are a culturally aware
            assistant who creates
            authentic replies from r/
            BlackPeopleTwitter."},
        {"role": "user", "content": (
            "Below is a Reddit comment.
                Generate the top 5
                likely replies from r/
                BlackPeopleTwitter, "
            "using authentic tone,
                cultural slang, and
                humor. Separate replies
                with <|reply|>.\n\n"
            f"### Comment:\n{comment}\n\
                n### Replies:"
```