Who Decides What's Toxic? How Player Identity Shapes Social Perception of Toxicity in Online Game Communities

Anonymous submission

Abstract

How people perceive the same online discourse as toxic or harmless depends not only on the content itself but on the perceiver's social group membership. We propose a study examining how *player identity*—allegiance to rival online game communities—shapes the social perception of toxicity in gaming discourse. Using posts from competing fan communities on Reddit and Bilibili, we plan to collect toxicity ratings from annotators belonging to opposing player factions. We hypothesize that identical posts will be perceived as significantly more toxic by out-group members than by in-group members, particularly for discourse that functions as identity performance (faction-specific nicknames, competitive banter, ironic insults). Drawing on Social Identity Theory, we frame toxicity perception as an instance of intergroup evaluative judgment, where the same linguistic behavior is interpreted through the lens of group membership. We outline a multilingual annotation study and propose computational models that condition on perceiver identity, arguing that modeling *who is perceiving* is essential for any NLP system that aims to capture the social dynamics of online toxicity.

Introduction

The perception of online toxicity is inherently social. Whether a given utterance is perceived as offensive, playful, or neutral depends not only on its lexical content but on the social relationship between the speaker and the perceiver (Sap et al. 2022; Kiritchenko et al. 2021). Research in social psychology has long established that identical behaviors are evaluated differently depending on whether the actor belongs to one's in-group or out-group (Tajfel and Turner 1979)—a phenomenon with direct implications for how toxicity is perceived, reported, and modeled in NLP systems.

Despite this, most computational approaches to toxicity detection treat the task as objective classification, training models on labels aggregated across annotators with diverse social positions (Plank 2022). Recent work has begun to examine how annotator demographics—including political orientation, racial identity, and gender—shape toxicity judgments (Sap et al. 2022; Mostafazadeh Davani et al. 2022). However, these studies focus primarily on broad social categories. We argue that *subcultural group membership*—specifically, allegiance to a particular online game community—offers an especially clear and tractable case for studying how social perception modulates toxicity judgments.

Competitive game ecosystems such as *Genshin Impact* vs. *Wuthering Waves*, or *League of Legends* vs. *Dota 2*, generate rival fan communities with distinct discursive norms. Within these communities, teasing labels, ironic insults, and faction-specific slang function simultaneously as identity markers and perceived slurs. For instance, in Chinese-language gaming discourse, terms like "*Mi Xuezhang*" (for *Genshin Impact* fans) and "*Ku Xuedi*" (for *Wuthering Waves* fans) carry valences that shift depending on who is speaking and who is interpreting. This creates a **natural laboratory** for studying how social group perception shapes language evaluation: player faction membership is categorical, self-selected, and directly relevant to the discourse being judged.

Study Design

Data Collection. We will collect 600–800 posts from rival game community forums on Reddit (English) and Bilibili (Chinese) containing inter-community references, factional rhetoric, competitive banter, or cross-community teasing. The sample will span a spectrum from clearly benign through ambiguous to clearly toxic, with emphasis on the ambiguous middle ground where social perception is most likely to diverge.

Perception Elicitation. Two groups of annotators (15–20 per group) will be recruited from opposing player communities via Prolific (English) and Wenjuanxing (Chinese). Each annotator rates each post on a 4-point scale: *not toxic*, *mildly toxic*, *toxic*, *severely toxic*. Critically, annotators also report their primary game allegiance, community engagement level, and years of participation. A set of control posts unrelated to game rivalries will serve as calibration items to establish each annotator's baseline strictness.

Analysis. Our analysis targets three questions: *(1) Does group membership predict perception divergence?* We compute inter-group vs. intra-group agreement using Krippendorff's α, expecting significantly lower agreement across factions than within factions. *(2) Which discourse types are most perception-sensitive?* Using CrowdTruth metrics (Dumitrache et al. 2018), we identify items with maximal factional disagreement and qualitatively analyze their discourse features (e.g., ironic faction labels, competitive ban-

ter, memes, direct insults). *(3) Does engagement depth amplify perceptual bias?* We test whether more deeply embedded community members (self-reported "hardcore" players) show greater in-group favorability in toxicity ratings than casual participants.

Perception-Aware Modeling. We compare three classifiers: (a) a **majority-vote baseline** trained on aggregated labels; (b) a **multi-annotator model** preserving individual ratings (Mostafazadeh Davani et al. 2022); and (c) a **perception-conditioned model** that receives annotator group identity as an input feature. We expect the perception-conditioned model to outperform aggregation on posts where social perception diverges across factions.

Theoretical Framework

Our study is grounded in **Social Identity Theory** (Tajfel and Turner 1979), which predicts that individuals evaluate in-group and out-group behaviors asymmetrically. Applied to language perception, this means identical discourse may be perceived as "playful banter" when attributed to an in-group member but "toxic provocation" when attributed to an out-group source. This asymmetry is not a failure of perception but a systematic consequence of social categorization.

We further draw on research in **language attitudes and sociolinguistic perception** (Eckert 2006), which demonstrates that language evaluation is never purely linguistic but always socially situated. Rival game communities develop distinct registers where shared vocabulary acquires community-specific connotations. The "meaning" of a faction label is not fixed in the text but emerges from the perceiver's social position—precisely the kind of phenomenon that NLPercep aims to model.

This framing has implications beyond gaming. Social media toxicity detection, content moderation, and platform safety systems all implicitly assume a universal standard for what counts as toxic. Our study provides empirical evidence that this assumption is untenable for discourse embedded in intergroup dynamics, and that perceiver identity must be modeled explicitly.

Expected Contributions

(1) Social group perception as a predictor of toxicity evaluation. We introduce game community allegiance as a novel social variable that shapes language perception, distinct from standard demographic predictors and directly relevant to the discourse under evaluation.

(2) Multilingual evidence for perception-dependent toxicity. By collecting data from both English and Chinese platforms, we provide cross-linguistic evidence that social perception effects on toxicity evaluation are not language-specific but reflect universal intergroup dynamics.

(3) Perception-aware computational models. We demonstrate that conditioning toxicity classifiers on perceiver identity improves performance for socially contested discourse, arguing that *who perceives* is as important as *what is said*.

References

Mostafazadeh Davani, A.; Díaz, M.; and Prabhakaran, V. 2022. Dealing with Disagreements: Looking Beyond the Majority Vote in Subjective Annotations. *Transactions of the ACL* 10:92–110.

Dumitrache, A.; Aroyo, L.; and Welty, C. 2018. Crowdsourcing Semantic Label Propagation in Relation Classification. In *Proc. 1st Workshop on Fact Extraction and VERification (FEVER)*, 16–21. ACL.

Eckert, P. 2006. Communities of Practice. In *Encyclopedia of Language and Linguistics*, 683–685. Elsevier.

Kiritchenko, S.; Nejadgholi, I.; and Fraser, K. C. 2021. Confronting Abusive Language Online: A Survey from the Ethical and Human Rights Perspective. *Journal of Artificial Intelligence Research* 71:431–478.

Plank, B. 2022. The "Problem" of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. In *Proc. EMNLP*, 10671–10682. ACL.

Sap, M.; Swayamdipta, S.; Vianna, L.; Zhou, X.; Choi, Y.; and Smith, N. A. 2022. Annotators with Attitudes: How Annotator Beliefs and Identities Bias Toxic Language Detection. In *Proc. NAACL*, 5884–5906. ACL.

Tajfel, H.; and Turner, J. C. 1979. An Integrative Theory of Intergroup Conflict. In Austin, W. G.; and Worchel, S., eds., *The Social Psychology of Intergroup Relations*, 33–47. Brooks/Cole.