4chan Toxicity Analysis
A Comparative Evaluation of OpenAI Moderation API and Google Perspective API

Dhinesh Sadhu Subramaniam Ponnarasan
State University of New York at Binghamton
dsadhusubram@binghamton.edu

Abstract

This project evaluates automated toxicity detection performance by comparing two widely used systems: OpenAI's Moderation API and Google's Perspective API. We collected a dataset of 1,352 posts from 4chan's "/pol/" board, processed them through both APIs, and conducted statistical analyses including correlation, agreement/disagreement checks, and independent t-tests. Results show a moderate correlation ($r = 0.337$) between the APIs' toxicity measures, with agreement in 74.33% of cases. However, systematic disagreements (25.67%) reveal sensitivity differences: OpenAI tends to flag more violence- and harassment-related content, while Perspective detects more general and subtle toxicity. These findings highlight the complementary nature of both APIs and support the case for multi-API moderation strategies.

Keywords: Automated content moderation, Toxicity detection, OpenAI Moderation API, Google Perspective API, Comparative analysis, Online communities, Harassment detection, Violence detection, Hate speech, Multi-API moderation

Code — DhineshPonnarasan/4chan-toxicity-analysis

1 Executive Summary

Automated moderation tools are essential for managing large-scale online discussions. This report presents a comparative analysis between:

- **OpenAI Moderation API** — category-specific toxicity scores (violence, sexual, harassment, hate, self-harm).
- **Google Perspective API** — attribute-based toxicity probabilities (TOXICITY, INSULT, THREAT, etc.).

Key findings:

- Moderate correlation between OpenAI violence and Perspective toxicity ($r \approx 0.34$).
- 74.33% agreement in binary classification.
- Significant disagreement (~26%), reflecting category sensitivity differences.
- OpenAI stricter on violence/harassment; Perspective broader on subtle toxicity.

2 Methodology

2.1 Data Collection

A Python pipeline (`main.py`) scraped 4chan "/pol/" posts:

- Posts saved in `posts.jsonl` with metadata (thread ID, timestamp).
- Deduplication ensured unique entries.
- Final dataset: 1,352 posts.

2.2 Toxicity Detection

- **OpenAI Moderation API**: Queried with

`omni-moderation-latest`, returned category scores and binary flags.

- **Google Perspective API**: Queried with TOXICITY attribute, returned probability scores.

2.3 Analysis

The analysis script (`plot.py`) implemented:

1. Pearson correlation between OpenAI violence and Perspective toxicity.
2. Binary agreement (threshold = 0.5).
3. Category score distributions.
4. Independent t-tests.
5. Scatterplots and bar plots for visual comparison.

3 Results

3.1 Overall Metrics

- Correlation: $r = 0.337$ ($p < 0.001$).
- Agreement: 74.33% (1005/1352 posts).
- Disagreement: 25.67% (347 posts).
- Statistical test: $t = -17.076$, $p < 0.001$.

3.2 Category Trends

Table 1: Mean OpenAI Category Scores

Category	Mean Score
Harassment	0.305
Hate	0.147
Violence	0.063
Sexual	0.036
Self-harm	0.012

3.3 Visualizations

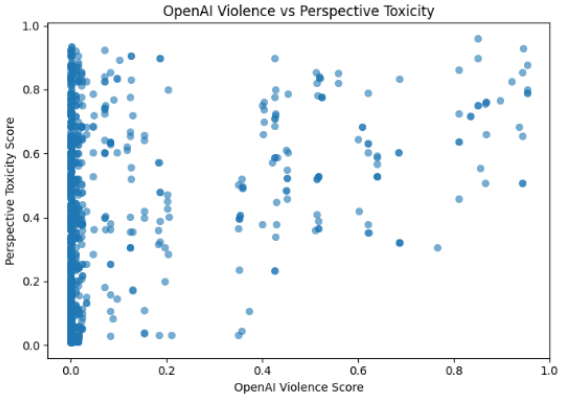

Fig 1: Scatterplot of OpenAI violence scores vs. Per- spective toxicity scores.

Fig 1, shows the relationship between OpenAI violence scores and Perspective toxicity scores. Most posts cluster at low values, while divergence appears in higher-toxicity regions.

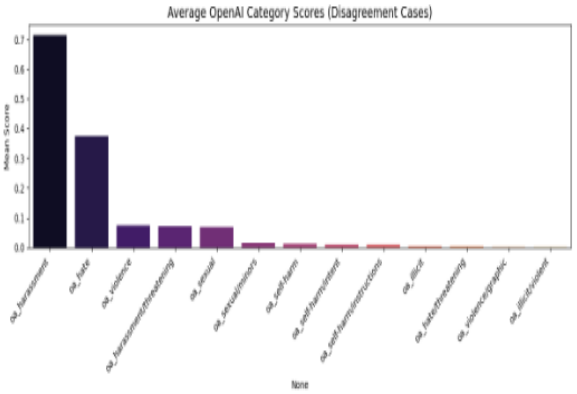

Fig 2: Average OpenAI scores for disagreement cases (where APIs disagreed).

Fig 2, illustrates the average OpenAI category scores across all posts. Harassment, hate, and violence emerge as the most frequently detected categories.

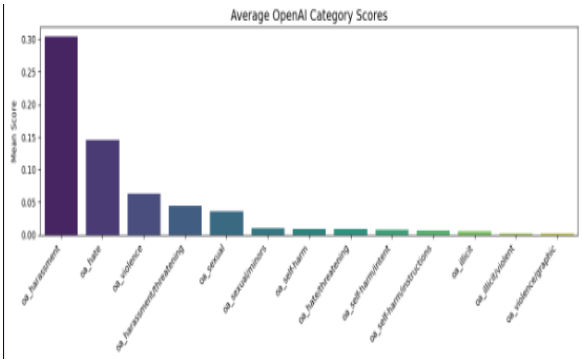

Fig 3: Average OpenAI category scores across the database

Fig 3, displays average OpenAI category scores for posts where the two APIs disagreed. Harassment dominates these disagreement cases, highlighting sensitivity differences., displays average OpenAI category scores for posts where the two APIs disagreed. Harassment dominates these disagreement cases, highlighting sensitivity differences.

4 Discussion and Implications

Complementary strengths: OpenAI excels in identifying harassment and violence; Perspective better captures general toxicity and subtle insults.

Implications:

- Multi-API integration improves robustness.
- Platform-specific tuning (e.g., harassment detection in gaming vs. broad toxicity in forums).
- Threshold calibration is critical in real deployments.

5 Transparency Statement

Generative AI tools supported this project:

- ChatGPT (OpenAI GPT-5) assisted in code drafting, debugging JSON handling, improving plots and structuring the report.

6 Acknowledgments

I thank OpenAI and Jigsaw for making their moderation APIs accessible for research.

References

Antonis Papasavva et al. (2020) – *"Raiders of the Lost Kek: 3.5 Years of Augmented 4chan Posts from the Politically Incorrect Board"*, ICWSM. Offers a massive dataset (3.3M threads, 134M posts) from 4chan's /pol/, annotated with toxicity scores.

MH Saeed et al. (2024) – *"Attributing Coordinated Hate Attacks on YouTube Videos…"*, ICWSM. Leverages data from 4chan's /pol/ to trace coordinated hate propagation across platforms.

Jennifer C. (2022) – *"Dissecting Automated Hateful Meme Detection Through the [Multimodal V&L Models]…"*, ICWSM Workshop. Investigates hateful memes from 4chan using multimodal classifiers

Mark Warner et al. (2025) – *"A Critical Reflection on the Use of Toxicity Detection…"*. A reflection on proactive moderation tools, design contexts, and socio-technical implications from an HCI perspective.

A Aleksandric et al. (2024) – *"Users' Behavioral and Emotional Response to Toxicity in…"*, ICWSM. Examines emotional impacts (anger, anxiety) on users receiving toxic replies.

A Papasavva et al. (2020) – *"Raiders of the Lost Kek…"*, ICWSM. (Same as citation 1, referenced again for emphasis.)

M Singhal (2023) – *"SoK: Content Moderation in Social Media, from Guidelines…"*, ICWSM. Covers moderation guidelines and even community moderation norms, including 4chan.

A Rajadesingan et al. (2020) – *"How Distinctive Toxicity Norms Are Maintained in Political…"*, ICWSM. Analyzes how toxicity norms vary and persist in political subreddits.

M Warner (2025) – *"A Critical Reflection on the Use of*

Toxicity Detection…" (same as citation 4; restated for clarity).

And We Will Fight for Our Race (2023) – *ICWSM measurement focusing on genetic testing discussions on Reddit and 4chan.* Highlights cross-community toxicity patterns