Urban Dictionary Brand Safety Specification

Executive Summary

Urban Dictionary is the world's largest crowdsourced slang dictionary, serving over 75 million monthly visitors and defining contemporary language and culture.

What we do: We screen every page for brand safety using AI, assigning each page a safety grade (A–D) aligned with Google advertiser policy. This protects advertisers from appearing next to harmful content while preserving authentic cultural expression.

What makes us different: Unlike legacy keyword filters that often flag harmless identity mentions as unsafe, our model was specifically chosen to reduce false positives on minority identity references — protecting both advertisers and marginalized voices.

Key features: Transparent methodology, 15+ harm categories, customizable safety tiers, and audit tools for brand safety teams.

Overview

Designed for compliance with brand safety standards — and built to protect marginalized voices.

Urban Dictionary applies a multi-layered moderation and classification system to ensure that ads appear only alongside suitable content. This document details the brand safety model used in our ad platform.

Model & Methodology

We use **unbiased-toxic-roberta**, a variant of RoBERTa fine-tuned to detect toxic, offensive, and harmful language with reduced bias against minority communities. The model is deployed via Detoxify, a widely-used ML library for content moderation.

How scoring works: The model classifies each published definition across 15+ harm categories and produces probability scores per category — a score between 0 and 1 indicating the likelihood of harmful content. For example:

- 0.0–0.2: Very low risk (unlikely to contain harmful content)
- 0.2–0.5: Low to moderate risk
- 0.5–0.8: High risk
- 0.8–1.0: Very high risk (likely contains harmful content)

For each page, we compute a composite safety grade based on the highest scoring definition on that page. Definitions with high scores in any harm category may lower a page's grade. Thresholds were selected based on internal testing against Google Ads safety guidelines.

Research Background

> **Key Point:** This model emerged from a $65,000 Google-funded research competition explicitly designed to reduce bias in toxicity detection. Urban Dictionary selected one of the top submissions for production use.

Our safety scoring is based on unbiased-toxic-roberta, a machine learning model trained on public data released through a Kaggle competition funded by Jigsaw, an Alphabet subsidiary that leads Google's Conversation AI research.

The training dataset was built from 500,000 real-world public comments labeled for toxicity and identity references by crowdsourced raters. The project's aim was to detect harmful content while reducing unintended bias — especially false positives triggered by benign mentions of identity (e.g. "I am a gay woman").

Key references:

- Kaggle: Jigsaw Unintended Bias in Toxicity Classification
- Paper: Nuanced Metrics for Measuring Unintended Bias in Text Classification
- Model card: unitary/unbiased-toxic-roberta

Jigsaw awarded $65,000 in prizes to spur innovation and open-source bias mitigation. Urban Dictionary selected one of the top-scoring models — fine-tuned for identity fairness — and calibrated its thresholds to align with Google Ads brand safety policy.

Screened Harm Categories

- **Toxicity**: General toxicity, severe toxicity, obscenity, insults
- **Identity & Hate**: Identity attack, race, gender, religion, sexual orientation
- **Harm & Violence**: Threats, sexual explicitness, self-harm, violence, illicit activity, legal risk content

ABCD Safety Grades

Each page receives a grade based on the worst scoring content on that page. Thresholds were calibrated against Google advertiser policies.

Grade	Description	Ad Eligibility
A	Minimal risk across all categories	Eligible for all advertisers
B	Minor concerns in some categories	Eligible for most advertisers
C	Elevated risk or multiple mid-level violations	Eligible for open-tier advertisers only
D	High risk or violates advertiser policies	No ads served

Note: Unreviewed or unanalyzed pages default to Grade D to ensure advertiser protection.

Moderation Workflow

1. **Content Review**: All definitions undergo editorial review for platform compliance.
2. **ML-Based Scoring**: Published content is re-evaluated using the ML model across harm categories.
3. **Safety Grade Assignment**: Worst definition score determines the page grade (A–D).
4. **Default Protections**: Pages lacking safety analysis are excluded from monetization.

Advertiser Targeting Tiers

Tier	Grade Inclusion	Use Case
Strict	A only	Family-friendly brands with low risk tolerance
Balanced	A and B	Standard brand-safe advertisers
Open	A, B, and C	Maximized reach, moderate tolerance

Custom tiering is available: advertisers may define their own thresholds per harm category.

Built for transparency

✓ Open methodology

Our safety system is fully documented — including the model used, scoring thresholds, and harm categories. We're happy to share the underlying code on request.

✓ **Fairness-first design**
We use a model fine-tuned for bias reduction and identity fairness, minimizing false positives for minority voices.

✓ **Customizable defaults**
Our default safety tiers align with Google Ads policies. We support custom thresholds, manual overrides, and CSV exports on request.

Contact

To request more information or a custom brand safety integration, contact ads@urbandictionary.com.

For live scoring examples, customizable filtering, and audit tools, visit urbandictionary.biz/brand-safety