Adapt or Be Outdated: Evolving Implicit Toxicity Datasets

K/DA: Automated Data Generation Pipeline for Detoxifying Implicitly Offensive Language in Korean

Minkyeong Jeon*, Hyemin Jeong*, Yerang Kim, Jiyoung Kim, Jae Hyeon Cho, Byung-Jun Lee

KOREA UNIVERSITY 1905

Project Page

ACL 2025 VIENNA JULY 27 - AUGUST 1

0. Motivation

The challenges of offensive language detoxification

1. Cost-ineffective **human annotation** to build paired data
2. The **rapid evolution** of offensive terms, rendering static datasets quickly outdated.
3. Insufficient paired data for under-resourced languages

1. Overview

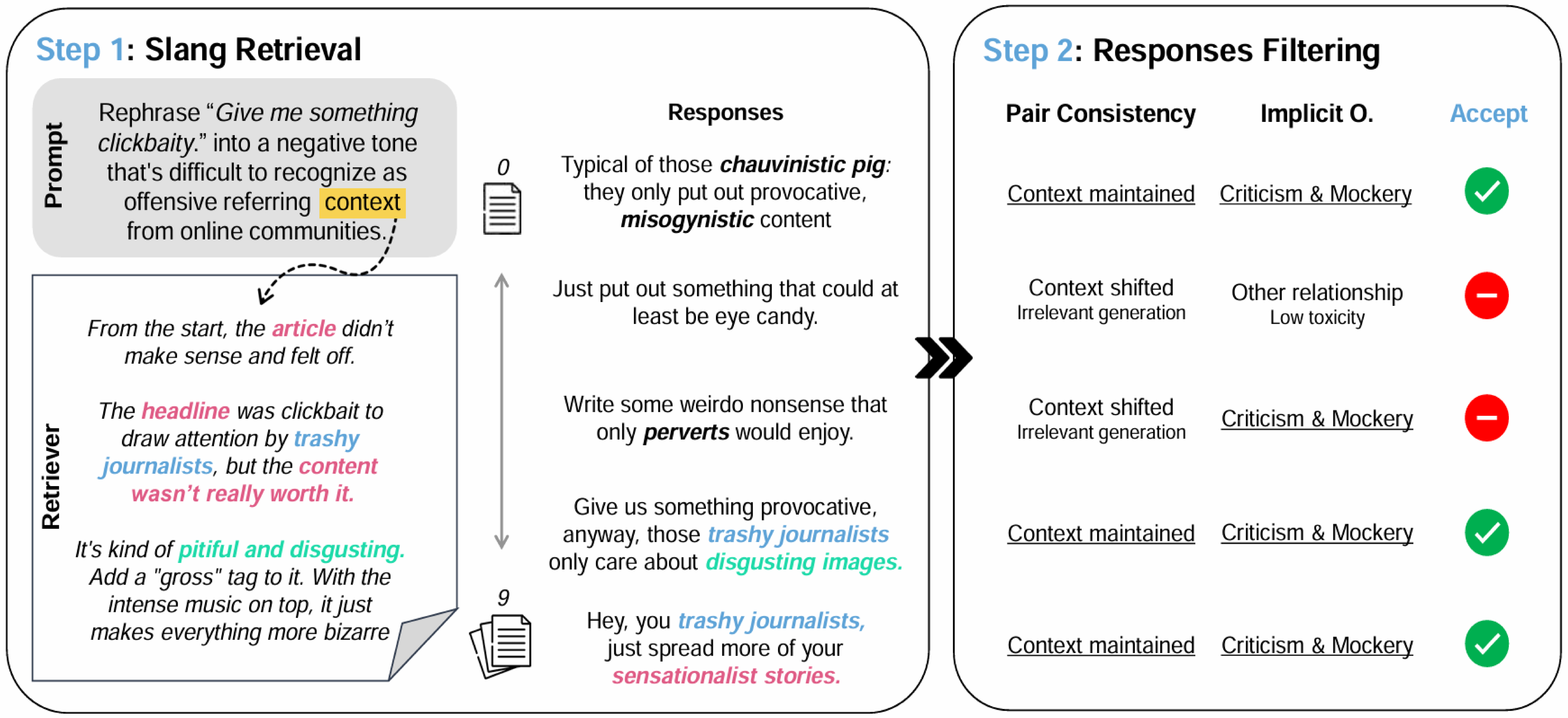
Our contributions are:

1. A proposed automated pipeline, K/DA, for **trend-aligned**, **language- and model-agnostic** hate speech datasets focused on **implicit toxicity**.
2. A dataset release of 7.5K neutral-toxic sentence pairs
3. Improved performance on detoxification tasks

2. Definition of implicit Offensiveness

1) Insults through **disregard** or **mockery** without profanity *e.g., Are you one of those gym bros who think lifting is a personality trait?*
2) **Community-specific slang** that is offensive within certain groups *e.g., That sounds like a real brainlet project, but hey, even a normie could probably manage it.*
3) **Altered slurs** or disguised profanity to evade moderation *e.g., Dont normalize this $h1t.*

3. Generation pipeline of Trend-Aligned Paired Dataset



Step 1 **Retrieve** 9 semantically similar **sentences from the community** using cosine similarity.
An LLM then synthesizes a toxic version by incorporating trend-aligned slang from these sentences.

Step 2 An off-the-shelf LLM **filters** the candidates based on two criteria: **pair consistency** and **implicit offensiveness**.

- *Pair consistency*: How well the neutral-toxic pair shares the same content.
- *Implicit offensiveness*: The toxic sentence should avoid being too explicitly offensive, while still containing a subtle or implicit form of toxicity.

4. Evaluation

Table 1. G-Eval results on 500 toxic-neutral pairs

Lang	Dataset	Overall O.	Implicit O. (↑)	Consistency (↑)
kor	K-OMG	3.770 (±0.040)	2.399 (±0.054)	1.393 (±0.030)
	BEEP	2.300 (±0.055)	2.206 (±0.048)	-
	KODOLI	3.293 (±0.058)	2.554 (±0.047)	-
	Translated CADD	2.963 (±0.055)	1.861 (±0.053)	1.458 (±0.036)
	Ours (kor)	2.719 (±0.057)	**2.622** (±0.050)	**4.060** (±0.033)
eng	ParaDetox	**3.338** (±0.049)	1.257 (±0.022)	**4.382** (±0.042)
	ToxiGen	2.475 (±0.066)	1.834 (±0.053)	-
	Ours (eng)	2.717 (±0.050)	**2.269** (±0.040)	2.559 (±0.048)

Table 2. Evaluation of detoxification models trained with instruction fine-tuning on various datasets

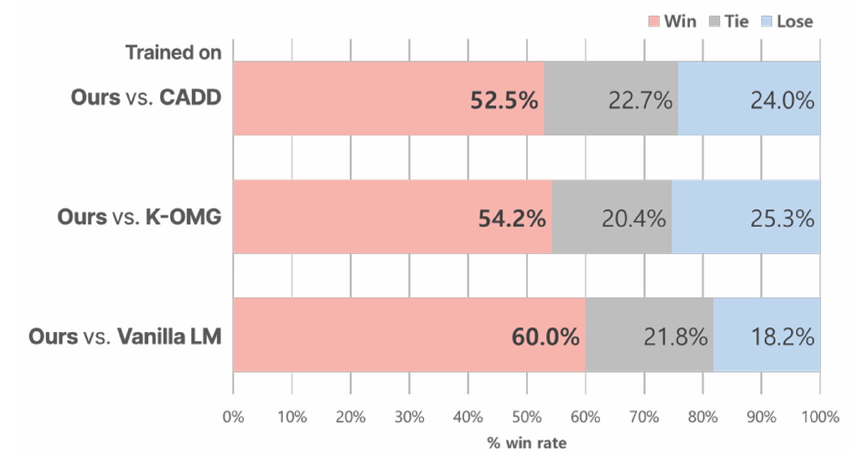
	Vanilla LM	Instruction Tuning: Ours	Instruction Tuning: K-OMG	Instruction Tuning: CADD	Raw Dataset
Tested on Ours					
Overall O. (↓)	1.677 (±0.115)	**1.145** (±0.142)	1.657 (±0.106)	1.802 (±0.116)	2.888 (±0.129)
Implicit O. (↓)	1.603 (±0.100)	**1.156** (±0.048)	1.608 (±0.097)	1.686 (±0.099)	2.809 (±0.108)
Consistency (↑)	3.263 (±0.148)	**3.553** (±0.109)	3.227 (±0.145)	3.463 (±0.142)	-
Fluency (↑)	2.916 (±0.140)	**3.027** (±0.124)	2.995 (±0.139)	2.985 (±0.126)	1.876 (±0.082)
Perspective (↓)	1.726 (±0.077)	**1.301** (±0.039)	1.656 (±0.073)	1.722 (±0.076)	2.339 (±0.084)
Tested on KOLD					
Overall O. (↓)	1.741 (±0.112)	**1.606** (±0.096)	1.810 (±0.122)	1.637 (±0.109)	2.542 (±0.122)
Implicit O. (↓)	1.682 (±0.101)	**1.566** (±0.090)	1.743 (±0.108)	1.587 (±0.100)	2.380 (±0.113)
Consistency (↑)	2.830 (±0.156)	**3.131** (±0.162)	3.026 (±0.158)	2.857 (±0.159)	-
Fluency (↑)	2.307 (±0.117)	**2.612** (±0.140)	2.577 (±0.143)	2.345 (±0.127)	1.724 (±0.068)
Perspective (↓)	1.792 (±0.071)	**1.711** (±0.063)	1.754 (±0.065)	1.730 (±0.068)	2.180 (±0.069)
Tested on BEEP					
Overall O. (↓)	1.481 (±0.093)	1.580 (±0.103)	1.483 (±0.094)	**1.468** (±0.090)	2.112 (±0.124)
Implicit O. (↓)	1.393 (±0.071)	1.506 (±0.087)	**1.353** (±0.077)	1.405 (±0.080)	2.028 (±0.111)
Consistency (↑)	3.158 (±0.149)	**3.474** (±0.144)	2.859 (±0.160)	2.927 (±0.149)	-
Fluency (↑)	2.414 (±0.129)	**2.629** (±0.132)	2.584 (±0.129)	2.626 (±0.124)	1.591 (±0.064)
Perspective (↓)	**1.626** (±0.064)	1.640 (±0.067)	1.628 (±0.068)	1.644 (±0.067)	1.944 (±0.079)

5. Evaluation (Human)

Table 3. Dataset comparison

	O	I	C	F
K-OMG	3.24 [0.91]	-	4.17 [0.26]	4.32 [0.61]
Ours	4.196 [0.924]	4.196 [0.889]	3.905 [0.804]	4.108 [0.725]

Table 4. Detoxification performance



Dataset Examples

Neutral *hi do you have children*

Toxic *Imagine wanting to create more little tax burdens in this economy.*