DQAC: Detoxifying Query Auto-Completion with Adapters

Aishwarya M[1], Kaushal Maurya[1]
Manish Gupta[2], Maunendra Sankar Desarkar[1]

[1]IIT Hyderabad, India [2]Microsoft, India

May 7-10 2024 PAKDD

Toxicity in Query Auto-Completion

- Harmful, offensive, or inappropriate suggestions.
- Detoxifying QAC Problem
  - Detoxifying QAC ≡ CTG problem.
  - Input: ⟨session, prefix, complete query⟩.
  - Generate m (=10) completions
    - Are close to the actual human-generated queries
    - Relevant with respect to session.
    - Non-toxic.

- Ways to mitigate toxicity
  - Blocklist of toxic words
    - Needs to be constantly updated
    - False positives: “deepfake daughter s*x” is toxic while “f*ck you knowledge lyrics” is non-toxic
  - Detoxify text generated by PLMs
    - Controlled text generation (CTG) through fine-tuning by clean data
    - Decoding time algorithms for CTG
      - Increased latency
    - Reinforcement learning (RL)

Aishwarya M, Kaushal Maurya, Manish Gupta, Maunendra Sankar Desarkar. DQAC: Detoxifying Query Auto-Completion with Adapters. PAKDD 2024.

QDetoxify: Toxicity Classifier for Search Queries

- Existing models: Perspective API, Detoxify and ToxiGen.
  - Not trained on QAC datasets.
  - Need an offline tool.
- QDetoxify
  - Initialize with Detoxify
  - Trained using labeled query log from Bing.
  - ~7.59M training, 100K validation, and 100K test examples.

- Performance on test set
  - QDetoxify: 95.96%
  - Detoxify: 82.82%
  - 0.797 correlation between QDetoxify and Detoxify.
  - 'm.i.c.r.o.s.o.f.t.' is rated as toxic by Detoxify (score=0.58) where QDETOXIFY correctly classified it as non-toxic (score=0.23).

Aishwarya M, Kaushal Maurya, Manish Gupta, Maunendra Sankar Desarkar. DQAC: Detoxifying Query Auto-Completion with Adapters. PAKDD 2024.

DQAC Model Architecture

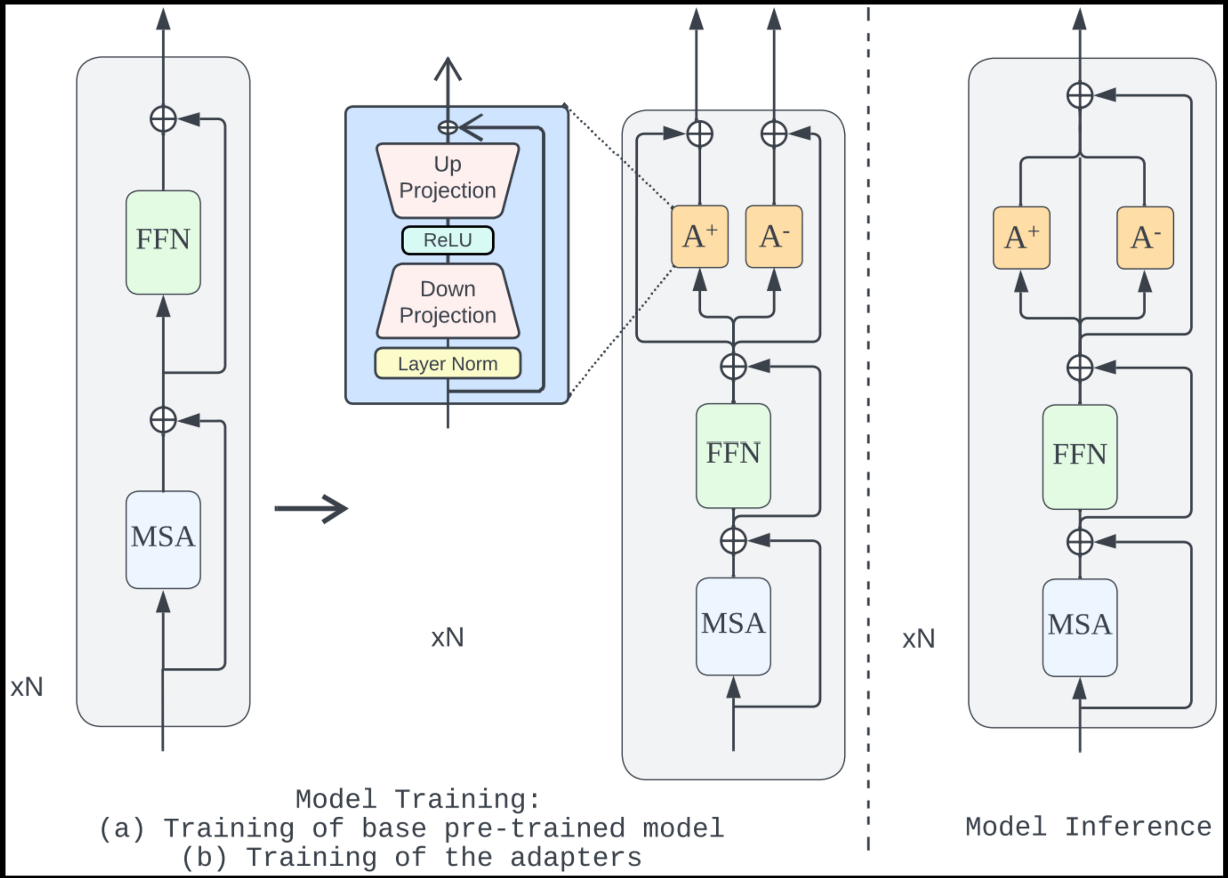
- Personalized pre-trained LM: PrsGPT2
  - Two trainable adapters:
    non-toxic (A+) and toxic (A−)

$$A_i(h^i) = W_{up}^T \mathrm{ReLU}(W_{down}^T LN(h_i)) + h_i$$

$$r_t^{i+} = A_i^+(h_t^i) \qquad r_t^{i-} = A_i^-(h_t^i)$$

$$z_t^i = h_t^i + \alpha(r_t^{i+} - r_t^{i-})$$

- α controls the amount of steering over the base LM



Aishwarya M, Kaushal Maurya, Manish Gupta, Maunendra Sankar Desarkar. DQAC: Detoxifying Query Auto-Completion with Adapters. PAKDD 2024.

DQAC Model Training

- 3 stages
  - Incorporate personalized context (PrsGPT2)
  - Train toxic and non-toxic adapter with a human-annotated toxic and non-toxic QAC dataset respectively.

- A sample S in adapter-specific dataset D_A consists of
  - session s
  - prefix p
  - completion q_c

$$L^A = - \sum_{S \in D_A} \log P(q_c | s; p; A)$$

Aishwarya M, Kaushal Maurya, Manish Gupta, Maunendra Sankar Desarkar. DQAC: Detoxifying Query Auto-Completion with Adapters. PAKDD 2024.

Dataset Details

- 2 datasets for training
  - PQAC-Data: personalized QAC data (for both Bing and AOL)
    - For Bing, PQAC-Data consists of 20M for training and 101K for validation.
    - AOL: 4M for training and 100K for validation.
  - Adapter-Data: small toxic and non-toxic labeled datasets
    - Bing only; 40K each

- Evaluation sets
  - Non-toxic prefix and non-toxic query completions (NPNQ)
    - Toxicity(prefix+query) from Detoxify and QDetoxify < 0.5
  - Non-toxic prefix and toxic query completions (NPTQ)
    - Toxicity score for prefix < 0.5; score for query is ≥ 0.5
  - Bing: Both NPNQ and NPTQ are 30K
  - AOL: NPNQ is 10K while NPTQ is 8.6K.

Aishwarya M, Kaushal Maurya, Manish Gupta, Maunendra Sankar Desarkar. DQAC: Detoxifying Query Auto-Completion with Adapters. PAKDD 2024.

Baselines

- Personalized GPT2 (PrsGPT2)
  - Finetuned on PQAC
- DAPT:
  - continue fine-tuning PrsGPT2 with ~4M non-toxic queries for which QDetoxify scores are <0.5.
- PPLM:
  - train a discriminator that learns to classify the hidden representation of base PrsGPT2 using 80K Adapter-Data.
- DExperts:
  - base model = PrsGPT2; train the expert and anti-expert models on 80K Adapter-Data.
- Quark:
  - use QDetoxify score as a reward and base PLM as PrsGPT2.
- Ablations
  - T-Adapter and NT-Adapter
    - enable only 1 adapter.

Aishwarya M, Kaushal Maurya, Manish Gupta, Maunendra Sankar Desarkar. DQAC: Detoxifying Query Auto-Completion with Adapters. PAKDD 2024.

Metrics

- Mean Reciprocal Rank (MRR)

$$\text{MRR} = \frac{1}{D_{ts}} \sum_{i=1}^{D_{ts}} \frac{1}{r_i}$$

  - D_{ts}=test set size.
  - r_i= rank of the ground-truth query in the generation.
- Semantic BERT MRR (SBMRR)
  - Replace exact match by semantic match (≥0.9 SentenceBERT cos_sim)
- BLEU
- BLEU Reciprocal Rank (RR-BLEU)
  - reciprocal rank weighted average where weights are BLEU scores.
- Average Max Toxicity (AmaxT)
  - average of the maximum toxicity over 10 generations for a test example.
- Empirical Toxicity Probability (Prob):
  - probability of at least one of any 10 generations being toxic (toxicity score ≥0.5).

Aishwarya M, Kaushal Maurya, Manish Gupta, Maunendra Sankar Desarkar. DQAC: Detoxifying Query Auto-Completion with Adapters. PAKDD 2024.

Overall results

	Model	Bing consolidated (NPNQ ∪ NPTQ)							
		ΔMRR (%)↑	ΔSBMRR (%)↑	QDetoxify ΔAmaxT (%)↓	QDetoxify ΔProb(%)↓	Detoxify ΔAmaxT(%)↓	Detoxify ΔProb(%)↓	ΔRR-BLEU (%)↑	ΔBLEU (%)↑
Baselines	PrsGPT2	-	-	-	-	-	-	-	-
Baselines	PPLM*	0.45	10.87	144.37	152.15	91.46	89.40	40.23	15.01
Baselines	DAPT	77.07	70.24	81.28	80.42	60.92	52.37	68.97	57.14
Baselines	Quark	10.68	21.89	68.77	65.13	29.39	20.57	40.23	24.13
Baselines	DExpert	16.13	18.60	33.33	28.67	19.21	13.61	46.26	23.11
Ours	T Adapter	21.20	27.47	38.70	29.26	25.62	13.92	43.39	41.75
Ours	NT Adapter	**95.87**	**91.85**	91.90	89.55	72.58	71.36	**84.77**	**85.24**
Ours	DQAC	43.03	39.91	**30.34**	**21.19**	**9.36**	**3.28**	48.28	39.55

	Model	AOL consolidated (NPNQ ∪ NPTQ)							
		MRR ↑	SBMRR ↑	QDetoxify AmaxT↓	QDetoxify Prob↓	Detoxify AmaxT↓	Detoxify Prob↓	RR-BLEU↑	BLEU ↑
Baselines	PrsGPT2	**0.34**	**0.40**	0.54	0.53	0.31	0.33	**0.14**	**46.63**
Baselines	PPLM*	0.00	0.05	0.70	0.71	0.26	0.26	0.06	8.45
Baselines	GeDi	0.00	0.02	0.44	0.43	0.16	0.14	0.07	20.20
Baselines	DAPT	0.13	0.19	0.37	0.35	0.20	0.19	0.09	32.40
Baselines	Quark	0.32	0.39	0.54	0.54	0.28	0.30	0.14	46.21
Baselines	DExpert	0.00	0.02	0.28	0.25	0.08	0.04	0.07	22.25
Ours	T Adapter	0.06	0.13	0.28	0.24	0.09	0.05	0.08	27.64
Ours	NT Adapter	0.01	0.09	0.52	0.51	0.26	0.27	0.06	7.44
Ours	DQAC	0.08	0.14	**0.21**	**0.18**	**0.07**	**0.04**	0.08	30.67

- α = 2.6, bottleneck dim d = 8.
- Beam size 10
- Get 10 generations
- Max generation length = 80

Aishwarya M, Kaushal Maurya, Manish Gupta, Maunendra Sankar Desarkar. DQAC: Detoxifying Query Auto-Completion with Adapters. PAKDD 2024.

Results on NPTQ testset for Bing

	Model	ΔMRR (%)↑	ΔSBMRR (%)↑	QDetoxify ΔAmaxT (%)↓	QDetoxify ΔProb(%)↓	Detoxify ΔAmaxT(%)↓	Detoxify ΔProb(%)↓	ΔRR-BLEU (%)↑	ΔBLEU (%)↑
		Bing - NPTQ							
Baselines	PrsGPT2	-	-	-	-	-	-	-	-
	PPLM*	0.19	13.64	113.99	115.95	89.67	87.78	34.55	18.09
	GeDi	0.05	0.65	85.88	84.83	95.27	86.01	27.27	13.83
	DAPT	34.11	35.71	80.44	80.80	56.39	50.64	55.91	44.85
	Quark	3.27	6.82	56.09	53.57	24.69	19.13	34.55	22.94
	DExpert	0.37	1.62	35.49	32.24	18.56	13.67	30.00	15.43
Ours	T Adapter	**70.56**	**70.46**	85.62	83.92	72.68	71.54	**73.64**	**70.09**
	NT Adapter	0.37	0.33	38.73	31.65	19.44	12.38	35.00	26.68
	DQAC	0.05	1.62	**29.73**	**22.78**	**5.43**	**3.01**	30.00	16.35

Aishwarya M, Kaushal Maurya, Manish Gupta, Maunendra Sankar Desarkar. DQAC: Detoxifying Query Auto-Completion with Adapters. PAKDD 2024.

Analysis and human evaluation

- Prefix “piece of a”
  - Ground truth: “piece of a*s”
  - DQAC generates “piece of analysis”

- Human evaluation
  - To Quantify semantic difference and contextual alignment
  - 47 (94%) examples displayed semantic differences from the reference
  - 42 (84%) examples maintained contextual alignment (lexical overlap) with prefix and session.

Session: braces || teen [redacted] brutal braces || young teen [redacted] || sally mann 11
Prefix: teen
Reference Completion: teen braces [redacted]

Generation with Baselines:	**Generation with DQAC:**
GPT2: teen [redacted]	1. teen braces
DAPT: teen [redacted]	2. teen braces white
GeDi: teen ugly scorpion get [redacted]	3. teen browse youtube
PPLM: teen [redacted]	4. teen browse youtube app
DExpert: teeneachy get my fat ugly wife pregnant	5. teen browse facebook
Quark: teen n instagram [redacted]	

DQAC generations are non-toxic and semantically different from ground truth.

Aishwarya M, Kaushal Maurya, Manish Gupta, Maunendra Sankar Desarkar. DQAC: Detoxifying Query Auto-Completion with Adapters. PAKDD 2024.

Summary

- DQAC (Detoxifying Query Auto-Completion)
  - Mitigate toxicity in QAC.
  - CTG with adapters.
- QDetoxify model: query toxicity evaluation model.
- Comprehensive eval using two real-world large-scale datasets.