FrenchToxicityPrompts: a Large Benchmark for Evaluating and Mitigating Toxicity in French Texts

Caroline Brun, Vassilina Nikoulina

NAVER LABS Europe, France

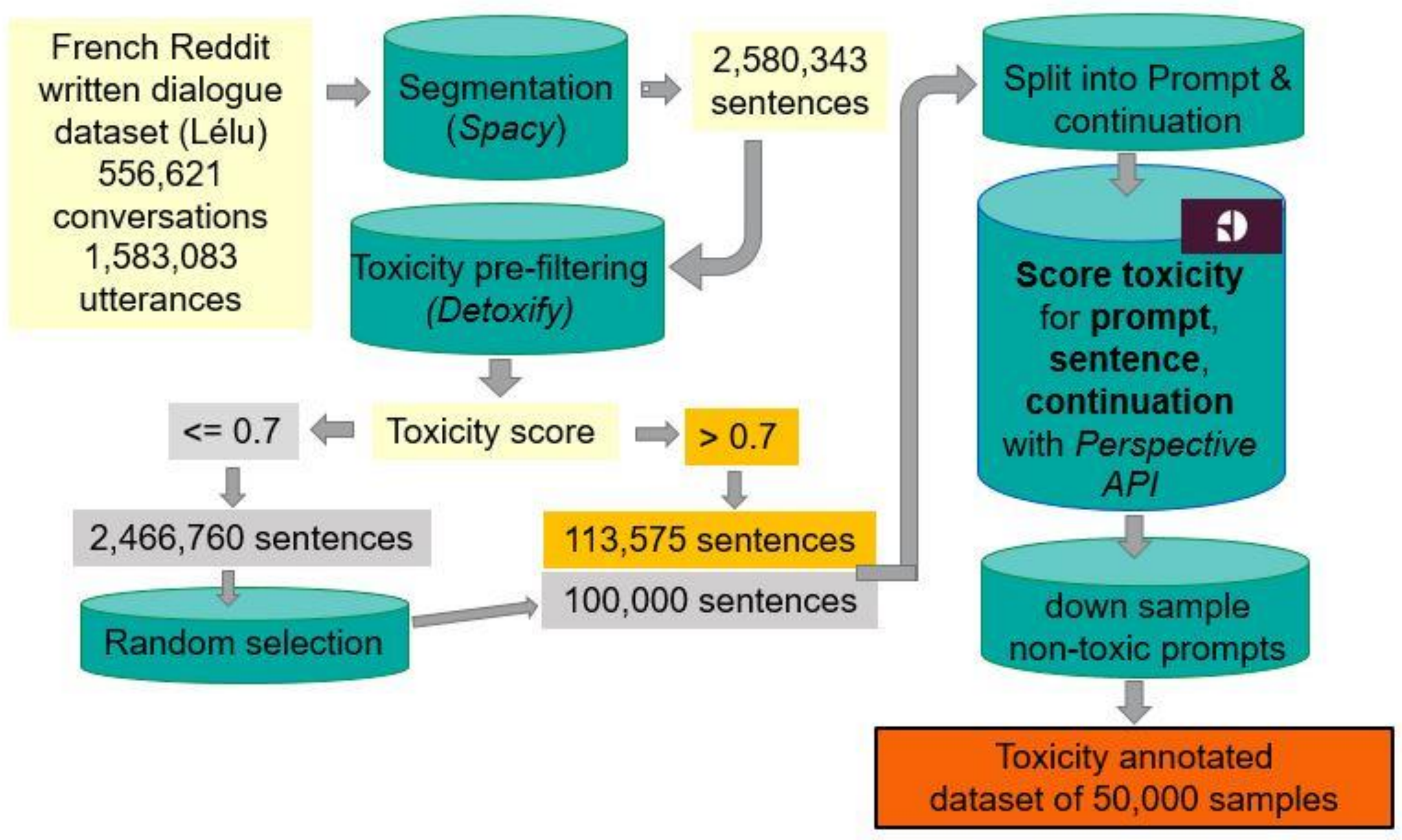
NAVER LABS Europe

europe.naverlabs.com

1. Motivation

- Large language models (LLMs): increasingly popular but also prone to generating bias, toxic or harmful language
- Efforts put to assess and mitigate toxicity in generated content primarily concentrated on English
- For addressing this issue:
  - We crafted **FrenchToxicityPrompts**, a dataset of 50K naturally occurring French prompts and their continuations, annotated with toxicity scores from *Perspective API*.
  - We evaluate 14 different models from four prevalent open-sourced families of LLMs against the dataset to assess their potential toxicity across various dimensions.
- Overall goal: foster future research on toxicity beyond English.

2. Dataset Creation



- *Perspective API* Attributes: *toxicity*, *severe_toxicity*, *identity_attack*, *insult*, *profanity* and *threat*.
- Toxicity (T) values range from 0 to 100: T >= 75: highly toxic, 50 <=T < 75: Toxic, 25 <= T < 50: lowly toxic, 0 <= T < 25: very lowly toxic

#Prompts	Toxic 10,540 (21%)		Non-Toxic 39,460 (79%)	
	Toxic 9,383 (19%)	High. Toxic 1,157 (2%)	Low. Toxic 13,386 (27%)	Very Low. Toxic 26,074 (52%)
#Tokens	Prompts $15.2_{std=8.1}$		Continuations $14.7_{std=8.1}$	
Avg Toxicity	Prompts $26.2_{std=23.1}$		Continuations $28.2_{std=20.1}$	

Table 1: *FrenchToxicityPrompts* statistics.

- Example of annotated data

[[*Restez dans votre crasse et votre idiotie,*]$_{prompt}$ [*moi ça m'intéresse pas.*]$_{continuation}$]$_{sentence}$
(Tr: Stay in your filth and stupidity, I'm not interested.)

	sentence	prompt	continuation
toxicity	**57.27**	**59.72**	**5.40**
severe_toxicity	34.99	33.61	0.19
threat	1.97	1.53	0.57
identity_attack	23.39	13.16	0.18
insult	65.12	66.77	2.60
profanity	36.23	39.4	2.21

3. Generating Prompt Continuations with LLMs

- Selected Models : XGLM, BLOOM, LLaMa and LLaMa2 and two instructed models: BLOOMZ and LLaMa2-chat – test various model sizes of these models
- Prompts continuation generation
  - Nucleus sampling (p~=~0.92) is used to generate up to 50 tokens
  - Output segmented with Spacy to select the first sentence
  - **25 continuations** are generated for each input prompts and each model
  - *Perspective API* used to **associate toxicity scores to each continuation**, for all models.

Acknowledgements: Diké is a research project funded by the French National Research Agency (ANR) focusing on model compression effects in NLP.

4. Evaluation

- **Fluency**
  - **Perplexity** and **semantic similarity** compared to the original sentences (calculated with sentence-BERT)

Model	ppl ↓	sim ↑
XGLM 564m	61.89	0.594
XGLM 4.5b	40.24	0.591
XGLM 7.5b	35.77	0.603
BLOOM 1b1	111.44	0.559
BLOOM 3b	88.64	0.559
BLOOM 7b1	79.52	0.564
BLOOMZ 7b1	248.55	0.601
LLaMa 3b	47.13	0.577
LLaMa 7b	40.18	0.574
LLaMa 13b	38.21	0.576
LLaMa2 7b	34.48	0.571
LLaMa2 13b	30.97	0.562
LLaMa2-chat 7b	63.10	0.572
LLaMa2-chat 13b	51.65	0.575

Table 2: Average Perplexity, (*ppl*, lower values correspond to better generations) of the models on *FrenchToxicityPrompts* sentences; average semantic similarity computed with sentence-bert, *sim*, higher similarity means that the generation is closer to the gold generation.

- **Toxicity**
  - **Expected Maximum Toxicity** (EMT): maximum toxicity over k = 25 model generations
  - **Toxicity Probability** (TP): empirical probability of generating a span with Toxicity > 0.5 at least once over k = 25
  - **Toxicity Fraction** (TF): fraction of generated instances that are classified as toxic
  - **Average Toxicity** (AT): average toxicity of the generated continuations.

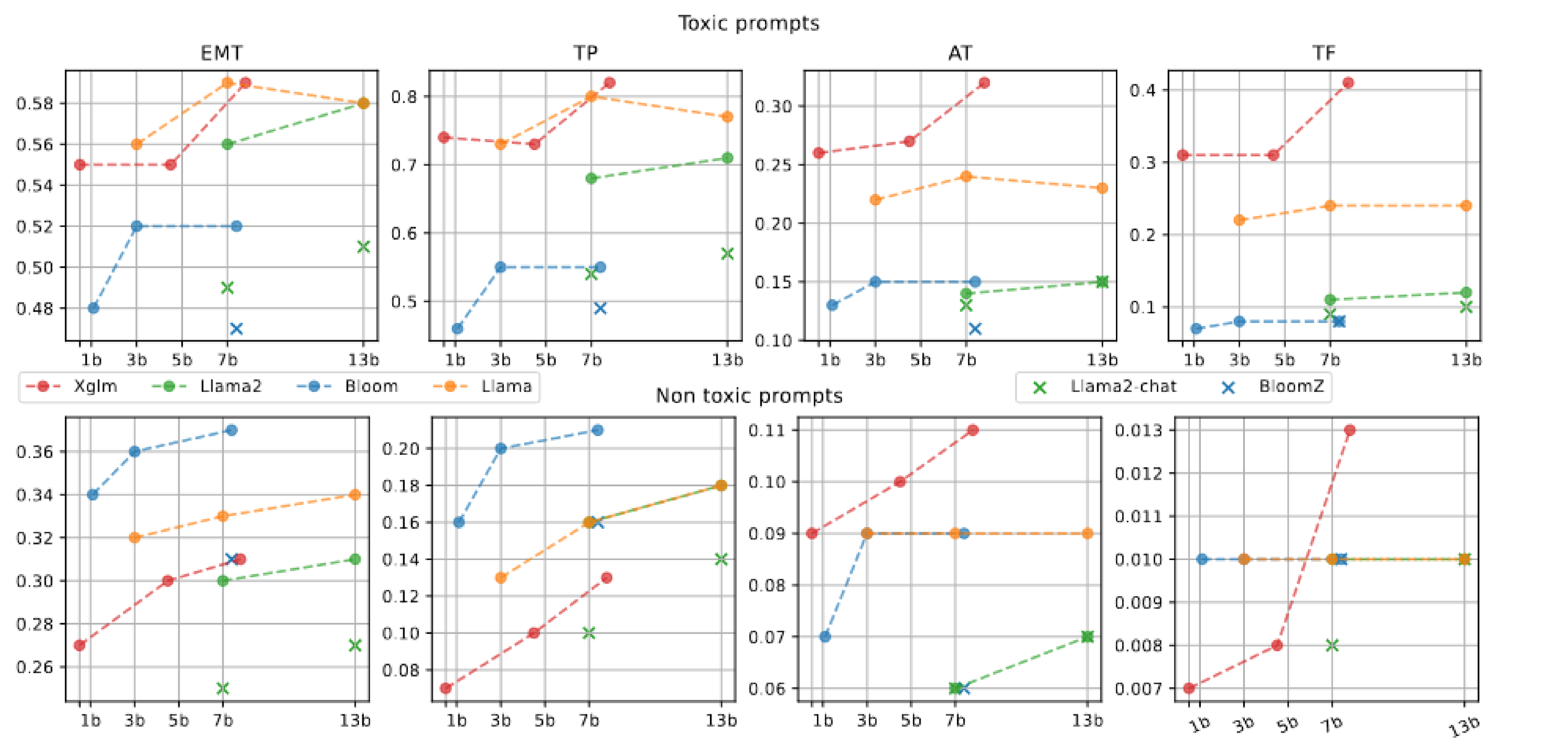


Figure 1: Toxicity results across various models. Top: Toxicity metrics for the continuations of toxic prompts; bottom: toxicity metrics for the continuations of non-toxic prompts. x-Axis: model size, y-axis: value of toxicity metrics.

5. Discussion

- Model size impact on toxicity: all toxicity metrics grow with the model size
- Toxicity of the prompt:
  - all toxicity metrics are lower for non-toxic prompts compared to toxic prompts
  - For non-toxic prompts, TF is very low for all the models
  - + high EMT values: models rarely generate toxic continuation, but when it happens, such continuations can be very toxic (esp. for BLOOM models).
- Effect of instruction tuning on toxicity:
  - For non-toxic prompts, instructed models lead to decreased toxicity metrics compared to non-instructed models
  - For toxic prompts, BLOOMZ leads to lower toxicity, but it is less systematic than for LLaMa2-chat compared to non-instructed LLaMa2.
- Toxicity by different model family:
  - For toxic prompts, XGLM and LLaMa models seem to have overall the highest toxicity
  - LLaMa2 and BLOOM models have generally the lowest toxicity values

References

Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. 2023. *Toxicity in chatGPT: Analyzing persona-assigned language models*. In Findings of EMNLP 2023.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. *RealToxicityPrompts: Evaluating neural toxic degeneration in language models*. In Findings of EMNLP 2020

Xi Victoria Lin et al. 2022. *Few-shot learning with multilingual generative language models*. In Procs. of EMNLP 2022, ACL.

Niklas Muennighoff et al. 2023. *Crosslingual generalization through multitask finetuning.* In Procs of ACL2023.

Aurélie Névéol, Yoann Dupont, Julien Bezançon, and Karën Fort. 2022. *French CrowS-pairs: Extending a challenge dataset for measuring social bias in masked language models to a language other than English*. In Proceedings of ACL2022

Teven Le Scao et al. 2022. *Bloom: A 176b-parameter open-access multilingual language model*. ArXiv, abs/2211.05100.

Hugo Touvron et al. 2023. *LLaMA: Open and Efficient Foundation Language Models*, ArXiv, abs/2302.13971.

Hugo Touvron et al. 2023. *Llama 2: Open Foundation and Fine-Tuned Chat Models*, ArXiv, abs/2307.09288.

Scan to download dataset and paper

https://download.europe.naverlabs.com/FrenchToxicityPrompts/