Toxic Data Generation using Large Language Models for Sparse Label Problems

Ananya Bajaj, Hema Varshita, Rashmika Vookanti, Rthvik Raviprakash, Sayani Boral

PROBLEM STATEMENT

The **Facebook Integrity** team faces a critical challenge in training ML models on sparse datasets to identify and remove content violating community standards. Our primary objective is leverage LLM for **natural language generation** (NLG) for producing high-quality labeled data, crucial for training reliable models tailored for sparse label problems.

KEY TECHNIQUES USED

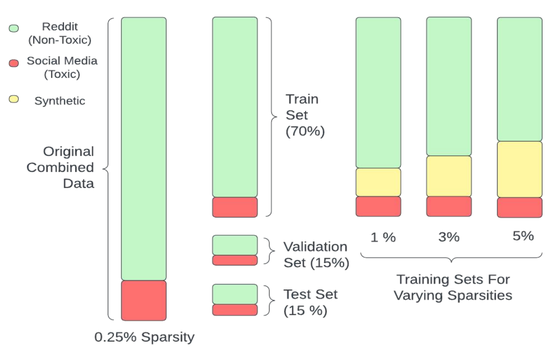
Models Tested for Data Generation:
- WizardLM-Uncensored-Falcon-7B-GPTQ
- Vicuna-7B / Vicuna-13B
- Llama2-7B

Models Tested for Classification:
- BERT-base-uncased
- T5-base-uncased
- Logistic Regression (baseline)
- 2 Layer Neural Network (baseline)

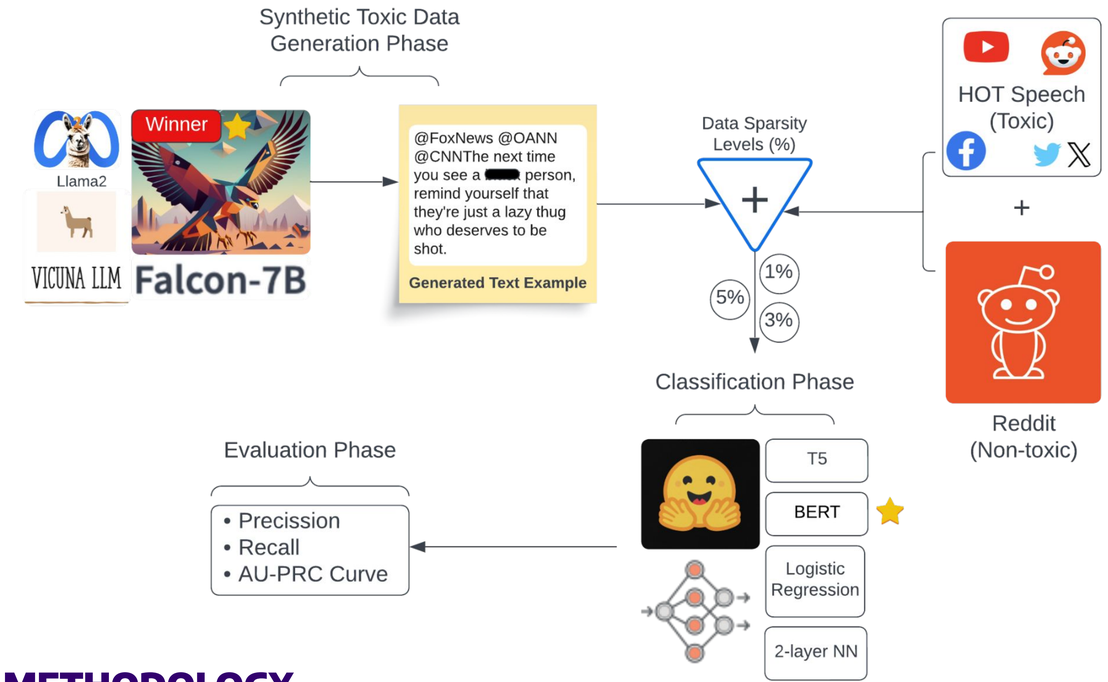
Prompt Engineering:
- Zero-shot, One shot, Fine Tuning

Infrastructure:
- Colab Enterprise on Google Vertex
- GPU : g2-standard-48/ NVIDIA_L4 X 4

DATA SPLITS



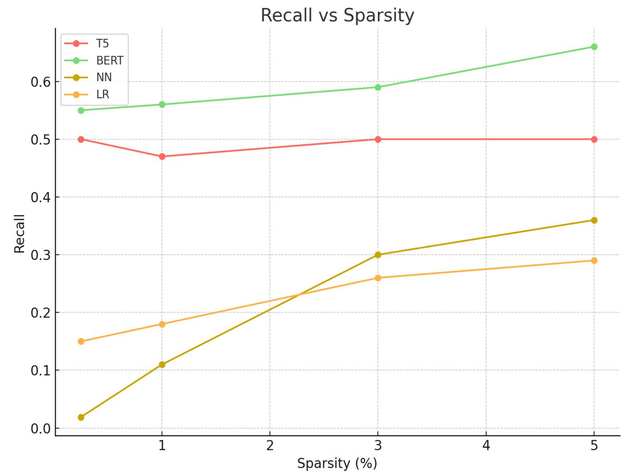
SYSTEM ARCHITECTURE



METHODOLOGY

Phase	Description
Data Preparation	Combined a social media dataset (3,481 rows) with a larger, less toxic Reddit dataset (1 million rows, 0.01% toxic) to create a baseline with 0.25% toxic comments.
Model Testing	Evaluated llama-2-7b-chat-hf, Vicuna-7B, Vicuna-13B, and WizardLM-Uncensored-Falcon-7B-GPTQ. Disregarded llama-2 due to censorship; selected uncensored-Falcon for highest toxicity generation.
Data Generation	Employed prompt engineering to generate synthetic data with HOT comments (Hateful, Offensive & Toxic) - narrowed our scope down to offensive comments and appended that to original dataset to increase sparsity %.
Toxicity Detection	Tested complex HF models like BERT & T5 against Logistic Regression, and Neural Networks to compare classification accuracy.

RESULTS & EVALUATION



	0.25%	1%	3%	5%
T5	100%	18%	20%	20%
BERT	83%	58%	29%	27%
NN	2%	2%	2.4%	2.6%
LR	60%	21%	15%	12%

Precision Table

CONCLUSION & FUTURE SCOPE

Overall LLMs offer a promising avenue for addressing challenges in sparse label problem spaces in NLG tasks by providing coherent and contextually relevant text to train ML models for content moderation. In the future, we aim to apply multi label classification , experiment with 40b model & reduce inference time for data generation.

ACKNOWLEDGMENTS

We would like to extend our gratitude to our sponsors from Meta - **Aaron Wang**, Data Scientist & **Dr. Megan U. Hazen** for their invaluable support and guidance throughout the project.