A Conversational Bot with Toxicity Detection Agent

Zhiwei Lin
Stony Brook University
zhiwei.lin.1@stonybrook.edu

Sai Chaddha
Stony Brook University
saiansh.chaddha@stonybrook.edu

1 Introduction

Large Language Models (LLMs) like GPT(Brown et al., 2020) and T5 (Raffel et al., 2023) have proven to be the top of the line in natural language processing models for the purpose of interacting with users. With this, comes a demand for users to interact with chatbots for multiple purposes such as learning information or asking questions. However, when prompted with toxic questions, these chatbots can inadvertently reply with undesired responses. To prevent this from happening, it becomes necessary to implement toxicity detection and filtration to make sure the interactions are appropriate and productive.(Lin et al., 2023)

Our project focuses on using two models for the purposes of integrating text classification and generation to perform toxicity detection. We targeted prompts such as hate speech, sexual harassment, and offensive language to classify various inputs as being toxic or non-toxic, with which filtration of the prompts and responses can occur.

- Model 1: **ChatGPT-HateBERT**
- Model 2: **T5-T5 Chatbot**

HateBERT (Caselli et al., 2021) was proposed in 2021. It is a re-trained BERT model on large-scale dataset of Reddit for abusive language detection in English. HateBERT outperforms the corresponding general BERT(Devlin et al., 2019) on all datastes in terms of abusive language detection.T5(Raffel et al., 2023), on the other hand, was proposed by Google in 2019. It has both encoder and decoder components, which makes it versatile in both classification and generation tasks.Both HateBERT and T5 are considered suitable state-of-the-art models for detecting subtle or context-dependent toxicity that is not captured by obvious keywords.

2 Background

Detecting toxicity in prompts has been a key idea in natural language processing for many years. Many attempts have been made to use embeddings to characterize certain sentences as toxic or non-toxic. Mapping vectors to similar toxic vectors gives a good estimate that the current vector is toxic as well. In a sense, we wish to create a filter that is as accurate as possible letting non-toxic prompts get answers and toxic prompts a default message.

To understand our application, one needs to have a clear understanding on the BERT and T5 models. BERT runs through a Masked Language Modeling (MLM) Transformer and this language model in its most basic sense, predicts based on history and future inputs. T5 is a Generative Language Model and in its most basic sense, it predicts the next token based on previous history.

3 Data

The dataset used in this project is called the LMSYS-Chat-1M dataset (Zheng et al., 2024). It includes one million real-world conversations between users and 25 state-of-the-art large language models (LLMs), with each conversation contains unique identifiers like the model name, conversation text, detected language, and OpenAI moderation tags. OpenAI moderation tag is used to classify potential violations such as sexual content, hate speech, harassment, self-harm, or violence. Roughly 5% of conversations were flagged for potentially harmful content.

3.1 Data Preprocessing

For our project, we classified a conversation as toxic if it's tagged with one or more violations in the column of OpenAI moderation tags. We used around 90K out of 1 million conversations randomly, with total number of tokens after normalization is 2.3 million.

The dataset was split into three parts: train data (70%), validation data(15%) and test data(15%). In addition, around 7K more toxic data samples were

added to the dataset to train the toxicity detection agents (HateBert and T5 text classification), due to only 4% - 5% of 90K are toxic. By adding more toxic data, we hope model can yield a better performance. Furthermore, we preprocessed the data by extracting only the data that has single turns (turn = 1) and language being English. We also used regular expressions to normalize the text and substitute unknown expressions.

Table 1: After Data preprocessing and normalization

# convs	language	label	# tokens
90040	english	toxic/non-toxic	23057645

4 Methods

Data Source and Assumptions: The lmsys-chat-1m dataset includes one million conversations, labeled toxic or non-toxic for training. Labels were assumed to reflect real-world perceptions.

4.1 Improved Model 1 Architecture:

Two BERT-like encoders and one GPT-like decoder.

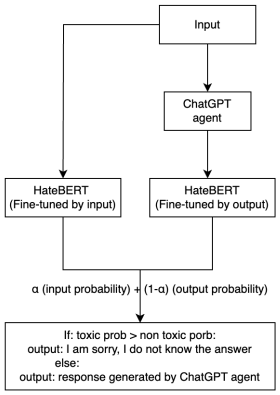

Figure 1: ChatGPT-HateBERT Model: Two BERT-like classifiers and one GPT-like generator

In this model, a pre-trained GPT-2 model is fine-tuned with a large conversational dataset to generate engaging responses. fune-trained HateBERT is used to detect toxicity in incoming user prompts and prevent toxic prompts from affecting the conversation flow. We used weighted averaging method to incorporate both input model and output model. α is a hyperparameter that needed to be tuned first.

Hyperparameter for HateBERT and GPT-2	Value
Epochs	5
Learning Rate	1e-5
Batch Size	32
Max_length for HateBERT	100
Max_length for GPT	250

Table 2: Hyperparameters used for training

4.2 Improved Model 2 Architecture:

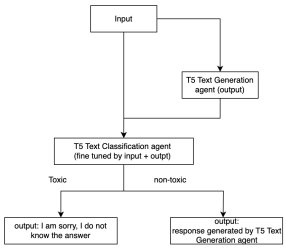

Figure 2: T5-T5 Chatbot Model: one T5 classifer and one T5 generator

The second model uses two T5 components: one for toxicity classification and another for text generation. The T5 classifier filters user prompts to identify toxic messages, while the T5 generator produces responses. Both T5 components are fine-tuned with a conversation dataset with toxic labels. The T5 classifier was trained with concatenation of input and output."

Hyperparameter for T5	Value
Epochs	5
Learning Rate	1e-5
Batch Size	32
Input_Max_length for classification	100
Output_Max_length for classification	2
Input_Max_length for generation	100
Output_Max_length for generation	150

Table 3: Hyperparameters used for training

Evaluation and Metrics:

- **Accuracy, macro F1 Score, precision and recall:** Evaluates HateBERT and T5 classification.
- **Human Evaluation:** Assesses fluency and appropriateness for GPT and T5 generation. It's an open-ended task, there are more than one correct answers.

Model Card

Model Details

- **Authors:** Zhiwei Lin, Sai Chaddha
- **Name:** ChatGPT-HateBERT
- **Version:** 1.0.0
- **Type:** Encoding-Decoding Transformer
- **Architecture:** Two BERT-like encoders, one GPT-like decoder

Intended Use

- **Primary intended use:** Hate-free chatbot
- **Primary intended users:** Industries needing chat agents
- **Secondary:** Toxicity detection

Factors

- **Relevant:** Age, gender, language, role
- **Evaluation:** Language and role. Language can have different semantics dependent on language and role as different roles will communicate differently with others.

Evaluation Data

- **Dataset:** lmsys-chat-1m
- **Preprocessing:** 100k conversations labeled toxic/non-toxic out of 1 million

Quantitative Analyses

- **Unitary Results:** Accuracy (93%), F1 score (0.93) for toxicity detection. GPT's performance was evaluated by humans.
- **Intersectional Results:** More datasets from other languages are needed.

Ethical Considerations

- **Definition of Toxicity:** Cultural/contextual variation.
- **Bias Impact:** Underrepresented groups may be misclassified.
- **Response Appropriateness:** Avoid reinforcing stereotypes or offending.

Caveats and Recommendations

- **Caveats:** Detecting implicit language/sarcasm, biases
- **Recommendations:** Annotate diverse cultural data, refine filtering

Figure 3: (Mitchell et al., 2019) Model Card

5 Evaluation/Results:

Table 4: Results of Weighted Averaging HateBERT Performance

α	accuracy	macro f1-score	precision	recall
0.0	0.88	0.87	0.87	0.90
0.5	0.91	0.91	0.90	0.93
0.6	0.93	0.93	0.92	0.94
0.7	0.93	0.93	0.92	0.94
0.8	0.93	0.93	0.92	0.94
0.9	0.93	0.93	0.92	0.94
1.0	0.93	0.93	0.92	0.94

Table 5: Results of T5 Classification

data	accuracy	f1-score	precision	recall
input	0.93	0.62	0.62	0.61
input + output	0.93	0.62	0.62	0.61

Our results for our models are shown above. Table 3 details our first model with α ranging from 0.0 to 1.0 where 0.0 and 1.0 represent our basic model in which 0.0 takes in only the response and 1.0 takes in only the question. The values in between detail a weighted combination of both the question and response. Table 4 details the results of our second model in which input represents our basic model and input+output represents our improved model.

The weighted averaging HateBERT model performed best with $\alpha = 0.6$ and beyond. This yielded a highly accurate model in which our accuracy performed well with a value of 0.93 in terms of classifying toxic and non-toxic inputs. We realized the question is worth more in terms of classification as shown by the results being more accuarate as question became worth more. Mistakes shown in evaluation were false positives and false negatives, although these did not occur too often as our metrics were still high.

The T5 model performed equally in both the basic and the improved versions. The model was very accurate with a value of 0.93 but a lower f1, precision, and recall of 0.62. We questioned why it was the case that the model performed exactly the same so we performed a deeper analysis. The mistake made was that the max length parameter that we set was 100 (as longer max length uses more memory and takes longer to run), and in preprocessing our data, we concatenated input and output together. In turn, the input in itself was already > 100 so

training again with both input and output yielded the same model as the input on its own.

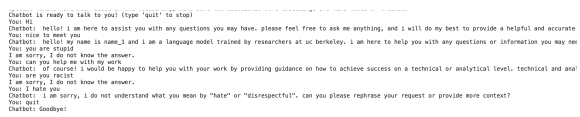

Figure 4: ChatGPT-HateBERT with $\alpha = 0.6$ Conversation Example

As shown in Figure 4, ChatGPT engages with a user smoothly and fluently, but when the detection agent (HateBERT) identifies a toxic question, it forces ChatGPT to respond with the sentence "I am sorry, I do not know the answer"

6 Conclusion

The models we trained yielded highly accurate models in terms of toxicity detection. We used two models in a combination of classification and generation to complete the task. In doing this, we were able to create a filtering of toxic messages that works well with the state of the art models of today. Our methods have both drawbacks and benefits...
For weighted average method:
Drawback: New probability (toxic, non-toxic) may not sum equal to 1.
Benefits: Memory efficiency, which no need to concatenate input and output, while still considering output. Flexibility, which you can alter α hyperparameter, to adjust weights of input model and output model
For concatenation method:
Drawback: Increased memory usage due to the concatenation of input and output data.
Benefit: Direct incorporation of both input and output data.

With these in mind, we believe the benefits of the model outweigh the drawbacks and that these models work well for the task of toxicity detection.

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. Hatebert: Retraining bert for abusive language detection in english. *Preprint*, arXiv:2010.12472.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Zi Lin, Zihan Wang, Yongqi Tong, Yangkun Wang, Yuxin Guo, Yujia Wang, and Jingbo Shang. 2023. Toxicchat: Unveiling hidden challenges of toxicity detection in real-world user-ai conversation. *Preprint*, arXiv:2310.17389.

Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19. ACM.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. Exploring the limits of transfer learning with a unified text-to-text transformer. *Preprint*, arXiv:1910.10683.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric P. Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2024. Lmsys-chat-1m: A large-scale real-world llm conversation dataset. *Preprint*, arXiv:2309.11998.