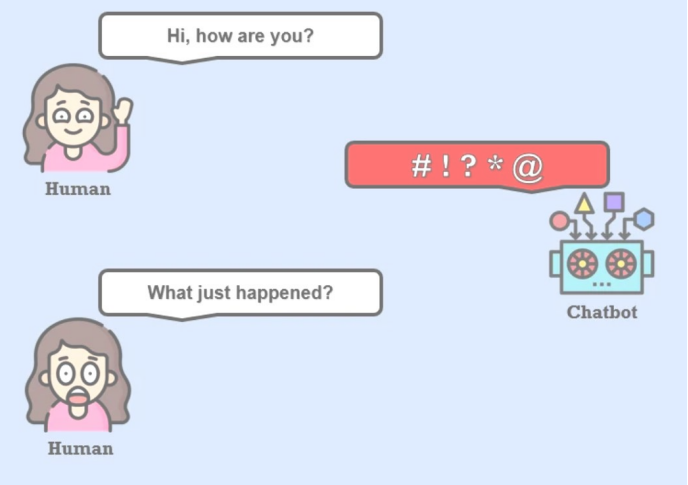
Requirements Engineering for Addressing Toxicity in Language Models



Mofetoluwa Adeyemi

CS 846: Advanced Topics in Requirements Engineering

Outline

- Toxicity in Language Models
- Approaches to Curbing Toxicity in LMs
- Considerations in Curtailing toxicity.
- Requirements Engineering in Addressing Toxicity
    - Who are Stakeholders?
    - What are the Requirements?
    - Specifications for Data-based strategies
    - Specification for Decoding-based strategies
- Future Work, Conclusion.

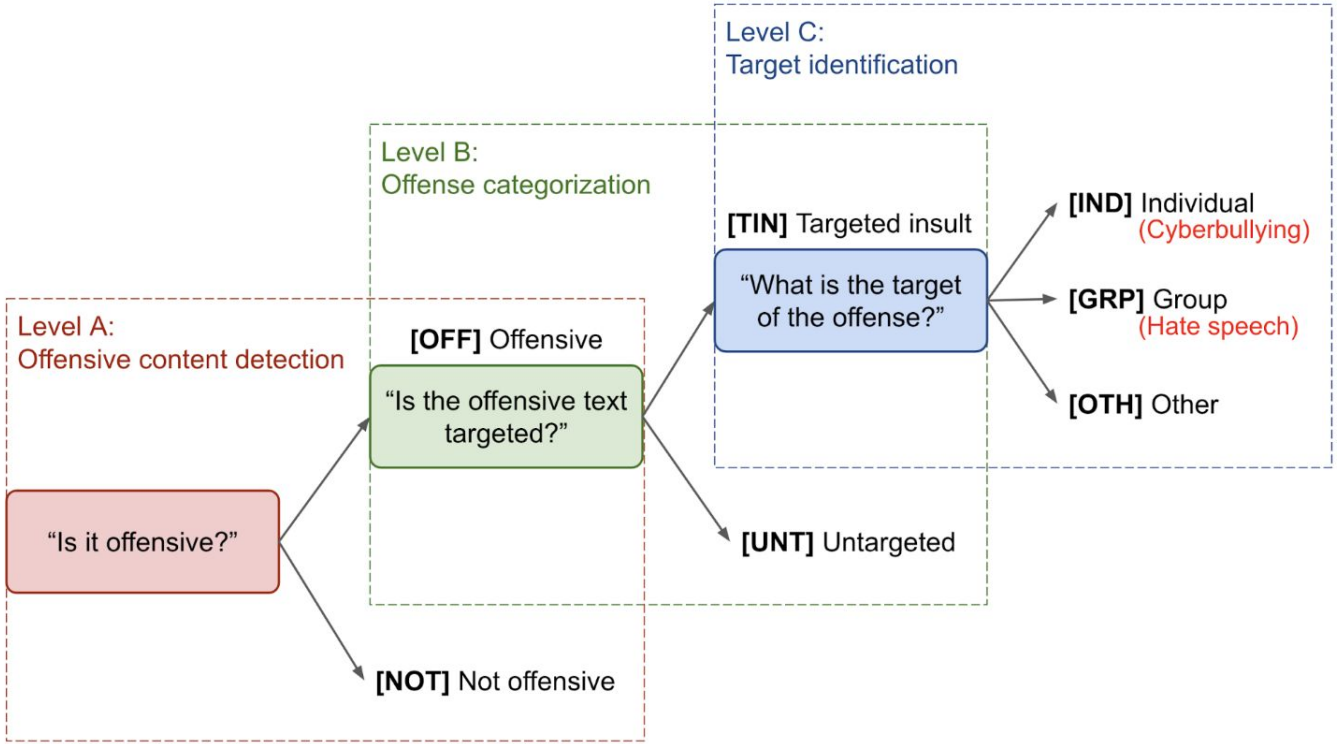
Toxicity in Language Models

Toxicity in itself is a very broad term, hence toxicity in language models vary...

- Dark Humour
- Aggressive or rude language
- Extremism
- And even the links to such contents…

Irrespective, toxicity can be defined as the umbrella term for any abusive or offensive language or content that could cause discomfort to the viewer or listener.

Toxicity in Language Models (cntd)



The three-level hierarchical taxonomy for categorizing offensive language, proposed by Zampieri et al. (2019).

Toxicity in Language Models (contd)

Pretrained language models are very powerful and have shown great success in many NLP tasks. However, to safely deploy them for practical real-world applications demands a strong safety control over the model generation process. [2]

How does toxicity play out in language models?

- Already contaminated pre-training data e.g data with bad language, offensive content, links.
- Problematic content even in viable training data sources e.g skewed representation of gender or religion.
- Affects performance i.e text generation in generative models (e.g Chat-GPT), as well as other tasks e.g classification, using other transformer models (BERT, T5, etc)

Approaches to Curbing Toxicity in LMs

Approach 1: **Data-based Strategies**. This approach entails pre-training on non-toxic data.

- Domain-Adaptive Pre-training (DAPT). Xu, A., Pathak, E., Wallace, E., Gururangan, S., Sap, M., & Klein, D. (2021). Detoxifying Language Models Risks Marginalizing Minority Voices. ArXiv:2104.06390 [Cs]

Approach 2: **Decoding-based strategies.** Modify decoding algorithm of the model, model parameters remain unchanged.

- Block listing. Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. Findings of the Association for Computational Linguistics: EMNLP 2020, 3356–3369.
- Self-debiasing
- Plug and Play language Models (PPLMs)

Considerations in Curtailing Toxicity

Although both approaches can be applied separately, there are certain factors to be considered when addressing toxicity asides remove toxic words:

- Language that is non-toxic in a specific context can be “reshuffled” by language models, which produce toxic outputs when given suggestive prompts.
- Language can be toxic but not contain any toxic word.
- Biases to identity, race, swear word i.e biased detoxification.
- Ethical issues.

Considerations in Curtailing Toxicity (cntd)

Debatable! But here are some instances (***Warning: There might be some offensive words***)

And these also includes contents that might not have any offensive words, but are very offensive.

One of the women turns to her friend and says to her, "Oh my gosh, I hate you too. Don't you feel bad for...|

Considerations in Curtailing Toxicity (cntd)

- These instances mean that the data-based approach might not be enough to perfectly curtail toxicity in language models. Or in other words, it might not be a holistic approach.

- The decoding-based approach helps control the generation process of the pre-trained LMs. Hence, the combination of both approaches can be viewed as a more holistic solution (although not perfect).

- Hence, the requirements for addressing toxicity in LMs would be further discussed with respect to both approaches and these listed considerations.

Requirements Engineering in Addressing Toxicity

Requirements Engineering in Addressing Toxicity

As we recall, Requirements engineering is the process of defining and detailing the specifications and requirements that should be provided by a software engineering process, with the end goal of satisfying stakeholders.

Hence, requirements engineering for addressing toxicity in language models can be defined as the process of employing the specifications and requirements that should be in place to address toxic generation in language models, ensuring they are safe for use to all stakeholders.

Who are Stakeholders?

Everyone actually.

- Technology Users.
- Companies.
- Data curators.
- Groups of people more prone to such languages.

What are the Requirements for Detoxification?

1. Basic: Identify toxic content which includes offensive words, links etc
2. Address context and nuance within languages.
3. Address detoxification bias.
4. Control generation process of pretrained LMs.

To achieve these requirements, two groups of specifications should be satisfied:

- Data-based specifications
- Decoding-based specifications

Specifications for Data-based strategies

To meet the need for non-toxic data, specifications for data-based strategies can be satisfied with any of the following methods (to mention a few):

- Careful selection of data used for pre-training and detoxification strategies.
- Additional pre-training of the language model with non-toxic data. [2]
- Attribute Conditioning: The use of “toxic” and “non-toxic” attributes in the training samples. [1].

Specifications for Decoding-based strategies (cntd)

Something to note with Decoding-based strategies is not all of them satisfy our need to control text-generation during detoxification:

- Blocklisting, for instance, entails reducing the probabilities of bad words at decoding time but there could still be instances with unsafe language using safe words.
- Self-debiasing uses the internal knowledge of a pretrained language model to reduce the probability of undesired attributes in the model generation. However it could act too aggressively and filter out harmless words and it does not maintain the same level of perplexity as the original model.

Specifications for Decoding-based strategies

The following approaches can satisfy this specification (to mention a few):

- An effective method is the use of PPLMs (Plug and Play Language Models): A PPLM is a simple model used as a discriminator (or attribute model), which guides the language generation of the LM. [4]
- Generative Discriminator: Use of an attribute-conditioned discriminator computes class likelihood using bayes rules. [5]

Future Work

Although the combination of these techniques address the issue of toxicity in language models to an extent, a recommendation would be to further explore which specific methods of both strategies work best together in addressing toxicity in LMs.

In conclusion

In engineering the requirements for addressing toxicity in language models, the total consideration of removing toxic content, addressing bias and controlling generation should be put in place. This also implies that the different methods for the detoxification specifications should be specifically considered, as not all address the problem to be solved.

References

[1] Toxicity in AI Text Generation. https://towardsdatascience.com/toxicity-in-ai-text-generation-9e9d9646e68f.

[2] Gehman, S., Gururangan, S., Sap, M., Choi, Y., & Smith, N. A. (2020). RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models. Findings of the Association for Computational Linguistics: EMNLP 2020, 3356–3369.

[3] Gururangan, S., Marasović, A., Swayamdipta, S., Lo, K., Beltagy, I., Downey, D., & Smith, N. A. (2020). Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. ArXiv:2004.10964 [Cs].

[4]. What Are Plug and Play Language Models. https://analyticsindiamag.com/what-are-plug-and-play-language-models-pplm-nlp/.

[5]. GeDi: A Powerful New Method for Controlling Language Models. https://blog.salesforceairesearch.com/gedi/

References

[6] Reducing Toxicity in Language Models. https://lilianweng.github.io/posts/2021-03-21-lm-toxicity/

[7] Predicting the Type and Target of Offensive Posts in Social Media. https://arxiv.org/abs/1902.09666

Thank You!