Battling bias and other toxicities in natural language generation

Kobielus, James . InfoWorld.com ; San Mateo (Mar 11, 2021).

ProQuest document link

ABSTRACT (ENGLISH)

[...]the author of record on an NLG-generated text may not realize if they are publishing distorted, false, offensive, or defamatory material. [...]recent research coauthored by scientists at the University of California, Berkeley; the University of California, Irvine; and the University of Maryland found that GPT-3 placed derogatory words such as “naughty” or “sucked” near female pronouns and inflammatory words such as “terrorism” near “Islam.” Slated to deliver a first iteration of this technology, known as GPT-Neo, as soon as August 2021, the intiative is attempting to, at the very least, match GPT-3’s 175 billion-parameter performance and even ramp up to 1 billion parameters, while incorporating features to mitigate the risk of absorbing social biases from training data. There’s a growing consensus that NLG professionals should rely on a set of practices that includes the following: * Avoid sourcing NLG training data from social media, websites, and other sources that been found to contain bias toward various demographic groups, especially historically vulnerable and disadvantaged segments of the population. * Discover and quantify social biases in acquired data sets prior to their use in developing NLG models. * Remove demographic biases from textual data so they won’t be learned by NLG models. * Ensure transparency into the data and assumptions that are used to build and train NLG models so that biases are always evident. * Run bias tests on NLG models to ensure that they are fit for deployment to production. * Determine how many attempts a user must make with a specific NLG model before it generates biased or otherwise offensive language. * Train a separate model that acts as an extra, fail-safe filter for content generated by an NLG system. * Require audits by independent third parties to identify the presence of biases in NLG models and associated training data sets.

FULL TEXT

NLG (natural language generation) may be too powerful for its own good. This technology can generate huge varieties of natural-language textual content in vast quantities at top speed.

Functioning like a superpowered “autocomplete” program, NLG continues to improve in speed and sophistication. It enables people to author complex documents without having to manually specify every word that appears in the final draft. Current NLG approaches include everything from template-based mail-merge programs that generate form letters to sophisticated AI systems that incorporate computational linguistics algorithms and can generate a dizzying array of content types.

The promise and pitfalls of GPT-3

Today’s most sophisticated NLG algorithms learn the intricacies of human speech by training complex statistical models on huge corpora of human-written texts.

Introduced in May 2020, OpenAI’s Generative Pretrained Transformer 3 (GPT-3) can generate many types of natural-language text based on a mere handful of training examples. The algorithm can generate samples of news articles which human evaluators have difficulty distinguishing from articles written by humans. It can also generate a complete essay purely on the basis of a single starting sentence, a few words, or even a prompt. Impressively, it can even compose a song given only a musical intro or lay out a webpage based solely on a few lines of HTML code.

With AI as its rocket fuel, NLG is becoming more and more powerful. At GPT-3’s launch, OpenAI reported that the algorithm could process NLG models that include up to 175 billion parameters. Showing that GPT-3 is not the only

NLG game in town, several months later, Microsoft announced a new version of its open source DeepSpeed that can efficiently train models that incorporate up to 1 trillion parameters. And in January 2021, Google released a trillion-parameter NLG model of its own, dubbed Switch Transformer.

Preventing toxic content is easier said than done

Impressive as these NLG industry milestones might be, the technology's immense power may also be its chief weakness. Even when NLG tools are used with the best intentions, their relentless productivity can overwhelm a human author's ability to thoroughly review every last detail that gets published under their name. Consequently, the author of record on an NLG-generated text may not realize if they are publishing distorted, false, offensive, or defamatory material.

This is a serious vulnerability for GPT-3 and other AI-based approaches for building and training NLG models. In addition to human authors who may not be able to keep up with the models' output, the NLG algorithms themselves may regard as normal many of the more toxic things that they have supposedly "learned" from textual databases, such as racist, sexist, and other discriminatory language.

Having been trained to accept such language as the baseline for a particular subject domain, NLG models may generate it abundantly and in inappropriate contexts. If you've incorporated NLG into your enterprise's outbound email, web, chat, or other communications, this should be ample cause for concern. Reliance on unsupervised NLG tools in these contexts might inadvertently send biased, insulting, or insensitive language to your customers, employees, or other stakeholders. This in turn would expose your business to considerable legal and other risks from which you might never recover.

Recent months have seen increased attention to racial, religious, gender, and other biases that are embedded in NLG models such as GPT-3. For example, recent research coauthored by scientists at the University of California, Berkeley; the University of California, Irvine; and the University of Maryland found that GPT-3 placed derogatory words such as "naughty" or "sucked" near female pronouns and inflammatory words such as "terrorism" near "Islam."

More generally, independent researchers have shown that NLG models such as GPT-2 (GPT-3's predecessor), Google's BERT, and Salesforce's CTRL exhibit larger social biases toward historically disadvantage demographics than was found in a representative group of baseline Wikipedia text documents. This study, conducted by researchers at the University of California, Santa Barbara in cooperation with Amazon, defined bias as the "tendency of a language model to generate text perceived as being negative, unfair, prejudiced, or stereotypical against an idea or a group of people with common characteristics."

Leading AI industry figures have voiced misgivings about GPT-3 based on its tendency to generate offensive content of various sorts. Jerome Pesenti, head of Facebook's AI lab, called GPT-3 "unsafe," pointing to biased and negative sentiments that the model has generated when asked to produce text about women, Blacks, and Jews.

But what truly escalated this issue with the public at large was the news that Google had fired a researcher on its Ethical AI team after she coauthored a study criticizing the demographic biases in large language models that are trained from poorly curated text datasets. The Google research found that the consequences of deploying those biased NLG models fall disproportionately on marginalized racial, gender, and other communities.

Developing techniques to detoxify NLG models

Recognizing the gravity of this issue, researchers from OpenAI and Stanford recently called for new approaches to reduce the risk that demographic biases and other toxic tendencies will inadvertently be baked into large NLG models such as GPT-3.

These issues must be addressed promptly, given the societal stakes and the extent to which very large, very complex NLG algorithms are on a fast track to ubiquity. Several months after GPT-3's launch, OpenAI announced that it had licensed exclusive use of the technology's source code to Microsoft, albeit with OpenAI continuing to provide a public API so that anyone could receive NLG output from the algorithm.

One hopeful, recent milestone was the launch of the EleutherAI grassroots initiative, which is building an open source, free-to-use NLG alternative to GPT-3. Slated to deliver a first iteration of this technology, known as GPT-

Neo, as soon as August 2021, the intiative is attempting to, at the very least, match GPT-3’s 175 billion-parameter performance and even ramp up to 1 billion parameters, while incorporating features to mitigate the risk of absorbing social biases from training data.

NLG researchers are testing a wide range of approaches to mitigate biases and other troublesome algorithmic outputs. There’s a growing consensus that NLG professionals should rely on a set of practices that includes the following:

* Avoid sourcing NLG training data from social media, websites, and other sources that been found to contain bias toward various demographic groups, especially historically vulnerable and disadvantaged segments of the population.

* Discover and quantify social biases in acquired data sets prior to their use in developing NLG models.

* Remove demographic biases from textual data so they won’t be learned by NLG models.

* Ensure transparency into the data and assumptions that are used to build and train NLG models so that biases are always evident.

* Run bias tests on NLG models to ensure that they are fit for deployment to production.

* Determine how many attempts a user must make with a specific NLG model before it generates biased or otherwise offensive language.

* Train a separate model that acts as an extra, fail-safe filter for content generated by an NLG system.

* Require audits by independent third parties to identify the presence of biases in NLG models and associated training data sets.

NLG toxicity may be an intractable problem

None of these approaches is guaranteed to eliminate the possibility that NLG programs will produce biased or otherwise problematic text in various circumstances.

Toxic and biased content will be a tough issue for the NLG industry to address with a definitive approach. This is clear from recent research by NLG researchers at the Allen Institute for AI. The institute studied how a dataset of 100,000 prompts derived from web text correlated with the toxicity (the presence of ugly words and sentiments) in the corresponding textual outputs from five different language models, including GPT-3. They also tested different approaches for mitigating these risks.

Sadly, researchers found that no current mitigation method (providing additional pretraining on nontoxic data, filtering the generated text by scanning for keywords) is “fail-safe against neural toxic degeneration.” They even determined that “pretrained language models can degenerate into toxic text even from seemingly innocuous prompts.” Just as concerning were their findings that toxicity “can also have the side effect of reducing the fluency of the language” generated by an NLG model.

No clear path forward

Well before the NLG industry addresses these issues from the technical standpoint, they may have to accept increased regulatory burdens.

Some industry observers have suggested regulations that mandate products and services to acknowledge when they generate text through AI. Under the Biden administration, we may see renewed attention to NLG debiasing under the broader heading of “algorithmic accountability.” It would not be surprising to see the reintroduction of the Algorithmic Accountability Act of 2019, a bill that was proposed by three Democratic senators and went nowhere under the prior administration. That legislation would have required tech companies to conduct bias audits on their AI programs, such as those that incorporate NLG.

OpenAI has admitted that there may be no hard-and-fast solution that eliminates the possibility of social bias and other toxic content in NLG-generated text, and the issue is not limited solely to implementations of GPT-3. Sandhini Agarwal, an AI policy researcher at OpenAI, recently said that a one-size-fits-all, algorithmic, toxic-text filter may not be possible because cultural definitions of toxicity keep shifting. Any given piece of content may be toxic to some people while innocuous to others.

Recognizing that algorithmic bias may be a dealbreaker issue for the entire NLG industry, OpenAI has announced that it won’t broadly expand access to GPT-3 until it’s comfortable that the model has adequate safeguards to protect against biased and other toxic outputs.

Considering how intractable this problem of algorithmic bias and toxicity is proving, it wouldn’t be surprising if GPT-3 and its NLG successors never evolve to that desired level of robust maturity.

Crédito: James Kobielus

DETAILS

Subject:	Language; Human bias; Data acquisition; Islam; Demographics; Iterative methods; Social networks; Websites; Terrorism; Researchers; Mathematical models; Algorithms; Parameters; Digital media; Data sets; Training; Speech recognition; Natural language; Accountability; Bias
Business indexing term:	Subject: Social networks
Location:	California
Company / organization:	Name: OpenAI; NAICS: 541715
Publication title:	InfoWorld.com; San Mateo
Publication year:	2021
Publication date:	Mar 11, 2021
Section:	analysis
Publisher:	Infoworld Media Group
Place of publication:	San Mateo
Country of publication:	United States, San Mateo
Publication subject:	Computers--Microcomputers, Computers--Computer Industry
Source type:	Trade Journals
Language of publication:	English
Document type:	News
ProQuest document ID:	2500164667
Document URL:	https://search.proquest.com/trade-journals/battling-bias-other-toxicities-natural-language/docview/2500164667/se-2?accountid=8330
Copyright:	Copyright Infoworld Media Group Mar 11, 2021
Last updated:	2021-03-12

Database: ProQuest One Academic

LINKS

Check for full text via 360 Link