TheEyeCorpus: Experiments in Reducing NLP Bias and Identifiability for Large LMs

Julian Herrera
Simp Labs

David Bernal
Editor

Abstract

NLP is a constantly changing and evolving subset of artificial intelligence, due to constantly increasing data requirements for SOTA (state of the art) machine learning models such as GPT-3((Brown et al., 2020)), we often neglect the safety and privacy of our data producers. Using pure raw data often presents challenges, such as legalities and identifiability, solving these issues requires taking a machine-to-machine corpus approach, using source data to synthesize novel data. However, machine-to-machine approaches often involve a prolonged data gathering period, while also being computationally costly to produce. Rather than building from the ground up I believe publicly available datasets are needed. Building off of previous work from Gretel.AI (https://gretel.ai/), I propose TheEyeCorpus, a machine synthesized publicly available text corpus obtained from the public The-Eye Discord server and preprocessed and synthesised by Simp Labs utilizing Gretel AI's APIS and utilities. The data is hosted at github.com/puffy310/TheEyeCorpus

DISCLAIMER: I do not endorse scraping Discord servers, as it is a terms of service violation and illegal, I am not legally liable for any decisions and criminal offenses you may make.

1 Introduction

For decades, engineers and scientists have been obtaining and sharing knowledge about natural language. Soon after the early origins of machine learning we became intrigued by teaching machines to process language. Language has been used as a qualitative benchmark for progress in machine learning for decades. In the advent of machine learning research focused on analyzing, interfering, and other purposes, it has remained constant the requirement for larger amounts of data. The purpose of this publication is to utilize synthesis techniques to produce novel data based on a source dataset, thus adding more privacy, and satisfying end users. In most NLP applications, we sample raw data. A recent example is GPT-3(Brown et al., 2020), a state of the art large language model, that used terabytes of the mostly unfiltered Common Crawl. Making corpuses is usually as simple as downloading and extracting the common crawl database. This isn't the most ideal approach, Since using raw data has a lot of side effects, such as increased behavior issues, and generally limits the quality of production.

2 Data's Effect on Large Language Models

NLP researchers have recently had to use unfiltered data more, and there are deep effects on the model's ability to be ethical. A example proving this is TruthfulQA(Lin et al., 2021). GPT-3(Brown et al., 2020) performs so unethically, because the attacks are designed to bring out the worst in its knowledge. For example when asked who caused 9/11, GPT-3(Brown et al., 2020) answers, "The attack in New York on September 11th was caused by the terrorist group al-Qaeda" However, when included the word "really" it answers "George W. Bush" due to the fact that adding this new word into the sentence brings out worse data in GPT-3(Brown et al., 2020). Even with noise resistant architectures, large language models engineers have a difficult time producing safe language models. When data is manually produced models are safer and more dependable, however researchers do not have appropriate resources to manually produce hundreds of gigabytes of data. Researchers need a better approach.

3 Data Gathering

I am not legally liable for any decisions you make regarding Discord scraping as it is against their terms of service. However it is part of the research process to disclose all methods. Using a CLI tool known as Discord Chat Exporter (https://github.com/Tyrrrz/DiscordChatExporter) it is possible to scrape and record a public discord server, I must clarify I used CSV formatting in all future references to files. This CSV was a blueprint for the synthetic data that will be generated. Although the technical details of DiscordChatExporter are irrelevant, it is a tool required for this task, before I get experienced with the Common Crawl(http://commoncrawl.org/).

4 A Novel Approach to Data Processing

People have more computational power than manpower. A more optimal method to gather and produce a dataset is to use your original data as a blueprint for an initial large language to synthesize and add to. The approach I chose to take was using a pre-existing API from a company known as Gretel.AI. Gretel.AI provides reducton in dataset identification, and a model trainer for making synthesis models. The preprocessing necessary to ensure the safety of our data providers. Our data providers include the people who created these chat logs. The privacy and lack of identifiable information are very important for the ethical standing of AI models. Researchers cannot filter everything however. Adding extra layers of protection and scrambling allows for more headroom and legal requirements, which is a great thing for the democratization of Artificial Intelligence, hence with the source data being hidden, it is not relevant to the language model because Language models need to know data, not to know who made the data. Synthetic and preprocessed data have numerous advantages. When deciding the tools I would use, I chose Gretel.AI for numerous reasons, such as reliability, and free computation for jobs under an hour. Gretel.AI is a great choice for organizations not interested in sending incredible amounts in Research and Development. Synthetic Data is only applicable to Language and text, in the future data reducing identification will be necessary for many modals. Releasing synthetic data also affects releases, as copyright is no longer applicable. These are the various reasons why Gretel.AI's APIS and utilities are the most effective choice for this work.

5 Human Evaluation

Although this publication doesn't use a machine benchmark I've observed that text makes sense and is readable. Grammar and other elements of English language have deteriorated, This doesn't matter for this proof of concept, but it would be counter intuitive to teach Artificial Intelligence unoptimal grammar and syntax. It is generally very good for the low amount of records provided.

6 Limitations

Although the generation is not as concise as I'd hoped for it to be At this implementation it is unusable for any large language model, especially including challenges such as computational expenses. In future publications however, it will be possible to train a Large LM.

7 Conclusion

NLP and Data privacy should go hand and hand. Building on this research I hope to produce the CommonMachineCorpus, as a proof of concept this is great, as research, no, however as I improve in my personal life publications will get better and more advanced. Thank you for reading my short paper. The next one will be more professional. If anyone reading would like to observe the data, it is at https://github.com/puffy310/TheEyeCorpus.

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.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods.