Creating a framework for a Benchmark Religion Dataset

Deepa Muralidhar[1], Ashwin Ashok[2]

[1]Georgia State University
[2]Georgia State University

Abstract

Development of Language Models (LM) such as OpenAI's GPT series generating natural language text is growing at a rapid pace. The LMs take a text prompt as input and generate text as output that represent the most probable sequence of words matching the prompt's context and pattern. Our preliminary investigations revealed bias but not much evidence for its cause. Our goal is, therefore, to build a benchmark dataset on various religions for evaluating this bias. We envision that our conceptual method of creating a dataset and developing a bias rating mechanism can serve as a fundamental tool establishing a process to measure bias. Comparing the Bias Indicator Value (BIV) for one religion against another should give us enough information to provide a holistic bias rating for the text generated.

Keywords

Large Language Models, religious bias, algorithmic bias, bias, metrics, Text mining, socioeconomic factors, mitigating,

1. Introduction and Motivation

Our research on religious bias targets two key challenges for AI text generators: (1) The need for a religion-based benchmark data set to evaluate an AI text generator for bias. The problem with existing data sets is that there are no clear documentation and other data set management practices in place [1]. A well-designed benchmark data set helps verify that the output data is unbiased across a diverse distribution of real-world contexts [2]. (2) To present indicators of bias,a quantitative value that can represent the implicit bias numerically. Our key observation is that it is challenging to create a bias metric in LMs as the bias changes depending on the context of the text.

2. Target design goals

We conduct experiments to measure the sentiment of the AI generated text and test for religious bias. For every prompt, we program GPT-3 to generate 200 tokens, approximately 178 words. For six sets of 20 prompts (five religions and one religion-neutral that acts as control value) the text has about 11100 words (12000 tokens). Using VADER [3], a rule-based tool that measures the sentiment within the text, we measure positive and negative values (between -1 to +1) and compare the sentiment value of each sentence. We use this to create a preliminary quantifiable metric,a Bias Indicator Value, to identify stereotypical bias with respect to a religion and interpret why LMs generate texts differently for different religions. This metric which acts as a indicator of religious bias in GPT-3 is computed for the generated data. We share our results of these experiments with the community through a preliminary religion dataset that includes textual prompts, the test data, the associated AI-generated output text and the analysis done of the text. The metrics and graphs calculated are part of this benchmark dataset.

3. Open Challenges

An open research question for future work is that one metric maybe insufficient to measure biases, instead look to develop a bias-reporting toolkit. This could include transparency measurement, an examination on the how and why of the decision-making process in an AI system, is useful in detecting systemic biases.[4]

References

[1] K. Peng, A. Mathur, A. Narayanan, Mitigating dataset harms requires stewardship: Lessons from 1000 papers (2021). URL: https://arxiv.org/abs/2108.02922.

[2] P. P. Liang, C. Wu, L. Morency, R. Salakhutdinov, Towards understanding and mitigating social biases in language models, 2021. URL: https://arxiv.org/abs/2106.13219.

[3] G. E. Hutto, C.J., Vader: A parsimonious rule-based model for sentiment analysis of social media text, 2015.

[4] J. Stoyanovich, Transfat: Translating fairness, accountability and transparency into data science practice, 2018. URL: http://ceur-ws.org/Vol-2417/paper1.pdf.

Human in the Loop Data Curation Workshop'22: CKIM '22, October 17–21, 2022, Atlanta, Ga

deepa.muralidhar@gmail.com (D. Muralidhar); aashok@gsu.edu (A. Ashok)