

Addressing Fairness in Large Language Models

Students: Erick Pelaez Puig, Elnaz Toreihi, Valeria Giraldo, Jeslyn Chacko

Mentor: Wenbin Zhang

Instructor/Faculty: Masoud Sadjadi, Florida International University

PROBLEM

Large Language Models (LLMs) have been widely adopted in various applications. However, inherent biases within these models can lead to unfair treatment of different demographic groups, reinforcing stereotypes and perpetuating inequalities.

To address this issue, my project focuses on evaluating biases framework for bias evaluation. using four key metrics: Categorical Bias Score (CBS), All Unmasked Likelihood (AUL), Counterfactual Sentiment, and Gender Polarity.

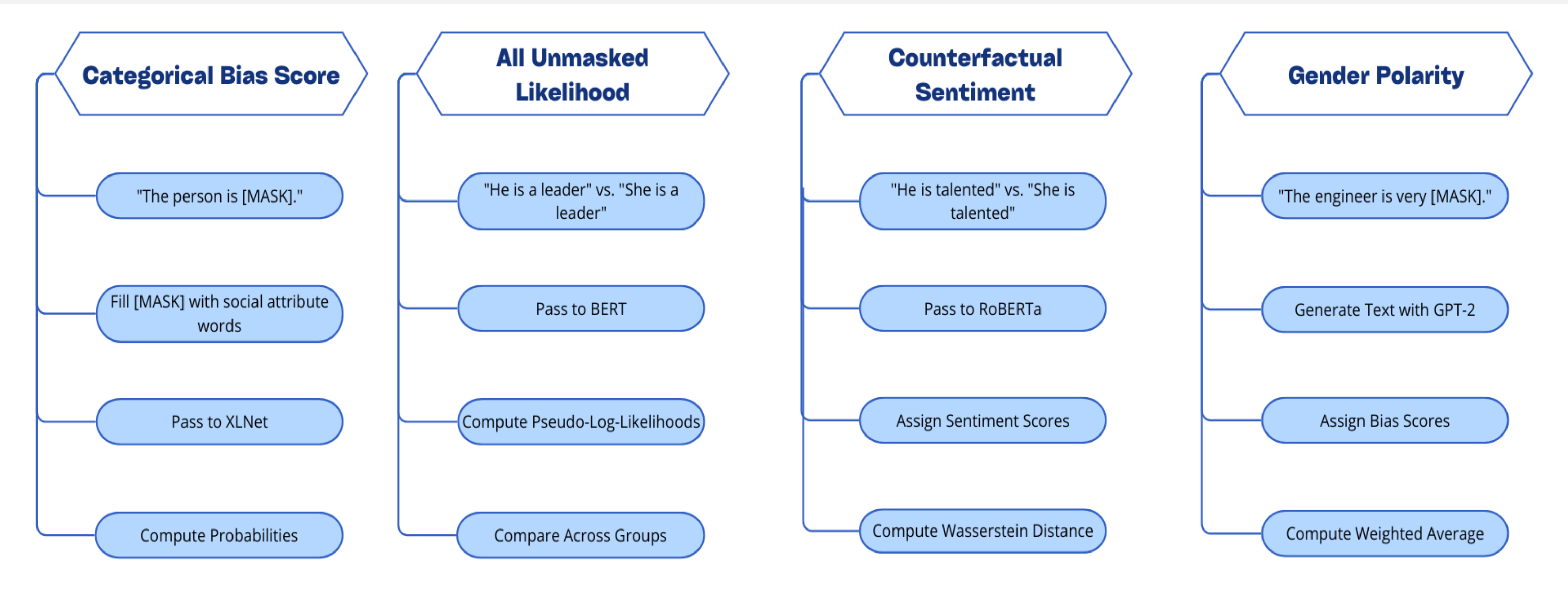
CURRENT SYSTEM

LLMs such as BERT, GPT, and RoBERTa are known to exhibit biases. Existing approaches focus on general fairness but lack metric-specific evaluations to identify nuanced biases in sentence generation, sentiment, and demographic association.

REQUIREMENTS

- Develop a framework for testing the four assigned metrics (CBS, AUL, Counterfactual Sentiment, and Gender Polarity).
- Use datasets like CrowS-Pairs and the Gender Lexicon to evaluate fairness across demographic groups.
- Conduct experiments to measure and compare biases across LLMs.
- Provide actionable insights for bias mitigation.

SYSTEM DESIGN



OBJECT DESIGN

- CBS:** Customized templates filled with demographic-related words, variance computed over CrowS-Pairs.
- AUL:** Sentence pairs differing only by protected attributes are tested for likelihood consistency.
- Counterfactual Sentiment:** Sentiment scores are derived using Wasserstein-1 distance to measure distribution shifts.
- Gender Polarity:** Gender bias is quantified by averaging bias scores for gendered words across generated text.

SUMMARY

This project provides a focused evaluation of fairness in LLMs using four specific metrics. Results reveal areas where models like GPT-2, BERT, and RoBERTa exhibit biases, particularly in sentiment, gender representation, and probability distributions. These findings offer pathways for improving fairness in LLMs.

IMPLEMENTATION



REFERENCES

Paleaz, E., Toreihi, E., Giraldo, V., & Chacko, J. (n.d.). ERICK0726/metrictesting. GitHub. <https://github.com/Erick0726/MetricTesting>

Wang, A. (n.d.). Sent-bias/tests at master · W4NGATANG/sent-bias. GitHub. <https://github.com/W4ngatang/sent-bias/tree/master/tests>

Zhang, W., Wang, Z., & Chu, Z. (n.d.). Fairness in large language models: A taxonomic survey. <https://arxiv.org/pdf/2404.01349>

ACKNOWLEDGEMENT

The material presented in this poster is based upon the work supported by Masoud Sadjadi. We thank Masoud Sadjadi for his assistance and mentorship that I/we received throughout the senior design project.