Individual Fairness in Generative Text Models

Brian Whelan, Master of Science in Computer Science

University of Dublin, Trinity College, 2024

Supervisor: Andrea Patanè

Generative text models, specifically Large Language Models (LLMs), have bridged the communication divide between humans and computers by enabling people to use natural language to interact with state-of-the-art models. This advancement raises concerns about the extent to which such models perpetuate biases and prejudices through their responses. However, in spite of this concern, research in bias and fairness for generative text models remains sparse. Hence, this dissertation aims to add to the existing literature by proposing a method for algorithmically evaluating the individual fairness of generative text models, where individual fairness captures the notion that similar individuals should be treated similarly.

This work formally defines the notion of an 'individual' within generative text model fairness and quantifies the similarity across individuals through the definition of a similarity metric. A fairness criterion is then defined which encodes the notion of individual fairness by specifying that the distance between the responses for a given generative text model given some input prompts, should be no greater than the distance between the respective input prompts, where the distance is quantified using the similarity metric. This fairness criterion is incorporated into existing dataset-based methods for identifying biases in NLP models to propose, to the best of the author's knowledge, the first method for assessing individual fairness in a generative text model.

Evaluating this method against two state-of-the-art generative text models with known biases using two different similarity metrics, the results offer positive evidence for incorporating additional context, through a similarity metric, into methods for evaluating individual fairness in generative text models. While the method proposed is not without limitations, it can hopefully serve as initial motivation for future work in individual fairness in generative text models.