Abstract

GENBIT: MEASURE AND MITIGATE GENDER BIAS IN LANGUAGE DATASETS

Kinshuk Sengupta, Global Data Science and Analytics, kisengup@microsoft.com
Rana Maher, Global Data Science and Analytics, rana.maher@microsoft.com
Declan Groves, Global AI, degroves@microsoft.com
Chantal Olieman, Global AI, chantal.olieman@microsoft.com

Keywords. Gender bias, responsible AI, AI for good, machine learning, language modeling, text analytics.

Natural Language Processing (NLP) systems have shown incredible results while solving business problems as part of many automated solutions. However, it has become clear that many NLP systems suffer from various biases often inherited from the data on which these systems are trained. The prejudice is exhibited at multiple levels spilling from how individuals generate, collect, and label the information leveraged into datasets. Datasets, features, and rules in machine learning algorithms absorb and often magnify such biases present in datasets. Therefore, it becomes essential to measure preferences at the data level to prevent unfair model outcomes. The paper introduces GenBiT, a tool to measure gender bias. The model is designed based on word co-occurrence statistical methods. In addition to measuring bias, a novel approach for mitigating gender bias is introduced based on contextual data augmentation powered by language models combined with random sampling, sentence classification, and filtering on targeted gendered pieces of data to eliminate unintended gender bias in multilingual training data. Our experiments demonstrate that this ensembled mitigation approach can ensure historical gender biases are reduced in conversational parallel multilingual datasets. This facilitates fairer machine learning model training over the augmented datasets to improve fairness and inclusiveness across a range of potential model applications.