Abstract

Gender bias is the inclination towards or unfair discrimination against individuals of one gender in favour of individuals of other gender(s). Natural language text that exhibits such prejudice or inclination is said to be gender biased. The presence of gender-biased text in textbooks has been widely established. However, attempts to detect gender bias in textbooks have been reliant on extensive manual inspection.

This thesis is an attempt to automate the process of gender bias identification in textbooks. Computationally, this involves addressing the issues of name gender inference, detection of linguistic markers of gender, and incorporating contextual information to identify gender-biased text. Towards this aim, we explore the resilience of named entity recognition when applied to text extracted from textbook PDFs, the problem of gender identification of named entities, and finally, detecting gender bias in the text by leveraging linguistic markers of gender.

We propose a data-driven adversarial evaluation for named entity recognition that is able to determine the extent to which named entity recognition relies on case and context. We incorporate named entity recognition with contextual information in a bidirectional cascading transformer network to identify the gender of named entities using supervised learning. We also explore unsupervised name gender inference using coreference resolution and multi-context aggregation using retrieval. We are able to achieve state-of-the-art results for name gender inference using the proposed methods when evaluated on five English language datasets.

We explore neural models that incorporate contextual information at different levels of abstraction to identify gender-biased text. Finally, we address the issues of global context dependence, locality of bias, and data sparsity using pooled memory augmented transformers for jointly learning to identify linguistic markers of gender and gender bias in a text.

The importance of high recall, in addition to a high F1 score for evaluating gender bias identification in textbooks, is explored. The results indicate that

automated gender bias identification in textbooks with high recall and F1 scores is computationally viable. We also assess the need for gender identification and explore the detection of linguistic markers of gender as a viable alternative. Our findings suggest that linguistic gender markers, rather than the presence of any specific gender, are a better indicator of the presence of gender bias in a text. Evaluation is carried out on four English language gender bias-labelled school textbook datasets to establish the generalizability of our proposed approaches for different subject domains.

To summarise our contributions, we create three sets of datasets for the tasks of named entity recognition, name gender inference, and gender bias detection from the textual content of textbooks. We design an adversarial evaluation of named entity recognition models. We present a supervised and an unsupervised technique for name gender inference. We describe a novel neural approach for gender bias detection. Finally, we propose a memory-augmented transformer-based approach for gender bias detection in textbooks.

Keywords: gender bias detection, gender identification, entity classification, text classification, memory augmented transformers.