Bias in Hate Speech and Toxicity Detection

Paula Reyero Lobo
paula.reyero-lobo@open.ac.uk
Knowledge Media Institute, The Open University
United Kingdom

ABSTRACT

Many Artificial Intelligence (AI) systems rely on finding patterns in large datasets, which are prone to bias and exacerbate existing segregation and inequalities of marginalised communities. Due to their socio-technical impact, bias in AI has become a pressing issue. In this work, we investigate discrimination prevention methods on the assumption that disparities of specific populations in the training samples are reproduced or even amplified in the AI system outcomes. We aim to identify the information from vulnerable groups in the training data, uncover potential inequalities in how data capture these groups and provide additional information about them to alleviate inequalities, e.g., stereotypical and generalised views that lead to learning discriminatory associations. We develop data preprocessing techniques in automated moderation (AI systems to flag or filter online abuse) due to its substantial social implications and existing challenges common to many AI applications.

CCS CONCEPTS

• **Social and professional topics** → *Hate speech*; *Codes of ethics*; • **Computing methodologies** → **Artificial intelligence**; *Knowledge representation and reasoning*.

KEYWORDS

Bias, Artificial Intelligence, Toxic Speech, Semantic Web

ACM Reference Format:
Paula Reyero Lobo. 2022. Bias in Hate Speech and Toxicity Detection. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES'22), August 1–3, 2022, Oxford, United Kingdom.* ACM, New York, NY, USA, 1 page. https://doi.org/10.1145/3514094.3539519

Hate speech and toxicity detection systems have limitations that may compromise their applicability to real-world scenarios. We draw attention to the three concerns in this research context. First, the inequalities of representation in the data of minority communities compared to the general population can lead to erroneous or less precise patterns learned by the system. Second, the lack of reliable ground truth due to the complexity of assigning labels to this type of subjective text limits the volume of available training data and the level of consistency across the available datasets. Third, the limitation of existing audits for measuring bias in this type of Natural Language Processing (NLP) application to only a list of a few words to represent groups with protected attributes can not account for the dynamic nature of vulnerable groups that become targets. Our crucial question focuses on reducing gaps in the representation and quality of vulnerable populations in the data to control bias in the models. Semantic Web (SW) technologies are a promising source, as their information is machine-understandable and time and domain-independent.

1 SEMANTICS AND BIAS IN ARTIFICIAL INTELLIGENCE

Given the ability of SW technologies to support different AI processes, we investigate how semantics can be a "tool" to address bias in different algorithmic scenarios. Our systematic review reveals that Knowledge Graphs (KGs), ontologies and lexical resources will play a more critical role in enabling and applying fairness, explainability and data preprocessing techniques. This research finds challenges and recommendations for assessing and addressing issues in the data and building resources that are useful in developing semantic techniques to counter bias in NLP and other AI applications.

2 SUPPORTING TOXICITY DETECTION WITH KNOWLEDGE GRAPHS

Our study uses background knowledge from frequently attacked gender and sexual orientation groups to better understand the problem of missing target information in toxic speech annotations. Using the Gender, Sex, and Sexual Orientation ontology reveals that 3% of 19k texts mentioned these groups, but annotators did not correctly identify them. This contribution focusing on validating baseline annotations intends to be a starting point. Using knowledge about specific demographic groups to assist in the annotation or to enrich and preprocess datasets in which their information remains unidentified are promising directions for building more consistent and reliable training datasets for the detection of online abuse.

3 FUTURE WORK

Three elements are central to dealing with the identified challenges: the data content, labels, and bias metrics. Our research contribution so far has shed light on content issues. We need to dive into how this content is being used as ground truth for the system, and how to integrate this set of tools and best practices to generate a bias-aware framework for designing automated detection systems.

ACKNOWLEDGMENTS

This work has received funding from the European Union's Horizon 2020 research and innovation programme under Marie Sklodowska-Curie Actions (grant agreement number 860630) for the project "NoBIAS-Artificial Intelligence without Bias".

AIES'22, August 1–3, 2022, Oxford, United Kingdom

ACM ISBN 978-1-4503-9247-1/22/08.
https://doi.org/10.1145/3514094.3539519