Can We Identify and Dismantle “ISMs” that Plague Our Society: An Online Approach

Keywords: Society, Bias, Discrimination, Artificial Intelligence, Machine Learning

Extended Abstract

As humans, we are predisposed to several biases, i.e., we have been imprinted with negative beliefs, social prejudices, and stereotypes about certain demographic groups that has led to discrimination, harassment, and abuse both offline and online. Online spaces provide a sense of sanctity to avoid offline abuse; however, when the underlying training data contains biases from user-experiences, the algorithms trained on them learn these biases and reflect them into their predictions. As a result, existing biases in data can affect the algorithms using the data, producing biased outcomes. Algorithms can even amplify and perpetuate existing biases in the data. In addition, algorithms, themselves, can display biased behavior due to certain design choices, even if the data itself is not biased, thus, making online spaces susceptible to *oppression.* Oppression is a systematic social phenomenon based on the differences between social groups e.g., *racism, sexism, ableism, ageism, size-ism, homophobia/heterosexism,* etc. Consequently, social media remains a hostile environment as there has been an alarming increase in hate speech, offensive and abusive language, towards certain demographic groups. To examine the presence of online oppression, we aim to investigate how large language models (LLMs) e.g., BERT [2] and GPT-2 [4] perform on texts containing several forms of oppression: (1) *Personal or Individual* – attitudes, behaviors, socialization, interpersonal interactions, self-interest, (2) *Cultural* – values, norms, language, standards of beauty, holidays, society’s expectations, music, aesthetics, religious values, and (3) *Institutional* – housing, employment, education, media, health care, politics, government, court system, non-profits, medicine, business, religion, family.

Motivation. Dismantling the “ISMs” that plague our society in the form of social bias and discrimination has a long history in philosophy and psychology, and recently in machine-learning (ML). However, to be able to fight against discrimination and we first need to identify *fairness*. We define fairness as the absence of any prejudice or favoritism towards an individual or a group based on their demographic, intrinsic or acquired traits in the context of decision-making of online algorithms. Even though fairness is an incredibly desirable quality in society, it can be surprisingly difficult to achieve in practice. By identifying these social biases in popular LLMs, we aim to highlight the need for real-world solutions to minimize online social issues and reduce feelings of online despotism and subjection. We posit it is vital for the development and deployment of AI technologies to account for minimal stereotypical systemic and sociodemographic biases against protected classes such as *individual*, *cultural, and institutional* forms of oppression in downstream tasks.

Method. We define *explicit bias* when a sentence containing intrinsic features such as *race, color, national origin or ancestry, religion, creed, sex, physical or mental disability, medical condition, marital status, age, sexual orientation, citizenship*, or status as a covered *veteran*, and any other category protected by applicable federal, state, or local laws produces a less favorable response than a person without demographic identity terms such as “gay” and “Muslim” or similar scenarios. To investigate explicit bias, we will examine several subgroups of “ISMs”, namely, *racism, sexism, ableism, ageism, size-ism, homophobia/heterosexism, classism, xenophobia, transphobia, and antisemitism*. Sentences in all groups will be in the form of several subgroup template variations: For example, (a)

A/an <race> person is *<verb>*, (b) *A/an <gender identity>* [person] is *<verb>*, & (c) *A <religion>* [person] is *<verb>, etc*. The [person] tag for each variation is populated from [1] which includes two of the largest benchmark datasets of possessive (gender-specific and gender-neutral) nouns and attribute (career-related and family-related) words datasets to date, and the *<verb>* tag is generated at random. We aim to answer the fundamental question, "*how we can actively work to dismantle 'isms' that plague our society in online environments and celebrate our diversity?*". We aim to generate ~500,000 sentences in attempts to maximize a diverse dataset, then we intend to implement Detoxify [3], a multi-headed BERT-based toxic comment detection model capable of detecting different types of toxicity such as *threats*, *obscenity*, *insults*, and *identity-based attacks* and discovering unintended bias in both English and multilingual toxic comments. Specifically, Detoxify[1] is created using Pytorch Lightning[2] and Transformers[3], and fine-tuned on datasets from 3 Jigsaw challenges, namely Toxic comment classification, Unintended Bias in Toxic comments and Multilingual toxic comment classification for multi-label classification to detect toxicity across a diverse range of conversations. Due to the nature of the generated text using GPT-2, we will employ Amazon Mechanical Turk (AMT) annotators to manually rate 10,000 randomly sampled generated samples from each subgroup to measure the effectiveness of the toxicity classifier, due to the examples of stereotypes, profanity, vulgarity, and other harmful languages towards certain subgroups.

To address the above question, in this paper, we aim to intersect AI and social issues by implementing an ML algorithm and two LLMs to generate sentences and readily identify toxic textual content. Firstly, as humans, and secondly as AI practitioners, scientists, and researchers we have a lifelong responsibility to unlearn the oppressive attitudes sustained in society and to dismantle stereotypical systemic and sociodemographic biases against protected classes offline and online. During the conference we intend on discussing the results, i.e., figures and tables where we investigate each subgroup in greater detail. Furthermore, we will detail information about figures on data breakdowns, distribution plots (*i.e.*, to depict the variation in the data distribution), and knowledge of which words constitute a "harmful" or "non- harmful" comment for each generated sample.

References

[1] Jamell Dacon and Haochen Liu. 2021. Does gen- der matter in the news? detecting and examining gender bias in news articles. In Companion Proceedings of the Web Conference 2021, WWW '21, page 385–392, New York, NY, USA. Association for Computing Machinery.

[2] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of NAACL-HLT, pages 4171–4186.

[3] Laura Hanu and Unitary team. 2020. Detoxify. Github. https://github.com/unitaryai/detoxify.

[4] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

[1] https://github.com/unitaryai/detoxify

[2] https://www.pytorchlightning.ai

[3] https://huggingface.co/docs/transformers/index