Document 1: Case study - gender bias in machine learning

Machine learning models work by identifying patterns in data and making predictions by attempting to replicate them (Wei, 2020). These patterns are assumed to reflect reality, because they emerge from data concerning actual events and phenomena that have already occurred. In Natural Language Processing (NLP), these patterns can be easily biased (Sun et al., 2019; Wei, 2020). Because NLP deals with the link between computer understandings of human language, it is subject to a complex combination of factors including ambiguity, context and nuance (Hirschberg and Manning, 2015). It is true that languages follow certain rules (grammar, syntax), however these rules are flexible due to cultural differences, local dialects, colloquialism and so forth (Chowdhury, 2003). Literary and lingual devices like irony and sarcasm further complicate understanding (Chowdhury, 2003). Finally, implicit cultural bias can result in patterns that cause bias in machine learning predictions (Zhao et al., 2018). Issues like systemic racial prejudice and gender role bias can and do make their way into machine learning predictions (Sun et al., 2019; Wei, 2020). An example of this machine learning bias is the association of ‘black’ to ‘criminal’ and ‘Caucasian’ to ‘police’, or ‘doctor’ to ‘man’ and ‘nurse’ to ‘woman’ (Ethayarajh, 2020; Wei, 2020).

Despite gender equity making great strides globally in the past several decades, there remains a well-documented gap in society when it comes to female participation in STEM careers (Beede et al., 2011; Robnett, 2016; Wu et al., 2020). Long-held public beliefs regarding gender roles, combined with social expectations associated with these roles has resulted in a lower participation rate of women than men in ‘STEM’ related fields (Robnett, 2016; Sun et al. 2019). In recent years, there has been a renewed push, driven by the accessibility of social media, to raise awareness of this gap and encourage women to enter these careers, and to work to identify hurdles to their progression into more senior roles (Australian Government, 2019). However, as this report show, there is still a significant bias towards men in STEM data, and studies show that a lack of representation can discourage women from pursuing certain careers (Wu et al., 2020).

This report will investigate gender bias in Natural Language Processing and machine learning. The data will be derived from Twitter, due to its role as a repository of global attitudes and sentiments over time (Pozzi et al. 2016). Specifically, it will investigate the association of gender to Tweets relating to ‘science’, ‘technology’, ‘engineering’ and ‘mathematics’ – the so-called ‘STEM’ fields.

NLP is an ideal tool for uncovering implicit bias and even understanding the breadth and potentially the source of these biases, and ensuring that these biases do not reinforce harmful prejudices (Sun et al., 2019; Wu et al., 2020). This analysis shows that it is important to first identify the source of the bias in order to effectively mitigate it.