Course: Natural Language Processing

Language (Technology) is Power: A Critical Survey of “Bias” in NLP

Oreen Yousuf

November 30, 2020

1 Abstract

An important aspect of most fields of Natural language Processing (NLP) is building a capable model that can ensure accurate, relevant and in specific cases - non-partisan information. However, as these models are created by people there will be oversight in their capabilities in handling sensitive information, or inadvertently and incorrectly categorizing elements of a text source. This is bias, and correcting a model to mitigate bias is seen by some as a "gold standard".

2 Approach

The relationship between language and social hierarchies, and measurable corrections are rarely addressed when academic studies refer to their own bias in academia or industry. Simply acknowledging a failure of oversight is insufficient in the overarching attempt to mitigate bias. Misrepresentation of people arises from models created with the intent of exuding an air of authority. For example: case studies featuring system preferences when illustrating limitations in tagging parts of speech for texts containing African American English fall short by simply mentioning constraints of the trained tagger. Racial bias was, however, never considered as an element leading to the limitations - limitations which in turn can lead to the "findings" of the paper being a misconstrued representation. This flow of information ultimately leads to tangible reality in the form of negative perceptions for groups that ultimately aid in the completion of the cycle of yet again influencing future NLP models.

3 References

Chen, M., Weinberger, K., and Blitzer, J. 2011. Co-Training for Domain Adaptation

Proceedings of the 7th Joint Conference on Lexical and Computational Semantics (*SEM). 2018.

April Baker-Bell (2019): Dismantling anti-black linguistic racism in English language arts classrooms: Toward an anti-racist black language pedagogy, Theory Into Practice