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

Oreen Yousuf

December 3, 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 could be considered as an unfair inclination against the represented peoples, and correcting a model to mitigate this potential prejudice 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: a case study performed by Kiritchenko, S. Mohammad, S.[1] found that a model used both for gender and race biases found that there was a significant difference in the calculated the difference between the average predicted score on the set of sentences with African American names and the average predicted score on the set of sentences with European American names for anger, fear, and sadness intensity predictions. Conversely, European American names yielded significant calculated differences for joy and valence. Evaluating these results, within the scope of this study, can allow one to note that African-Americans are more associated with negative emotions. The literature guiding thinking in and outside of the industry gives way to models such as these. Another study by Baker-Bell, A.[2] found that hierarchy for language variation corresponding to race is innately set at early ages and is difficult to ascertain why that is. Gathering case studies such as this one can illustrate that system preferences are unintentionally limited in mitigating biases. 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. An admirable strive in NLP technology can be seen Chen, M., Weinberger, K., Blitzer, J. in their CODA algorithm, which is driven by "[bridging] the gap between source and target domains by slowly adding to the training set both the target features and instances in which the current algorithm is the most confident.[3]" Methodology such as this can attempt to diminish the discrepancies between banks of information that associate more with less thought of or unfairly treated racial or gender related language.

3 References

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

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

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

Proposal Paper: Blodgett, S., Barocas, S., Daumé III, H., Wallach, H. 2020. Language (Technology) is Power: A Critical Survey of "Bias" in NLP