DISCOVERING AND CATEGORISING LANGUAGE BIASES IN REDDIT

Xavier Ferrer,
Tom van Nuenen,
Jose Such,
Natalia Criado

Department of Informatics,
King’s College London

{xavier.ferrer_aran, tom.van_nuenen,
jose.such, natalia.criado}@kcl.ac.uk

KING'S College LONDON

EPSRC
Engineering and Physical Sciences
Research Council

This paper proposes a data-driven approach to automatically discover language biases encoded in the vocabulary of online discourse communities on Reddit. In our approach, protected attributes are connected to evaluative words found in the data, which are then categorised through a semantic analysis system.

Through the use of word embeddings and similarity metrics which leverage the vocabulary used within specific communities, we are able to discover biased concepts towards different social groups when compared against each other. This allows us to forego using fixed and predefined evaluative terms to define language biases, which current approaches rely on.

- We verify the effectiveness of our method by comparing the biases we discover in the Google News dataset with those found in previous literature. We then successfully discover gender bias, religion bias, and ethnic bias in different Reddit communities such as r/theRedPill, r/Atheism and r/TheDonald.

- Our method can help in understanding and measuring social problems and stereotypes towards certain populations in communities with more precision and clarity and using the community's own language. Most importantly, it can help identify biases and dangerous stereotypes in language models before deployment.

- Our method could be used to underpin tools to help administrative bodies of web platforms to discover and trace negative and dangerous biases in online communities to decide which do not conform to content policies.

- Even though our method is automated, it is designed to be used with a human in the loop. It is necessary a minimum knowledge of the community to correctly understand the wider context surrounding the conceptual biases found.