Introduction to Accountability, Evaluation and Obscurity of AI Algorithms Minitrack

Radmila Juric
ALMAIS Consultancy, UK
radjur3@gmail.com

Robert Steele
Capitol Technology University
Laurel, MD 20708, USA
rjsteele@captechu.edu

This Minitrack has attracted very interesting submissions and we have chosen to accept two.

It is obvious that we are still puzzled with the evaluation and obscurity of AI algorithms, and research communities across the world are clear that have not conquered AI accountability. The co-chairs have agreed that it is quite difficult to address the visible shortcomings of AI and guarantee that we will have a means of evaluating its algorithms very soon, but we are also very excited, with the submitted papers, because they pave the way towards AI algorithms which address all three: evaluation, obscurity and accountability.

The accepted submissions are very different and illustrate various ways of thinking and ideas which are hidden behind our concerns on AI. One paper gives a comprehensive explanation and possibilities of removing human bias when using natural language processing methods and the other is focused on ethical principles in AI which have been overlooked, and potentially harmed marginalized communities.

The paper entitled "Comparing Methods for Mitigating Gender Bias in Word Embedding" gives a comprehensive comparison of debiasing methods, built on top of GloVe, in order to determine which method is the best in removing bias. It would be of interest to natural language processing and AI evaluation enthusiasts, because word embedding, which transforms texts into numeric vectors and reproduces similar words with similar vector representations, can carry forward prejudices of societies and thus bring along human biases. The authors have defined HSR-RAN-GloVe method and compared it with different debiasing methods with the aim of determining a) which methods perform the best at removing bias and b) if these methods truly remove bias from word embedding. The results are interesting. There is no method that outperforms throughout their analysis and no method leads to a decrease in performance for either similarity or semantics. However, RAN-GloVe is probably the method that on average gets better results in debiasing word embedding. In addition, it is worthwhile extracting from this paper that, for their SemBias dataset, the RAN-GloVe word embedding has remarkable accuracy in identifying the gender-definition pairs. It does not confuse the gender-definition with the gender-stereotype pairs, which can be extremely useful in any real-life applications.

The paper entitled "Applying Reflexivity to Artificial Intelligence for Researching Marginalized Communities and Real-World Problems" has been created by US social scientists who claimed that the problems of ethical AI are due to a lack of critical insight into the complex positionality of the researcher, power dynamics between scholars and the communities being studied, and the structural impact on real-world problems when AI systems appear to be accurate but ethically fail. When addressing AI ethics, the authors apply 5 stages of "reflexivity", a process that yields a better understanding of community-specific nuances, areas requiring local expertise, and the potential consequences of scholastic interventions for real-world problems. They claim that as the focus of AI shifts to more practical, real-world interventions, it begins to enter the space that social work has inhabited. Therefore they advocate for AI researchers to be flexible by engaging in a more reflexive and ethical approach with the understanding that their AI model may produce unintended consequences that affect lives and livelihoods of vulnerable/ marginalized communities. We encourage readers to pay attention to the last paragraph of the paper, where the authors say that they "believe that the application of social work ethics and approaches to data science can possibly prevent future mistakes in our research with communities".

URI: https://hdl.handle.net/10125/102718
978-0-9981331-6-4