A Story of Discrimination and Unfairness

Implicit Bias Embedded in Language Models

Aylin Caliskan-Islam
Princeton University
aylinc@princeton.edu

Joanna Bryson
Princeton University
University of Bath
jjb@alum.mit.edu

Arvind Narayanan
Princeton University
arvindn@cs.princeton.edu

1. TALK OVERVIEW

Bias can be present individually in humans or collectively in corporations, groups, or cultures. We observe collective and individual implicit bias through analyzing writing in an automated way. Automating bias observations is possible through incorporating machine learning and natural language processing techniques to text analysis. The use of such an automated method makes large scale analysis possible with a variety of settings to compare and contrast bias in different conditions, such as subject of interest, time, location, culture, and language. Our proposed approach is a step towards a principled method for quantifying bias and fairness in language models that are used digital communications.

Machine learning algorithms and models have been criticized for incorporating bias from the data they train on. Eliminating bias in machine learning has been limited to controlling parameters at the algorithmic level to avoid overfitting, which does not prevent implicit bias from getting embedded to the model and revealing itself at the contextual level. Based on this knowledge, we train language models on writings of subjects of interest to generate a semantic space represented with word embeddings. Each word embedding quantifies a word used by the subject as a vector, where the dimensions of the vector represent a combination of contexts. We focus on the numeric vectors of concepts that have been used in bias studies in the literature, such as gender, racism, religion, and age. Then, we measure associations between concepts and potentially biased terms to observe implicit bias through spatial relations.

We investigate bias in famous individuals, the Enron corporation, Wikipedia, Twitter, and Google News. We uncover bias at different levels, even when the data comes from Wikipedia, which has a neutrality and objective writing policy. We conclude with a discussion of the implications of bias that is present in large language models which are being widely used in digital communications for text generation and summarization, automated speech, and machine translation. How can we engage policymakers and developers to enable algorithmic transparency and fairness?

1.1 Machine learning decides for all of us

Machine learning models are widely used for various applications that end up affecting billions of people and Internet users everyday. Random forest classifiers guide the US drone program to predict couriers that can lead to terrorists in Pakistan[1]. Employers use algorithms, which might be racist[2], to aid in employment decisions. Insurance companies determine health care or car insurance rates based on machine learning outcomes. Internet search results are personalized according to machine learning models, which are known to discriminate [2] against women by showing ads with lower salaries, while showing higher paying job ads for men. On the other hand, natural language processing models are being used for generating text and speech, machine translation, sentiment analysis, and sentence completion, which collectively influence search engine results, page ranks, and the information presented to all Internet users within filter bubbles. Given the enormous and unavoidable effect of machine learning algorithms on individuals and society, we attempt to uncover implicit bias embedded in machine learning models, focusing particularly on language models.

1.2 Human semantics as language models

Linguists and machine learning researchers have been trying to generate word embeddings that closely represent human semantics similar to how the human brain operates. Recent advances in neural networks and availability of big data and computational power has led to high quality language models such as word2vec [7] and GloVe [8]. These language models, which consist of up to half a million unique words, are trained on billions of documents from sources such as Wikipedia, CommonCrawl, GoogleNews, and Twitter. The words in the models are quantified as numeric vectors whose dimensions correspond to a combination of contexts. Such numeric word representations are able to accurately solve semantic and syntactic analogy tasks. For example, after presenting the relation between Rome and Italy, we can ask the language model to find the word that would generate the same relation for France. The language model is able to understand that Rome is Italy's capital and for France, the model outputs Paris to complete the country to capital analogy. Such language models form the basis of most accurate state-of-the-art named entity recognizers, sentiment classifiers, machine translators, and text generators. Consequently, the models are widely distributed and incorporated into a large variety of applications. The ideal of machine-based perfection overlooks the fact that most recent progress

[1] https://www.theguardian.com/science/the-lay-scientist/2016/feb/18/has-a-rampaging-ai-algorithm-really-killed-thousands-in-pakistan

[2] https://technical.ly/philly/2016/05/12/solon-barocas-hiring-racism-big-data/

in artificial intelligence (AI) is not derived by reasoning from first principles, but by mining human intelligence. Recently, there has been considerable concern that such an approach could introduce an unacceptable amount of bias into AI. Based on human implicit bias studies in psychology [6], we propose a method to uncover implicit bias in language models which demonstrates that semantics derived from natural language texts contains biases endemic in the culture that created the texts.

The Implicit Association Test[3] (IAT) measures attitudes and beliefs that people might have or might not be aware of, which they are not able to or unwilling to report. For example, you might believe that men and women should be equally associated with science but the implicit associations might reveal that you associate men with science more than you associate women with science. To replicate this experiment on language models, we analyze words and concepts that have been used in IATs which have been taken by millions of individuals. We investigate gender, race, religion, and age bias with valence classifications and association analysis of potentially biased concepts. We are able to perform this analysis through spacial associations in the high dimensional space via clustering, classifying and projecting context dimensions to human understandable dimensions. For example, we find associations between females and domestic terms when men are associated with career terms. We observe negative valence towards the elderly, certain races, and certain religions.

1.3 Uncovering implicit bias

Researchers have not agreed on a perfect strategy for evaluating word vector quality in language models for several reasons. Different methods generate vectors that are better in different tasks [3]. Some models mainly capture semantic similarity, some can easily find related words, and some are better at identifying syntactic similarity [4]. On the other hand, word frequency in the training data causes more frequent words to have similar vectors. Also, the majority of word embeddings do not handle homonyms. Given these shortcomings, vectors are still evaluated by correlation of word similarity rankings of humans and word cosine similarity, and accuracy in syntactic and semantic analogies, sentence completion tasks, and sentiment analysis [10]. Given these four evaluation strategies, we modify them to measure bias instead of quality in language models that reach state-of-the-art accuracy in quality evaluations. In word cosine similarity, we calculate spearman's correlation coefficient between Harvard IAT results from two million people and our cosine similarity for the same tasks. For syntactic and semantic analogies, we generate analogies to test IAT concepts. For example, if father is to doctor, mother is to what? And the model associates mother with nurse, based on the relation between father and doctor. For sentence completion tasks, we come up with representative IAT questions such as: '______ is career oriented.' and '______ is family oriented.' Given two options, man and woman, how would the model complete each sentence? And again, men end up being associated with career while women are associated with family. For sentiment analysis, we perform valence tests. For example, the two sentences 'This person is young.' and 'This person is old.' are classified as being good or bad. Besides these evaluation techniques, we convert the language models to sparse vectors in order to make context dimensions more easily interpretable [9] to allow for fine grained analysis of bias associations.

1.4 Need for transparency and policy

In this work, we demonstrate that semantic understanding derived from large-corpus linguistics contain the history and prejudices of the cultures that created the texts. Mining culture comes with an associated cost. Being based on search as much as reasoning, culture is a product of its own history, and therefore necessarily caries with it assumptions and precedents that may not be current, efficient, or even acceptable. Used as a basis of for example algorithmic decision making, mined bias unaddressed could conserve discrimination, even masking intentional prejudicial corporate behavior behind the veneer of neutral machine learning [1, 5]. We show that the implicit biases known to be displayed by ordinary humans can be replicated by this same mechanism. These findings amplify the immediate need for transparency in algorithmic decisions and fairness and non-discrimination policy for machine learning applications.

[3] https://implicit.harvard.edu/

2. REFERENCES

[1] S. Barocas and A. D. Selbst. Data's disparate impact. *California Law Review*, 104, 2015.

[2] A. Datta, M. C. Tschantz, and A. Datta. Automated experiments on ad privacy settings. *Proceedings on Privacy Enhancing Technologies*, 2015(1):92–112, 2015.

[3] M. Faruqui, Y. Tsvetkov, P. Rastogi, and C. Dyer. Problems With Evaluation of Word Embeddings Using Word Similarity Tasks. *ArXiv e-prints*, May 2016.

[4] M. Faruqui, Y. Tsvetkov, D. Yogatama, C. Dyer, and N. Smith. Sparse overcomplete word vector representations. In *Proceedings of ACL*, 2015.

[5] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 259–268, New York, NY, USA, 2015. ACM.

[6] A. G. Greenwald, B. A. Nosek, and M. R. Banaji. Understanding and using the implicit association test: I. an improved scoring algorithm. *Journal of personality and social psychology*, 85(2):197, 2003.

[7] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119, 2013.

[8] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543, 2014.

[9] F. Sun, J. Guo, L. Yanyan, X. Jun, and X. Cheng. Sparse word embeddings using l1 regularized online learning. In *Proceedings of international joint conference on artificial intelligence (IJCAI)*, 2016.

[10] D. Yogatama, M. Faruqui, C. Dyer, and N. A. Smith. Learning word representations with hierarchical sparse coding. *arXiv preprint arXiv:1406.2035*, 2014.