Conceptual and methodological problems with bias detection and avoidance in natural language processing

Alicja Dobrzeniecka

2021-06-19

Contents

Chapter 1

Introduction

Natural language processing (NLP) is a subfield of computer science that processes and analyzes language in text and speech with the use of modern programming methods. It has practical applications in everyday life as it concerns tasks such as email filters, smart assistants, search results, language translations, text analytics and so on. Models used to accomplish these tasks need a lot of data to learn from. This data originates from human activities and historical recordings such as texts, messages or speeches. It turns out that in the learning process these models can learn implicit biases that reflect harmful stereotypical thinking still present in modern societies. One can find methods that aim at identifying and measuring hidden biases and/or try to remove them by modifying the models. There are many different types of models in NLP depending on a task that they are supposed to solve. However, all of them need as an input words represented by means of numbers and this is accomplished with word embedding models. The models usually assign the values based on the context in which the words appear. This means that the input data can have enormous influence on the outcome. The biases seem to have their primary source in the way the words are assigned numerical values.

One can find various definitions trying to capture what bias and fairness actually are. With the choice of the definition, implications into the real-life applications may change as well. Mehrabi, Morstatter, Saxena, Lerman, & Galstyan (2019) mark out that there exist different types of biases such as historical bias, representation bias or measurement bias (the list is long). This indicates how complex the issue of bias is. Without the proper understanding and awareness of the problem, people are prone to unconsciously sustain the bias. Mehrabi, Morstatter, Saxena, Lerman, & Galstyan (2019) also distinguish different types of discrimination, some of them will be briefly described. By protected attributes we mean those qualities, traits or characteristics that one should not discriminate against. Direct discrimination occurs when protected attributes of individuals explicitly result in non-favorable outcomes toward them. Such discrimination takes place for example when one uses the information about someone's religious beliefs to evaluate whether they are competent for the work they apply for. In contrast, in indirect discrimination individuals appear to be treated equally, but anyway they end up being treated unjustly due to the hidden effects of biases towards their protected attributes. An example of such situation is when in a job advert there is a requirement of 10 years experience instead of the list of a specific type of experience and knowledge. This job advert can discriminate indirectly against young people who can have the required skills, but not yet so much work experience. Systemic discrimination takes place when policies, customs or behaviors that result from certain culture or organizational structure lead to discrimination against some groups of people. Finally, statistical discrimination consists in relying on group statistics to judge person belonging to a given group. The topic of discrimination is entangled with another concept; fairness. It is essential to grasp some concepts of fairness to take them into consideration while designing implementation of some machine learning model. In Mehrabi, Morstatter, Saxena, Lerman, & Galstyan (2019) one may notice that depending on the context and application different definitions may be applied.

There is a considerable amount of literature available on the topic of bias detection and mitigation in NLP models. Bolukbasi, Chang, Zou, Saligrama, & Kalai (2016) focus on gender biases that may be observable while investigating the representation of job occupations and gender in terms of their assigned numerical values. The authors apply cosine similarity measurement to investigate the phenomenon where (the vectors corresponding to) words related to jobs that are stereotypically associated with a given gender are in fact in the model situated closer to this gender. They also use analogy tasks to evaluate if bias is present in a word embedding model. They check analogies by comparing pairs of word vectors. For example they search for the word complementing the puzzle: man is to doctor as woman is to …? First they subtract the word "man" from the word "woman"[1] and then they search for the ranked list of other words pairs that have similar vector difference. They also include in the formula a threshold to ensure that the resulting pairs could not be randomly picked.

However, as Nissim, Noord, & Goot (2019) points out, there are some limitations to this approach. In practice, most of analogies implementations do not return any input words. This means that it does not make sense to expect the algorithm to return the same profession for both woman and man. Therefore, this method seems to be limited in terms of bias detection. Other problems are related for example to the choice of pairs and words that are used to detect the presence of discrimination, as it is often subjective and without proper justification. Additionally, the choice of the parameter used in Bolukbasi, Chang, Zou, Saligrama, & Kalai (2016) formula to ensure that word pairs are not picked by random, is also not justified and changing it drastically influences the results.

Islam, Bryson, & Narayanan (2016) touch upon the topic of biases regarding race and gender. They apply knowledge from well-known psychological studies such as Implicit

[1]This is pointwise subtraction of vectors, but for the simplicity we will often talk about words while really meaning vectors. What is meant should we clear from the context.

Association Test to investigate the relation between human stereotypical thinking and model-learnt biases to discover a close relationship between these two. For the evaluation they use Word Embedding Association Test (WEAT) and the Word Embedding Factual Association Test (WEFAT).

Manzini, Lim, Tsvetkov, & Black (2019) propose a novel way of using cosine similarity to verify if the assumed resemblance between protected words and harmful attributes exists. They investigate an approach that enables them to measure the bias for a class (such as gender, religion or race) and express the final result with a single metric.

Islam, Bryson, & Narayanan (2016) refer to Implicit Association Test (Greenwald et al., 1998) that measures the strength of associations between concepts or stereotypes by measuring human reaction time for special tasks. Humans naturally exhibit some biases which do not always cause social concern. One can imagine the intuitive associations between for example insects and flowers, and the feelings of pleasantness or unpleasantness. In general, people would rather associate flowers with feeling pleasant than insects, and this preference in a sense is a bias or prejudice in some direction. However, this type of preference does not cause an uproar and is a rather morally neutral case. This example is rather used to show that the methodology used in IAT makes sense. Islam, Bryson, & Narayanan (2016) were able to obtain similar results to the ones that measured reaction time with the use of WEAT metric. Unfortunately, further studies discover other biases and prejudices that directly influence the quality of other people's lives and therefore they should be taken care of.

The most common bias detection methods in the literature focus on comparing the similarity between words from protected groups and those that are considered to be stereotypical or harmful in some way. One can find in this group methods such as euclidean distance or cosine similarity (which is equivalent to dot product if the vectors are normalized). There are also other ways to detect the effects of biases. One can in-

vestigate the model performance on certain downstream tasks and validate if the model uses the gender or race information to complete the task. The currently used methods employing cosine similarity, often aggregate the similarity values often aggregated in a way that may lead to hasty conclusions. The averaging of values and the lack of attention to uncertainty involved in such aggregation may lead to the incomplete picture of the situation.

In the paper we indicate how current methods used to detect biases in natural language models are limited, if we investigate them from the perspective of Bayesian analysis. Our research enhances the current way in which the bias detection is performed to make sure that it is methodologically valid. The key hypothesis is that greater understanding of data and bias implications can be achieved when Bayesian methods are applied to issue.

Chapter 2

Cosine similarity and bias detection

Placeholder

2.1 Word embeddings

2.2 Cosine similarity and distance

2.3 Cosine distance in a one-class bias detection

2.4 Cosine distance in a multi-class bias detection

2.5 Limitations of the approach

Chapter 3

Walkthrough with the religion dataset

Placeholder

3.1 Loading and understanding the dataset

3.2 First look at the empirical distributions

3.3 Looking at the islam-related words

3.4 Bayesian model structure and assumptions

3.5 Choosing predictors

3.6 Dataset-level coefficients

3.7 Model structure and assumptions

3.8 Protected classes in Reddit and Google embeddings

3.9 Dataset-level coefficients after debiasing

3.10 Protected classes after debiasing

Chapter 4

Discussion

Placeholder

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. *CoRR*, *abs/1607.06520*. Retrieved from http://arxiv.org/abs/1607.06520

Islam, A. C., Bryson, J. J., & Narayanan, A. (2016). Semantics derived automatically from language corpora necessarily contain human biases. *CoRR*, *abs/1608.07187*. Retrieved from http://arxiv.org/abs/1608.07187

Manzini, T., Lim, Y. C., Tsvetkov, Y., & Black, A. W. (2019). Black is to criminal as caucasian is to police: Detecting and removing multiclass bias in word embeddings. Retrieved from http://arxiv.org/abs/1904.04047

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. *CoRR*, *abs/1908.09635*. Retrieved from http://arxiv.org/abs/1908.09635

Nissim, M., Noord, R. van, & Goot, R. van der. (2019). Fair is better than sensational: Man is to doctor as woman is to doctor. *CoRR*, *abs/1905.09866*. Retrieved from http://arxiv.org/abs/1905.09866