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

Alicja Dobrzeniecka

2021-06-09

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 humans 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. It 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 the numerical values.

There is considerable amount of literature available on the topic of bias detection and mitigation in NLP models. Bolukbasi, Chang, Zou, Saligrama, & Kalai (2016) focuses 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 the bias is present in the 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 word "man" from word "woman" and then they search for the ranked list of other words pairs that have similar vectors' difference. They also include in the formula a threshold to ensure that the resulting pairs could not be randomly picked.

However, as in Nissim, Noord, & Goot (2019) it is pointed out, there are some limitations of this approach. According to the authors 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. The other problems regard for example 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 parameter set 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) touches upon the topic of biases regarding race and gender. They apply knowledge from well-known psychological studies like Implicit Association Test to research the relation between human stereotypical thinking and

model learnt biases to discover 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) proposes a novel way of using cosine similarity to obtain the information on assumed resemblance between words. They investigate an approach that enables them to measure the bias for a class (like gender, religion, race) and express the final result with a single metric.

It is worth noticing the general distinction of biases mentioned in Islam, Bryson, & Narayanan (2016). They refer to the publication concerning Implicit Association Test Greenwald et al., 1998) where certain baseline of bias phenomenon was introduced. Namely it seems that humans naturally exhibit some biases and that not always bring 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 could be named a bias or prejudice in some direction. However this type of preference does not cause an uproar and it is rather morally neutral case. Unfortunately there are other biases and prejudices that directly influence the quality of other people's lives and therefore they should be taken care of.

One can find a bunch of 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 as it was pointed out in Mehrabi, Morstatter, Saxena, Lerman, & Galstyan (2019). They mark out that there exist different types of biases, the list if long but among others there are historical bias, representation bias, measurement bias. This indicates how complex the process itself is. Without the proper understanding and awareness of the problem, people are prone to unconsciously sustain this phenomenon.

In the article one can also find the distinction on different types of discrimination,

some of them will be shortly described. It is worth first mentioning that protected attributes are those qualities, traits or characteristics that one cannot, according to the law, discriminated against. Direct discrimination refers to the situation when protected attributes of individuals explicitly result in non-favorable outcomes toward them. 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. 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, very common statistical discrimination refers to using average group statistics to judge person belonging to the group.

The topic of discrimination is entangled with another concept which is fairness. It is essential to grasp some concepts of fairness to take them into consideration while designing implementation of some machine learning model. In Mehrabi2019Survey one may notice that depending on the context and application different definitions may be applied.

The most popular methods 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, dot product or cosine similarity. There are also other ways to detect the effects of biases. For example through the investigation of the model performance on certain tasks that validate if the model returns some values independently on gender or race or not.

In the currently used methods (like cosine similarity) the values of similarity are often aggregated in a way that may lead to false conclusions. For example due to the averaging of values and the lack of confidence interval information.

One can find a number of articles on negative real-life implications resulting from the presence of unaddressed biases in the machine learning models.

In the paper we indicate how current methods used to detect biases in natural language models are limited in terms of confidence interval.

Our research tries to answer the question of how to enhance the current way in which the bias detection is performed to make sure that it is methodologically valid.

Our hypothesis is that there can be greater understanding of data and bias implications when confidence interval and Bayesian method are applied to the methodology.

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 Selection and number of attributes

2.6 No control groups

2.7 Means of means

2.8 Interpretability issues

2.9 The curse of dimensionality

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

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