Bias in Word Embeddings

Armando Bellante
William Bonvini

April 2020

Contents

1 Introduction

1.1 Research Direction

Here we provide the reader with the research direction as presented to us at the time of the project:
"The overall goal of the work is to further analyze the gender bias dimension in the given word embedding context, to then pair it with other dimensions that are commonly sources of other stereotypes and biases, such as ethnicity, sexual orientation, and the like. Based on this, we then aim at proposing a way to incorporate in the analysis an overall "ethic" dimension, which should suitably encompass all the others mentioned above and possibly more. [...]. We first aim at analyzing how the categories with which the initial data-set in [Boulkbasi et al., 2016] is labelled treats agentic and communal terms, and whether they share common structure with the gender-connotated ones, Then, using the "debiasing" method proposed in [Boulkbasi et al., 2016] ,we want to analyze how such terms move across the gender dimension as a result of the "neutralization" phase, and understand if the re-positioning is linguistically meaningful. Should the changes in the position be not consistent from the linguistic point of view, we want to understand if and how it is possible to modify the proposed methodology in order to improve the results. Secondly, we will focus on using the dataset to expand the issue of stereotypes to other issues related to "diversity", such as: religion and ethnicity, and understand if the proposed method can be applied in a conceptually similar way in these contexts. Issue of intersectionality, i.e., the intersections of more dimensions, will be object of particular attention, as they are the most challenging from a conceptual viewpoint in summing many different biases flavours. Finally, we aim at how to define, by combining the aspects indicated above and possibly other relevant ones, a subspace that identifies "ethical" issues, studying if the neutralization procedures can be applied in this area, and possibly which other transformations might be interesting in this context."

1.2 Purpose

The aim of this project is to study [Boulkbasi et al., 2016] and [Pietraszkiewicz et al., 2019], and further investigate the methodologies for debiasing word embeddings. In addition to the two before-mentioned papers, we have identified and studied other two relevant papers: [Gonen et al., 2019] and [H. Zhang et al., 2020].
This report:

- overviews studies on bias identification and debiasing methods, focusing on their limitations and proposing promising research directions
- summarizes studies on the linguistic manifestations of two important psychological traits (agency and communion) that can be helpful to identify biased relationships among words

1.3 Fairness in Word Embeddings

The most popular way to represent words semantic is to define a word embedding.
A word embedding is a mapping of words of a vocabulary to an n-dimensional space such that words with similar meanings are close to each other and dissimilar ones are distant.
This representation of natural language is very powerful, though, studies carried on in the last few years have shown that it is not always fair. In recent years, a lot of focus has been put on fairness issues in machine learn tasks. [Boulkbasi et al., 2016] first showed how word embeddings could accidentally incorporate bias present in the text corpus used to generate them and may serve as a mean to propagate it. Indeed, machine learning models trained on biased embeddings could reflect bias in sensitive applications such as medicine or law.
It is then important to address this problem by finding out ways to identify the bias and mitigate it. The papers that we discuss approach bias identification in different ways depending on the particular design of the embeddings and on the real-world applications they are meant for. We have noticed that there are two approaches to bias elimination: postprocessing the embedding and modifying the loss function of the model that produces the embedding. In particular [Boulkbasi et al., 2016] and [H. Zhang et al., 2020] respectively adopt the above different approaches. However, the research on how to identify and remove the bias, as pointed out in [Gonen et al., 2019] is still open.

1.4 Sections Outline

We present a summary of each of the papers we have read, providing you with the goal of the author's research, a summary of the steps they have followed, and the conclusions they came up with. In the final section of the report we will discuss our point of views on the research state and highlight possible research directions.

2 Debiasing word embeddings

This paper discusses how bias is present in word embeddings, linking it to the geometry of the embedding space. The main embeddings taken into account are $word2vecNEWS$ and $GLoVe$.

2.1 Preliminary knowledge

2.1.1 Notation

We use w to identify a word and W for a Vocabulary. We denote as $N \subset W$ the set of words neutral w.r.t. the bias taken into account . We denote the cardinality of a set S as $|S|$. We define the corresponding vector in the word embedding space as $\vec{w} \in \mathbb{R}^d$, where $\mathbb{R}^d = \prod_{i=1}^{d} \mathbb{R}$. When writing a vector $\vec{v}$ we mean it as a column vector, $\vec{v}^T$ is the corresponding row vector. The i^{th} component of vector $\vec{v}$ is denoted as v_i. We indicate the scalar product in its usual notation as $\vec{v}_1 \cdot \vec{v}_2 = \vec{v}_1^T \vec{v}_2$. We define the Dirac Delta as δ_{ij}, which is equal to 1 when $i = j$ and 0 otherwise. We denote the subspace spanned by k vectors as $span(\vec{v}_1, ..., \vec{v}_k)$.

2.1.2 Geometry and linear algebra

This section is a brief recall to some linear algebra that will be useful to understand the debiasing methodology presented in this paper.
A vector is called unitary if its norm is one: $||\vec{v}|| = 1$, where $||\vec{v}|| = \sqrt{\sum_{i=1}^{d} v_i^2}$.
A vector can be always transformed in a unitary vector: $\vec{v} = \dfrac{\vec{v}}{||\vec{v}||}$. We deal with unit vectors. For two unit vectors $\vec{v}_1, \vec{v}_2 : cos(\vec{v}_1, \vec{v}_2) = \vec{v}_1 \cdot \vec{v}_2$, the two vectors are parallel if $cos(\vec{v}_1, \vec{v}_2) = 1$, anti-parallel if $cos(\vec{v}_1, \vec{v}_2) = -1$ and orthogonal if $cos(\vec{v}_1, \vec{v}_2) = 0$.

Let $B \subset \mathbb{R}^d$ be a subspace of $\mathbb{R}^d$ such that $B = span(\vec{b}_1, ..., \vec{b}_k)$, $||\vec{b}_i|| = 1$ and $\vec{b}_i \cdot \vec{b}_j = \delta_{ij}$. We can define its orthogonal space as $B^\perp = \mathbb{R}^d / B$.
A generic vector $\vec{v} \in \mathbb{R}^d$ can be written as $\vec{v} = \vec{v}_B + \vec{v}_{B^\perp}$, where $\vec{v}_B = \sum_{i=1}^{k} (\vec{v} \cdot \vec{b}_i)\vec{b}_i$, and $\vec{v}_{B^\perp} = \dfrac{\vec{v} - \vec{v}_B}{||\vec{v} - \vec{v}_B||}$. Note that $\vec{v}_B \cdot \vec{v}_{B^\perp} = 0$.

2.1.3 Word Embeddings

A word embedding is a representation of words of a vocabulary W on an euclidean space $\mathbb{R}^d$. Each vector $\vec{w} \in \mathbb{R}^d$ is a unit vector and therefore the similarity between two words can be computed as $cos(\vec{w}_1, \vec{w}_2) = \vec{w}_1 \cdot \vec{w}_2$.
Embedding models such as $word2vec$ receive as input a corpus of text $T = \{D_1, ..., D_n\}$ of n documents, and output the embedding $\vec{w} \in \mathbb{R}^d$ of each term in the corpus.

2.1.4 Associations in Word Embeddings

It is possible to play with the geometry of the space of the embedding to discover associations among words. In the embedding space it is possible to satisfy equations such as $\vec{King} - \vec{Man} + \vec{Woman} = \vec{Queen}$, which is basically equivalent to $cos(\vec{Woman} - \vec{Man}, \vec{Queen} - \vec{King}) \sim 1$.
In general we can say $\vec{a} : \vec{x} = \vec{b} : \vec{y}$ if $cos(\vec{a} - \vec{b}, \vec{x} - \vec{y}) \sim 1$.

2.2 Discovering the bias

2.2.1 Generating associations

The authors have defined a method to generate associations automatically. Given a pair of words $\vec{a}, \vec{b}$ it is possible to iterate over the couple of words $\vec{x}, \vec{y} : ||x - y|| \leq \delta$ for a certain δ and score the association by computing $S_{ab_{xy}} = cos(\vec{a} - \vec{b}, \vec{x} - \vec{y})$. After having generated several associations with gender related pair seeds, they have validated the method and the presence of bias by asking 10 crowd-workers to manually go through the associations. Out of 150 associations generated, 72 were reported as appropriate and 29 as stereotyped by 5 or more people.

2.2.2 Identifying gender neutral words

The main focus of the paper is on gender bias. An important task is the one of distinguishing neutral words such as doctor, nurse or professor from gender-words like mother, father and grandmother. In this research paper they show that it is possible to separate neutral words from gender words in the $w2v$ representation by training an SVM and achieving an F-score of about 0.627 ± 0.102.

2.2.3 Bias metric

The main intuition behind this research is that there is some bias reflected along the direction of the vectors. If we want to consider gender bias and we are able to identify a gender direction $\vec{g}$ (almost parallel to $(\vec{she} - \vec{he}), ..., (\vec{mother} - \vec{father})$, we can measure the bias of the embedding as:

$$\frac{1}{|N|} \sum_{w \in N} |cos(\vec{w}, \vec{g})|^c$$

Where c is a parameter, usually set to 1.

2.3 Removing the Bias

2.3.1 Identifying the Bias space

The first step to perform debiasing is to identify the subspace in which the bias lies. The main idea is to compute PCA over category of vectors that express the bias.

Procedure: (Identifying the Bias space) [Boulkbasi et al., 2016]
Inputs: word sets W, defining sets $D_1, D_2, ..., D_n \subset W$ as well as embedding $\{\vec{w} \in \mathbb{R}^d\}$ and integer parameter $k \geq 1$. Let

$$\vec{\mu}_i := \frac{1}{|D|} \sum_{w \in D_i} \vec{w}$$

be the centroid for each defining set. The bias subspace B is spanned by the first **k** rows of $SVD(C)$, where:

$$C := \sum_{i=1}^{n} \frac{1}{|D|} \sum_{w \in D_i} (\vec{w} - \vec{\mu}_i)(\vec{w} - \vec{\mu}_i)^T$$

In case of gender debiasing k=1 suffices, and the subspace is a spanned by a single vector.
After identifying the bias subspace B, one can proceed either with Hard Debiasing or with Soft Debiasing.

2.3.2 Hard Debiasing

The Hard Debiasing consists of two steps: Neutralization and Equalization. Neutralization removes the biased components from each neutral word. In Equalization, after defining some sets of equality (eg. $\{Man, Woman\}$, $\{Father, Mother\}$, etc.), each neutral word is set to be equidistant to each word in the equality sets, for each equality set.

Procedure: (Neutralization) [Boulkbasi et al., 2016]
Inputs: words to neutralize $w \in N$ and the bias subspace B. For each word in N:

$$\vec{w} := \vec{w}_{B^\perp}$$

Procedure: (Equalization) [Boulkbasi et al., 2016]
Inputs: equality sets $E_1, ..., E_m \subset W$. For each equality set compute the centroid and its neutral component:

$$\vec{\mu}_i := \frac{1}{|E_i|} \sum_{w \in E_i} \vec{w}$$

$$\vec{v}_i := \vec{\mu}_{B^\perp}$$

For each word inside E_i:

$$\vec{w} = \vec{v}_i + \sqrt{1 + ||v||^2} \frac{\vec{w}_B - \vec{\mu}_i}{||\vec{w}_B - \vec{\mu}_i||}$$

In this way distances among the bias space are preserved and distances of neutral words to biased words inside the same equality set are made equal.

2.3.3 Soft Debiasing

It is possible to reformulate the problem so that it is reduced to finding a linear transformation $T \in \mathbb{R}^d$ to apply to the matrix of terms $W \in \mathbb{R}^{d \times |W|}$. To solve the problem it is necessary to solve the semi-definite program

$$min_T ||(TW)^T(TW) - W^T W||_F^2 + \lambda ||(TN)^T(TB)||_F^2$$

Where λ is a hyperparameter.
As always, the output embedding is normalized to have unit length $\hat{W} = \{\frac{Tw}{||Tw||}, w \in W\}$.

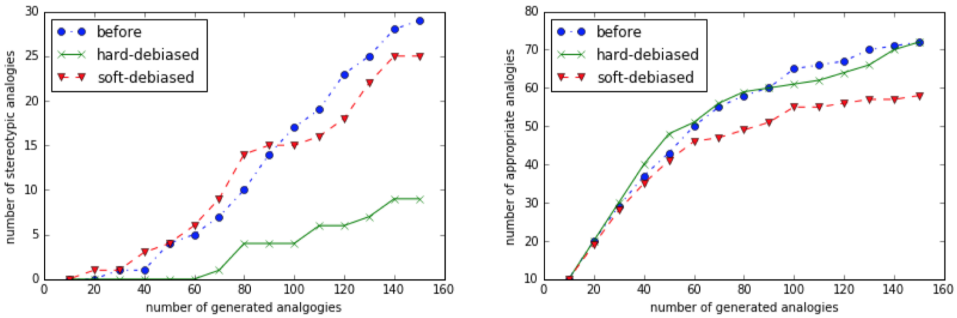

Figure 1: Debiasing Results

3 Lipstick on a Pig

This paper analyzes the contributions of [Boulkbasi et al., 2016] and [Zhao et al., 2018], discussing their measure of bias and claiming that those debiasing methods only cover the bias and do not remove it. This motivates further research on the essence of the bias.

3.1 Settings

Both [Boulkbasi et al., 2016] and [Zhao et al., 2018] offer methodologies to perform debiasing of word embeddings. The first paper removes the bias via post-processing of the embedding, the second one modifies the training function of the model that outputs the embedding. In this paper the authors discuss Hard Debiasing on $word2vecNEWS$ [Boulkbasi et al., 2016] and Gn-GLoVe (GenderNeutral-GLoVe) [Zhao et al., 2018], with a particular focus on gender-bias. Both the debiasing methods rely on the definition of $bias-by-projection$, as defined in [Boulkbasi et al., 2016]. However, machine learning models are able to model complex relationships among words. Then the main question is: Is bias more profound than what discovered till now?

3.2 Uncovering the Bias

The first experiment of this research work consists in trying to see whether it is possible to cluster together, in the debiased space, words that were associated with a particular bias direction in the original space. The authors select the top 500 male-biased and 500 female-biased words according to the $bias - by - projection$ bias definition. Both for Hard Debiasing and Gn-GLoVe they run k-means with $k = 2$ before and after debiasing. It turns out that yes, words that were biased towards males are still close to each other and the same goes for female-biased words.

	Before Debiasing	After Debiasing
Hard Debiasing	99.9%	92.5%
Gn-GLoVe	100%	85.6%

Table 1: 2-Means accuracy

3.2.1 Bias-by-Neighbours

This findings lead the authors to define another plausible metric to measure bias. The measure is called $bias - by - neighbours$ and, in the case of gender bias, it can be the percentage of male/female socially-biased words among the k-nearest neighbors of the target word. This measure is shown to be correlated to the $bias - by - projection$ on both the studied embeddings, with a Pearson correlation of 0.686 (compared to a correlation of 0.741 when checking neighbors according to the biased version) and 0.736 (compared to 0.773) respectively in case of $HardDebiasing$ and $Gn - GLoVe$.

3.3 Generalizing Bias Detection

After discovering that the information on the bias of neutral words is still present in the embeddings, the authors try to automatically separate male-biased words and female-biased words. The new experiment goes on on a dataset of 5000 gender-biased words (2500 male-biased, 2500 female-biased), split into a training set of 1000 samples and a test set of 4000 samples, both balanced. They train a RBF-kernel SVM and achieve great scores on the embeddings both before and after debiasing. This further shows the need to research for deeper and more complex bias definitions.

	Before Debiasing	After Debiasing
Hard Debiasing	98.25%	88.88%
Gn-GLoVe	98.65%	96.53%

Table 2: RBF-Kerenel SVM Accuracy

4 Hurtful Words: Quantifying Biases in Clinical Word Embeddings

This paper deals with bias identification and removal in contextual word embeddings models.

4.1 Introduction

Contextual models can encode marginalized populations differently, consequently perpetuating biases and performing better in certain prediction tasks on a population subgroup with respect to another. It is important to investigate on the quality of contextual word embeddings models both because of their popularity and because little studies have dealt with such topic successfully.
Healthcare is one of the fields in which a disparity on population subgroups could translate into different prediction outcomes with respect to the subgroup considered, so it makes sense to study bias in this context.
The authors identify such biases, discuss the implications of their existence and attempt to remove them, highlighting the limitations of their approach.
The steps they go through are the following:

1. Pretrain a deep embedding model on clinical notes (MIMIC-III dataset)
2. Identify dangerous latent relationships using a fill-in-the-blank method
3. Evaluate performance gaps with respect to population groups on clinical prediction tasks
4. Present their Adversarial Debiasing solution with its limitations

4.1.1 BERT Settings

The authors trained their own clinical BERT model. There are two main reasons:

1. None of the existing model use whole-word masking, a recent amendment to BERT pretraining which generally improves performance.
2. Existing model are not initialized on SciBERT, which outperforms all other models in many downstream tasks.

4.2 Identifying the Bias

4.2.1 Identify dangerous relationships via log Probability Bias Scores

The authors use *log probability bias scores*, an effective method for exposing bias in language models.
BERT receives as training input pairs of sentences and learns to simultaneously predict their ordering and a masked token for each sentence given the context

(we'll denote such tokens as [MASK]). The considered score takes into consideration the probability with which a sensitive attribute (e.g. gender = male) is predicted by BERT given a prompted sentence (e.g. [MASK] is a programmer).

For example, to compute the association between the target *male gender* and the attribute *programmer* we feed the masked sentence "[MASK] is a programmer" to BERT, and compute the probability assigned to "*he* is a programmer" (p_{tgt}).
To measure the association, however, we need to measure how much *more* BERT prefers the male gender association with the attribute *programmer*, compared to the female gender.
We thus re-weight this likelihood p_{tgt} using the prior bias of the model towards predicting the male gender.
To do this, we mask out the attribute *programmer* and query BERT with these sentence "[MASK] is a [MASK]", then compute the probability with which BERT fills the first mask with "he" (p_{prior}).
Intuitively, p_{prior} represents how likely the word *he* is in BERT, given the sentence structure and no other evidence.
In practice we compute the quantity

$$\text{Log Probability Bias} = \log \frac{p_{tgt}}{p_{prior}}$$

for each target (gender) and we call the absolute difference *log probability bias score*.

As expected, the mean score is statistically significantly high for some categories. Therefore training SciBert on clinical notes effectively transfers gender-related associations from the notes into the model.

4.2.2 Performance Gaps on Prediction Tasks

Three classification tasks are performed on the dataset:

1. **In-hospital mortality**
   Predict whether a certain patient has died during its stay in the hospital given the notes charted during its first 48 hours in the intensive care unit.

2. **Phenotyping using all notes**
   The classification task is to predict patient membership in one of the 25 HCUP CCS code groups [Harutyunyan et al., 2019].

3. **Phenotyping using first note**
   Same task as above but based only on the first note within the first 48 hours of stay in the Intensive Care Unit.

A binary classifier is trained for each of the categories appearing in the tasks above.
The sensible groups taken into considerations are gender, language, ethnicity,

and insurance status.
Such classifiers are evaluated using three definitions of fairness:

- **Demographic Parity**
  The proportion of each segment of a protected class (e.g. gender) should receive the positive prediction at equal rates.
- **Equality of opportunity for the positive class**
  Each group should get the positive prediction at equal rates assuming that people in this group qualify for it.
- **Equality of opportunity for the negative class**
  Each group should get the negative prediction at equal rates assuming that people in this group qualify for it.

To evaluate a fairness gap between two groups, the authors examine the difference between the performance in relevant tasks for such groups. Experiments show that performances are different and that better performances correspond, as expected, to majority groups in most of the protected classes examined.

4.3 Removing the Bias

To explore how the baseline clinical BERT model can be debiased, the authors use an established adversarial debiasing approach [Beutel et al., 2017].
Adversarial debiasing consists in training simultaneously two neural networks to compete with each other. The goal of the first one (predictive network) is to predict a certain token of the input sequence given the context (the words close by) while the second network (adversarial networks) wants to simultaneously learn the value of a sensitive attribute (e.g. race) starting by the context as well.
Denote X to be the input of BERT (the clinical notes) and Z to be the sensitive attribute we don't want to encode in the embedding (e.g. race). The intuition for adversarial debiasing is as follows: if our original model produces a representation of X that primarily encodes information about Z, an adversarial model could easily recover and predict Z using that representation.
Conversely, if the adversary fails to recover any information about Z, then we must have successfully learned a representation of the input that is not substantially dependent on our protected attribute. A schema taken from [Beutel et al., 2017] of adversarial debiasing is in Figure 2.

The authors train a debiased version of their clinical BERT with the above mentioned framework and they observe that it performs strongly overall, but even if debiasing slightly reduces the number of tasks for which there is a significant performance gap, this reduction is not sufficient for deployment in a high stake medical setting. Moreover, classifiers applied post-hoc can still extract information about sensitive attributes.

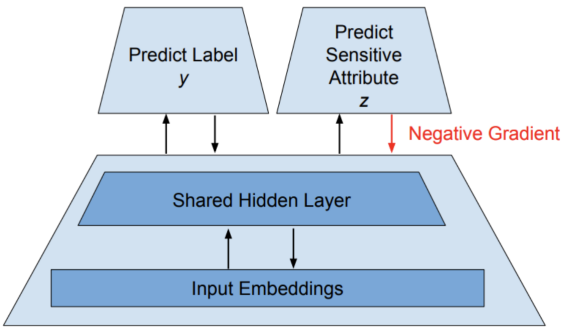

Figure 2: Adversarial Debiasing Approach

5 The Big Two Dictionaries

This paper aims at developing (study 1) and validating (studies 2,3,4) a dictionary (set of words) for the psychological traits identified as "Agency" and "Communion".

5.1 Introduction

We will report definitions as presented in [Abele, Uchronski et al., 2008].
"Agency" refers to a person's striving to be independent, to control one's environment, and to assert, protect and expand one's self.
"Communion" refers to a person's striving to be part of a community, to establish close relationships with others, and to subordinate individual needs to the common good. We refer to the traits described above as the "Big Two" psychological traits.
The authors first create two dictionaries respectively containing terms for Agency and Communion and then validate them. They compare their results with dictionaries that embed similar semantic constructs and with subjective scores of agency and communion that have been provided in previous studies.
The study of such constructs is interesting since they frequently appear in self descriptions and in descriptions of others. Consequently they could offer an approximation of people's opinions on others and we can exploit such information to investigate bias.

5.1.1 LIWC: Word Counting Approach

The developed dictionaries have been validated also with the help of a text analysis software called Language Inquiry and Word Count (LIWC).
This tool is based on the assumption that language is inevitably linked to human psychology so it exploits wording to define psychological processes scores.

Examples of psychological processes are *Anger, Friends, Positive, Emotions* but even biological ones such as *Body, Health, Sexuality, Ingestion.*
LIWC contains a vocabulary of words and a set of words to represent each LIWC dimensions (e.g. the psychological processes mentioned above). We'll call these sets of words "word categories".
LIWC takes as input a sentence and outputs a score for each word categories. It does so by increasing the score of word categories every time that it encounters a word that is both in the vocabulary and in the considered word category.

5.2 Study 1

Development of Big Two dictionaries by a committee of judges.

1. Generation of a list of words psychologically relevant to agency and communion.
2. Evaluation of psychometric properties.
3. Removal and addition of words after re-evaluation.

5.3 Study 2

Testing the validity of the developed dictionaries with respect to LIWC word categories. Two similarity measures have been taken into account:

- Direct Overlap:
  Shared number of words over total number of words in both dictionaries.
- Semantic Similarity:
  Computing the cosine distance among the vectors that represent the two dictionaries and the LIWC categories in the space obtained via Latent Semantic Analysis.

The authors have obtained meaningful results confirming the validity of the two dictionaries.

5.4 Study 3

The authors used a dataset of evaluations by american adults on common american professions on the dimension of agency and communion.
They continued the validation process by

1. Computing the semantic similarity between the Big Twos and each profession (similarly to Study 2).
2. Comparing the obtained scores with the subjective ratings of the participants of the original study.

A significant correlation between the computed scores and the self-reported ratings has been found.

5.5 Study 4

The authors used a dataset of 20,000 job advertisements, sampled from the site Monster.com, in order to further validate their dictionaries:

1. They selected male and female dominated jobs from the dataset by analyzing previous studies on the matter.
2. They computed agency and communion scores for each advertisement as the percentage of total agentic and communal words in it.

Agency and Communion dictionaries caught well-documented differences in how female and male-dominated jobs are advertised.
It is important to remark that results suggested an association in the form

- Agency - male-dominated job
- Communion - female-dominated job

However, these associations do not imply causal relationship.

5.6 Shortcomings

- LIWC is inadequate for handling *polysemy*.
- The frequency count approach embraced by LIWC does not allow a monitoring of the linguistic function of a word in the sentence ("he is a *strong* man" vs "he walks through a *strong* winds").

These limitations makes LIWC categories not completely sound for dictionary validation.

6 Conclusions

6.1 Measuring Bias

[Boulkbasi et al., 2016] tackle the bias identification problem as a search in the word embedding space of a "bias subspace", while [H. Zhang et al., 2020], since dealing with a masked language model, look for dangerous relationships among words and performance gaps in certain tasks with respect to the bias groups.
[Gonen et al., 2019] criticize the approach used by [Boulkbasi et al., 2016] and propose a *bias-by-neighbors* measure, that in the case of gender bias can be seen as the percentage of male/female socially biased words among the k-nearest neighbors of the target word [cons of this approach?]. This is to say that it is still unclear what metric to use and suggests it to be dependent on the field of application and on the model used to generate the original (biased) word embedding.
Moreover it would make sense to disentangle the source of bias and tackle the problem by directly acting on the root when it's possible. If for example we

discover a data imbalance problem, we could act on it by adding data of the neglected class. If the reason of the unwanted behavior of our model is strictly connected to the quality of the training data we should look up for more accurate and correct datasets. If the problem is of inherent social nature we should take it into consideration when analyzing results.

6.2 Removing Bias

As shown in [Gonen et al., 2019], [Boulkbasi et al., 2016] fails to remove bias since prejudiced words are still linearly separable from non prejudiced ones in the destination space.
The adversarial approach used in [H. Zhang et al., 2020] has shown to be problematic for several reasons.

- It is unfeasible to train a discriminator model to the overall BERT model, which raises a central issue in using adversarial methods to debias contextual embeddings.
- Applying adversarial debiasing to only the [CLS] token would in theory debias classification tasks, but not sequence-based tasks such as named entity recognition or question answering.
- Finally, adversarial debiasing during pretraining might not be conceptually desirable in the first place: if the model is encouraged to not encode information about the protected group Z in the representation, the predict label $\hat{Y}$ would be independent of Z, resulting demographic parity,which, is a problematic definition of fairness in healthcare.

As brought up by the authors of [H. Zhang et al., 2020], little work has been done for debiasing contextualized word embedding so it is imperative to act on it.

In conclusion, bias removal is a non trivial problem and these findings suggest to consider application-specific definitions of bias to construct application-specific debiased spaces.

6.3 Future works

[Pietraszkiewicz et al., 2019] create and validate dictionaries for the two psychological traits *Agency* and *Communion*. Such dictionaries could be exploited in the identification of bias in future works, especially because such traits are very common in self-descriptions and descriptions of others.
Moreover, these dictionaries could be enriched, considering a contextualized word embedding, by identifying neighbor vectors to words already in the dictionaries.
It is important though to be aware that the dictionaries presented are not complete. In fact, the main validation approach used by the authors of [Pietraszkiewicz et al., 2019]

does not take into account both the context and the linguistic function of words in a corpus.
Both [Boulkbasi et al., 2016] and [H. Zhang et al., 2020] manage to find biased semantic association. However, they limit themselves in finding bias with respect to one and only one dimension at a time, as gender ([Boulkbasi et al., 2016], [H. Zhang et al., 2020]), ethnicity, language and insurance status ([H. Zhang et al., 2020]). It could be interesting to identify bias with respect to a combination of attributes, i.e. gender and ethnicity.
The approach used in [Boulkbasi et al., 2016] is not trivially generalizable to other languages that instead include gender in neutral word. For example, in Italian, the word "doctor" is translated in "dottore" or "dottoressa" dependently on the doctor's gender (male and female respectively). Moreover, [Boulkbasi et al., 2016] do not consider contextualized word embeddings, which would surely give a more complete semantic representation of the corpus used during training.
[H. Zhang et al., 2020] uses an adversarial debiasing approach, that, shortly, minimizes the prediction performance of a discriminator with respect to the value of a sensitive attribute (e.g. gender). This approach makes the model learn a new embedding that favours demographic parity, which is a dangerous definition of fairness in the clinical context as already discussed. Consequently, a possible direction of research for contextualized word embedding is to use a debiasing approach that directly optimizes the embedding with respect to the most suitable definition of fairness for the application considered. In the case of [H. Zhang et al., 2020] it would be Equality of Opportunity for the Positive Class.

References

[Boulkbasi et al., 2016] Boulkbasi et al. (2016). Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings.

[Gonen et al., 2019] Gonen et al. (2019). Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings But do not Remove Them.

[H. Zhang et al., 2020] H. Zhang et al. (2020). Hurtful Words: Quantifying Biases in Clinical Contextual Word Embeddings.

[Pietraszkiewicz et al., 2019] Pietraszkiewicz et al. (2019). The big two dictionaries: Capturing agency and communion in natural language.

[Harutyunyan et al., 2019] Harutyunyan et al. (2019). Multitask learning and benchmarking with clinical time series data.

[Abele, Uchronski et al., 2008] Abele, Uchronski et al. (2008) Towards and operationalization of fundamental dimensions of agency and communion.

[Zhao et al., 2018] Zhao et al. (2018). Learning gender-neutral word embeddings.

[Beutel et al., 2017] Beutel et al. (2017). Data Decisions and Theoretical Implications when Adversarially Learning Fair Representations.