Neural Computing and Applications (2021) 33:3371–3384
https://doi.org/10.1007/s00521-020-05211-z

ORIGINAL ARTICLE

Extensive study on the underlying gender bias in contextualized word embeddings

Christine Basta[1] · Marta R. Costa-jussà[1] · Noe Casas[1]

Received: 29 April 2020 / Accepted: 13 July 2020 / Published online: 24 July 2020

Abstract
Gender bias is affecting many natural language processing applications. While we are still far from proposing debiasing methods that will solve the problem, we are making progress analyzing the impact of this bias in current algorithms. This paper provides an extensive study of the underlying gender bias in popular contextualized word embeddings. Our study provides an insightful analysis of evaluation measures applied to several English data domains and the layers of the contextualized word embeddings. It is also adapted and extended to the Spanish language. Our study points out the advantages and limitations of the various evaluation measures that we are using and aims to standardize the evaluation of gender bias in contextualized word embeddings.

Keywords Gender bias · Contextualized embeddings · Natural Language processing

1 Introduction

Gender bias, among many other social biases, in natural language processing (NLP), is negatively affecting the user experience [4]. Examples of these biases are when certain professions are associated with males (computer programming) and others with females (housekeeper) [3]. Evidence of gender bias appears in word embeddings, which are the representation of words in a multidimensional space, and this is at the core of many natural language processing applications [8]. Scientists define bias to happen when the model outcomes differently, given pairs of individuals that only differ in a targeted concept, like gender [19].

While word embeddings can be retrained in any particular task, currently, most natural language processing applications tend to benefit from sizeable pre-trained word embeddings. Beyond the classical techniques of Word2vec [22] or Global Vectors (GloVe) [23], most modern applications use the latest contextualized embeddings of Embeddings from Language Models(ELMo) [24] and Bidirectional Encoder Representations from Transformer (BERT), which provide different vector representations for the same word in different contexts.

While gender bias has been studied, detected and partially addressed for the classical word embeddings techniques [3, 9, 33] , this is not the case for the latest techniques of contextualized word embeddings. Recently, Zhao et al. [35] presented the first analysis on the topic based on the methods proposed by Bolukbasi et al. [3]. In this paper, we further analyze the presence of gender biases in contextualized word embeddings employing the proposed methods in Gonen and Goldberg [9] and extend the previous work Basta et al. [1]. This study aims to analyze five evaluation measures that have been proposed in both contextualized and standard word embeddings. Our final objective is to make progress toward finding standard evaluation measures, which are essential to propose debiasing methods in the future. Among the different evaluation measures that we are using (detecting gender direction, direct bias, clustering, classification and K-nearest neighbors (KNN)), we conclude that the first two measures are better at generalizing across languages and domains.

✉ Christine Basta
christine.raouf.saad.basta@upc.edu

Marta R. Costa-jussà
marta.ruiz@upc.edu

Noe Casas
noe.casas@upc.edu

[1] Universitat Politècnica de Catalunya, Barcelona, Spain

The rest of the paper is organized as follows. Section 2 provides an overview of the relevant work on which we built our analysis; Sect. 3 presents the related work; Sect. 4 states the specific questions addressed in this work, while Sect. 5 describes the experimental framework that is proposed to address these questions, and Sect. 6 discusses the results; finally, Sect. 7 discusses the conclusions of our work and proposes some further research.

2 Background

In this section, we describe the relevant natural language processing techniques used in the study, including word embeddings and contextualized word representations.

2.1 Word embeddings

Word embeddings are distributed representations in a vector space. These vectors are generally learned from large corpora and used in downstream tasks such as machine translation. Several approaches have been proposed to compute those vector representations, with Word2vec Mikolov et al. [22] being one of the dominant options. Word2vec introduces two variants: continuous bag of words (CBoW) and Skipgram. Both variants consist of a single hidden layer neural network trained on predicting a target word from its context words for CBoW, and the opposite for the Skipgram variant. The outcome of Word2vec is an embedding table where a numeric vector is associated with each of the vocabulary words.

Another dominant technique is the Glove model, which learns word embeddings using the co-occurrence matrix Pennington et al. [23]. Each row of the matrix represents a word, while the column represents the context where the word can appear. The value represents the frequency of this word in the context. The target of learning is minimizing the error between the co-occurrence statistics predicted by the model and actual corpus statistics. Computing statistics on a broader context enables Glove to capture longer dependencies. No evidence for advantage has appeared for either Word2vec or GloVe, as the overall performance depends on the training data and the considered evaluation task Khattak et al. [16].

These vectors' representations are computed on co-occurrence statistics; they exhibit geometric properties resembling the semantics of the relations between words. Therefore, by subtracting the vector representations of two related words and then adding the result to a third word, one may obtain a representation close to the application of the semantic relationship between the first two words and the third one. This application of analogical relationships has been used to showcase the bias that is present in word embeddings. The prototypical example of this is when the vector representation of *man* is subtracted from that of *computer*, and then adding it to *woman*, we obtain *homemaker*.

2.2 Contextualized word embeddings

Pre-trained language models (LM) such as ULMfit [13], ELMo [24], OpenAI GPT [26, 27] and BERT [6], have proposed different neural language model architectures and made their pre-trained weights available to ease the application of transfer learning to downstream tasks. Accordingly, they have enhanced the performance of several state-of-the-art benchmarks, including question answering on SQuAD, (cross-lingual) natural language inference, and named identity recognition. As ELMo and BERT have gained a high reputation in many NLP tasks, we are giving more details about them.

ELMo Peters et al. [24] progressed using arbitrary-length histories and directly incorporating the recurrent neural networks effective language models. ELMo was novel in training recurrent neural networks as language models, then using the context vectors they provide for each word token as pre-trained word (token) vectors Smith [28]. The neural architecture proposed in ELMo consists of a character-level convolutional layer that processes the characters of each word and creates a word representation. Consequently, this representation is fed into a 2-layer bi-directional long-short term memory (LSTM; [12]) that is trained on the language modeling task of a large corpus. Since it uses a bi-directional architecture, the embedding relies on both the next and previous words in the sentence. ELMo provides word-level representations. Peters et al. [25] and Liu et al. [18] confirmed the viability of using ELMo representations directly as features for downstream tasks without retraining the full model on the target task.

BERT Devlin et al. [6] is multilayer bi-directional transformer-encoder model for learning contextualized embeddings, adopting the transformer architecture with self attention layers Vaswani et al. [31]. BERT was originally pre-trained for masked language modeling and next token prediction tasks. BERT uses sub-word-level tokens and the learnt representations are at sub-word-level, giving it advantage for dealing with out of vocabulary words. However, the model needs special fine-tuning and sub-word embeddings requires particular handling when employed with tasks that deal with whole words Huang et al. [14].

Unlike the standard vector representations, which are constant regardless of their context, ELMo and BERT representations depend on the sentence where the word appears, and therefore, the whole sentence must be used in the model to obtain the word representations.

3 Related work

In a broader context, gender bias started to attract researchers to study the bias in word embeddings. The first work in this field was introduced by Bolukbasi et al. [3], which showed that bias is inherited in word embeddings. Bolukbasi et al. [3] studied the presence of gender bias in word embeddings from a geometrical point of view. To do this, they computed the subspace where the gender information concentrates by calculating the principal components of the difference of the vector representations of male and female gender-defining word pairs. Within the gender subspace, the authors identified direct and indirect biases in words relating to the profession. Finally, they mitigated the bias by nullifying the information in the gender subspace for words that should not be associated with gender and equalizing the distance to both elements of the gender-defining word pairs.

Zhao et al. [34] proposed an extension to GloVe embeddings, where the loss function used to train the embeddings is enriched with terms that confine the gender information to a specific portion of the embedded vector. The authors refer to these pieces of information as protected attributes. Once the embeddings are trained, the gender can be removed from the vector representation, eliminating any gender bias present in it.

Kaneko and Bollegala [15] mitigated the bias in the pre-trained embeddings using particular parameters in training, reserving non-discriminative gender-related information, while removing stereotypical discriminative gender biases from pre-trained word embeddings. Another technique was proposed by Zhang et al. [32], which used an adversarial network to mitigate bias in word embeddings. Dev et al. [5] demonstrated a reduction in invalid inferences via bias mitigation strategies on static word embeddings (GloVe) and explored adapting them to contextualized word embeddings (ELMo). However, Gonen and Goldberg [9] noted that gender bias has a more profound existence in word embeddings. Even when the embedded gender information is removed, gender information remains in the vector representation. All of these studies focus on the English language; however, a stimulating work by Zhou et al. [36] examined the gender bias in gendered languages and introduced another direction that determines gender known as the grammatical direction, which determines the direction between feminine and masculine nouns. Furthermore, they proposed mitigation techniques to shift the bias along the semantic gender direction (same direction of Bolukbasi et al. [3]) and an alignment technique for the gendered language with the bias-reduced English language.

The contextualized embeddings helped begin the debate about performing more evaluations of gender bias within these embeddings. Works in Tan and Celis [30], Kurita et al. [17], Guo and Caliskan [11] focused on assessing social biases in contextualized embeddings, especially BERT. They presented that human biases are encoded in contextual word models. May et al. [20] created the Sentence Encoder Association Test (SEAT) to test sentence encoders (e.g., ELMo) for human biases. Zhao et al. [35] showed that contextualized embeddings inhibit gender bias and delegate its effect on downstream tasks. Basta et al. [1] analyzed the gender bias in contextualized embeddings using different measures. Some measures demonstrated that bias is less inhibited in contextualized embeddings. All of these studies worked with the English language, and recently, Zhou et al. [36] extended the study to traditional Spanish embeddings only.

4 Research questions

Contextualized word embeddings still exhibit gender bias Basta et al. [1] and Zhao et al. [35]. Different measures were used to quantify the bias, and each measure had its own set of words. Each technique gives different insights and raises new and different questions. To begin our exploration of contextualized embeddings, we began by focusing on the following questions:

1. Does the effect of gender bias propagate from the corpus level to the contextualized word-level across different domains?
2. Is gender bias more represented in the contextualized word embeddings of professions?
3. What evaluation measures can be easily applied to gendered languages such as the Spanish language?
4. Can we rely on particular measures of evaluation more than others?

5 Experimental framework

For our study to understand the influence of bias in contextualized embeddings, experiments were performed with two languages, English as a nongendered language and Spanish as a gendered language. A gendered language is a language that has gender distinctions for all nouns Zhou et al. [36], for example, *la manzana* (feminine)—*el plátano* (masculine), which are translations for *apple* and *banana*, respectively, and for each, there is a gender associated with the noun. Spanish is considered a highly-gendered morphological language compared to English, where the professions and adjectives also have gender associations. For example, *I am a nurse* is translated to *soy enfermero* for a male speaker, and *soy enfermera* for a female speaker.

In order to evaluate and quantify the presence of bias in contextualized word embeddings, we apply the established methodologies for classical word embeddings by Bolukbasi et al. [3], Zhao et al. [33] and Gonen and Goldberg [9], reformulating them appropriately for contextual representations. They rely on direct intrinsic measures based on probes on different gender-predicting tasks. We focus on intrinsic measures, as opposed to other bias detection on extrinsic measures, where WinoMT Stanovsky et al. [29] is the main representative example. WinoMT measures bias in machine translation (MT) systems, using a crafted test set with coreference challenges. However, given the tight coupling of the test to the downstream task (i.e., MT) implies multiple problems: the impact of pre-trained debiased embeddings in the resulting translations cannot be measured in isolation; apart from that, pre-trained embeddings are seldom used in MT due to the importance of learning them along with the task; furthermore, the word-level token granularity in our contextual word embeddings is not appropriate for neural MT systems, where sub-word token granularity is needed to achieve good translation quality. Therefore, we understand that intrinsic measures, like those under study in this work, are the most appropriate testing framework for learned word-level representations like ELMo's contextualized word embeddings.

5.1 Contextualized word embeddings toolkit

Contextualized representations have proven to be essential in improving the results compared to non-contextual representations (i.e., classical word embeddings) on a wide range of tasks. Among the different contextualized representation learning approaches, tokenization is a differential factor. Some approaches, like BERT Devlin et al. [6], use sub-word-level tokens. This makes the association between word-level information, like semantics, and tokens hard to establish. On the other hand, contextualized word representation learning, like ELMo Peters et al. [24], enables connecting these word-level representations with their semantic traits (e.g., gender) and to reason about such a connection. This is the motivation to choose ELMo representations over BERT or other sub-word-level ones.

ELMo produces multiple forms of word embeddings for every single word, which is different from traditional word embeddings. ELMo produces three layers of word embeddings for a single word, where the higher layers capture different context-dependent aspects of word embeddings, and the lower-level layers capture syntax-dependent aspects. We can use only one layer of the embeddings or use the concatenation of all of the layers to obtain the benefits of the different representations of the layers. After experimenting with different representations from the three layers, there was no much variance in performance between layers. No distinguished information was explained by trying the evaluation measures on the three different layers. Therefore, we demonstrate the results of the representations of the third layer, which is more related to the context and its semantics, and the concatenation layer, which concatenates the representations of the three layers. The ELMo embeddings for Spanish were computed with the corresponding library.[1]

5.2 Experiments on the English language

5.2.1 Data and lists

We selected four different domains to explore the effect of diversity on the contextualized representations. We chose a medical domain (Pubmed[2]), a political domain (Europarl[3]), a social domain (TEDx[4], and a news domain (WMT[5]). From the statistics in Table 1, we can observe that there is a difference in the size of each corpus, the number of existing professions and the number of existing biased words. The TEDx and WMT corpora are more general and smaller in size, while Pubmed and EuroParl are more specific domains and larger concerning the size. We found from the statistics that each domain considers certain professions. For example, the most dominant professions in TEDx are *student* and *teacher*, which appear 160 and 153 times, respectively, while those that appear the least or only once are *mechanic, waitress, receptionist* and *firefighter*. For those professions that appear once, the gender of their appearance will surely prevail. Therefore, the cluster and classification will treat them with the gender of their appearance. One interesting fact about TEDx is that it contains a wide diversity of professions (Fig. 1). Europarl has evident examples of the diversity of the domain where *citizen* appears 2408 times, *advocate* occurs 1157 times, *judge* occurs 1071 times, and *minister* and *president* appear 1038 and 841, respectively. The large occurrence of certain political professions influences the computation of any measure having such a profession. As expected, in the Pubmed corpus, the highest occurring professions are *physician, nurse* and *doctor*.

To perform our English analysis, we used a set of lists from previous works Bolukbasi et al. [3] and Gonen and Goldberg [9]. We refer to the list of definitional pairs[6] as

[1] https://github.com/HIT-SCIR/ELMoForManyLangs.
[2] https://github.com/biomedicaltranslationcorpora/corpora.
[3] http://opus.nlpl.eu/Europarl.php.
[4] http://opus.nlpl.eu/TED2013.php.
[5] http://www.statmt.org/wmt13/translation-task.html.
[6] https://github.com/tolga-b/debiaswe/blob/master/data/definitional_pairs.json.

Table 1 Domain-specific data summary figures

Domain	TEDx	WMT	PubMed	EuroParl
No. lines in direct bias	2894	2866	6507	19821
No. lines of corpus	157,895	174,441	287,811	1,965,734
Total professions in KNN	144	114	68	142
Female in KNN %	48.61	42,98	54.41	45.07
No. of biased words (Cluster)	700	640	409	740
Females in biased clustering %	51.29	42,18	53.79	46.08
No. of biased words (Classify)	3637	3447	2223	3923
Females in biased classify%	49.46	41.77	53.08	45.73

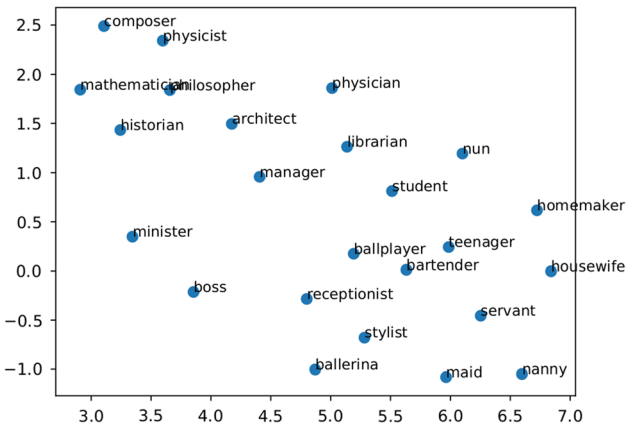

Fig. 1 TEDx annotations of the professions

the 'definitional list' (e.g., *she-he*, *girl-boy*). We refer to the list of all of the definitional pairs added to other gendered words (e.g., *lady-gentleman*, *niece-nephew*)[7] as the 'equivalent list.' We refer to the list of female and male professions[8] as the 'professional list' (e.g., *accountant*, *surgeon*). The 'biased list' is the list used in the clustering experiment, and it consists of biased male and female words (500 female-biased tokens and 500 male-biased tokens). This list is generated by taking the most biased words, where the bias of a word is computed by taking its projection on the gender direction ($\overrightarrow{he}$-$\overrightarrow{she}$) (e.g., *breast-feeding*, *bridal* and *diet* for female and *hero*, *cigar* and *teammates* for male). The 'extended biased list' is the list used in the classification experiment and contains 5000 male and female-biased tokens, 2500 for each gender, generated in the same way as the biased list.[9] The lists we used in our experiments were obtained from Bolukbasi et al. [3] and Gonen and Goldberg [9]. However, since we used words in sentences, our corpora may not contain examples of all of the words in the lists, which prevented us from obtaining their contextualized embeddings.

5.2.2 Evaluation measures

The English experiments are considered an extensive study for evaluating the gender bias in contextualized embeddings. The study was done in different domains. For each domain, five experiments were conducted to understand more about the bias in each perspective. As mentioned, in our analysis, we used a set of lists from previous works Bolukbasi et al. [3] and Gonen and Goldberg [9]. According to these works, Bolukbasi et al. [3] have referred that gender bias can be detected when one can determine the gender of non-explicitly gendered words, by looking at its projection on gendered pair in the definitional list. Gonen and Goldberg [9] have demonstrated that there is still gender bias when these non-explicitly gendered words have a direct relation with gendered or biased words.

The measures used in the first two experiments are: determining the gender direction and computing the direct bias between the profession's neutral words and this direction. The other three measures reformulated Gonen and Goldberg [9] measures to study if bias is deeply encoded in embeddings.

Word embedding association test (WEAT), the most common association test for word embeddings, has been proven to overestimate bias systematically Ethayarajh et al. [7]. Additionally, the work in Kurita et al. [17] has implied that WEAT cannot be considered as a useful measure for bias in contextual embeddings. Additionally, WEAT was used on sentence embeddings in Kurita et al. [20] of ELMo and BERT, but no evidence of bias was found. These conclusions guided us not to adopt WEAT in our experiments.

The main experiments carried out in our evaluation are illustrated as follows:

– *Detecting gender direction (Exp.1)* To compute the gender subspace, we followed the state-of-the-art

[7] https://github.com/tolga-b/debiaswe/blob/master/data/equalize_pairs.json.

[8] https://github.com/tolga-b/debiaswe/blob/master/data/professions.json.

[9] Both the 'biased list' and 'extended biased list' were kindly provided by Hila Gonen to reproduce experiments from her study Gonen and Goldberg [9].

method in Bolukbasi et al. [3] in a manner that was suitable for the contextualized embeddings. For a given corpus, we generated the corresponding gender-swapped variants, for sentences that had any instance of equivalent pairs in the equivalent list (changing he to she and vice versa, business-man to business-woman and vice versa, etc.). Thus, we had a sentence pair, each with a different gender for the definitional word.

In order to compute the gender subspace, the representations of words were selected from randomly sampled sentences that contain words from the definitional list. The ELMo representations of the definitional word in each sentence pair were obtained, and their difference was computed. On the set of difference vectors, we computed their ten principal components, using principal components analysis (PCA), to get the gender direction and its value from the top component.

– *Direct bias computation (Exp.2)* Direct bias is a measure of how close a specific set of words are to the gender vector. To compute it, we extracted the sentences that contained professional words in the professional list from the training data. We excluded the sentences that had both a professional token and a definitional gender word to avoid the influence of the latter over the presence of bias in the former, e.g., *he was my doctor*. Sentences with other equivalent words from the equivalent list, which are not definitional, were excluded, e.g., *I listened to the congressman*. We applied the definition of direct bias (see Eq. 1) from Bolukbasi et al. [3] on the ELMo representations of the professional words in these sentences.

$$\frac{1}{|N|}\sum_{w\epsilon N}|cos(\mathbf{w},g)| \tag{1}$$

where N is the amount of gender neutral words, g the gender direction, and $\mathbf{w}$ is the word vector of each profession. In our case, N is the number of sentences with professional words.

For the next experiments, one representation for each word was considered, to avoid dealing with a word as male-biased and female-biased simultaneously.

– *Male- and female-biased clustering approach (Exp.3)* To study how biased male and female words from biased lists cluster together when applying contextualized embeddings, we used k-means to generate two clusters of the token embeddings from the biased list and computed accuracy of clustering with the original biased version as a measure of bias. The higher the accuracy was, the more the clusters aligned with gender.

– *Classification approach (Exp.4)* To study if contextualized embeddings learn to generalize bias from a set of gendered words to others based only on the contextualized representations and how the classifier learns from being trained on a subset of the extended biased list. We trained a radial basis function-kernel support vector machine (SVM) classifier on the ELMo embeddings of 1000 random male and female-biased words from the extended biased list. Next, we evaluated the generalization of the other 4000 biased tokens. The accuracy of classification was taken as a measure of bias. The higher the accuracy was, the more the words were classified according to the gender.

– *K-nearest neighbors approach (KNN) (Exp.5)* We applied the KNN on the professional list to obtain the nearest 100 neighbors to each profession. For each token on the profession list, a randomly sampled sentence is used to get a contextualized representation. After applying the KNN algorithm on each profession, we computed the percentage of female and male stereotyped professions among the 100 nearest neighbors of each profession target token. Then, we computed the Pearson correlation of this percentage with the original bias of each profession.

One key factor of the experiments is randomization, as it has a considerable influence on the experiments. ELMo provides a different representation to a word according to its context in a sentence. We randomized the sentence chosen for such representation to choose a particular representation for a word. Moreover, experiments 3-5 were repeated ten times and averaged to guarantee this randomness.

There is a difference that should be noted between the professional list used in the direct bias experiments and the list used in the KNN experiment: Exp.2 and Exp.5. In Exp.2, we considered all of the sentences that contained the words of the professions. However, in Exp.5, we only considered 200 professions, including the 100 top female-biased professions and 100 top male-biased professions, and one random representation for each profession was considered.

5.3 Extension to the Spanish language

5.3.1 Data and lists

The corpus used in our evaluation is the Spanish version of the news corpus in WMT13, translation task from Spanish to English, as the English version is used in our English experiments. Consequently, the Spanish corpus has 174,441 lines.

The definitional list, equivalent list, and professional list were translated from English to Spanish. Native speakers were asked to revise all of the translations. The biased and

extended biased lists were created from scratch by including the top biased female and male words concerning the grammar and semantic directions, as explained in Sect. 5.3.2.

5.3.2 Evaluation measures

Extending our evaluation to the Spanish language had considerable challenges. To begin with, we had to swap gender in sentences containing the equivalent pairs of the equivalent list. Given the properties of the Spanish language, in addition to swapping the equivalent pairs, adjectives and professions had to be swapped to the other gender, and the articles also had to be considered in the swapping procedure. The articles and the equivalent-pairs swapping were done automatically, but the rest was done manually to ensure that the whole sentence had the same gender. This manual check consumed time and resources, which prevented us from applying the experiments to different domains. We had to check all of the swapped sentences (5848 lines) to make sure each sentence was grammatically correct. We made sure not to swap the gender in sentences with a proper name. For example, 'presidente Barack Obama,' was not swapped to 'presidenta Barack Obama.'

Answering positively to research question 4 from Sect. 4, we adapted the experiments to be suitable for the Spanish language and to give us insight about the bias in it. We applied the following:

- *Gender directions (Exp.6)* We adopted the idea of obtaining different gender directions, including the semantic direction Bolukbasi et al. [3] and the grammar direction Zhou et al. [36].

  Semantic direction $\overrightarrow{d_{pca}}$: We followed the same procedure previously mentioned in Exp.1 by using the PCA approach over the differences between male and female definitional contextualized word embeddings, from the sentences that have these definitional words and their swapped variants.

  Grammar direction $\overrightarrow{d_g}$: We extracted the nouns (feminine and masculine) from the corpus, approximately 7000 nouns for each gender, using Spacy parts of speech library[10] to extract these nouns. Next, in order to learn the grammar direction, since there are no equivalent pairs, linear discriminant analysis (LDA) for dimension reduction was applied. We applied LDA on 3000 random sets of nouns of each gender multiple times. We tried random contextualized representations for these nouns. The range of the accuracy of learning the grammar direction was between 0.45-0.65. When the ELMo representations of these nouns were plotted, they were scattered in the subspace, as shown in Fig. 8. There is no discrimination between a feminine subspace and a masculine subspace with these nouns.

  Following the literature, the grammatical gender component in the computed gender direction is projected out to make the semantic gender direction $\overrightarrow{d_s}$ orthogonal to the grammatical gender direction:

$$\overrightarrow{d_s} = \overrightarrow{d_{pca}} - \langle \overrightarrow{d_{pca}}, \overrightarrow{d_g} \rangle \overrightarrow{d_g}, \tag{2}$$

  where $\overrightarrow{d_s}$ is the semantic gender direction, which will be used in our experiments.
- *Direct bias (Exp.7)* For the professional list in Spanish, we obtained two translations for each profession, a male-gendered and a female-gendered translation. The number of lines with professions is 4987, with 2198 being feminine, and the rest masculine. We computed the direct bias on the male and female lists, separately, and then on their concatenated version. We computed the direct bias on the semantic direction $\overrightarrow{d_s}$ computed from Exp.6.
- *Clustering and classification experiments (Exp.8 and Exp.9)* With respect to the biased list and the extended biased list, we performed the following procedure to obtain the 500 and 5000 masculine and feminine biased words for these experiments:
  - We downloaded the Spanish Word2vec embeddings[11], which is trained on one billion words.
  - For these embeddings, the semantic gender direction was derived using the PCA method on the definitional standard word embeddings, and then, the grammar gender direction was derived following the previously described method in Exp.6. Following Eq. 2, the semantic gender direction was computed.
  - We obtained the top male and female biased words, with respect to the grammar direction and the semantic direction, respectively. Top 500 male-biased and 500 female-biased for clustering, both female and male biased words are considered the 'biased list,' semantic biased list, and grammar biased list. For classification experiment, 5000 female-biased and 5000 male-biased were gathered, and both together were considered the 'extended biased list,' semantic biased list, and grammar biased list.

  The clustering experiment, following the description in Exp.3, was applied to the semantic biased list and the grammar biased list. The classification experiment,

[10] https://spacy.io/models/es.

[11] https://github.com/dccuchile/spanishwordembeddings.

following the description in Exp.3, was applied to the grammar and semantic extended biased list. As the Word2vec embeddings were trained on different words, not all the words in the lists are available. For the semantic biased lists, only 254 were available from the biased list, and 1881 were available from the extended biased list. Whereas for the grammar biased lists, 457 were available from the biased list, and 4339 were available from the extended biased list.

6 Discussion

We will discuss the English and Spanish experiments separately in order to focus on different aspects.

6.1 English results

Figures 2, 3, 4 and 5 show results of Exp.1 for all domains and layers. Tables 2, 3, 4 and 5 show the results of experiments Exp.2, Exp.3, Exp.4 and Exp.5. Regarding the last three experiments, the average of ten experiments is shown for each domain for the third and concatenation layers. The numbers in brackets show the difference between the maximum and the minimum of the ten experiments.

6.1.1 Propagation of gender bias from the corpus to the contextualized word representations

The variability in the numbers of lines of the corpus, the diversity of the professions, the existing biased words and the percentage of feminine biased words are factors influencing our analysis and conclusions. Accordingly, our analysis is not based on the comparison between domains but relies on deriving conclusions about the gender bias propagation across the domains from the corpus level to the contextualized word representations level. Understanding the effect of gender bias on a particular domain leads to awareness of its impact on training neural models on such domain in different tasks.

From Exp.1, shown in Figs. 2, 3, 4 and 5 for the plots for the percentage of variance explained by the ten gender pairs of the definitional list, PCA can derive a dominant subspace, as the subspace of gender-flipped vectors contain less informative dimensions. After using the PCA, the first component appears to have dominant information, as it explains more variance than the other components, whereas, in the Europarl domain, the first two components explain more variance, not only the first.

As shown in Tables 2, 3, 4 and 5, the direct bias of professions computed with gender direction demonstrates the propagation of gender bias in professions across the domains. To understand the impact of gender bias propagating from the corpus to the contextual word representation, we applied the clustering and classification techniques which associates male and female nouns as concept words and their stereotypical clustering and classification. Clustering experiments, illustrated in Fig. 6 and Tables 2, 3, 4 and 5, show that male-biased words cluster together and so do female-biased words with accuracy more than 60% for all domains. Therefore, clusters seem to align with gender across domains. Classification experiments (Tables 2, 3, 4 and 5) demonstrate that bias is generalized from some gendered words to others, based only on their contextualized representations, with >80% accuracy across the four domains. Therefore, the classifier learns bias from gendered biased words. Accordingly, bias tends to propagate from the corpus level to the encoding level, which directly answers the first question in research questions (Sect. 4).

6.1.2 Layer concatenation versus layer 2

The main objective of experimenting on different layers was to deduce which layer results in less biased representations. Experimenting on the three different layers lead to slight differences; thus, we cautioned against definitive conclusions. By experimenting on the different two layers; the layer concatenation and the last ELMo layer (layer 2), we observe varying results. Considering that the last ELMo layer captures different semantic aspects of word embeddings, layer 2 appears to encode less bias in the case of more general domains (TEDx and WMT) in experiments 2-5. Along with, the concatenation layer benefits from the syntax and semantic aspects of the three ELMo layers in the more specific domains (Pubmed and Europarl). However, the difference between the results of the two layers in case of Europarl is not significant. The difference of means is 0.002 in direct bias, 0.2% in clustering experiment, 0.1% in classification and 0.013 in KNN. From the different results, we can conclude that using the representation of words from different ELMo layers are not distinguished concerning gender bias. Accordingly, choosing the layer should depend on other factors other than gender bias.

6.1.3 Professions perpetuates serious bias

Responding to the second research question (and also to first), Tables 2, 3, 4 and 5, direct bias computation and KNN experiments show obvious kinds of bias in most domains, except for TEDx, where less bias from KNN experiment is demonstrated. The bias of professions is evident in Pubmed that frequently associates medical occupations with male gender pronouns.

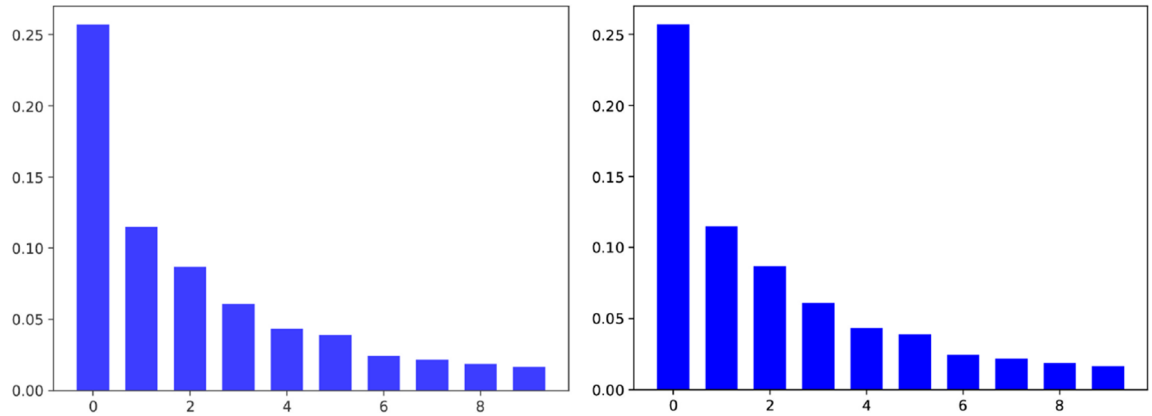

Fig. 2 X-axis refers to the ten PCA components and Y-axis refers to the percentage of variance explained by the ten principal components in TEDx Exp.1 in layer 2 (left) and layer concatenation (right)

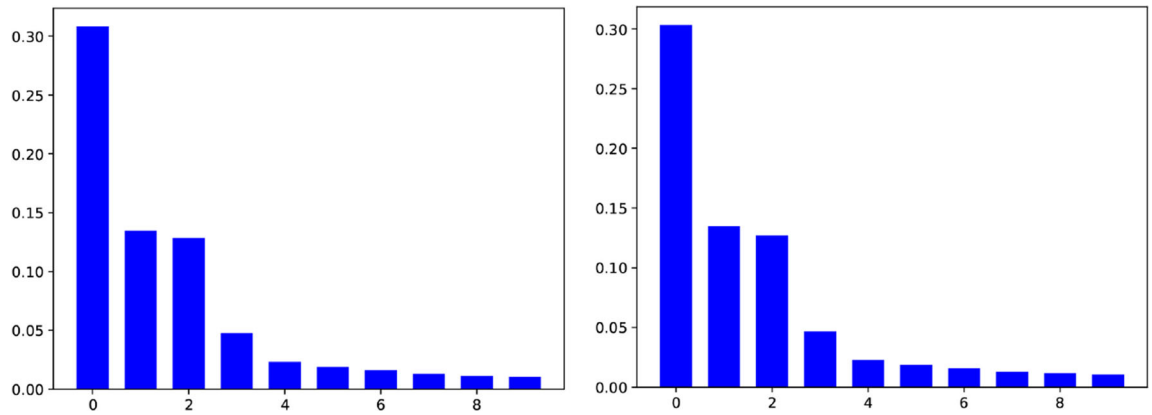

Fig. 3 X-axis refers to the ten PCA components and Y-axis refers to the percentage of variance explained by the ten principal components in WMT Exp.1 in layer 2 (left) and layer concatenation (right)

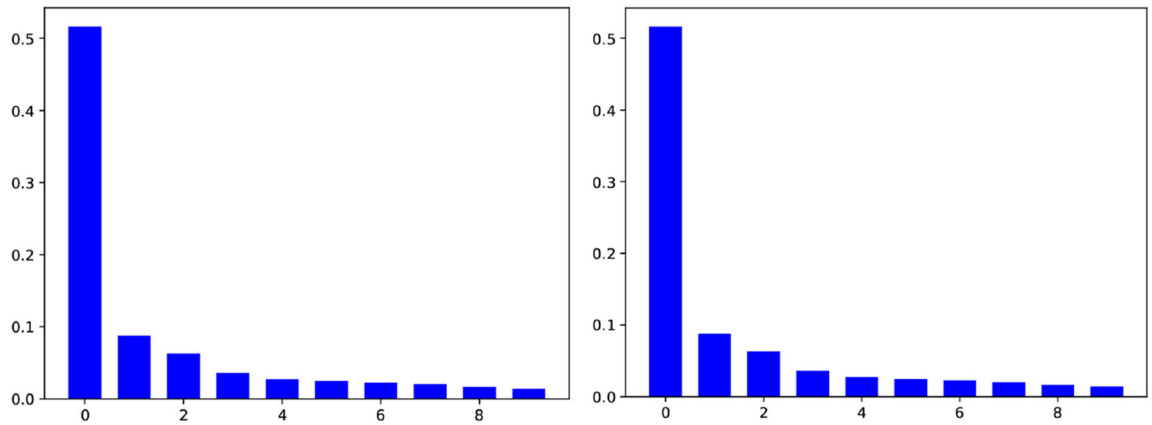

Fig. 4 X-axis refers to the ten PCA components and Y-axis refers to the percentage of variance explained by the ten principal components in Pubmed Exp.1 in layer 2 (left) and layer concatenation (right)

6.1.4 Effect of randomization is higher for clustering and KNN

Randomization of the ELMo embeddings of used words has led to the most varying results when repeating the clustering and KNN experiments. The corpus size and the number of word occurrences can also affect the randomization. Within the ten experiments, some have yield a wide range between the maximum and minimum results. The differences between the minimum and maximum in

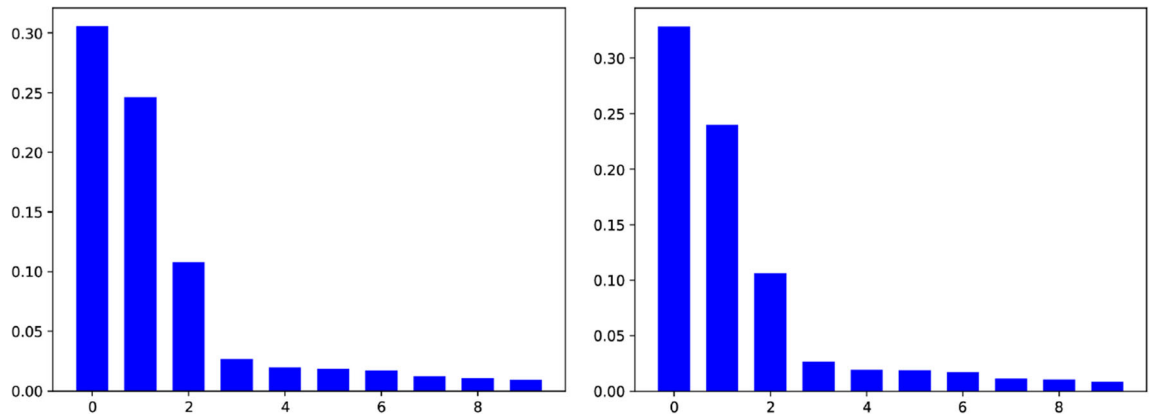

Fig. 5 X-axis refers to the ten PCA components and Y-axis refers to the percentage of variance explained by the ten principal components in Europarl Exp.1 in layer 2 (left) and layer concatenation (right)

Table 2 Results of TEDx experiments 2–5

TEDx	Layer 2	Layer Concatenation
Direct bias	0.031	0.031
Clustering (Acc.%)	67.9% (2%)	68.4% (2%)
Classification (Acc.%)	87.1% (2%)	87.3% (3%)
KNN (Pearson Cor.)	**0.160 (0.3)**	0.501 (0.35)

Less biased in bold, the higher, the worse. Numbers between brackets show the difference between the maximum and the minimum numbers acquired from the ten experiments

Table 3 Results of WMT experiments 2-5, the higher, the worse

WMT	Layer 2	Layer concatenation
Direct bias	0.028	0.026
Clustering (Acc.%)	**66.4% (3%)**	66.9 (3%)
Classification (Acc.%)	83% (2%)	85.4% (3%)
KNN (Pearson Cor.)	0.971 (0.02)	0.975 (0.01)

Numbers between brackets show the difference between the maximum and the minimum numbers acquired from the ten experiments

Table 4 Results of Pubmed experiments 2–5,the higher, the worse

Pubmed	Layer 2	Layer concatenation
Direct bias	0.021	0.021
Clustering (Acc.%)	79.4% (22%)	**77.3% (17%)**
Classification (Acc.%)	85.2% (3%)	**84.9% (4%)**
KNN (Pearson Cor.)	1	1

Numbers between brackets show the difference between the maximum and the minimum numbers acquired from the ten experiments

Exp.3 in Pubmed has reached 22% in layer 2 and 17% in layer concatenation. The wide range of differences can be attributed to the randomized representations of the words, which has resulted in different clustering. Pubmed is a large corpus that has a different context for the biased words used in the clustering experiment. Similarly, Exp.5 in TEDx is highly affected by randomization, showing differences of 0.3 in layer 2 and 0.35 in layer concatenation.

Table 5 Results of Europarl experiments 2–5, the higher, the worse

Europarl	Layer 2	Layer concatenation
Direct bias	0.05	**0.048**
Clustering (Acc.%)	**68.4 (1%)**	68.6 (2%)
Classification (Acc.%)	86% (2%)	**85.9 (2%)**
KNN (Pearson Cor.)	0.919 (0.03)	0.906 (0.03)

Numbers between brackets show the difference between the maximum and the minimum numbers acquired from the ten experiments

Conclusively, though the mean of the experiments sometimes implies a similar bias between corpora, lower numbers are shown in the results in one corpus than the other.

6.1.5 Measures to be considered and more reliable than others

Exp.1 and Exp.2 can be considered the most reliable measures across the four domains because they give a direct observation about gender nature in the domains. Exp.3 and Exp.4 can be regarded as related measures. They are synchronized and can give a reflection of the bias of the representations in the four domains.

KNN can be discarded as a measure when dealing with domains of low representation of professions. It is not reasonable to compute less than 100 KNN for each profession to understand how neighbors go together.

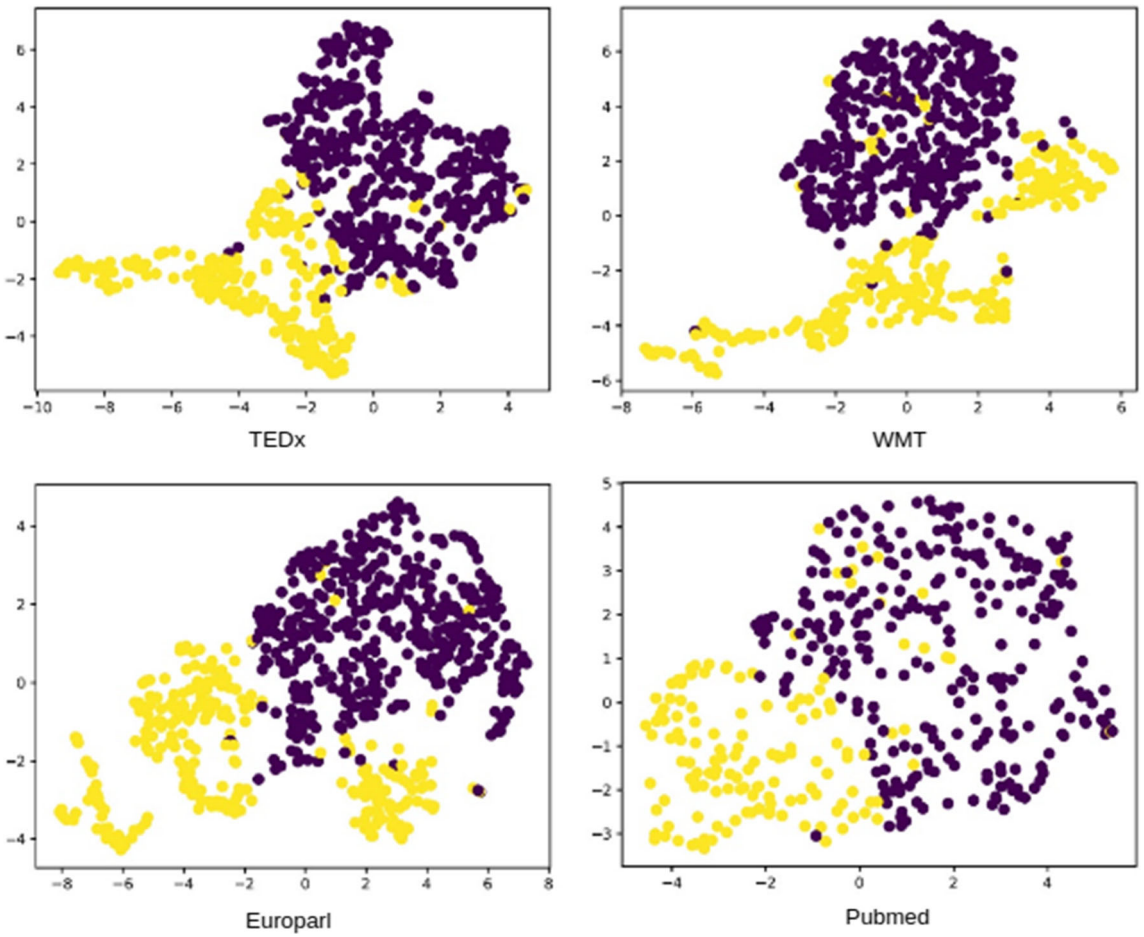

Fig. 6 Clustering experiments for TEDx, WMT, Europarl and Pubmed for representations from layer 2, male clusters are in violet and female clusters are in yellow

Therefore, the KNN measure is inconsistent in the case of the corpus with less biased professions. This can be applied on the Pubmed corpus, where only 64 professions out of 200 are present, and the correlation is always 1. These conclusion directly answers the third question from Sect. 4.

6.2 Spanish results

Again, Exp.8 and Exp.9 mainly used randomization and were repeated ten times, and their mean was calculated. Randomized representations were also used in Exp.6 to extract the nouns for the grammar direction.

6.2.1 Spanish semantic direction (results from Exp.6)

By applying the PCA experiment on the embeddings of gender definitional words in original and swapped sentences, the percentage of variance represented from the PCA components of definitional vector difference was obtained (see Fig. 7).

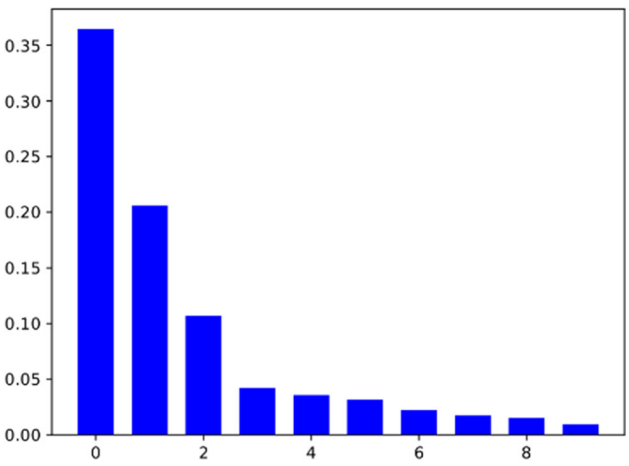

Fig. 7 X-axis refers to the ten PCA components and Y-axis refers to the percentage of variance explained by the ten principal components of definitional pairs' embeddings of Spanish

Additionally, for Spanish, we observe that the first component represents the most significant percentage of the variance of the ten PCA component, reaching 0.36, and

this top component determines the gender direction. After projecting out the grammar direction from the semantic direction, we found a slight decrease in the vector's percentages of variance determining the semantic direction.

6.2.2 Direct bias is higher for female professions

Direct bias is studied in both the grammar and the semantic gender directions, as described in Exp.7. Figures 8 and 9 and illustrated that plotting the professions, rather than the nouns, appears more segregated with gender (feminine vs. masculine). After calculating the direct bias on feminine professions and masculine professions separately, as shown in Table 6, the former case shows higher direct bias. This means that the feminine professions are closer to the semantic direction and, consequently, more biased.

6.2.3 Clustering and classification have to be performed on semantic biased words

Grouping of the embeddings of masculine and feminine words does not always indicate bias due to grammatical gender Zhou et al. [36]. The clustering and classification accuracy, noted in Table 7, is higher with words that are grammar biased. This is normal because nouns will be clustered and classified according to their gender. On the other hand, the accuracy is still high in clustering and classifying the semantic biased words. Thus, clustering according to gender and generalization of learning bias occur too.

6.3 Bias conceptualization

Motivated by the recent work Blodgett et al. [2], researchers have recommended to conceptualize bias in future work. One relevant suggestion is to understand the harm that bias of embeddings can cause to the NLP systems and to whom. This understanding is crucial for moving forward in the right direction with the bias problem.

In this paper, we have studied stereotypical associations between male and female gender with professional occupations and words in contextual word embeddings. Associating certain professions and activities with specific gender creates representational harm by perpetuating inappropriate stereotypes about which activities men and women are capable of. It gives a false impression about what women are allowed or expected to perform, e.g. leading to less professional females in STEM (Science, technology, engineering, and mathematics) McGuire et al. [21]. When such word representations are used in downstream NLP applications, there is an additional risk of unequal performance across genders Gonen and Webster [10]. While the observed correlations between genders and occupations in word embeddings may be a symptom of an inadequate training process, the dissociation between genders and occupations will enable systems to counteract the existing gender imbalances.

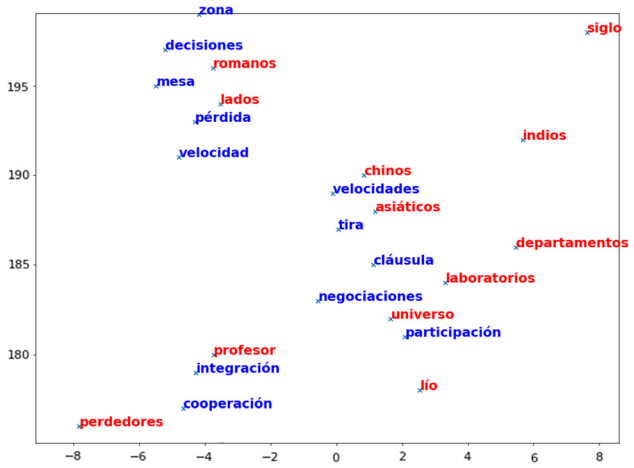

Fig. 8 Plotting Spanish representations of nouns on gender direction

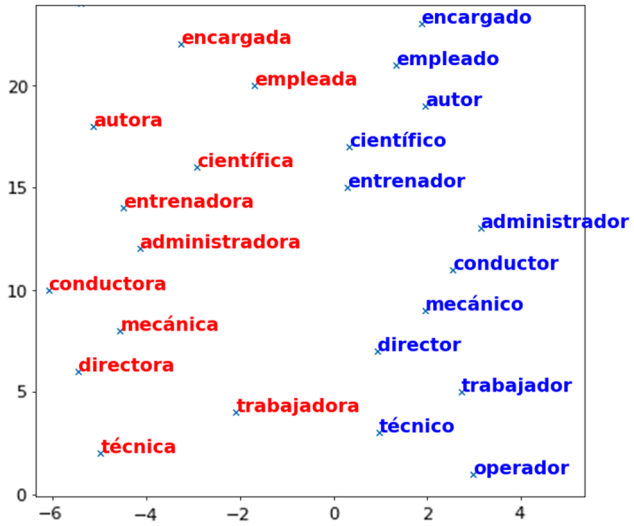

Fig. 9 Plotting Spanish representations of occupations on gender direction

Table 6 Direct bias of Spanish professions with semantic direction

Direct bias	Semantic direction
Female-version of professions	0.1215
Male-version of professions	0.0572
Male and Female together	0.098

Table 7 WMT Spanish clustering and classification experiments

	Classification (Acc.%)	Clustering (Acc.%)
Semantic biased words	94.25%	84.48%
Grammar biased words	99.05%	93.27%

7 Conclusions and further work

This paper makes the following contributions: first, we have extended existing analyses of gender bias to state-of-the-art ELMo contextual word models and indicate that such bias exists in these models. This highlights the scope of the problem of fairness in state-of-the-art models for language processing. We have provided evidence for how gender bias is encoded strongly in contextual word models in professions and stereotypical nouns.

Second, our paper understands the effect of domains on the contextualized word representations. Domains differ in statistics and nature, also differ in representing gender bias in contextualized word embeddings. This shows that such unsupervised methods perpetuate bias to downstream applications and our work forms the basis of evaluation. Additional contribution is analyzing the gender bias represented in Spanish contextualized word embeddings. This study reminds us that languages other than English have different properties that needs different kinds of treatment. Finally, we have compared between different measures to understand which ones to rely on to help mitigating gender bias in these embeddings.

Our extensive analysis is consistent with previous studies Dev et al. [5]. The techniques used to measure or mitigate bias in normal embeddings, do not necessarily succeed for contextualized embeddings.

One advantage of the gender direction and direct bias evaluation measures is being more generalized and based on less specific lists that are not domain or language dependent. On the other hand, direct bias seems to be less discriminating (see Tables 2, 3, 4 and 5).

While the clustering and classification seem more discriminating (again, responding to forth research question from Sect. 4, see Tables 2, 3, 4 and 5), the disadvantage is being strongly dependent on the existing vocabulary and less generalized to different domains. This can be attributed to studying the clustering and classification of embeddings of biased words that are biased in the original Word2vec embeddings, while each corpus may have its own set of different biased words. Again, applying clustering and classification of the biased words of each corpus would not be comparable from one domain to another; therefore, obtaining the biased set from the original embeddings is still more reasonable.

KNN can be neglected as a measure when there are few biased professions. As professions do not have enough neighbors, it is difficult to evaluate whether they are truly biased or just a matter of lacking neighbors.

Limitations of present work include focusing on binary type of bias (male/female). Also, representations are trained on natural data that contain bias without bias control mechanisms. Therefore, evaluations can only confirm the existence of bias not its absence. One other limitation is the dependence of the current methods used for evaluating bias on limited list of predefined templates.

Future work may include learning the grammar direction in gendered languages (e.g., Spanish) by using a wide range of pairs of adjectives. As shown, nouns are scattered in the subspace and thus it is difficult to have a discrimination line separating the feminine and masculine nouns. We think adjectives will perform better as applying PCA on pairs is feasible. The main challenge is gathering 3000–5000 adjectives in both feminine and masculine forms.

Our proposed study of using contextual word embeddings to assess bias represents an important step in considering fair context in NLP system. Debiasing technique for contextual word models remain a crucial direction for future work, though methods for debiasing remains a challenging issue.

Acknowledgements The authors want to thank Christian Hardmeier, Kellie Webster and Will Radford, for their enriching discussions on the bias conceptualization. This work is supported in part by the Catalan Agency for Management of University and Research Grants (AGAUR) through the FI PhD Scholarship and the Industrial PhD Grant. This work also is supported in part by the Spanish Ministerio de Economía y Competitividad, the European Regional Development Fund, the Agencia Estatal de Investigación through the postdoctoral senior grant Ramón y Cajal and the Projects EUR2019-103819, PCIN-2017-079 and PID2019-107579RB-I00.

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Basta C, Costa-jussà MR, Casas N (2019) Evaluating the underlying gender bias in contextualized word embeddings. arXiv:190408783
2. Blodgett SL, Barocas S, Daumé III H, Wallach H (2020) Language (technology) is power: a critical survey of "bias" in NLP. In: Proceedings of the 58th annual meeting of the association for computational linguistics, pp 5454–5476
3. Bolukbasi T, Chang KW, Zou JY, Saligrama V, Kalai AT (2016) Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In: Lee DD, Sugiyama M, Luxburg

UV, Guyon I, Garnett R (eds) Advances in Neural Information Processing Systems 29, Curran Associates, Inc., pp 4349–4357
4. Costa-jussà MR (2019) An analysis of gender bias studies in natural language processing. Nature Machine Intelligence 1
5. Dev S, Li T, Phillips J, Srikumar V (2019) On measuring and mitigating biased inferences of word embeddings. arXiv:1908.09369
6. Devlin J, Chang MW, Lee K, Toutanova K (2018) Bert: pre-training of deep bidirectional transformers for language understanding. arXiv:181004805
7. Ethayarajh K, Duvenaud D, Hirst G (2019) Understanding undesirable word embedding associations. In: Proc. of the ACL
8. Font JE, Costa-jussà MR (2019) Equalizing gender biases in neural machine translation with word embeddings techniques. arXiv:1901.03116
9. Gonen H, Goldberg Y (2019) Lipstick on a pig: Debiasing methods cover up systematic gender biases in word embeddings but do not remove them. arXiv:1903.03862
10. Gonen H, Webster K (2020) Automatically identifying gender issues in machine translation using perturbations. arXiv:2004.14065
11. Guo W, Caliskan A (2020) Detecting emergent intersectional biases: contextualized word embeddings contain a distribution of human-like biases. arXiv:200603955
12. Hochreiter S, Schmidhuber J (1997) Long short-term memory. Neural Comput 9(8):1735–1780. https://doi.org/10.1162/neco.1997.9.8.1735
13. Howard J, Ruder S (2018) Universal language model fine-tuning for text classification. In: Proc. of the ACL (Volume 1: Long Papers), Melbourne, Australia, pp 328–339
14. Huang L, Sun C, Qiu X, Huang X (2019) GlossBERT: BERT for word sense disambiguation with gloss knowledge. In: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), Hong Kong, China, pp 3509–3514
15. Kaneko M, Bollegala D (2019) Gender-preserving debiasing for pre-trained word embeddings. In: Proc. of the ACL, Florence, Italy, pp 1641–1650, https://doi.org/10.18653/v1/P19-1160
16. Khattak FK, Jeblee S, Pou-Prom C, Abdalla M, Meaney C, Rudzicz F (2019) A survey of word embeddings for clinical text. J Biomed Inf X 4:100057
17. Kurita K, Vyas N, Pareek A, Black AW, Tsvetkov Y (2019) Measuring bias in contextualized word representations. In: Proceedings of the first workshop on gender bias in natural language processing, pp 166–172
18. Liu NF, Gardner M, Belinkov Y, Peters ME, Smith NA (2019) Linguistic knowledge and transferability of contextual representations. In: Proceedings of the conference of the north american chapter of the association for computational linguistics: human language technologies
19. Lu K, Mardziel P, Wu F, Amancharla P, Datta A (2018) Gender bias in neural natural language processing. arXiv:1807.11714
20. May C, Wang A, Bordia S, Bowman SR, Rudinger R (2019) On measuring social biases in sentence encoders. arXiv:190310561
21. McGuire L, Mulvey KL, Goff E, Irvin MJ, Winterbottom M, Fields GE, Hartstone-Rose A, Rutland A (2020) Stem gender stereotypes from early childhood through adolescence at informal science centers. J Appl Develop Psychol 67:101109
22. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: Burges CJC, Bottou L, Welling M, Ghahramani Z, Weinberger KQ (eds) Advances in Neural Information Processing Systems 26, pp 3111–3119
23. Pennington J, Socher R, Manning CD (2014) Glove: Global vectors for word representation. In: Empirical Methods in Natural Language Processing (EMNLP), pp 1532–1543
24. Peters M, Neumann M, Iyyer M, Gardner M, Clark C, Lee K, Zettlemoyer L (2018) Deep contextualized word representations. In: Proc. of the ACL(Long Papers), New Orleans, Louisiana, pp 2227–2237, https://doi.org/10.18653/v1/N18-1202
25. Peters M, Ruder S, Smith NA (2019) To tune or not to tune? adapting pretrained representations to diverse tasks. arXiv:190305987
26. Radford A (2018) Improving language understanding by generative pre-training. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf
27. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I (2019) Language models are unsupervised multitask learners
28. Smith NA (2020) Contextual word representations: putting words into computers. Commun ACM 63(6):66–74
29. Stanovsky G, Smith NA, Zettlemoyer L (2019) Evaluating gender bias in machine translation. In: Proc. of the ACL, Florence, Italy, pp 1679–1684
30. Tan YC, Celis LE (2019) Assessing social and intersectional biases in contextualized word representations. In: Advances in Neural Information Processing Systems, pp 13209–13220
31. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008
32. Zhang BH, Lemoine B, Mitchell M (2018) Mitigating unwanted biases with adversarial learning. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, ACM, pp 335–340
33. Zhao J, Wang T, Yatskar M, Ordonez V, Chang KW (2018a) Gender bias in coreference resolution: evaluation and debiasing methods. arXiv:180406876
34. Zhao J, Zhou Y, Li Z, Wang W, Chang KW (2018b) Learning gender-neutral word embeddings. arXiv:180901496
35. Zhao J, Wang T, Yatskar M, Cotterell R, Ordonez V, Chang KW (2019) Gender bias in contextualized word embeddings. In: Proc. of the Conference of the NAACL
36. Zhou P, Shi W, Zhao J, Huang KH, Chen M, Cotterell R, Chang KW (2019) Examining gender bias in languages with grammatical gender. arXiv:190902224