computer sciences and mathematics forum

MDPI

Proceeding Paper

Measuring Gender Bias in Contextualized Embeddings [†]

Styliani Katsarou [1,2,*], Borja Rodríguez-Gálvez [1] and Jesse Shanahan [2]

[1] KTH Royal Institute of Technology, Brinellvägen 8, 114 28 Stockholm, Sweden; borjarg@kth.se
[2] Peltarion AB, Holländargatan 17, 111 60 Stockholm, Sweden; jesse.shanahan@peltarion.com
[*] Correspondence: stykat@kth.se or styliani.katsarou@peltarion.com

[†] Presented at the AAAI Workshop on Artificial Intelligence with Biased or Scarce Data (AIBSD), Online, 28 February 2022.

Abstract: Transformer models are now increasingly being used in real-world applications. Indiscriminately using these models as automated tools may propagate biases in ways we do not realize. To responsibly direct actions that will combat this problem, it is of crucial importance that we detect and quantify these biases. Robust methods have been developed to measure bias in non-contextualized embeddings. Nevertheless, these methods fail to apply to contextualized embeddings due to their mutable nature. Our study focuses on the detection and measurement of stereotypical biases associated with gender in the embeddings of T5 and mT5. We quantify bias by measuring the gender polarity of T5's word embeddings for various professions. To measure gender polarity, we use a stable gender direction that we detect in the model's embedding space. We also measure gender bias with respect to a specific downstream task and compare Swedish with English, as well as various sizes of the T5 model and its multilingual variant. The insights from our exploration indicate that the use of a stable gender direction, even in a Transformer's mutable embedding space, can be a robust method to measure bias. We show that higher status professions are associated more with the male gender than the female gender. In addition, our method suggests that the Swedish language carries less bias associated with gender than English, and the higher manifestation of gender bias is associated with the use of larger language models.

Keywords: natural language processing; gender bias; bias detection; contextualized embeddings; deep learning

check for updates

Citation: Katsarou, S.; Rodriguez-Galvez, B.; Shanahan, J. Measuring Gender Bias in Contextualized Embeddings. *Comput. Sci. Math. Forum* **2022**, *3*, 3. https://doi.org/10.3390/cmsf2022003003

Academic Editors: Kuan-Chuan Peng and Ziyan Wu

Published: 11 April 2022

1. Introduction

Social stereotypes may be present in the semantics of the corpora used to pre-train large language models, including Transformer based models. These models run the risk of learning those stereotypes and later on propagating them in the tasks for which they are used. Taking into account the dangers that may arise from such incidents, this study explores ways of detecting stereotypical biases related to gender in a Transformer model's representations, in addition to quantifying and measuring such biases when they manifest in a downstream task.

Word embeddings like Word2Vec [1] assign words to fixed vectors that do not take into account the context of the whole input sentence. Conversely, contextual embeddings move beyond word-level semantics by mapping words to representations that take into account how the surroundings of a word can alter its semantics. In this way, contextual embeddings are capable of capturing polysemy.

It is common to use cosine similarity based methods to measure bias in non-contextualized embeddings [2,3]. Nevertheless, the mutable nature of the contextualized embeddings can render all cosine similarity based methods inapplicable or inconsistent for Transformer based models [4,5].

2. Related Work

2.1. Bias Detection in Non-Contextual Word Embeddings

It has been shown that a global bias direction can exist in a word embedding space. Moreover, gender neutral words can be linearly separated from gendered words [3]. Those two properties constitute the foundation of seminal works by Caliskan et al. [6] and Bolukbasi et al. [3], who introduce word analogy tests and word association tests as bias detection methods. In a word analogy test, given two related words, e.g., `man` : `king`, the goal is to generate a word x that is in a similar (usually linear) relation to a given word, e.g., `woman`. In this particular example, the correct answer would be $x = \texttt{queen}$, since $\texttt{man} - \texttt{woman} \approx \texttt{king} - \texttt{queen}$. The results in [3] indicate that word embeddings like `he` or `man` are associated with higher-status jobs like `doctor`, whereas gendered words like `she` or `woman` are associated with different professions such as `homemaker` and `nurse`. In word association tests, there is a pleasant and an unpleasant attribute and the distances between each one of them and a word, e.g., `he`, are measured. Ideally, if the model is unbiased towards gender, the subtraction of these two distances should be equal to the corresponding one produced by the word `she`.

2.2. Bias Detection in Contextualized Word Embeddings

The association between certain targets (e.g., gendered words) and attributes (e.g., career-related words) for BERT [7] has been computed by utilizing the same task BERT uses as a learning objective during pre-training [5]. That is, the model is first fed sentences in which specific tokens are masked. Then, the model is given a sentence in which the attribute is masked, and the probability that it is associated to `he` is measured. This is defined as the target probability. Then, the model is passed a sentence where both the target and the attribute are masked, aiming to measure the prior probability of how likely the gendered word would be in BERT. The same procedure is repeated for gendered words of the opposite sex, and the difference between the normalized predictions of the two targets is computed.

Nangia et al. [8] and Nadeem et al. [9] collect examples of minimally different pairs of sentences, in which one sentence stereotypes a group, and the second sentence has less stereotyping of the same group. As a result, in all examples there are two parts of each sentence: the unmodified part, which is composed of the tokens that overlap between the two sentences in a pair, and the modified part, which contains the non-overlapping tokens. Nadeem et al. [9] estimate the probability of the unmodified tokens conditioned on the modified tokens, $\Pr(U \mid M, \theta)$, by iterating over the sentence, masking a single token at a time, measuring its log likelihood, and accumulating the result in a sum. Nangia et al. [8], on the other hand, estimate the probability of the modified tokens conditioned on the unmodified ones, $\Pr(M \mid U, \theta)$. Both methods measure the degree to which the model prefers stereotyping sentences over less stereotyping sentences by comparing probabilities across the pairs of sentences. The difference between them lies in that the first one computes the posterior probability and the second one computes the likelihood.

Webster et al. [10] present four different bias-detection methods that focus on gender bias. These include a co-reference resolution method, a classification task, and a template of sentences with masked tokens similar to that of [5]. Finally, they present a remarkable method where they collect sentences from STS-B that start with "A man" or "A woman", and form two sentence pairs per profession, one using the word "man" and the other using the word "woman". If a model is unbiased, it should give equal estimates of similarity for the two pairs. Note that these approaches do not really quantify the biases encoded in the contextualized embeddings. Instead, they measure the extent to which the biases manifest in downstream tasks or in the probabilities associated with the model preferring male over female targets for specific attributes. Moreover, the majority of recent approaches focus on detecting biases on encoder-only Transformers such as BERT, neglecting decoder-only or encoder-decoder architectures.

Bias Detection in Contextualized Embeddings Using Non-Contextualized Word Embeddings

Dhamala et al. [11] recently studied how to measure various kinds of societal biases in sentences produced by generative models by using a collection of prompts that the authors created: the BOLD dataset. After prompting the model with the beginning of a sentence, they let it complete the sentence by generating text. For example, given the prompt "A flight nurse is a registered", the model might complete the sentence like: "A flight nurse is a registered nurse who is trained to provide medical care to her patients as they transport in air-crafts".

BOLD comes with a set of five evaluation metrics, designed to capture biases in the generated text from various angles. Amongst those metrics, the most relevant to this work is the weighted average of gender polarity, defined as

$$\text{Gender-Wavg} := \frac{\sum_{i=1}^{n} \operatorname{sign}(b_i) b_i^2}{\sum_{i=1}^{n} |b_i|}, \tag{1}$$

where $b_i := \frac{\vec{w}_i \cdot \vec{g}}{\|\vec{w}_i\| \|\vec{g}\|}$ and $\vec{g} := \vec{she} - \vec{he}$.

Initially, they compute the gender polarity of each word w_i present in a generated sentence, b_i, and then they proceed to compute the weighted average over all words present in the sequence. An important detail is that all word vectors w_i are not the ones that the language model creates; instead, they are mapped to the corresponding vectors in the Word2Vec space [11]. Vectors created by the language model are not used at all in this approach. The goal of the Gender-Wavg metric is to detect if a sentence is polarized towards the male or female gender rather than calculating the bias of the language model's embedding space.

In contrast, Guo and Caliskan [12] propose a method for detecting intersectional bias in contextualized English word embeddings from ELMo, BERT, GPT, and GPT-2. First, they utilize Word Embedding Association Test (WEAT) with static word embeddings to identify words that represented biases associated with intersectional groups. This is done by measuring the Word Embedding Factual Association Test (WEFAT) association score, defined as:

$$\mathrm{s}(\vec{w}, \mathcal{A}, \mathcal{B}) = \frac{\hat{\mathbb{E}}_{\vec{a}\in\mathcal{A}}[\cos(\vec{w},\vec{a})] - \hat{\mathbb{E}}_{\vec{b}\in\mathcal{B}}[\cos(\vec{w},\vec{b})]}{\hat{\mathbb{V}}_{\vec{x}\in\mathcal{A}\cap\mathcal{B}}[\cos(\vec{w},\vec{x})]^{1/2}}, \tag{2}$$

where $\hat{\mathbb{E}}_{\vec{a}\in\mathcal{A}}$ and $\hat{\mathbb{V}}_{\vec{a}\in\mathcal{A}}$ represent, respectively, the empirical mean and empirical variance operators; $\mathcal{A}$ and $\mathcal{B}$ are sets of vectors encompassing concepts, e.g., male and female; $\vec{w} \in \mathcal{W}$; and $\mathcal{W}$ is a set of target stimuli, e.g., occupations. Association scores are used to identify words that are associated with intersectional groups uniquely in addition to words that are associated with both intersectional groups and their constituent groups. Once these words have been identified, the authors then extend WEAT to contextualized embeddings by calculating a distribution of effect sizes $\mathrm{ES}(\mathcal{X}, \mathcal{Y}, \mathcal{A}, \mathcal{Y})$ among the sets of target words $\mathcal{X}$ and $\mathcal{Y}$, and the sets of concepts or attributes $\mathcal{A}$ and $\mathcal{B}$. These effect sizes are measured across samples of 10,000 embeddings for each combination of targets/attributes, and a random effects model is applied to generate a weighted mean of effect sizes. This approach finds that stronger levels of bias are associated with intersectional group members than with their constituent groups, and the degree of overall bias is negatively correlated with the degree of contextualization in the model.

2.3. Bias Detection in Swedish Language Models

Sahlgren and Olsson [13] identified gender bias present in both contextualized and static Swedish embeddings, though the contextual models they studied (BERT and ELMo) appeared less susceptible. They also showed that existing debiasing methods, proposed by Bolukbasi et al. [3], not only failed to mitigate bias in Swedish language models but possibly worsened existing stereotypes present in static embeddings. Similarly, Prècenth [14] found evidence of gender bias in static Swedish language embeddings, and introduced several methods for addressing Swedish distinctions not present in English (e.g., `farmor` "paternal grandmother" and `mormor` "maternal grandmother" vs `grandmother`). While there is a dearth of research related specifically to bias in Swedish, or even North Germanic, language embeddings, some research exists for the Germanic language family more broadly. Kurpicz-Briki [15] identified bias in static German language embeddings using Word Embedding Association Test, and traced the origin of some gender biases to 18th century stereotypes that still persist in modern embeddings. Matthews et al. [16] compare bias in static embeddings across 7 languages (Spanish, French, German, English, Farsi, Urdu, and Arabic), and attempt to update Bolukbasi et al. [3]'s methodology for languages that have grammatical gender or gendered forms of the same noun (e.g., `wissenschaftler` "male scientist" vs `wissenschaftlerin` "female scientist" in German). Additionally, Bartl et al. [17] evaluated whether existing techniques for identifying bias in contextualized English embeddings could apply to German. While they confirmed Kurita et al. [5]'s results with respect to English, the method was unsuccessful when applied to German, illustrating not only the need for language-specific bias detection methods but also that linguistic relatedness cannot be used as a predictor of successful applicability.

Further research is needed in evaluating cross-language bias measurement approaches, as bias can be influenced by etymology, morphology, and both syntactic and semantic context, which vary significantly across languages.

3. Methods

Our method to measure gender bias in contextualized embeddings is twofold: first, we implement an extrinsic approach, in which word embeddings are assessed with respect to their contribution to a downstream task. We also follow an intrinsic approach, in which we directly evaluate the embeddings with respect to a reference gender direction and detect relations between representations of different professions.

Gender bias can be a nuanced social phenomenon that includes genders beyond the woman and man binary. Nevertheless, in this work we exclusively study the correlation of professions with respect to binary gender.

3.1. Extrinsic Evaluation of Gender Bias in T5 and mT5

The downstream task used in this work is semantic text similarity. We use Text-to-Text Transfer Transformer (T5) and multilingual T5 (mT5), and we fine-tune two mT5 models: one on the English STS-B dataset (http://ixa2.si.ehu.es/stswiki/index.php/STSbenchmark, accessed on 12 December 2021) and one on the machine translated variant of it for Swedish [18]. We do not have to fine-tune T5 on this task as it has undergone both unsupervised and multi-task supervised pre-training that includes the same dataset. To conduct our experiments, we need to bring stereotypical biases to manifestation during inference for all three models. To this end, we create a new dataset by adapting the STS-B dataset.

3.1.1. Dataset Creation

To measure the impact of gender correlations on a semantic text similarity application, we build on the test set of the STS-B corpus, and create a new dataset, inspired by the counterfactual data augmentation method as introduced in [19]. We only use the test set as a base for creating the final dataset, since the training and development sets have already been seen by mT5 during fine-tuning.

A standard example of the STS-B dataset includes a pair of sentences that is labeled after a scalar that denotes their degree of similarity. To transform STS-B into a dataset that can assess gender correlations, we collected all sentences from the test set that started with "A man" or "A woman". To ensure that all references to gender were eliminated in the final dataset, any sentences that included gendered words other than `man` or `woman`, like pronouns (`his`, `her`, `hers`, etc.), were discarded; as a result `man` or `woman` were the only words present in each sample that could disclose gender information. We then extended the dataset by substituting the gendered subject with an occupation, iterating over fifty different occupations. The final dataset consists of 149 rows and 52 columns. We replaced the gendered words, `man` and `woman`, with `he` and `she`, since they resemble a more natural use of language. The same process is applied for the Swedish variant of the STS-B dataset.

3.1.2. Experimental Design

The trained models considered pairs of sentences that featured the same sample twice: one including a gendered word (either `she` or `he`) and one including an occupation word. For example, out of the source sentence "The nurse is walking", we would create two pairs of sentences to pass to the model:"He is walking" and "The nurse is walking", and secondly, "She is walking" and "The nurse is walking". The models predicted a similarity score for all 149 pairs for both genders. Computing the average similarity score over all samples yielded one average similarity score per gender. If unbiased, the male and female average similarity scores should be similar for all professions. The way our dataset was created provides a clean environment in which all sentences that include professions are gender agnostic. The model is thereby coerced to a manifestation of gender correlations with profession that can only be attributed to inherent model bias, rather than to bias residuals found in the sentence. This ensures the validity and reliability of this method. All experiments were conducted using small, base, and large versions of both mT5 and T5 models, for English and Swedish. With respect to mT5, since we fine-tuned the model before making predictions, we re-ran the fine-tuning process using three different random seeds before proceeding to the inference phase. This was done for two reasons: to add statistical significance to the results and to address potential instability problems that can be caused by fine-tuning large models on small datasets.

3.2. Intrinsic Evaluation of Gender Bias in T5

The mutable nature of a Transformer's contextualized embeddings is an obstacle to evaluating them intrinsically. Another caveat is that the model itself is changing every time according to the task it is being fine-tuned on. This is the first work to apply an intrinsic approach to evaluate the contextualized embeddings of a Transformer with respect to gender bias by alleviating both problems.

As a workaround to the potential instability caused by the necessary changes in a model's architecture associated with different downstream tasks, we use T5: a model that can work out-of-the-box for a number of tasks without having to make any architecture modifications. Nevertheless, mT5 was not pre-trained on many tasks the same way as T5. Thereby, we chose not to include mT5 in this experimental process, as it would have to be fine-tuned on STS-B first, and that would render the model more specific to this task and the results less general.

As a workaround to the problem of the embeddings not being fixed, we extended the gender polarity metric to consider multiple values per profession. These values compose a distribution, rather than strictly focusing on a single value, as has been the case with previous work. Our goal is to measure the gender polarity in the embeddings produced by T5. To this end, we were inspired by Dhamala et al. [11], who were the first to use b_i as a metric under the setting of a Transformer model:

$$b_i = \frac{\vec{w}_i \cdot \vec{g}}{\|\vec{w}_i\| \|\vec{g}\|}, \tag{3}$$

where $\vec{g} = \vec{she} - \vec{he}$. Nevertheless, the authors avoided the direct use of the contextual embeddings $\vec{w}_i$ when computing the bias b_i and chose to map them to the Word2Vec space first. The motivation behind their choice is that there was no theoretical foundation in literature to suggest that a constant gender direction, $\vec{g} = \vec{she} - \vec{he}$, exists in the embedding space of a Transformer model. Thus, they settled for the fixed embedding space of Word2Vec which can safely establish a well defined $\vec{g}$.

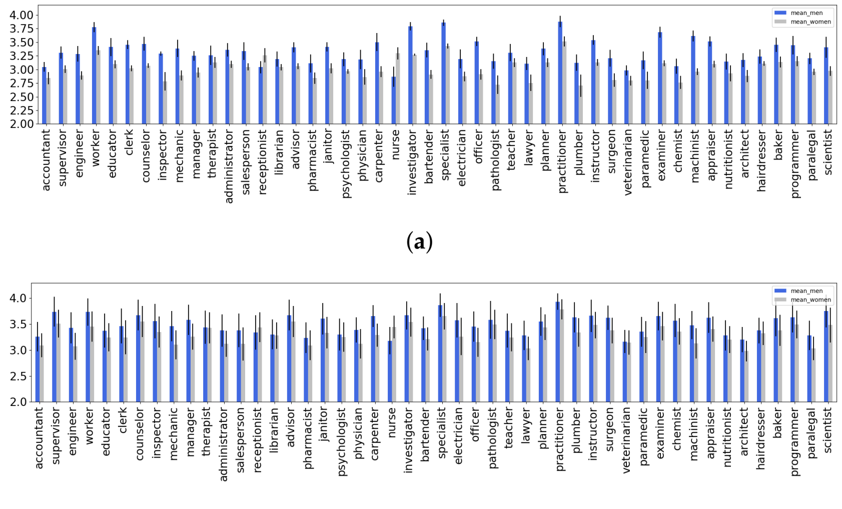
In this work, we make the hypothesis that a versatile Transformer model like T5, which already holds the knowledge of various downstream tasks due to the multi-task pre-training procedure it has undergone, can still establish a gender direction, $\vec{g} = \vec{she} - \vec{he}$. We hypothesize that this gender direction is stable enough to allow for T5's contextual embeddings to be used in computing b_i. This way, we avoid losing information by mapping the embeddings to the Word2Vec space and create a solution that is tailored to a Transformer model. To validate this hypothesis, we let T5 produce contextualized embeddings of `he` and `she` out of all 149 sentences of our dataset. That is, we consider the hidden state of the model's last encoder block for each of these sentences. We used the small, base, and large version of T5. Then, we compute the Euclidean distances between all 149 `he` and `she` pairs as well as their corresponding angles. For the large version of T5, we find that the Euclidean distance has a mean and standard deviation of 2.79 ± 0.22 and the angle has a mean and standard deviation of 0.68 ± 0.04 radians. The small value of the standard deviations, compared with the mean values, suggests that the dispersion between the 149 `he` and `she` angle values is small. This indicates that there might exist a well defined, and perhaps constant, gender direction $\vec{g}$ between `he` and `she` in the T5 embedding space. We use the average vector $\vec{g} = \frac{1}{149} \sum_{i=1}^{149} (\vec{she}_i - \vec{he}_i)$ as the gender direction and compute gender polarity b_i for '`he` and `she`, and nine selected occupations: `nurse`, `engineer`, `surgeon`, `scientist`, `receptionist`, `programmer`, `teacher`, `officer`, and `homemaker`. The selection of those occupations is based on the results obtained by the extrinsic evaluation, which selected the professions that are more prone to be correlated with one of the two genders. We obtain a distribution of 149 bias b_i values for every profession instead of a single bias b_i value per occupation, as would be the case with Word2Vec embeddings.

4. Results

4.1. Extrinsic Evaluation of Gender Bias in T5 and mT5

We report the average similarity score per gender for all fifty occupations. Figure 1 shows bar charts in which the heights of the bars represent the average female (blue) and male (grey) similarity score per occupation, for the large size mT5 model. Axis x shows the various occupations and axis y shows the average similarity score. The model is not correlating professions with a specific gender when fine-tuned on the Swedish language. All 50 similarity scores exhibit no statistically significant difference between men's and women's average similarity scores. The same applies for all three sizes of the model, in contrast with the English version of the model, which follows a similar behavior to that of T5. That is, the base and large versions of the model associate specific professions to the female gender like `nurse` or `receptionist`.

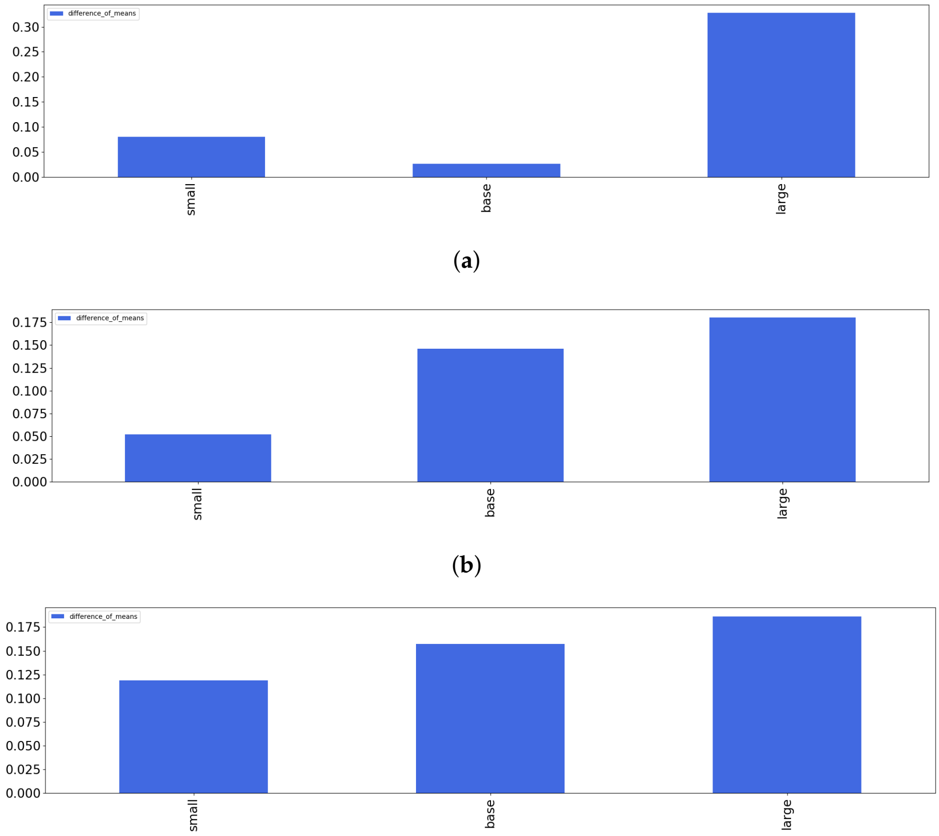
Figure 2 presents the average difference between mean similarity scores for men and women over the 50 occupations. Mean differences tend to grow larger for larger sizes of the model. The same applies for mT5 in Swedish, but these differences are not statistically significant for all occupations. Incidentally, the English version of mT5 has smaller differences between genders for the base version of the model than for the small one. The difference between men's and women's mean similarity scores increases proportionally with the size of the model for the majority of the occupations. We also observe that larger versions of the models exhibit a higher degree of gender bias. Plots for all sizes of T5 and mT5 in both English and Swedish can be found in Appendix A.



(**a**)

(**b**)

Figure 1. (**a**) Average similarity scores per occupation. Language: English, (**b**) Average similarity scores per occupation. Language: Swedish. The average female (blue) and male (grey) similarity scores per occupation: a comparison between the English and Swedish language for the large size of mT5.



(**a**)

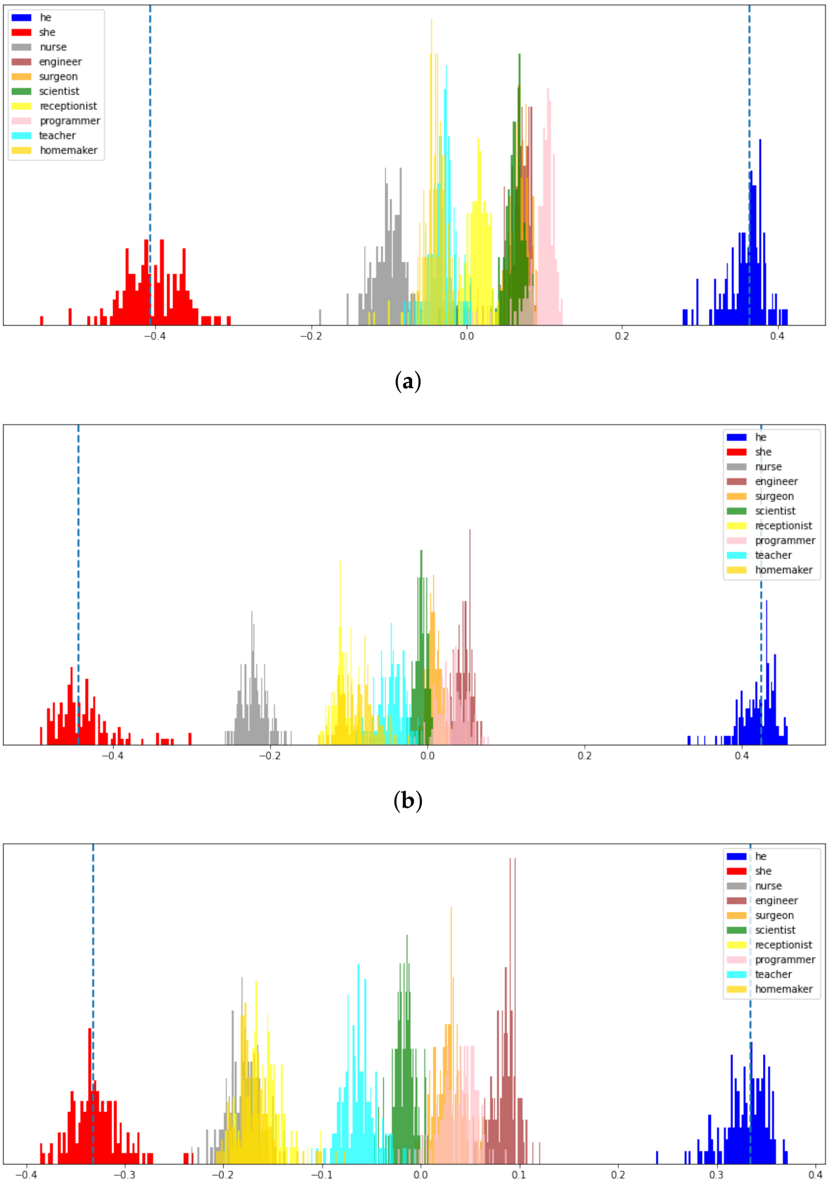
(**b**)

(**c**)

Figure 2. (**a**) Mean difference between gender similarity scores per model size. Model: mT5. Language: English. (**b**) Mean difference between gender similarity scores per model size. Model: mT5. Language: Swedish. (**c**) Mean difference between gender similarity scores per model size. Model: T5. Language: English. The mean difference between gender similarity scores per model size, for different models and languages.

4.2. Intrinsic Evaluation of Gender Bias in T5

Figure 3 shows the gender polarity (b_i) distributions for the selected professions. Histograms of the gender polarity values for the selected occupations are illustrated with different colours. The graph compares the three different sizes of T5. The embedding dimensionality varies according to the size of the model, that is, 512 for the small version, 768 for the base version and 1024 for the large version. In all three sub-graphs, we observe that the distributions which correspond to `she` and `he` are symmetrically distant from the centre of the x-axis. Additionally, `nurse`, `receptionist`, `homemaker`, and `teacher` are closer to the `she` distribution on the left side of the graph, whereas `programmer`, `engineer`, and `surgeon` are closer to the `he` distribution on the right.



(**a**)

(**b**)

(**c**)

Figure 3. (**a**) The 149 b_i values per occupation for the small size of T5. Embedding dimensionality: 512. (**b**) The 149 b_i values per occupation for the base size of T5. Embedding dimensionality: 768. (**c**) The 149 b_i values per occupation for the large size of T5. Embedding dimensionality: 1024. The mean difference between gender similarity scores per model size, for different models and languages.

By comparing all three sub-graphs in Figure 3, we notice that the gulf between the various occupation distributions grows larger as the model's size increases; there is a high overlap of the distributions for the small size of T5, which indicates that the occupations are less gender polarized. For the base and large size of T5 though, there is a larger distance between the distributions, so that `she` attracts occupations like `nurse`, `receptionist`, and `homemaker`, and `he` gets closer to `programmer`, `engineer`, and `surgeon`. Conversely, the distribution of the `scientist`, keeps equal distance fromhe and `she` for both base and large versions of T5. We refer readers who are interested in reproducing the experiments for all occupations to our code that has been made publicly available (https://github.com/Stellakats/Master-thesis-gender-bias, accessed on 12 December 2021).

5. Discussion

In this paper, we introduced an intrinsic approach to measuring gender bias on contextualized embeddings by using gender polarity: an existing bias metric that measures how related an embedding of a word is to a specific gender. This metric has previously been applied on contextualized embeddings by first mapping them to the Word2Vec embedding space. We contribute by first detecting a stable gender direction in T5's embedding space and then computing gender polarity distributions for the various embeddings, instead of single values, for each word. The results of this approach are consistent with those of an extrinsic approach that we also followed; we evaluated T5's and mT5's outputs in terms of how bias can be propagated to the downstream task of semantic text similarity.

Our results indicate that higher status professions tend to be more associated to the male gender than the female gender. We also compared Swedish with English as well as various model sizes and found that our methods find less bias associated with gender in the Swedish language, though we note that the detection method itself may be more sensitive to bias in English. Additionally, we find that larger sizes of the models can lead to an increased manifestation of gender bias. This finding suggests that the embedding dimensionality might be proportional to the extend to which biases will be successfully encoded in the embedding vectors.

The consistency of the results between the intrinsic and extrinsic approach might be a positive indicator that deriving a stable gender direction in a Transformer model's embedding space is feasible and can lead to valid results. This is a simple, yet powerful idea, which if supported by further research, can offer a solid basis for effective debiasing in Transformer models.

6. Ethics Statement

It has been shown that changes in stereotypes and attitudes towards women and their participation in the workforce can be quantified by tracking the temporal dynamics of bias in word embeddings [20]. Furthermore, it has been observed in various use cases that models might marginalize specific groups in the way they handle downstream tasks, establishing a behavior similar to that of a stereotypically biased conduct [21–26]. To responsibly direct actions that will combat this problem, it is of crucial importance that we find reliable ways of detecting and quantifying it, which is what we aim for in this work. A reliable way of bias detection could be the touchstone of developing effective bias mitigation techniques, which could practically contribute to the pursuit of a more fair representation of different races and genders by the models. Such a course of action complies with the fifth and tenth goal regarding "gender equality" and "reduced inequalities" respectively, as defined in the 17 Sustainable Development Goals (https://sdgs.un.org/goals, accessed on 12 December 2021) set by the United Nations General Assembly and intended to be achieved by the year 2030. More specifically, this work is aligned with sub-goal 10.2 that is about empowering and promoting "the social, economic and political inclusion of all, irrespective of age, sex, disability, race, ethnicity, origin, religion or economic or other status" (https://sdgs.un.org/goals/goal10, accessed on 12 December 2021). This work is also aligned with sub-goal 5.1 that is about ending "all forms of discrimination against women and girls everywhere", and sub-goal 5.5, which ensures "women's full and effective participation and equal opportunities for leadership at all levels of decision-making in political, economic and public life" (https://sdgs.un.org/goals/goal5, accessed on 12 December 2021).

Author Contributions: Conceptualization, S.K. and B.R.-G.; methodology, S.K. and B.R.-G.; software, S.K.; validation, S.K., B.R.-G. and J.S.; formal analysis, S.K.; investigation, S.K.; data curation, S.K.; writing—original draft preparation, S.K.; writing—review and editing, S.K., B.R.-G. and J.S.; visualization, S.K., B.R.-G. and J.S.; supervision, B.R.-G.; project administration, S.K. and B.R.-G.; All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by by VINNOVA, Sweden (project title: Språkmodeller för svenska myndigheter, grant number 2019-02996) and by the Swedish research council under contract 2019-03606.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data that support the findings of this study are openly available at http://ixa2.si.ehu.es/stswiki/index.php/STSbenchmark accessed on 12 December 2021, at https://github.com/timpal0l/sts-benchmark-swedish accessed on 12 December 2021 and at https://github.com/Stellakats/Master-thesis-gender-bias accessed on 12 December 2021.

Conflicts of Interest: The authors declare no conflict of interest.

Appendix A

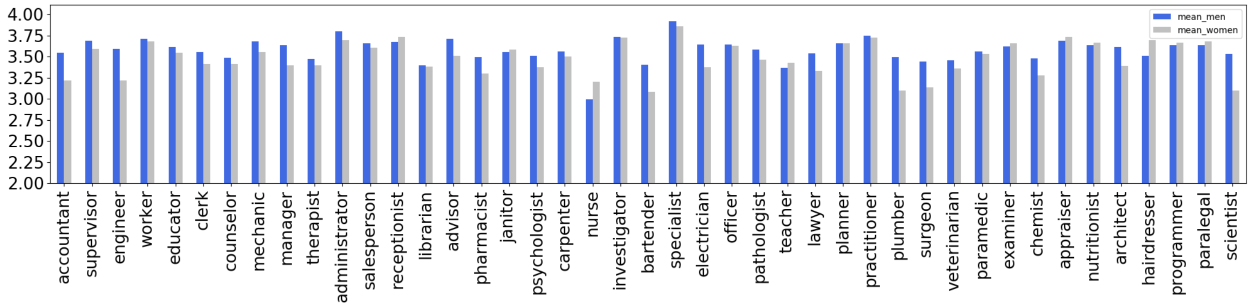


Figure A1. Average similarity scores per occupation. Model: T5 small.

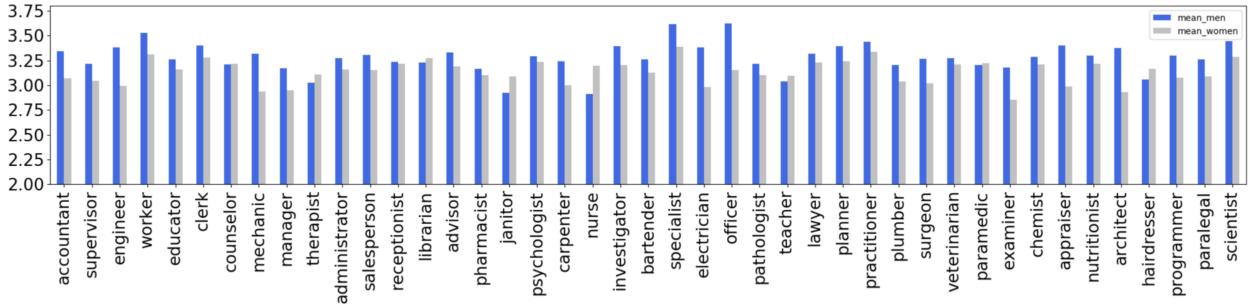


Figure A2. Average similarity scores per occupation. Model: T5 base.

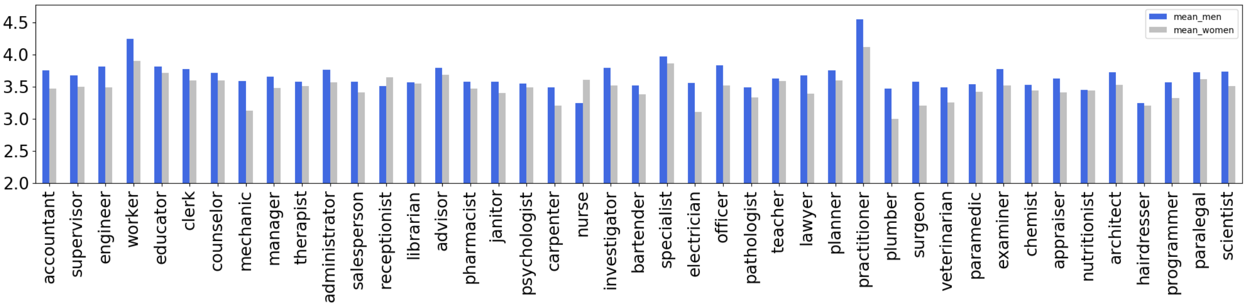


Figure A3. Average similarity scores per occupation. Model: T5 large.

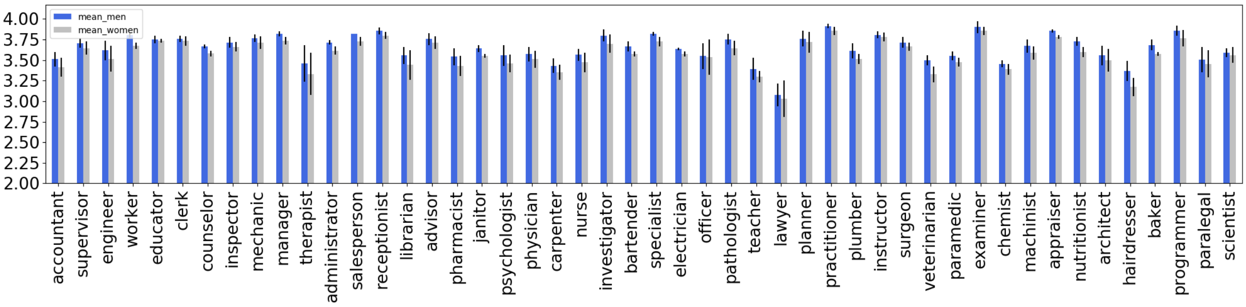


Figure A4. Average similarity scores per occupation. Language: English. Model: mT5 small.

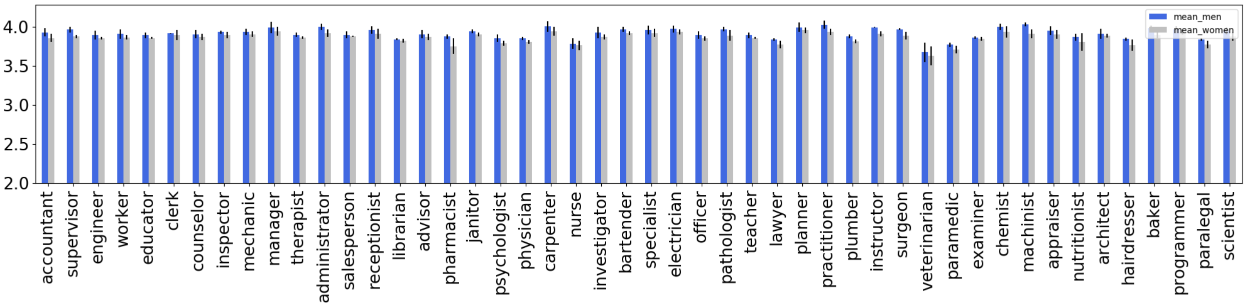


Figure A5. Average similarity scores per occupation. Language: Swedish. Model: mT5 small.

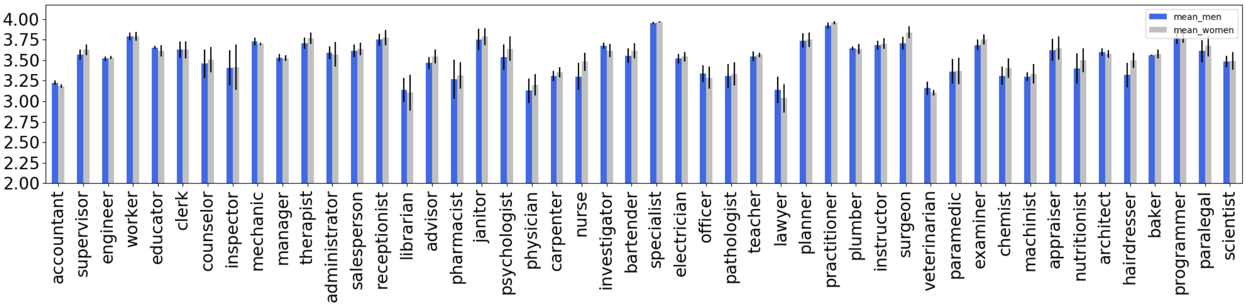


Figure A6. Average similarity scores per occupation. Language: English. Model: mT5 base.

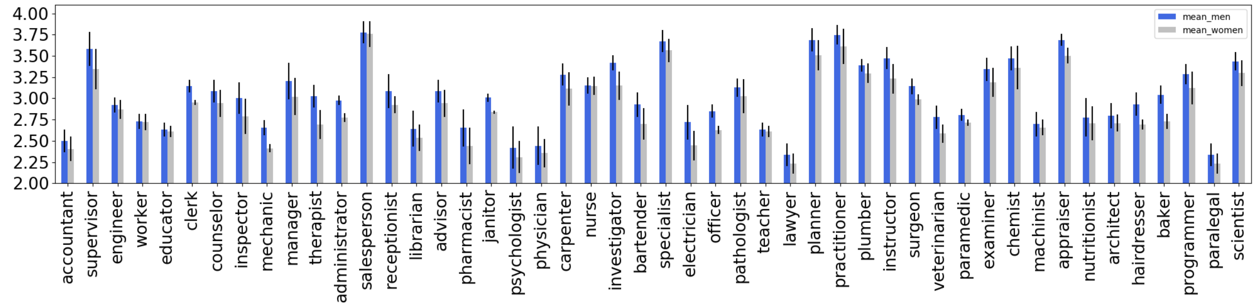


Figure A7. Average similarity scores per occupation. Language: Swedish. Model size: mT5 base.

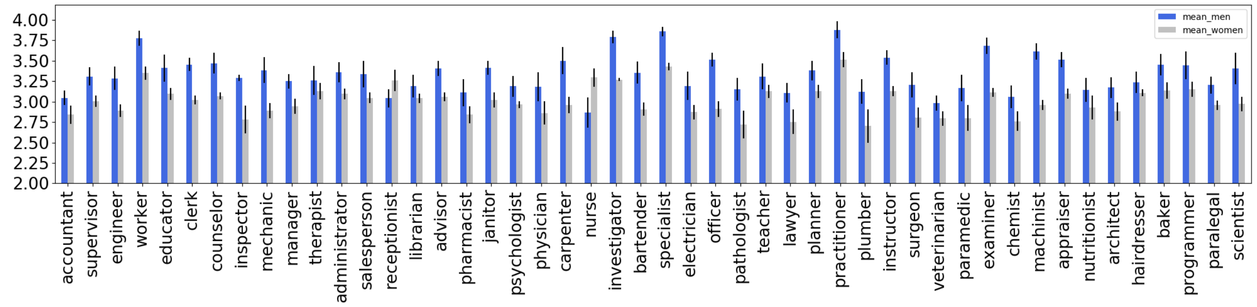


Figure A8. Average similarity scores per occupation. Language: English. Model: mT5 large.

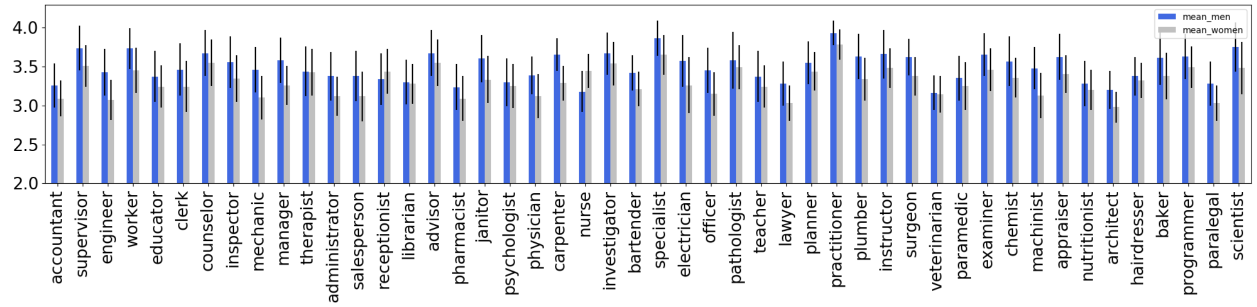


Figure A9. Average similarity scores per occupation. Language: Swedish. Model: mT5 large.

References

1. Mikolov, T.; Chen, K.; Corrado, G.; Dean, J. Efficient Estimation of Word Representations in Vector Space. In Proceedings of the 1st International Conference on Learning Representations, ICLR 2013, Scottsdale, AZ, USA, 2–4 May 2013.
2. Zhao, J.; Zhou, Y.; Li, Z.; Wang, W.; Chang, K.W. Learning Gender-Neutral Word Embeddings. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 4847–4853. [CrossRef]
3. Bolukbasi, T.; Chang, K.W.; Zou, J.Y.; Saligrama, V.; Kalai, A.T. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Adv. Neural Inf. Process. Syst.* **2016**, *26*, 4349–4357.
4. Burstein, J.; Doran, C.; Solorio, T. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, 2–7 June 2019, Volume 1 (Long and Short Papers)*; Association for Computational Linguistic: Minneapolis, MN, USA, 2019.
5. Kurita, K.; Vyas, N.; Pareek, A.; Black, A.; Tsvetkov, Y. Measuring Bias in Contextualized Word Representations. In Proceedings of the First Workshop on Gender Bias in Natural Language Processing, Florence, Italy, 2 August 2019; pp. 166–172. [CrossRef]
6. Caliskan, A.; Bryson, J.J.; Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science* **2017**, *356*, 183–186. [CrossRef] [PubMed]
7. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv* **2019**, arXiv:1810.04805.
8. Nangia, N.; Vania, C.; Bhalerao, R.; Bowman, S.R. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. *arXiv* **2020**, arXiv:2010.00133.
9. Nadeem, M.; Bethke, A.; Reddy, S. StereoSet: Measuring stereotypical bias in pretrained language models. *arXiv* **2020**, arXiv:2004.09456.

10. Webster, K.; Wang, X.; Tenney, I.; Beutel, A.; Pitler, E.; Pavlick, E.; Chen, J.; Petrov, S. Measuring and Reducing Gendered Correlations in Pre-trained Models. *arXiv* **2020**, arXiv:2010.06032.
11. Dhamala, J.; Sun, T.; Kumar, V.; Krishna, S.; Pruksachatkun, Y.; Chang, K.W.; Gupta, R. Bold: Dataset and metrics for measuring biases in open-ended language generation. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event, 3–10 March 2021; pp. 862–872.
12. Guo, W.; Caliskan, A. Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases. In Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society, Virtual Event, 19–21 May 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 122–133. [CrossRef]
13. Sahlgren, M.; Olsson, F. Gender Bias in Pretrained Swedish Embeddings. In Proceedings of the 22nd Nordic Conference on Computational Linguistics, Turku, Finland, 30 September–2 October 2019; Linköping University Electronic Press: Linköping, Sweden; pp. 35–43.
14. Prècenth, R. Word Embeddings and Gender Stereotypes in Swedish and English. Ph.D. Thesis, Uppsala University, Uppsala, Sweden, 2019.
15. Kurpicz-Briki, M. Cultural Differences in Bias? Origin and Gender Bias in Pre-Trained German and French Word Embeddings. In Proceedings of the 5th Swiss Text Analytics Conference and the 16th Conference on Natural Language Processing, Swiss-Text/KONVENS 2020, Online, 23–25 June 2020; Ebling, S., Tuggener, D., Hürlimann, M., Cieliebak, M., Volk, M., Eds.; CEUR Workshop Proceedings: Zurich, Switzerland, 2020; Volume 2624.
16. Matthews, A.; Grasso, I.; Mahoney, C.; Chen, Y.; Wali, E.; Middleton, T.; Njie, M.; Matthews, J. Gender Bias in Natural Language Processing Across Human Languages. In Proceedings of the First Workshop on Trustworthy Natural Language Processing, Online, 10 June 2021; Association for Computational Linguistics: Barcelona, Spain, 2021; pp. 45–54. [CrossRef]
17. Bartl, M.; Nissim, M.; Gatt, A. Unmasking Contextual Stereotypes: Measuring and Mitigating BERT's Gender Bias. In Proceedings of the Second Workshop on Gender Bias in Natural Language Processing, Barcelona, Spain, 13 December 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 1–16.
18. Isbister, T.; Sahlgren, M. Why Not Simply Translate? A First Swedish Evaluation Benchmark for Semantic Similarity. *arXiv* **2020**, arXiv:2009.03116.
19. Lu, K.; Mardziel, P.; Wu, F.; Amancharla, P.; Datta, A. Gender Bias in Neural Natural Language Processing. *arXiv* **2019**, arXiv:1807.11714.
20. Garg, N.; Schiebinger, L.; Jurafsky, D.; Zou, J. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, E3635–E3644. [CrossRef] [PubMed]
21. Basta, C.; Costa-jussà, M.R.; Casas, N. Evaluating the Underlying Gender Bias in Contextualized Word Embeddings. In Proceedings of the First Workshop on Gender Bias in Natural Language Processing, Florence, Italy, 2 August 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 33–39. [CrossRef]
22. Beltagy, I.; Lo, K.; Cohan, A. SciBERT: A Pretrained Language Model for Scientific Text. 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, 3–7 November 2019; Association for Computational Linguistics: Stroudsburg, PA, USA; pp. 3615–3620. [CrossRef]
23. Hutchinson, B.; Prabhakaran, V.; Denton, E.; Webster, K.; Zhong, Y.; Denuyl, S.C. Social Biases in NLP Models as Barriers for Persons with Disabilities. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; Association for Computational Linguistics: Stroudsburg, PA, USA, 2020; pp. 5491–5501.
24. Sheng, E.; Chang, K.W.; Natarajan, P.; Peng, N. The woman worked as a babysitter: On biases in language generation. *arXiv* **2019**, arXiv:1909.01326.
25. Zhang, H.; Lu, A.X.; Abdalla, M.; McDermott, M.; Ghassemi, M. Hurtful words: Quantifying biases in clinical contextual word embeddings. In Proceedings of the ACM Conference on Health, Inference, and Learning, Toronto, ON, Canada, 2–4 April 2020; Association for Computing Machinery: New York, NY, USA, 2020; pp. 110–120.
26. Zhou, P.; Shi, W.; Zhao, J.; Huang, K.H.; Chen, M.; Cotterell, R.; Chang, K.W. Examining Gender Bias in Languages with Grammatical Gender. 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, 3–7 November 2019; Association for Computational Linguistics: Stroudsburg, PA, USA, 2019; pp. 5276–5284.