CS230

Studying Gender Bias in News Using BERT Masked Language Models

Myra Deng, Justine Breuch
Department of Computer Science
Stanford University
myradeng@stanford.edu, jbreuch@stanford.edu
Github Repo: justinebreuch/cs230-final

Abstract

Model bias has become an increasingly urgent area for deep learning NLP researchers, given its far-reaching consequences. This paper attempts to quantify a major dimension of BERT's social biases – gender – on news data. Specifically, we evaluate bias through 1) word embedding similarity analysis between gender keywords and target concept identifiers in gendered domains (e.g., career and family) and 2) a word prediction study with masked gender keywords in order to evaluate baseline and fine-tuned BERT's predispositions. We perform analysis at the aggregate level and stratified by publisher political ideology and/or target word categories. Overall, we find that baseline and fine-tuned BERT more closely associate female embeddings to that of weakness and family and male embeddings to strength and career. In addition, baseline and fine-tuned BERT more accurately predict male keywords, with fine-tuned being slightly less biased.

1 Motivation and Problem Description

Bias is not a mere externality of deep learning but a facet of models themselves – even as the size, complexity, and overall effectiveness of models evolve [1],[2],[3]. Deep learning NLP models can encode and even exacerbate biases present in the training data, leading to unfair predictions. In particular, news media is a domain that is qualitatively analyzed day-to-day for its potential biases. The deep learning NLP field has made progress towards detecting quantitative patterns of bias in news data, particularly using recent advanced models like BERT – a state-of-the-art transformer-based NLP model with far-reaching applications, from powering Google's search algorithm to predicting hospital re-admission rates [4].

Our work aims to extend current research by analyzing BERT bias through word embedding and masked gender word prediction analyses on a news dataset. Uniquely, we examine specific subdomains of data which may encode gender biases differently to understand overall model behavior. First, we analyze how a publication's political affiliations impact bias, which carries import in an increasingly siloed, partisan media environment. Then, we determine how gendered categories of strength/weakness, career/family, intelligence/appearance affect outcomes. In order to investigate avenues for bias mitigation, we deployed two bias reduction techniques: dropout and counterfactual data augmentation (CDA) [6].

2 Related Work

There is significant prior work in both quantifying and mitigating bias for BERT models, though the latter has proven far more challenging. Most recent endeavors include counter-factual data augmentation (CDA), dropout regularization, and algorithms such as Self-Debiasing [5][6][7][8]. We attempt CDA and dropout to evaluate both a data pre-processing technique and model hyperparameter

CS230: Deep Learning, Spring 2022, Stanford University, CA. (LateX template borrowed from NIPS 2017.)

change. Doing so may contribute to extant discussions around gender bias mitigation. Measuring bias also varies across different projects. For example, the Sentence Encoder Association Test (SEAT) builds upon the Word Embedding Association Test (WEAT) to measure associations between two oppositional attribute identifiers and two target concepts ({"she", "he"} and {"family", "work"}, for example (see *Appendix 8.0.6*)). We draw inspiration from SEAT to analyze word embedding similarity between gender identifiers (male, female) and target concept pairs. In addition, we follow precedent for the masked language gender identifier modeling task by using the difference in normalized conditional probabilities for male vs. female masked prediction as a simple yet effective measure of gender bias (see *Section 5.3* for more details).

3 Dataset

We used a public news dataset, randomly sampling around 6 million tokens of news content from 320,000 rows across 10k articles[1][9]. The dataset provided the publication name per article which helped stratify the dataset. We extracted article text for finetuning and inference, where the features became tokenized words and the labels the words hidden by BERT's "[MASK]" token.

4 Data Processing

4.1 Pre-processing Data

We performed basic data cleaning (e.g., translating special encoding like "&apos") and converted contractions to ensure gender pronouns were captured consistently (e.g., "she'll" -> "she will"). We reshuffled the data randomly and then split into 80-10-10 train, dev, and test sets.

4.2 Gender Context Construction for Evaluation

Without target words: Existing research indicates that the surrounding 50 token ids are the most important context for BERT when making masked token predictions [11]. We parsed the corpus into 128 token contexts (after experimenting with 56, 128, 256 and choosing the best value in terms of loss and conditional probability metrics). We also identified proper nouns to replace names with "someone" in order to remove variables for bias. For the prediction task, we masked the center-most gender pronoun identifier in a context (see *Appendix 8.0.1* for all gender words).

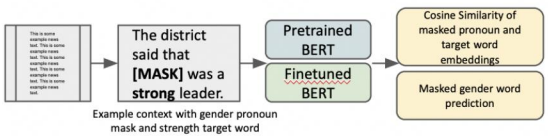
With target words: We employed the same strategies as above, but only on contexts where gender identifiers and target words co-occur (see *Appendix 8.0.2* for target word categories).

5 Methods

5.1 Models

Baseline: We used pretrained "bert-base-uncased" as a benchmark since retraining BERT is unfeasible [16]. BERT uses a transformer architecture and attention mechanism to encode relationships between words in a certain context. BERT's default tokenizer randomly masks 15% of input tokens for prediction.

Fine-tuned: In order to analyze how the news corpus might potentially alter these metrics, we fine-tuned BERT on our news corpus with an Adam Optimizer and an 80-10-10 train-dev-test split and ran against the same test dataset as in the baseline experiment. The fine-tuned model used a block size of 128, a batch size of 32, a learning rate of 2e-5, 1000 warm-up steps and a weight decay rate of 0.01 for 3 epochs after we experimented with various hyperparameters.



5.2 Word Embedding Cosine Similarity Study

For each gender-target word context, we modeled and measured the mean cosine similarity between a gender word group and target word embeddings within the same category (See: *Appendix 8.0.2*).

[1]Existing studies using finetuning demonstrate that 6 million tokens is more than sufficient for model tuning.

5.3 Gender Word Prediction Study

We ran the test masked gender contexts through pretrained and fine-tuned BERT. BERT outputs all predicted tokens with corresponding probabilities (See *Appendix 8.0.3* for example). We used the output to compute average conditional probabilities for female vs. male predictions – across the test set data for (1) the entire news corpus, (2) grouped by political ideology, and (3) gender-target concept examples. We chose target keywords related to strength/weakness, intelligence/appearance, career/family based on prior work from Chaloner et al. (2019) and grouped publications into political categories using a 2020 Knight Foundation and Gallup study [10][15].

The conditional probabilities equaled the total probability BERT predicted a female identifier (or male identifier) when the model output gendered identifiers among the top (K=100) most likely masked word predictions – given the true masked label was female (or male). These conditional probabilities were then normalized by the total probability of all gender word predictions. The differences between these values indicated a model's gender bias. The following is an example of female conditional probability (the formula is analogous for male):

$$\frac{p(\hat{y}{=}F|y{=}F)}{\Sigma_A(p(\hat{y}))}$$

F = Female word, A = All gender words

$\hat{y}$ = Predicted [MASK] word, y = Actual [MASK] word

5.4 Debiasing Attempts

Dropout Webster et al. attempted to de-bias BERT models by using dropout regularization ($a = .15$ and $h = .20$) [6]. We similarly experimented with dropout values and arrived at setting BERT's dropout layers to .2 (optimizing for loss and conditional predictivity).

Counterfactual Data Augmentation (CDA) We experimented with counterfactual data augmentation (CDA) on the training dataset. For each row in the corpus with a gender identifier, we inverted the identifier to its opposite using a map of gender identifiers published by Zhao et al. in the WinoBIAS study [12]. For example, "[he] is a neurologist" would be replaced with: "[she] is a neurologist".

6 Results and Analysis

The cross-entropy loss and perplexity on the held-out test set are lower after fine-tuning on the news data corpus (see Figure 1 below), indicating that the fine-tuned BERT model successfully learned the news data and is able to provide contextual embeddings specific to the news data corpus.

Model	Cross-entropy loss	Perplexity
Baseline	3.38	29.44
Fine-tuned	2.19	8.90

Figure 1

6.1 Word Embedding Cosine Similarity Study

Our analysis indicates that baseline and fine-tuned BERT do in fact learn biased embedding representations when evaluating man and woman embeddings vs. strength/weakness and career/family embeddings, respectively.

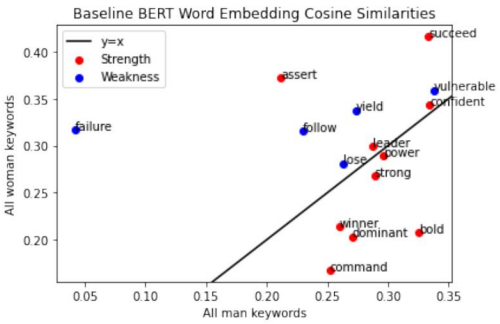


Figure 2

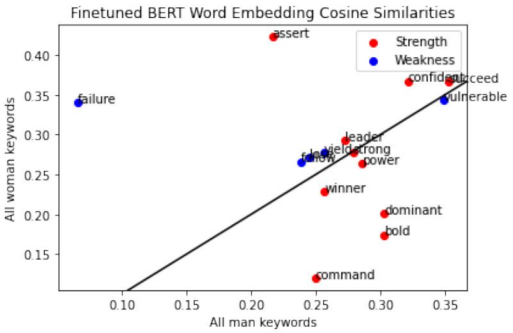


Figure 3

Result 6.1.1 Woman embeddings are more similar to weakness embeddings and man embeddings are more similar to strength embeddings in baseline (Figure 2). Woman embeddings are more similar to weakness embeddings in fine-tuned BERT – though on average similarity differences are smaller. Man embeddings are again more similar to strength embeddings in fine-tuned BERT (Figure 3).

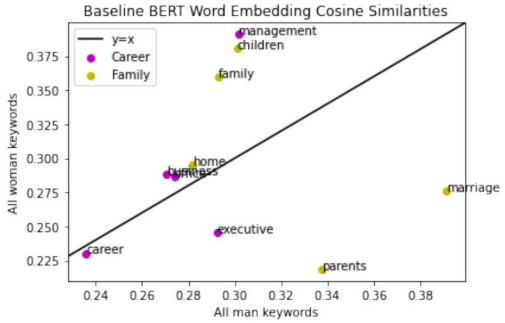


Figure 4

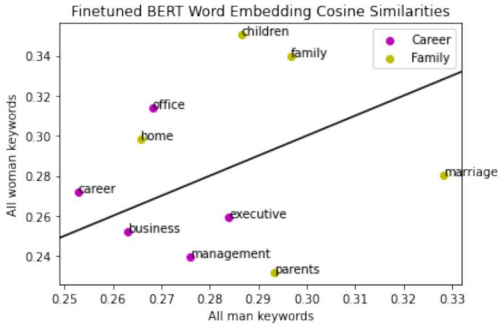


Figure 5

Result 6.1.2 Female word embeddings are more similar to most (3/5) family word embeddings in baseline BERT. This trend becomes more stark in fine-tuned BERT (Figure 5). Additionally, male embeddings are more similar to most (3/5) career embeddings in fine-tuned BERT (Figure 5).
Comment: There are too few data points to analyze intelligence / appearance. See *Appendix 8.0.3*.

6.2 Gender Word Prediction Study

Overall, baseline and fine-tuned predict much better for man keywords, with fine-tuned being slightly less biased (see Figure 6 below). One hypothesis as to why is described in Result 6.2.3.1. The overall skew towards male probability for all target word contexts is discussed in Result 6.2.2.1. We speculate this may be due to gender word imbalance in training data and idiosyncratic female gender evaluation contexts.

6.2.1 Results Over All News Corpus Contexts

Model	Male Probability	Female Probability	Difference
Baseline	0.832	0.584	0.248
Fine-tuned	0.826	0.604	0.222
Fine-tuned, dropout $(a, h = 0.2)$	0.815	0.527	0.288
Fine-tuned, CDA	0.476	0.696	0.22

Figure 6: Average Conditional Gender Probability Metrics per BERT Model

Result 6.2.1.2 Dropout was a relatively ineffective method at closing gender disparities, slightly increasing the disparity between male vs. female probabilities whereas CDA reduced the disparity but pointed the bias in the opposite direction. This could suggest dropout was contributing to less useful information captured on female gender contexts and CDA overcorrected keyword imbalance.

6.2.2 Results Over News Corpus – Breakdown by Target Word Categories

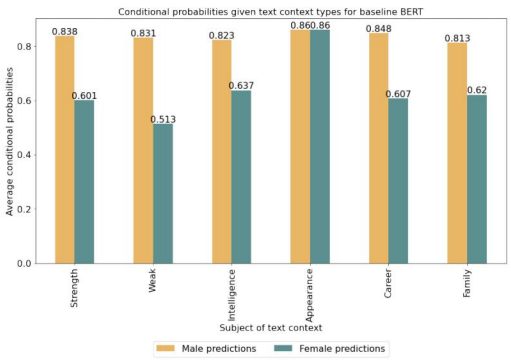


Figure 7

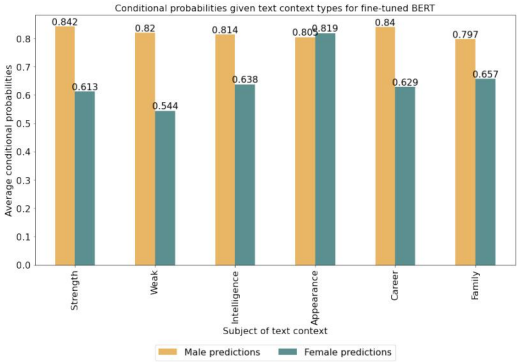


Figure 8

Result 6.2.2.1 Baseline BERT more frequently predicts man words across all contexts. The disparities grow largest when interest word is in the strength, weakness and career category. There were not enough data points in the appearance context to draw conclusive results.
Result 6.2.2.2 Fine-tuned BERT also reflects higher probabilities for men in all categories. It reflects similar biases to baseline with slight increase in parity in career and family.

6.2.3 Results Over News Corpus – Breakdown by Publication Ideological Alignment

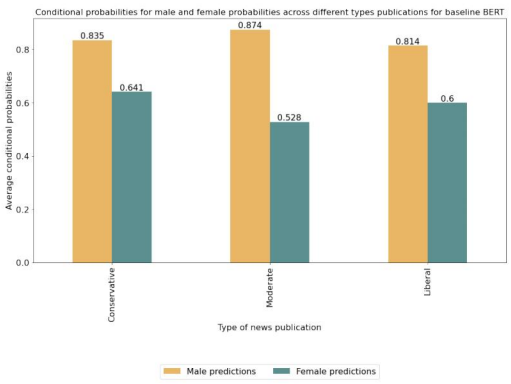


Figure 9

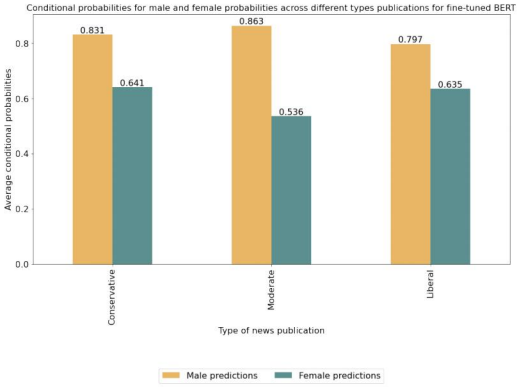


Figure 10

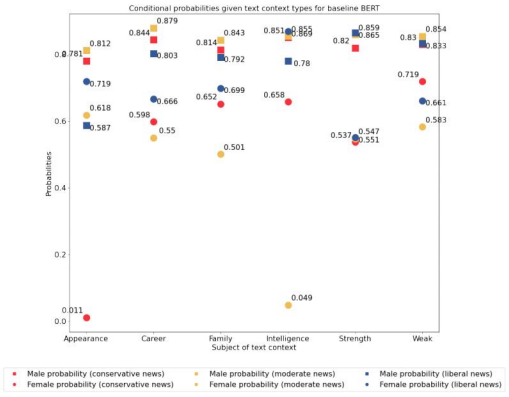


Figure 11

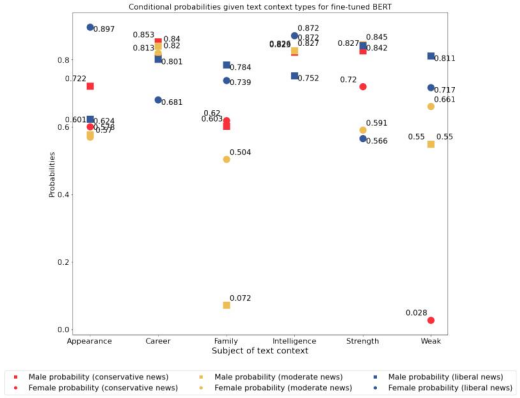


Figure 12

Result 6.2.3.1 Though average male probabilities were higher across baseline/fine-tuned, disparities between male and female reduced across all types on fine-tuned. The greatest increase in parity appeared in fine-tuned liberal, suggesting BERT learned more fair treatment from liberal publications.
Result 6.2.3.3 For baseline, conservative publications demonstrated considerable discrepancies between male and female probabilities (favoring men) across all contexts. However, probabilities diverged most widely for appearance, career, and strength contexts. Moderate publications presented a significant disparity between female and male probabilities for career and family [2] Liberal publications offered the best parity between predictions for appearance, career, family and intelligence.

7 Conclusion

We have performed a comprehensive analysis into how and why BERT (both pretrained and fine-tuned on news) may exhibit gender bias. We envision that this analysis could be extended to other data domains and types of bias – or applied when researching new bias mitigation techniques.

8 Contributions

Both authors contributed to all parts of the project – including literature research, data pre-processing, cosine similarity word embedding analysis, and model set-up, fine-tuning and evaluation.

[2]The gap seen in intelligence probabilities (0.049 for female and 0.869 for males respectively) is a statistical outlier due to lack of data in the category.

9 Appendix

9.0.1 Gender Identifiers

Gender	Keywords
Female	"woman", "women", "female", "she", "her", "hers"
Male	"man", "men", "male", "he", "his", "him"

9.0.2 Target Keywords Determining Contexts by Category

Context type	Target keywords
Strength	"power", "strong", "confident", "dominant", "potent", "command", "assert", "loud", "bold", "succeed", "triumph", "leader", "shout", "dynamic", "winner"
Weakness	"weak", "surrender", "timid", "vulnerable", "weakness", "wispy", "withdraw", "yield", "failure", "shy","follow", "lose", "fragile", "afraid", "loser"
Intelligence	"precocious", "resourceful", "inquisitive", "genius", "inventive", "astute", "adaptable", "reflective", "discerning", "intuitive", "inquiring", "judicious", "analytical", "apt", "venerable", "imaginative", "shrewd", "thoughtful", "wise", "smart", "ingenious", "clever", "brilliant", "logical", "intelligent"
Appearance	"alluring", "voluptuous", "blushing", "homely", "plump", "sensual", "gorgeous", "slim", "bald","athletic", "fashionable", "stout", "ugly", "muscular", "slender"
Career	"executive", "management", "professional", "corporation", "salary", "office", "business", "career"
Family	"home", "parents", "children", "family", "cousins", "marriage", "wedding", "relatives"

9.0.3 Example BERT Output for a Simple Toy Example with a Single Masked Token

```
[{'score': 0.434, 'token': 2002, 'token_str': 'he'},
{'score': 0.566, 'token': 2016, 'token_str': 'she'}]
```

9.0.4 Cosine Similarity Word Embedding Results

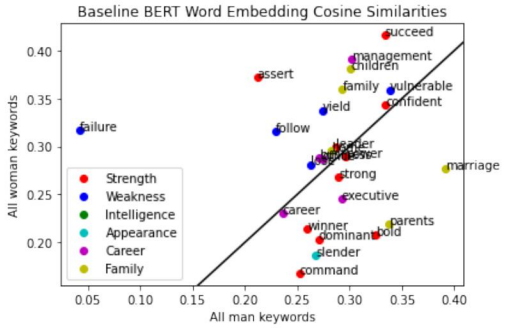


Figure 13

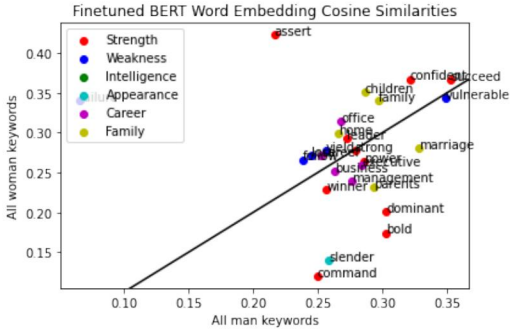


Figure 14

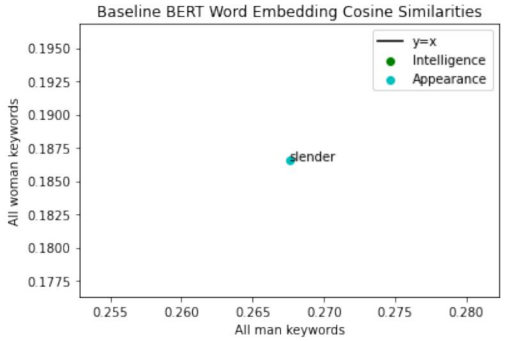


Figure 15

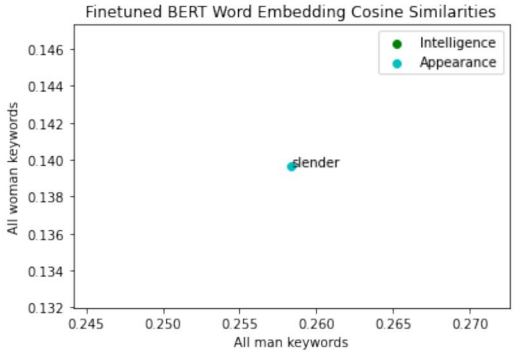


Figure 16

9.0.5 Gender Word Prediction

Context	Baseline BERT male-female probability differences	Fine-tuned BERT male-female probability differences	Difference between models
Strength	0.237	0.23	-0.007
Weakness	0.318	0.276	-0.042
Intelligence	0.185	0.175	-0.01
Appearance	0.024	0.014	-0.01
Career	0.241	0.210	-0.031
Family	0.193	0.140	-0.053

9.0.6 SEAT Formula [9]

$$s(X, Y, A, B) = \sum_{x \in X} s(x, A, B) - \sum_{y \in Y} s(y, A, B)$$

X, Y = target concept embeddings
A, B = attribute embeddings

$$s(w, A, B) = \mu_{a \in A} cos(w, a) - \mu_{b \in B} cos(w, b)$$

$$d = \frac{\mu_{x \in X} s(x, A, B) - \mu_{y \in Y} s(y, A, B)}{\sigma_{w \in X \cup Y} s(w, A, B)}$$

10 References

[1] Tony Sun, et al. "Mitigating gender bias in natural language processing: Literature review." arXiv preprint arXiv:1906.08976 (2019).
[2] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings". In: *Proceedings of the Annual Conference on Neural Information Processing Systems (NIPS)* (2016).
[3] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. "Semantics Derived Automatically From Language Corpora Contain
Human-Like Biases". In: *Science 356(6334):183–186* (2017).
[4] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. "Clinicalbert: Modeling clinical notes and predicting hospital readmission." arXiv preprint arXiv:1904.05342 (2019).
[5] Ran Zmigrod, et al. "Counterfactual data augmentation for mitigating gender stereotypes in languages with rich morphology." arXiv preprint arXiv:1906.04571 (2019).
[6] Webster, Kellie, et al. "Measuring and reducing gendered correlations in pre-trained models." arXiv preprint arXiv:2010.06032 (2020).
[7] Timo Schick, Sahana Udupa, and Hinrich Schütze. "Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp." In: *Transactions of the Association for Computational Linguistics 9* (2021).
[8] Paul Pu Liang, et al. "Towards debiasing sentence representations." arXiv preprint arXiv:2007.08100 (2020).
[9] Chandler May, et al. "On measuring social biases in sentence encoders." arXiv preprint arXiv:1903.10561 (2019).
[9] All The News dataset
[10] Knight Foundation and Gallup Foundation, "American Views 2020: Trust, Media, and Democracy", pg. 57 (2020).
[11] Olga Kovaleva, et al. "Revealing the dark secrets of BERT." arXiv preprint arXiv:1908.08593 (2019).
[12] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. "Gender Bias in Coreference Resolution: Evaluation and Debiasing Methods" In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2* (2018).
[13] Moin Nadeem, Anna Bethke, and Siva Reddy. "Stereoset: Measuring stereotypical bias in pretrained language models." arXiv preprint arXiv:2004.09456 (2020).
[14] Nikita Nangia, et al. "CrowS-pairs: A challenge dataset for measuring social biases in masked language models." arXiv preprint arXiv:2010.00133 (2020).
[15] Kaytlin Chaloner and Alfredo Maldonado. "Measuring gender bias in word embeddings across

domains and discovering new gender bias word categories." In: *Proceedings of the First Workshop on Gender Bias in Natural Language Processing* (2019).
[16] Hugging Face bert-base-uncased