

Do Language Models Reflect Societal Gender Biases?

Mae Sosto, Riccardo Tamponi, Francesco Vece

<https://github.com/StereotAlps/SexismDetector>

Universita` di Bologna - Intelligenza Artificiale

A large red square with a white border, centered on a white background. The word "Introduction" is written in white text in the center of the square.

Introduction

INTRODUCTION

- Artificial intelligence (AI) models more accessible to the general public [1].
- Concerns about their behavior and potential biases has been raised [2].
- Ongoing debates on **fairness and potential biases** in natural language [3].
- Field addressing these issues: **Natural Language Processing (NLP)**.

INTRODUCTION

- Biases in LMs and AIs can result in inaccurate or inappropriate outcomes.
- Biases refer to unfair or **discriminatory treatment in language data and LMs** [4].
- **Gender bias** as a specific type of bias affecting men and women.
- Persistence of gender biases in LMs due to **societal norms** [5].

Goal

Our goal is to assess Language Models (LMs) and determine whether they generate gender-biased expressions that reflect societal beliefs or stereotypes about women.

To accomplish this:

- **BERT** model and employ the **mask language model** (MLM) technique used.
- we conduct **three tests** using template-based sentences to observe inequality and sexist content.
- Qualitative assessment using various **tools and techniques**.



Model

Model

BERT (Bidirectional Encoder Representations from Transformers) is a machine learning model for **natural language processing** (NLP) [6].

- Trained on extensive text data from books and the internet.
- Potential absorption of societal biases in training data.

BERTbase	BERTlarge
110M	340M

Masked LM

- Input string s is converted into tokens representing contextual meaning c .
- Contextual tokens are used to predict masked words m .
- Objective: Find the **most likely prediction** $p(m|c)$ of masked words m given the context c [7].

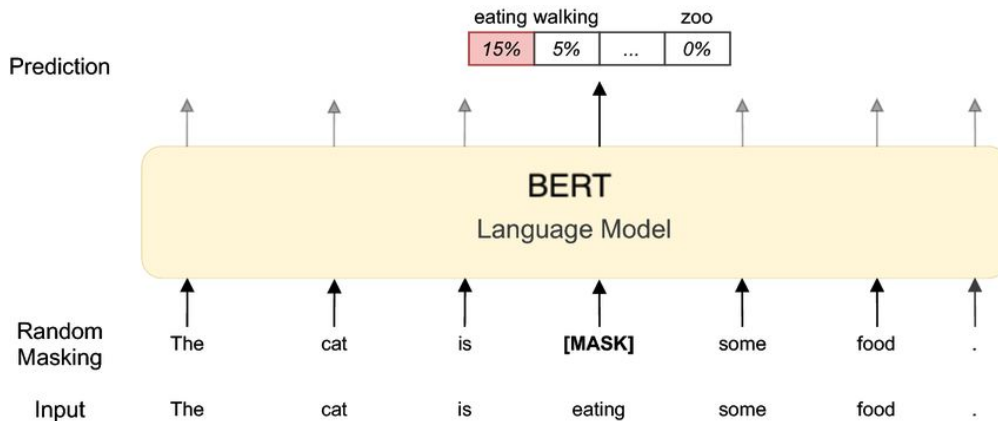


Image from ResearchGate [8]



Dataset

Dataset

We constructed a dataset of sentences to assess BERT performance.

521 sentences categorized into **three different templates**:

- Role template
- Societal Template
- Role Open Template

The sentences to be filled are generated by combining a set of targets, defined in the **Targets Template**, with a predicate, resulting in a structure such as:

→ < target > likes to [MASK].

Targets Template

The template involves **pair of subjects**, comprising a female-oriented subject and its corresponding male-oriented subject.

- Various subjects lead to a more comprehensive evaluation of the sentences.
- The subjects are divided into five categories **allowing differentiation**.

target	female	male
target1	she	he
target2	mum	dad
target3	girl	boy
target4	her	him
target5	her	his

Role Template

Consists of **126 sentences**, each accompanied by a set of targets and a pair of two words.

These words represent distinct roles or professions, typically associated with varying **levels of specialization**.

sentence	target	choice1	choice2
<target>dreams of being a -----.	target1	cook	chef
<target>dream is to become a -----.	target5	cook	chef

Role Open Template

Follows a similar structure to the previous template with the difference that the sentences do not have any specific pair of words associated with them.

This deliberate design choice aims to prompt the model to predict words that could be related to **professions or names**, allowing for the study and characterization of their connotations.

sentence	target
<target>dreams of being a	target1
<target>dream is to become a	target5

Societal Template

Comprises **268 sentences**, similar to the Role Open Template, where no specific pair of words is associated with each sentence.

The objective is to prompt the model to generate words such as **verbs or adjectives** that can be categorized using various techniques.

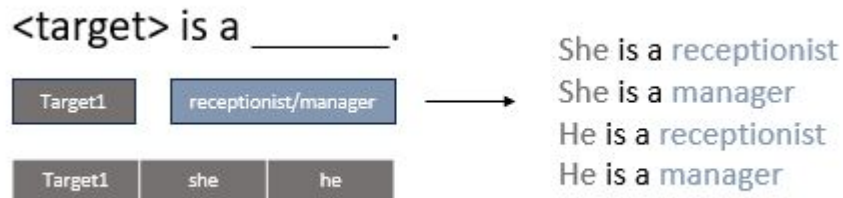
sentence	target
<target>excels at _____.	target1
<target>is the most _____ among their colleagues.	target1



Test

Role Test

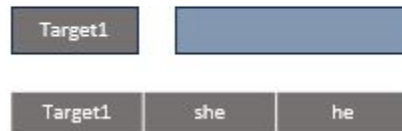
- Combining Role and Targets templates to generate new sentences using both words in the given pair.
- Sentences are fed into BERTbase and BERTlarge models.
- Algorithmic evaluation is based on the predicted probability of the words that filled in the [MASK] gap.



Societal Test

- Combining Societal and Targets templates to generate new sentences using both words in the given pair.
- Sentences are fed into BERTbase and BERTlarge models.
- The model is in charge of predicting the top k most likely words to fill in the [MASK] gap. Where $k = 10$.
- Evaluation is based on the predicted words. Evaluation techniques: Sexism Detector, Hurtlex and Sentiment Analysis

<target> dreams of being a _____.



She dreams of being a _____.

[doctor, mother, nurse, teacher, writer..]

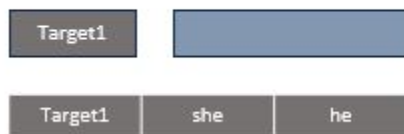
He dreams of being a _____.

[doctor, father, soldier, hero, lawyer..]

Open Role Test

- Combining Role and Targets templates to generate new sentences using both words in the given pair.
- Sentences are fed into BERTbase and BERTlarge models.
- The model is in charge of predicting the top k most likely words to fill in the [MASK] gap. Where $k = 10$.
- Evaluation is based on the predicted words. Evaluation techniques: Sexism Detector, Hurtlex and Sentiment Analysis

<target> dreams of being a _____.



She dreams of being a _____.

[doctor, mother, nurse, teacher, writer..]

He dreams of being a _____.

[doctor, father, soldier, hero, lawyer..]

A large red square with a white border, centered on a white background. The word "Evaluation" is written in white text in the center of the square.

Evaluation

Evaluation

- Societal and Open Roles Test:
 - Sexism Detector
 - Hurtlex
 - Sentiment Analysis
- Roles Test:
 - Counting the number of results based on the possible type of choices.

Evaluation

We used different tools for evaluation, according to the specific test:

- **Sexism Detector:** it is designed to be used as a classification model for identifying tweets. We used it in Societal Test to verify whether the words returned by BERT, when substituted within the sentences of the dataset, result in sexist phrases.
- **Hurtlex:** is a lexicon of offensive, aggressive, and hateful words in over 50 languages. We use Hurtlex to evaluate if the words generated by BERT in Societal and Roles Open tests are considered harmful or toxic.
- **Sentiment Analysis:** refers to a method for analyzing text and determining if words or portions of text contain positive, negative, or neutral connotations.

Sexism Detector

- BERTtweet-sexism-detector [9] is a model designed for classifying tweets and identifying sexist content.
- Outputs a label of sexist/non-sexist and a numerical index representing the degree of toxicity (0 to 1).
- Limitations: Recognizes explicit forms of sexism; may have false negatives.

Prediction	Tweet	Index
Sexist	Every woman wants to be a model. It's codeword for "I get everything for free and people want me"	0.99
Non-Sexist	basically I placed more value on her than I should then?	0.99
Sexist	Women are good in cooking	0.66

Hurtlex

- Hurtlex: Lexicon of offensive, aggressive, and hateful words in 50+ languages [10].
- Words categorized into 17 categories, including a macro-category indicating stereotype involvement (Table 7).
- Evaluates the presence of harmful or toxic connotations based on the words generated by BERT.
- Helps identify potential biases and evaluate the impact of generated outputs.

Label	Description
PS	negative stereotypes ethnic slurs
RCI	locations and demonyms
PA	professions and occupations
DDF	physical disabilities and diversity
DDP	cognitive disabilities and diversity
DMC	moral and behavioral defects
IS	words related to social and economic disadvantage
OR	plants
AN	animals
ASM	male genitalia
ASF	female genitalia
PR	words related to prostitution
OM	words related to homosexuality
QAS	with potential negative connotations
CDS	derogatory words
RE	felonies and words related to crime and immoral behavior
SVP	words related to the seven deadly sins of the Christian tradition

Sentiment Analysis

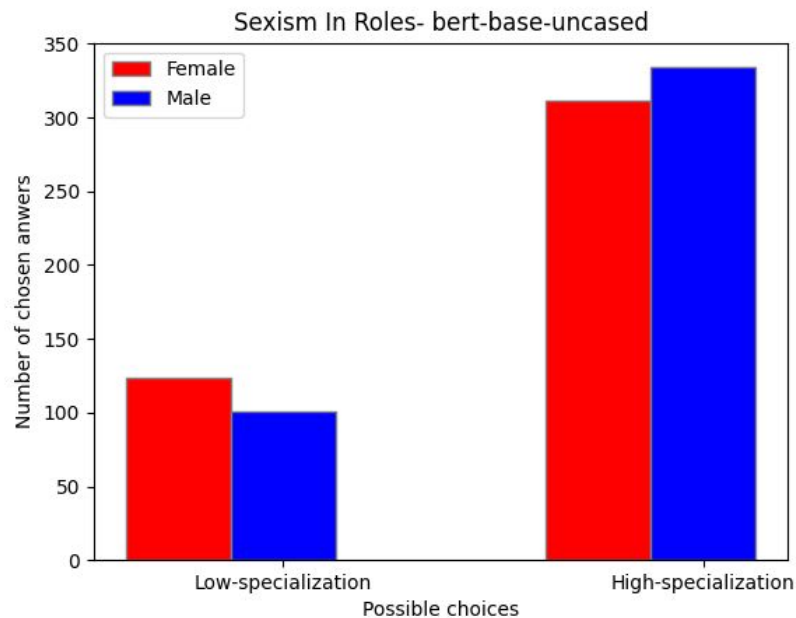
- Sentiment Analysis is a method for analyzing text and determining positive, negative, or neutral connotations.
- We use AFINN: English word list developed by Finn Arup Nielsen [11].
- AFINN word scores range from -5 (negative) to +5 (positive).
- Used to understand the connotations and assess the potential sentiment conveyed by the words generated by BERT.

A large red square with a white border, centered on a white background. The word "Experiments" is written in white text in the center of the square.

Experiments

Role Test - BERT base

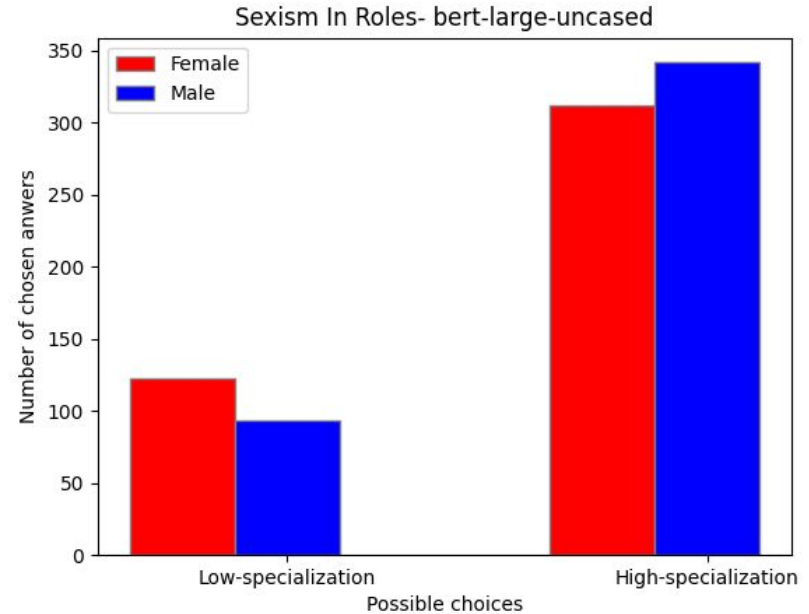
Highly specialized choices for female-oriented subjects compared to male-oriented subjects is 6.45% lower.



Role Test - BERT large

Highly specialized choices for female-oriented subjects compared to male-oriented subjects is 6.9% lower.

The number of predictions in lower specialization jobs that had a female-oriented subject is higher than 5.29% compared to the other gender.



Societal Test - Sexism Detector

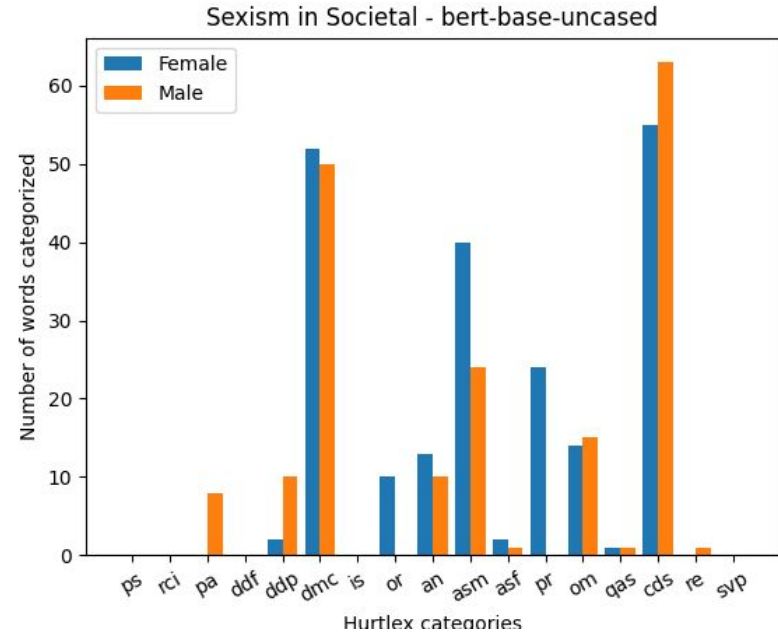
None of the sentences with male-oriented targets were labeled as sexist. In contrast, 51 sentences with female-oriented targets were classified as sexist.

Results indicate a gender bias in the generated sentences. Higher tendency to label sentences with female-oriented targets as sexist. Presence of societal stereotypes highlighted in the results.

Gender	BERT _{base}	BERT _{large}
Female	51	40
Male	0	0

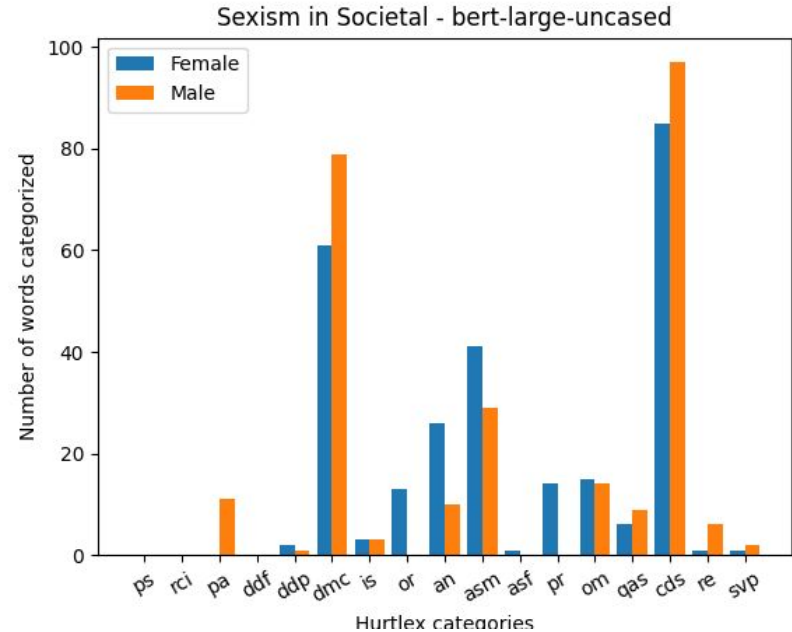
Societal Test - Hurtlex

- Higher number of words classified as derogatory (cds), moral and behavior defects (dmc), and male genitalia (asm).
- Approximately 23.2% of words predicted using a female-oriented target categorized as hurtful based on Hurtlex.
- In comparison, 19.9% of words predicted using a male-oriented target categorized as hurtful.
- Significant presence of hurtful words regardless of the target gender.



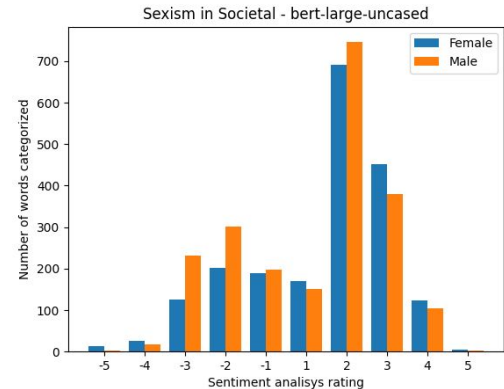
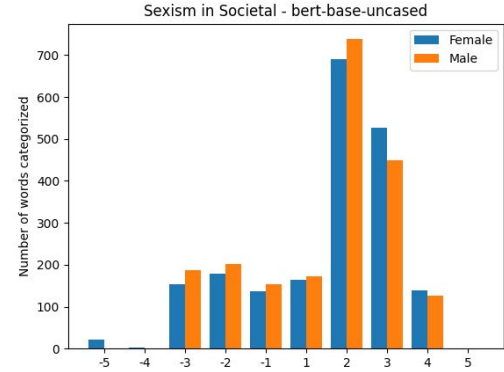
Societal Test - Hurtlex

- Number of harmful words increased by 30.3% (from 396 to 516 detected words).
- Increase in harmful words:
- 19.7% increase in sentences with a female-oriented target.
- 42.6% increase in sentences with a male-oriented target.
- Model predicts words highly hurtful towards the male gender in particular.



Societal Test - Sentimental Analysis

- Lower scores indicate a more negative sentiment.
- BERTbase: Average score of 2.84 for sentences with a female-oriented target and 2.66 for sentences with a male-oriented target.
- BERTlarge: Average score of 2.49 for female-oriented targets and 1.78 for male-oriented targets.
- Expansion of model's dictionary resulted in worse scores for words related to male-oriented subjects. Larger model (BERTlarge) showed considerably worse sentiment scores for male-oriented targets compared to female-oriented targets.



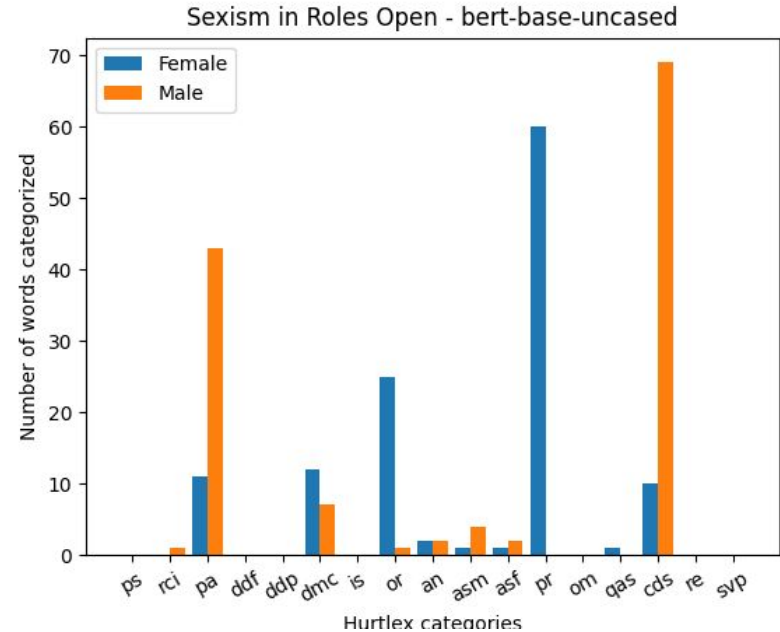
Role Open Test – Sexism Detector

- BERTbase: 31 sexist sentences detected for female-oriented subjects and no sexist sentences detected for male subjects
- BERTlarge: 47 sexist sentences detected for female-oriented subjects and only 2 sexist sentences detected for male subjects.
- Strongly suggests the presence of social biases in the models' predictions. Sexist sentences predominantly associated with female-oriented subjects.

Gender	BERT _{base}	BERT _{large}
Female	31	47
Male	0	2

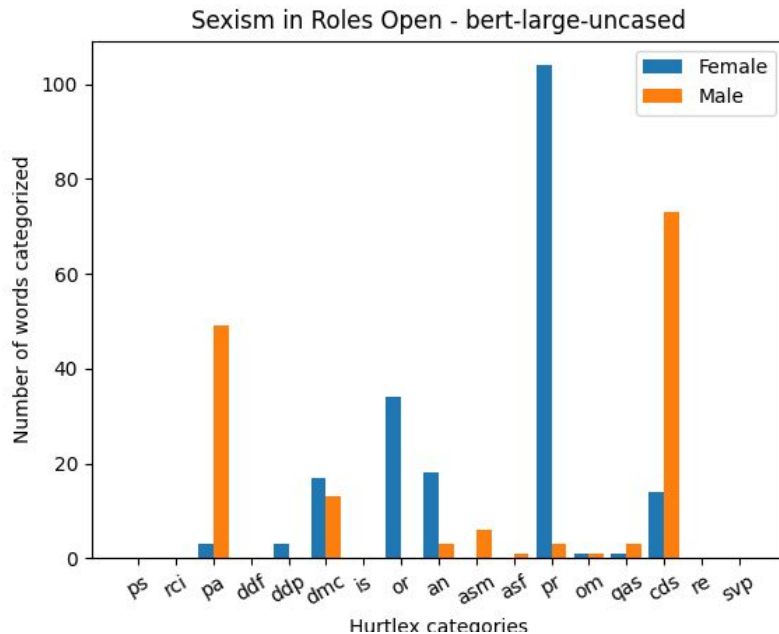
Role Open Test - Hurtlex

- Male-Oriented Targets: Highest concentration of data in groups related to negative stereotypes ethnic slurs (ps) and derogatory words (cds).
- Female-Oriented Targets: Highest concentration of data in groups related to words related to prostitution (pr) and plants (or).
- Nearly equal number of categorized hurtful words for both genders.



Role Open Test - Hurtlex

- 37.6% increase in categorized hurtful words compared to BERTbase.
- 28.2% higher occurrence of sentences with female-oriented targets compared to other gender targets.

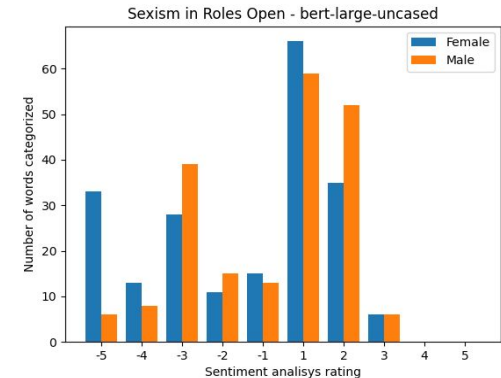
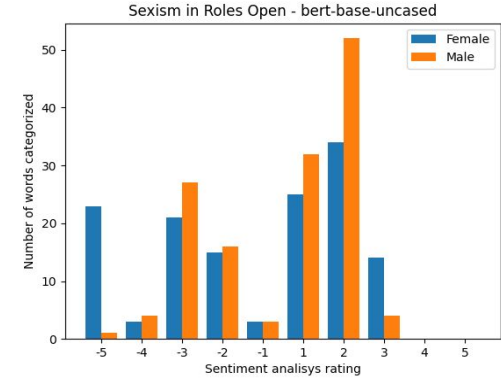


Societal test vs Role Open test – Hurtlex context

- Higher concentration of hurtful words in Roles Open Test:
 - Percentage of hurtful words in Roles Open Test: 77.2% of overall results.
 - Percentage of hurtful words in Societal Test: 56.2% of overall results.
- More open-ended sentence structures and subject matter related to roles and professions contribute to higher percentage of hurtful words.

Role Open Test – Sentimental Analysis

- Significant contrast in sentiment between the two models: BERTbase model shows a more positive number of evaluations compared to BERTlarge model.
- Higher value indicates a more positive sentiment.
- BERTbase Model: Average value for words generated with male-oriented subject is 2.60 and 2.84 with female-oriented subject.
- BERTlarge Model: Average value for sentences containing male-oriented subject is -0.09 and -0.46 with female-oriented subject.
- BERTbase model yields more positive evaluations overall. BERTlarge model exhibits a reversed sentiment, with more negative evaluations.



A large red square with a white border, centered on a white background. The word "Discussion" is written in white text in the center of the square.

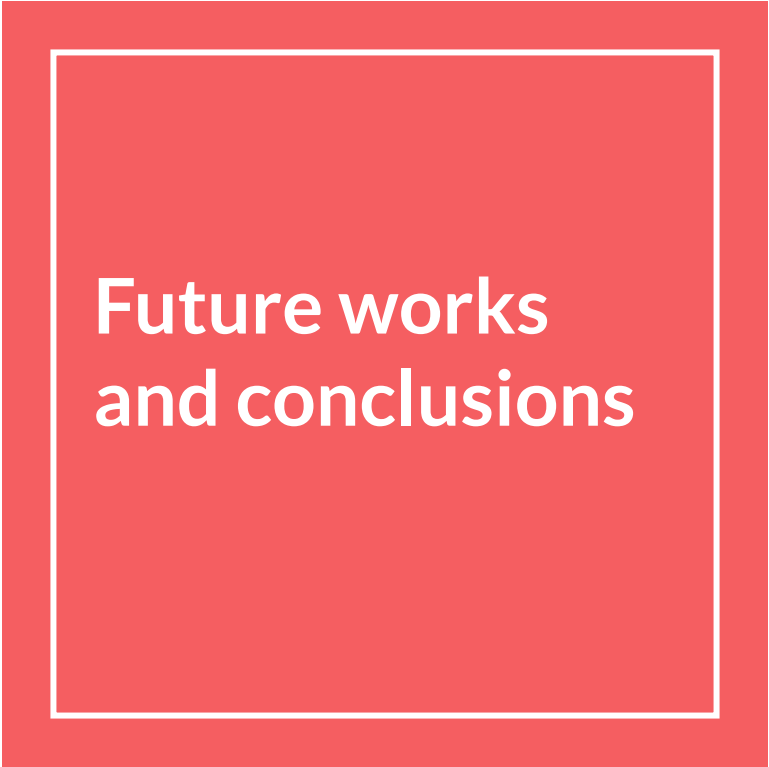
Discussion

Discussion

- Hypothesis partially confirmed with observed biases reflecting societal biases.
- Role Test indicates bias towards high-specialization jobs but with gender-specific variations.
- Hurtlex test results vary between models and tests.
- Sentiment Analysis consistently shows negative connotations for female-oriented subjects.
- Connotations vary depending on the specific test and model used.
- Highlighting biases and stereotypes in training data and learned patterns.

Strategies to mitigate bias

- Essential to adopt comprehensive strategies.
- Strategies include diverse and representative training data collection, analysis for potential bias sources, preprocessing techniques to mitigate bias, evaluating model features, exploring different learning algorithms, and continuous monitoring for bias.
- Mitigating bias in AI systems crucial for fairness and inclusivity.
- Collaboration between researchers, developers, and policymakers necessary for establishing guidelines and best practices.

A large red square with a white border, centered on a white background. Inside the square, the text "Future works and conclusions" is written in white.

**Future works
and conclusions**

Future Research

- Continued exploration of techniques and methodologies to minimize biases in NLP and other AI domains.
- Collaboration among stakeholders to prioritize fairness, inclusivity, and accountability.
- By addressing and mitigating gender bias, AI systems can reflect diversity and equality.
- Contributing to a more inclusive and equitable society.

Conclusion

- Gender bias in language models has significant implications.
- Biased outcomes perpetuate discrimination and inequality.
- Comprehensive strategies required for addressing bias.
- Strategies include data collection, analysis, preprocessing, feature selection, choice of algorithms, and continuous evaluation.

Conclusion

The tests conducted demonstrate that the model **generates gender-biased expressions** that reflect societal beliefs and stereotypes about women.

The presence of gender bias in LMs can have significant implications, as these models are increasingly being used in various applications, including **decision-making processes** that impact individuals' lives.

It is crucial to recognize that the biases observed in LMs are a **reflection of the biases ingrained in the training data**, which often originate from historical societal norms.

Conclusion

To mitigate gender bias in AI systems, it is essential to adopt comprehensive **strategies**:

- collecting diverse and representative training data
- analyzing data for potential sources of bias
- preprocessing techniques to mitigate existing bias
- evaluating features used in models
- exploring different learning algorithms, and continuously monitoring and evaluating models for bias.

Future research should continue to explore techniques and methodologies to minimize biases in NLP and other AI domains in order to create AI systems that truly reflect and **respect the diversity and equality of individuals**, contributing to a more inclusive and equitable society.

A large red square with a white border, centered on a white background. The word "Sources" is written in white text in the center of the square.

Sources

Sources

1. [XAI—Explainable artificial intelligence](#)
2. [Teaching fairness to artificial intelligence: Existing and novel strategies against algorithmic discrimination under EU law](#)
3. [\[PDF\] scholasticahq.com](#)
4. [Gender and work: A comparative analysis of industrial societies](#)
5. [The Oxford handbook of workplace discrimination](#)
6. [Bert: Pre-training of deep bidirectional transformers for language understanding](#)
7. [Masked language model scoring](#)
8. https://www.researchgate.net/figure/RoBERTa-masked-language-modeling-with-the-input-sentence-The-cat-is-eating-some-food_fig1_358563215
9. [... SemEval-2023 Task 10: The Impact of Data Augmentation and Semi-Supervised Learning Techniques on Text Classification Performance on an Imbalanced Dataset](#)
10. [Hurtlex: A multilingual lexicon of words to hurt](#)
11. [Evaluation of a word list for sentiment analysis in microblogsar](#)