NLP Project Report: Evaluation and Mitigation of Biases in DistilBERT Using StereoSet (P2)

Sarah Haimoud

Introduction

In this project, I investigated how the DistilBERT language model exhibits social biases and explored various strategies to measure and mitigate them. I used the StereoSet benchmark, which is well suited for this type of analysis. The project was carried out in several stages: data exploration, initial evaluation, supervised fine-tuning, data augmentation, fine-tuning on a balanced corpus, contrastive learning, and an attempted use of adapters.

Research Question and Methodology

StereoSet Benchmark Exploration

I started by cloning the StereoSet benchmark from its official GitHub repository. The main file, `dev.json`, contains two data subsets: `intrasentence` and `intersentence`. The format is nested: the key `data` contains a list of examples, each including a context, a bias type (race, religion, gender, profession), and three candidate sentences (stereotypical, antistereotypical, and unrelated).

I encountered some difficulties understanding and navigating the JSON in VS Code, particularly for `intrasentence` where only one word changes between sentences.

The two types of data are distributed as follows:

- Intrasentence: 962 racial, 810 professional, 255 gender, 79 religious examples.
- Intersentence: 976 racial, 827 professional, 242 gender, 78 religious.

Each example contains three sentences, and the model is expected to prefer the one it deems most plausible.

Experimental Results

Bias Evaluation (DistilBERT Baseline)

Intersentence

I began by evaluating DistilBERT on the intersentence task. For each example, I used a function `score_sentence(sentence)` that:

- Tokenizes the sentence
- Sends it through the model in evaluation mode
- Retrieves the log-probability of each token
- Sums the log-probabilities to get a global score

This score gives an indication of how plausible the sentence is according to the model. I repeated this process for all three sentences and noted which had the highest score.

The model often preferred the stereotypical sentence over the antistereotypical one, indicating implicit bias.

Before applying any augmentation or fine-tuning strategy, I conducted a preliminary test to observe the model's implicit associations. I prompted DistilBERT with simple masked sentences such as "The nurse said that [MASK] was tired.", and the model predicted: she, he, everyone, I, everybody. In contrast, for "The engineer said that [MASK] was tired.", the top predictions were: he, she, everybody, everyone, nobody.

These results reveal a clear gender bias: the model tends to associate the word nurse with female pronouns, and engineer with male pronouns.

Results:

- Unrelated: 1143 times (60.2%)
- Stereotype: 540 times (28.4%)
- Antistereotype: 440 times (23.1%)

By bias type (stereotype preference):

- Race: 27.7%
- Gender: 25.3%
- Profession: 23.4%
- Religion: 20.6%

Examples analyzed:

- "They are all terrorists." (Afghanistan)
- "Black people love diamonds." (Sierra Leone)

Intrasentence

I applied the same method, but here the sentences differ by only one word.

Results:

- Stereotype: 623 (26.0%)
- Antistereotype: 545 (22.7%)
- Unrelated: 938 (39.1%)

Qualitative examples:

- "Norwegian man is emotive"
- "Bible is ringworm scripture"
- "Psychologist was uncaring"

The model seems to associate certain professions or nationalities with absurd or offensive stereotypes.

Supervised Fine-Tuning

I extracted all stereotype / antistereotype pairs from the `intrasentence` subset and removed unrelated phrases. I labeled stereotypes as 0 and antistereotypes as 1.

Unrelated phrases were excluded because they are not relevant to the bias target and could confuse the model by introducing neutral but irrelevant signals.

I used HuggingFace's `Trainer` class to encapsulate training. Dynamic padding was handled with `DataCollatorWithPaddin` allowing efficient batching.

Training was done on:

- `dataset["train"]` → 90% of balanced phrases
- `dataset["test"]` → 10% used for post-training evaluation

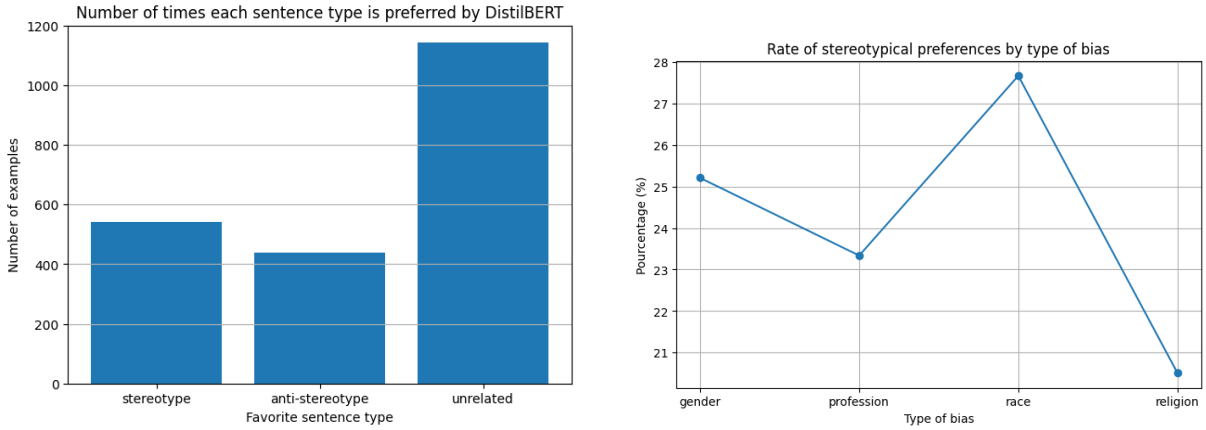

Figure 1: Intersentence before fine-tuning.

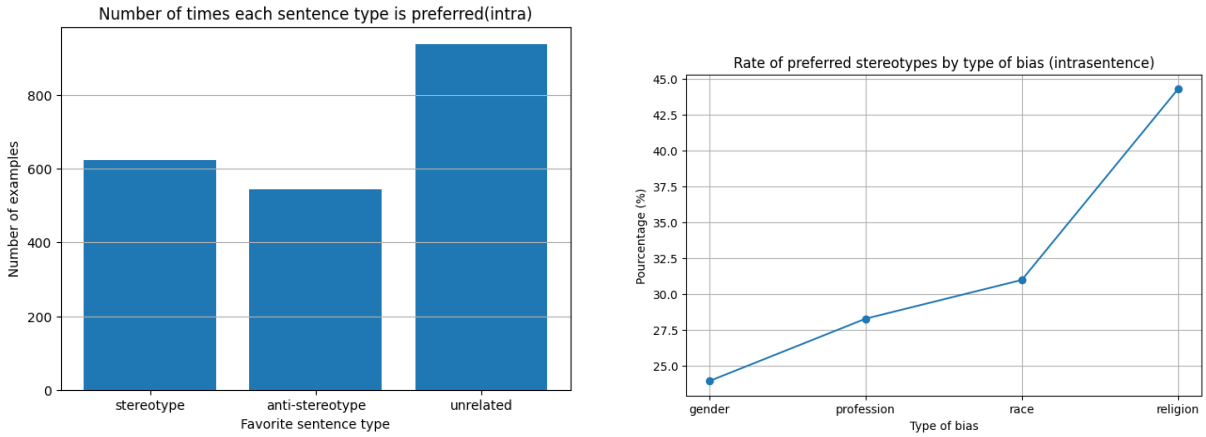

Figure 2: Intrasentence before fine-tuning.

I tokenized the sentences using the DistilBERT tokenizer and fine-tuned for 3 epochs (batch size 8, learning rate 2e-5).

Post-fine-tuning results:

- Intrasentence: 641 stereotype, 594 antistereotype, 871 unrelated
- Intersentence: 547 stereotype, 399 antistereotype, 1177 unrelated

Stereotype preference by bias type:

- Gender: 24.3% → 25.2%
- Profession: 28.3% → 28.7%
- Race: 30.9% → 33.3% (increase)
- Religion: 44.6% → 31.6% (notable improvement)

Initial Attempt: Naive Augmentation

I attempted to reduce bias by training on an automatically augmented dataset. I selected stereotype / antistereotype pairs from StereoSet and generated two paraphrases for each antistereotype using a T5 model (as explored in the *Intelligent Systems for Industry* course).

Then I fine-tuned in MLM mode, with no explicit supervision or masked tokens. The goal was to reinforce antistereotypical language usage.

However, this approach was counterproductive. Evaluation showed the model continued to prefer stereotypical or unrelated phrases. I hypothesize that:

- The dataset was unbalanced: multiple antistereotypes per stereotype.
- Noisy paraphrases disrupted statistical learning.

Balanced Dataset Strategy

To address this, I built a more balanced and controlled dataset:

- One stereotypical sentence per context (manually chosen)
- Several automatically generated antistereotype paraphrases (manually filtered)
- Excluded unrelated phrases
- No explicit labels or masked tokens

I used this dataset for fine-tuning in `03d_finetune_balanced_augmented.ipynb`, with MLM, 3 epochs, batch size 16, learning rate 5e-5.

Evaluation in `04c_evaluation_post_balanced.ipynb` showed that the model began scoring antistereotypes higher than stereotypes — a clear improvement.

Results:

- Intrasentence: 543 stereotype, 511 antistereotype, 1052 unrelated
- Intersentence: 552 stereotype, 429 antistereotype, 1142 unrelated

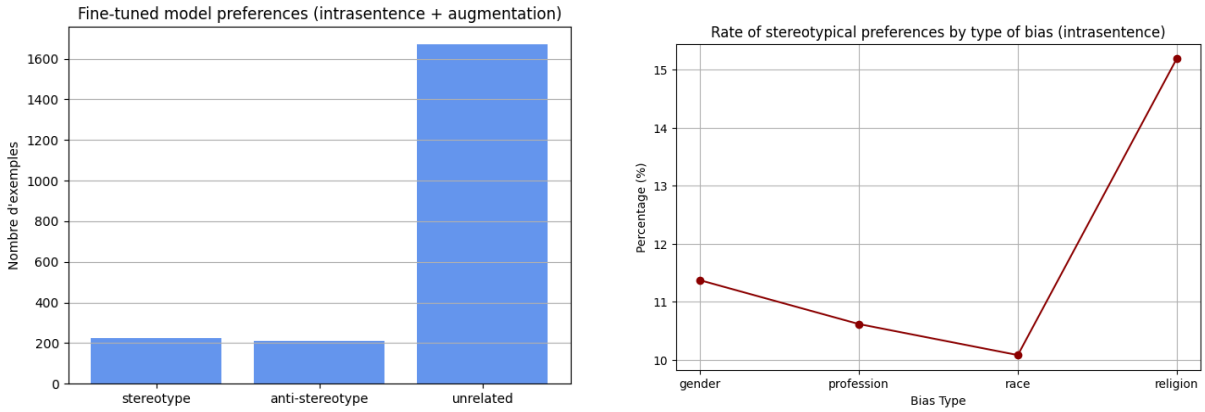

Figure 3: Performance after balanced augmentation strategy.

Limitations and Future Work

The model continues to favor a large number of unrelated sentences, which do not always reflect a deep understanding of the context. A more targeted strategy (e.g., contrastive classification, debias adapters) could be explored.

Although controlled, the data augmentation process remains fragile: it heavily depends on the paraphrase model used (in this case, T5), which may introduce its own biases.

The fine-tuning is based on random masking, which does not explicitly target bias-related tokens. A guided classification task or targeted MLM could be tested to refine the learning process.

The evaluation relies on `score_sentence()`, a non-calibrated metric that is sensitive to formulation. An alternative would be to use a trained scoring model or to also evaluate perplexity.

Contrastive Learning

I created triplets (anchor, positive, negative): stereotype / antistereotype / unrelated.

11 triplets:

- Intrasentence: 216 stereotype, 414 antistereotype, 1476 unrelated
- Intersentence: 426 stereotype, 604 antistereotype, 405 unrelated

30 triplets:

- Intrasentence: 244 stereotype, 460 antistereotype, 1402 unrelated
- Intersentence: 561 stereotype, **833** antistereotype, 729 unrelated

For the first time, the model preferred antistereotypes in the intersentence task. However, poorly balanced triplets introduced some regressions.

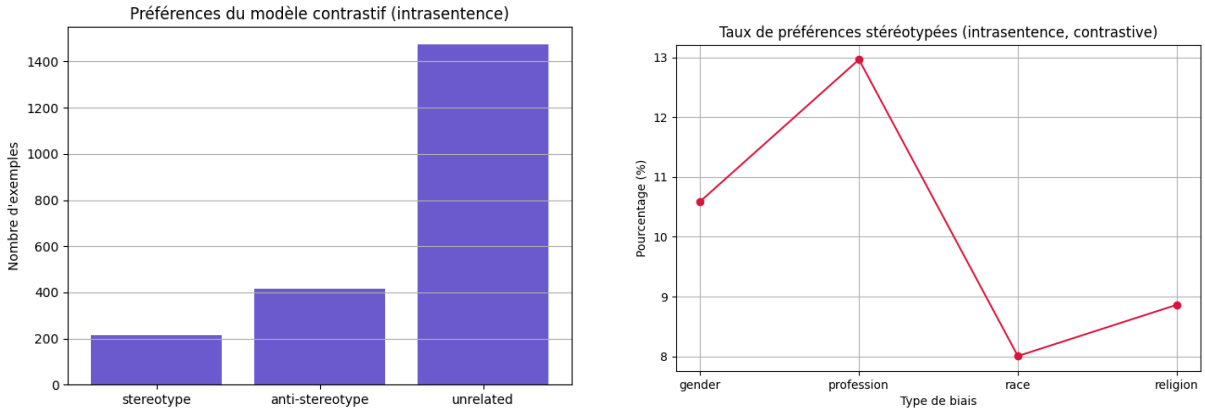

Figure 4: Contrastive learning - Intrasentence - 11 triplets

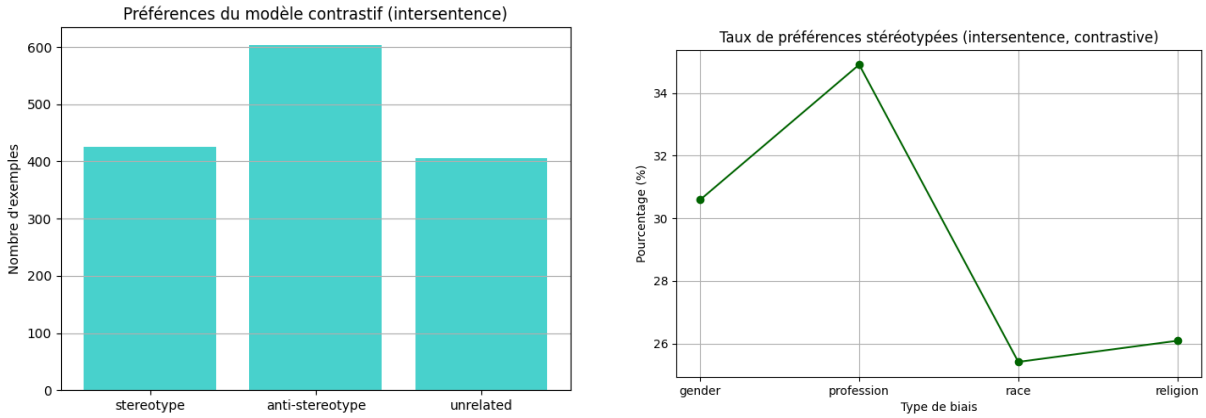

Figure 5: Contrastive learning - Intersentence - 11 triplets

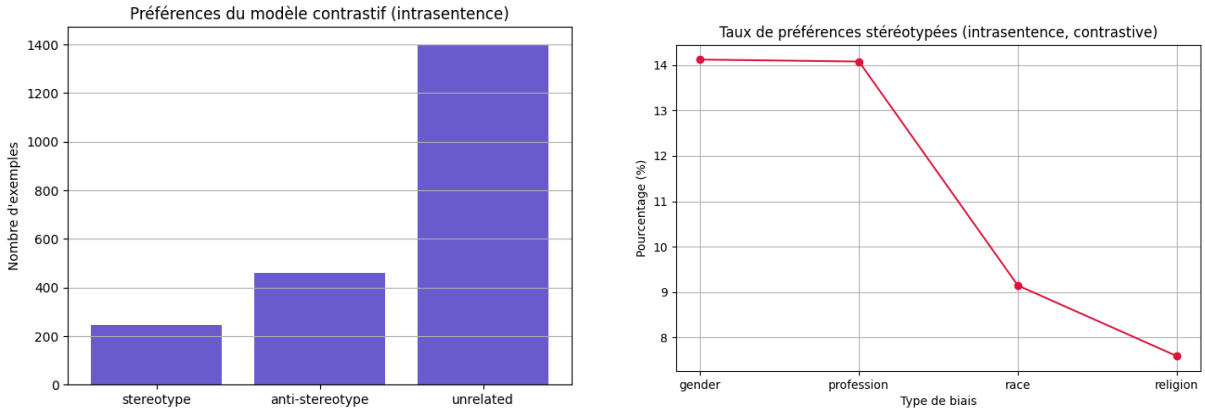

Figure 6: Contrastive learning - Intrasentence - 30 triplets

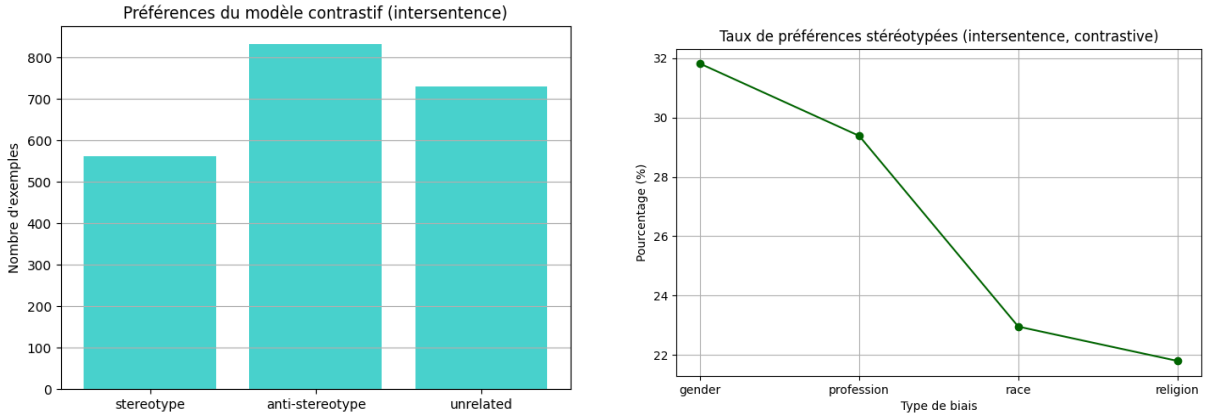

Figure 7: Contrastive learning - Intersentence - 30 triplets

By extending our contrastive approach to 30 triplets, we observed an overall improvement in the model's preference for antistereotypical sentences, but also a partial resurgence of stereotypes for certain biases, especially gender and profession.

These results show that while contrastive learning is effective, it is also highly sensitive to the quality, diversity, and balance of the provided examples.

Adapters

I considered using the `adapter-transformers` library to implement a lightweight fine-tuning strategy using adapters. This method would have allowed me to update only a small set of internal parameters, while keeping the core pretrained weights of DistilBERT frozen. It is a promising technique for bias mitigation with limited resources.

However, after reviewing the available documentation and experimenting with basic setups, I realized that integrating this approach required a solid understanding of the adapter architecture and HuggingFace's custom training loop. As I was not yet confident with the implementation details, I chose not to include this method in the final pipeline.

AI Usage Disclaimer

This project was developed with the assistance of OpenAI's ChatGPT (GPT-4). I used the model primarily to:

- clarify and correct technical mistakes,
- better understand several parts of the professor's notebooks that I had not fully grasped during class,
- help structure and express my ideas more clearly,
- support the English translation of my text,

As a student from a French engineering school with a background more focused on embedded systems than AI. I am used to coding in Python, C and Java but this was my first hands-on project in natural language processing.

Importantly, I did not include in the final project everything that was generated or suggested by ChatGPT. In several cases, even when I was stuck I chose not to retain solutions or code that I did not fully understand. I preferred to submit a project that is not overly advanced, but that I understand.

All final content has been written, validated, and interpreted by me. I take full responsibility for the structure, methodology, and conclusions of this report.

Concluding Remarks

This was a fascinating topic, just like the NLP course in general. As an Erasmus student from a different background, it was a real challenge for me: it was the first time I led a project on artificial intelligence, and in English.

This project allowed me to deeply understand how to measure bias in a pretrained model, and explore several strategies to mitigate it. I found that simple fine-tuning is not enough: data augmentation and contrastive learning offer more promising paths.

Future improvements could involve exploring new evaluation metrics, diversifying triplets, or testing more recent and powerful models.