With the accessibility of artificial intelligence models to the general public, there has been a significant rise in their usage. However, concerns have been raised regarding the behaviour and arguments supported by these models. Ongoing debates surround the fairness of these models and the potential for them to produce or perpetuate discriminatory biases through natural language.

A central belief in combating discrimination is the fight against sexism and gender bias. Unfortunately, these biases continue to persist in NLP algorithms and AIs, as they often reflect the societal norms where gender discrimination remains deeply ingrained.

Our objective is to investigate how language models, such as BERT, handle gender biases and stereotypes. Specifically, we aim to determine whether BERT generates sexist predictions. To achieve this, we plan to utilize BERT's masked language model feature to predict masked words in predefined sentences, using a template. We will evaluate how the predictions change when the subject of the sentence shifts from a male-associated gender to a female-associated gender.

In order to obtain qualitative results, we will assess the predictions using an existing sexism detector model. We will compare the detector's output with the actual labels assigned to the data, indicating whether the predictions are considered sexist or not.

The difficulties we found in the last project’s idea comes from the latest Twitter’s update. The developers decided to block definitely all the scrapers and make tweets accessible only through non-free API. As a consequence, we have been not able to build an adequate database as a base of our tests.