Group 05 24.01.2025

# Detection of Online Sexism - Management Summary

## The Task:

The goal of this project is to perform binary classification of short utterances on social media to determine whether they are sexist or non-sexist.

## The Challenges:

There are a lot of challenges we faced in this project. First, defining sexism proved to be a significant challenge due to its subjective nature, as there is no universally agreed-upon definition. Differentiating sexism from related concepts like abuse, offensiveness, or racism was particularly difficult, and our team had varying opinions on where to draw the line. Balancing the risk of not detecting sexist statements with the potential harm of incorrectly labeling non-sexist statements as sexist further complicated the process.

Second, the relabeling process itself was challenging, as aligning on consistent labels required extensive discussion and compromise. Even after agreeing on a definition, differences in perspective made achieving consensus difficult.

Third, the nature of the Twitter dataset introduced unique obstacles. Tweets are inherently short, often lack context (e.g., the gender of the author or the intended recipient), and frequently include slang, making them harder for the model to interpret. The absence of contextual information also makes it difficult to discern subtleties such as sarcasm, sugar-coated language, or hidden meanings like praise or appreciation.

Finally, class imbalance in the dataset, a common issue in tasks like this, posed additional difficulties, as sexist tweets are likely underrepresented compared to non-sexist tweets, making it challenging to train a balanced and fair model.

## Ressources used:

For our project, we relied on Google Colab as the primary platform for training our models. Additionally, we used GitHub as a centralized repository to manage and organize our code, document our results, and share key findings.

## Implemented Solutions:

For this task, we used the EDOS Dataset ("Explainable Detection of Online Sexism"), specifically focusing on the subset that was already annotated. This subset contains approximately 20,000 samples of short utterances labeled as sexist or non-sexist.

To solve this classification task, we built and compared two models:

1. Logistic Regression: A non-deep learning approach serving as a baseline.
2. RoBERTa: A state-of-the-art transformer-based deep learning model.

We analyzed the results of both models and looked for patterns and key words the models use to detect sexist samples. In this analysis we realized that we had different opinions regarding the labels for some samples.  So, we decided to relabel a random subset of the dataset. Everybody wrote their own guidelines to define sexism and relabeled the subset of samples. We sat together, discussed our results, agreed on a guideline and a relabeled dataset.

We retrained the RoBERTa model, on our relabeled dataset, as expected the performance was only slightly better. Out of curiosity, we generated a new dataset with ChatGPT, which has obvious sexist and non-sexist samples. Unfortunately, the model struggled to classify the new dataset.

## Limitations:

A key limitation of this project lies in the inherent subjectivity of the dataset. While efforts were made to establish a shared definition of sexism and relabel a subset of the data accordingly, the annotations remain influenced by the personal perspectives, backgrounds and biases of our group members. This subjectivity introduces potential biases into the relabeled dataset, which may affect the model's fairness and reliability. Additionally, ethical concerns arise, as the model's decisions could perpetuate unintended biases or lead to misclassifications that carry social implications, particularly in real-world applications. Another limitation is the challenge of generalization: the model, trained on the EDOS dataset, might not perform well on other datasets or in different linguistic or cultural contexts, limiting its broader applicability.

## Next Steps:

To advance the project, we consider multiple steps that can be taken. A first possible step is to train the model additionally to the twitter dataset, on a more obvious dataset such as the GPT generated on that we used, which offers a solid foundation for detecting sexist content.

Additionally, the model could be provided with a lexicon of specific words and phrases commonly associated with sexism. Each term should include a definition and context to aid in accurate classification. For example, "thots" is slang for "that ho over there," "SMV" refers to "sexual market value," and "tranny" is an offensive term used to describe transgender individuals.

Incorporating emojis into the dataset is another possible step, as they often carry nuanced meanings in digital communication. Each emoji should be defined and annotated based on its typical use in sexist contexts to enhance the model's ability to interpret non-verbal cues. Similarly, abbreviations and acronyms frequently used in such language should be identified, and their meanings documented, to ensure comprehensive understanding.

We also thought that establishing a baseline that differentiates sexism, racism, and offensive language could be helpful. By defining clear criteria for each concept, the model can be refined to classify content accurately and address the overlap between these categories.