Developing Transparent Methods for Identifying Sexism in Social Media: Combining Explainability with Human Rationalizations

Young Complexity Researchers Utrecht (YCRU) Meeting

Hadi Mohammadi Anastasia Giachanou Ayoub Bagheri

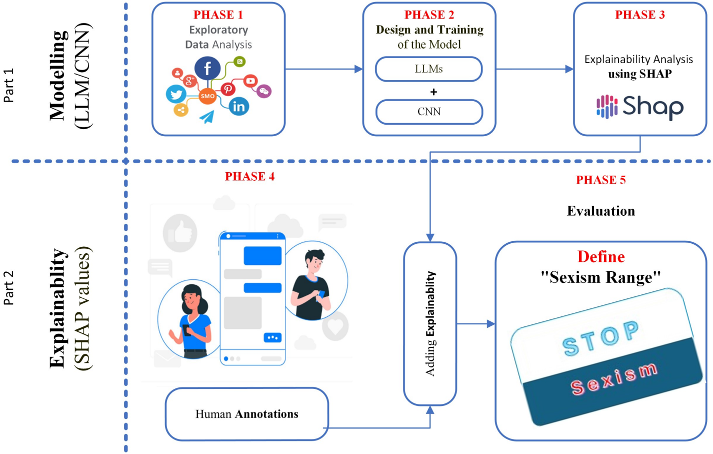
Department of Methodology and Statistics,
Utrecht University,
The Netherlands.

November 28, 2024

Universiteit Utrecht

Research Methodology

In this study, we propose a novel methodology that consists of explainability and rationalization. This approach is structured into two parts with various phases.



Part 1

Modelling (LLM/CNN)

PHASE 1
Exploratory Data Analysis

PHASE 2
Design and Training of the Model

LLMs
+
CNN

PHASE 3
Explainability Analysis
using SHAP

Shap

Part 2

Explainablity (SHAP values)

PHASE 4

Human **Annotations**

Adding **Explainablity**

PHASE 5

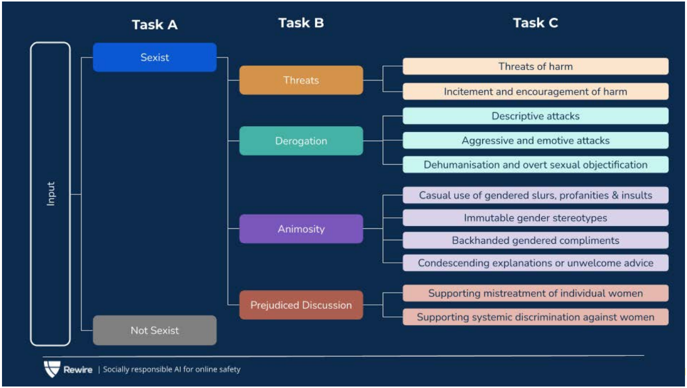
Evaluation

Define
"Sexism Range"

Universiteit Utrecht

Tasks Overview & Dataset

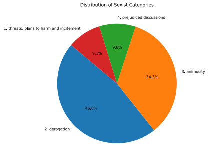
- Explainable Detection of Online Sexism (*EDOS*) dataset.
- 20,000 labeled posts from Gab and Reddit:
  - Subtask A: binary classifier for categorizing posts as sexist or non-sexist.
  - Subtask B: four-class for sexist posts.
  - Subtask C: 11-class for more specific labels of sexism .


Task A
Task B
Task C
Input
Sexist
Not Sexist
Threats
Derogation
Animosity
Prejudiced Discussion
Threats of harm
Incitement and encouragement of harm
Descriptive attacks
Aggressive and emotive attacks
Dehumanisation and overt sexual objectification
Casual use of gendered slurs, profanities & insults
Immutable gender stereotypes
Backhanded gendered compliments
Condescending explanations or unwelcome advice
Supporting mistreatment of individual women
Supporting systemic discrimination against women
Rewire

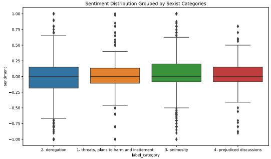
Socially responsible AI for online safety

Exploratory Data Analysis (EDA)

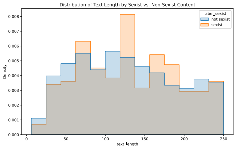
- Distribution of each Class.
- Text Distribution of each Class.
- Text Length Distribution.
- Top 3 Grams in Sexist Content.



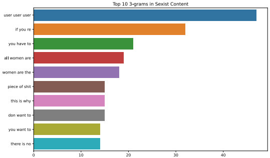
Distribution of Sexist Categories



Sentiment Distribution Grouped by Sexist Categories



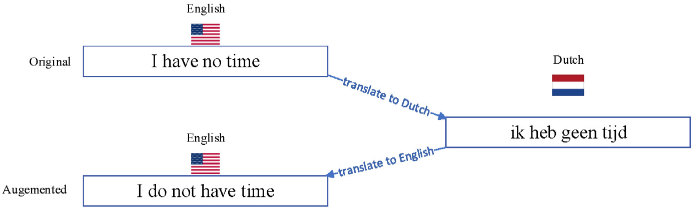
Distribution of Text Length by Sexist vs, Non-Sexist Content



Top 10 3-grams in Sexist Content

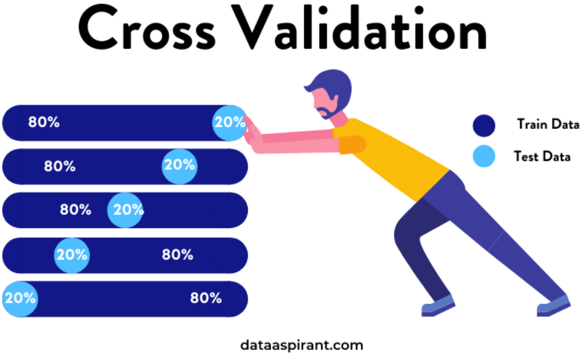
Data Augmentation Techniques

- Synonym augmentation
- Word swapping, insertion, substitution, deletion
- Introducing spelling variations
- Back translation techniques: English ¡-¿ Dutch
- ~~Paraphrasing by free generative AI models (such as GPT-2)~~



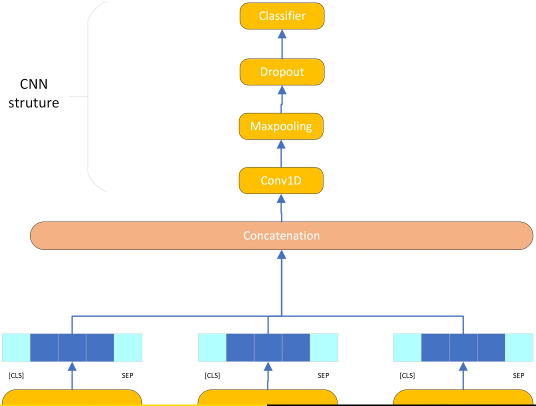
Cross-Validation Data Preparation

- Utilize `StratifiedKFold`
- Address class imbalance:
  - `RandomOverSampler`
  - `SMOTE`
- Compute class weights for balanced training.



Ensemble Model Design

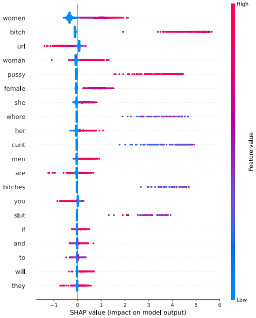
- Build and fine-tune sequence classification BERT model using:
  - *bert-base-multilingual-cased*
  - *xlm-roberta-base*
  - *distilbert-base-multilingual-cased.*
- CNN structure with a Conv1D and a Classification layer.
- Adding Explainablity with `SHAP` and `Human ranking`.

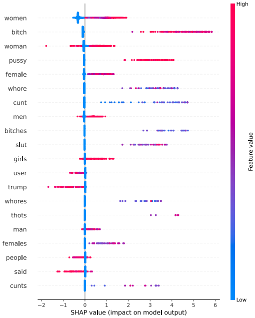


Universiteit Utrecht

Integration of Human Rationalization

- Combines SHAP insights with human analysis.
- Re-ranking of tokens based on their impact on sexism classification.
- Use of a 0 to 1 scale to blend human judgment with algorithmic insights.
- Pre and post-human rationalization comparison:

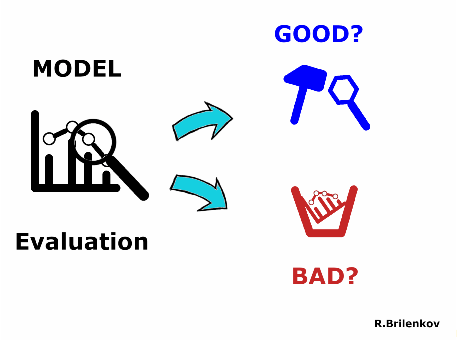




Universiteit Utrecht

Model Training & Optimization

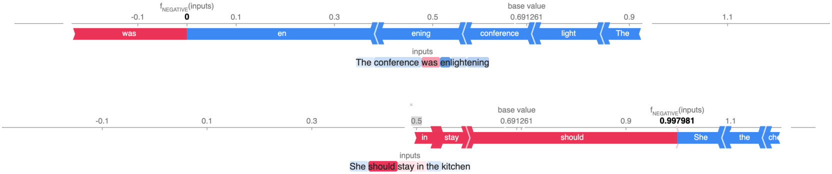
- Used Adam Optimizer with a learning rate of 3×10^{-5}
- Incorporated a 200-step warm-up and early stopping to prevent overfitting.
- Mixed precision training for efficiency.
- Tokenization limited to 512 tokens.
- Hyperparameters determined by random search and Keras Tuner.
- Learning rate followed a cosine decay schedule.



Universiteit Utrecht

Experimentation & Results

- Accurate non-sexist classification in both scenarios.
- Detected subtle sexism effectively with human rationalization.
- Model closely mirrors human logic in detecting sexism.



Evaluation on Test Data

- Evaluated on tasks A, B, and C before and after human rationalizations.
- After rationalization, all tasks showed a decline in performance metrics.
- Task-specific sensitivity to human rationalizations observed.

heatmap (1).pdf

Universiteit Utrecht

Thank You!

Thank you for your attention!

For further questions or details, please contact:

h.mohammadi@uu.nl

https://hadimohammadi.info/

Universiteit Utrecht