Towards Explainable Sexism Detection in Social Media: An Ensemble Approach with Human Rationalizations

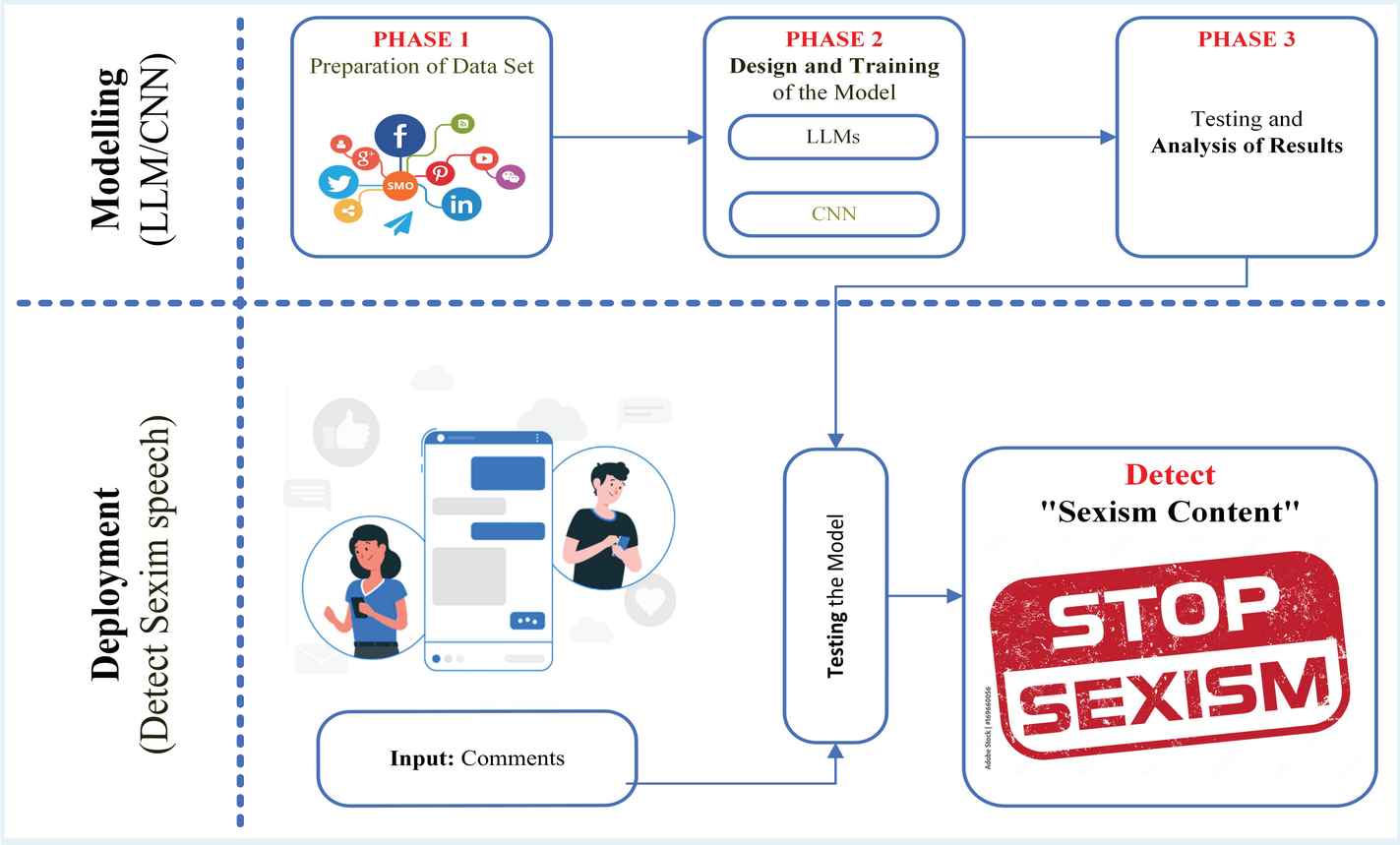
Hadi Mohammadi[1,*], Ayoub Bagheri[1], and Anastasia Giachanou[1].

1. Department of Methodology and Statistics, Utrecht University, Utrecht, the Netherlands

Introduction

Sexist speech is a pervasive issue on online social media platforms. Despite the ongoing development of increasingly sophisticated models for detecting sexist speech, the focus on the explainability and interpretability of these models remains limited. Automated tools can significantly aid in identifying sexism on a large scale. However, binary detection often overlooks the multifaceted nature of sexist content and fails to provide clear reasoning behind the classification of content as sexist. To address this issue, we present an ensemble model that is based on diverse pre-trained models. Our composite model leverages the power of its constituents by extracting relational information from the text by using different pre-trained models (*BERT, XLMRoBERTa, DistilBERT*) and passing it through a distinct Convolutional Neural Network (*CNN*) architecture. The design facilitates the integration of multiple techniques to improve the model's robustness and generalizability. To foster transparency, we apply explainable artificial intelligence techniques to discern the influence of individual tokens and various model components on the decision-making process. Moreover, we incorporate a feedback loop that involves human validation, where humans review and validate the model's predictions and explanations. This iterative approach enhances the model's rationality and alignment with human cognition. To continually evaluate the model's performance, we utilize explainability metrics. The result is a more reliable and interpretable model that aligns with human evaluation and contributes a comprehensive dataset in the field of sexism detection. This dataset, enriched with annotation and interpretability features, offers valuable insights into decision-making processes and serves as a robust foundation.

Methods



This task aims to detect and explain online sexist content through three hierarchical classification sub-tasks.

Task A: A binary classification task to determine if a post is sexist or not.

Task B: A multi-class classification task to further classify sexist posts into four categories: threats, derogation, animosity, and prejudiced discussions.

Task C: Another multi-class classification task that delves deeper into the classification of sexist posts into vectors of sexism. These vectors provide more specific categories under threats, derogation, etc.

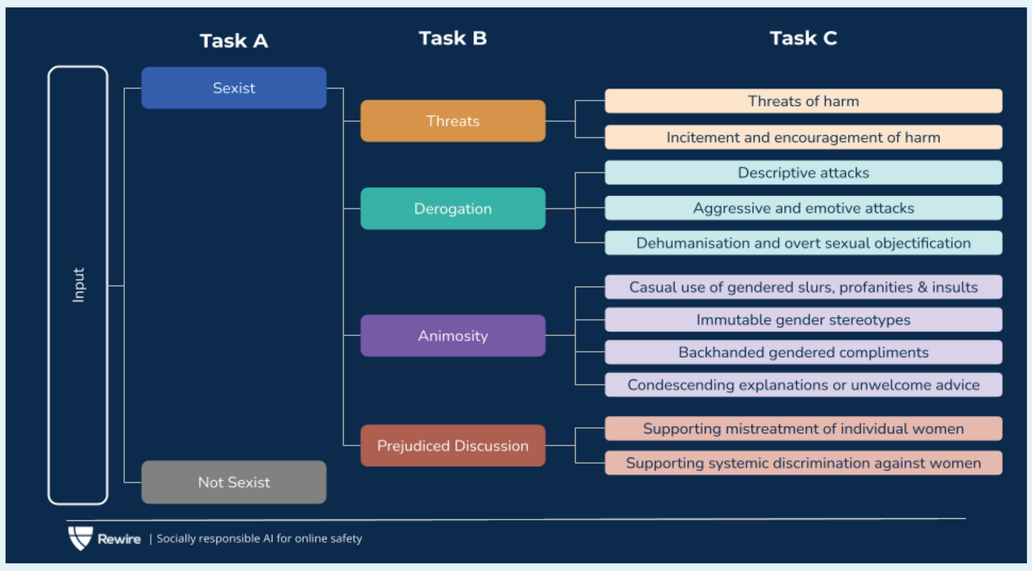


Figure 1: Model Architecture Visualization (Left) and structure of Dataset (Right) [1]

Results

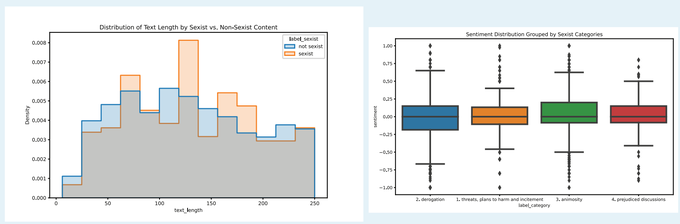
Research Contributions and Results:

- **Ensemble Model Design:**
  - Developed a custom model combining various *BERT* versions with *CNN*.
  - Captures relationships between pretrained model outputs.
  - Enhances robustness, especially for multilingual challenges.
- **Explainability and Human Rationalizations:**
  - *Used SHAP for result transparency.*
  - *Included human judgment to adjust token impact on outcomes.*
  - *Achieved model decisions aligned with human reasoning.*
- **Future Research Direction:**
  - Plan for in-depth analysis of model components.
  - Objective: Improve explainability of large language models.
  - Aim for models that excel across datasets and align with human thought.

Table 1: Performance in different tasks

Performance Before Applying Human Rationalizations				
Task	**Accuracy**	**Precision**	**Recall**	**F1-Score**
A	0.82	0.67	0.73	0.68
B	0.65	0.42	0.65	0.49
C	0.62	0.53	0.66	0.55
Performance After Applying Human Rationalizations				
Task	**Accuracy**	**Precision**	**Recall**	**F1-Score**
A	0.78	0.62	0.67	0.64
B	0.58	0.33	0.58	0.42
C	0.60	0.48	0.60	0.52

A) Text Length Distribution B) Text Distribution of each Class



C) Distribution of each Class D) Top 3 Grams in Sexist Content

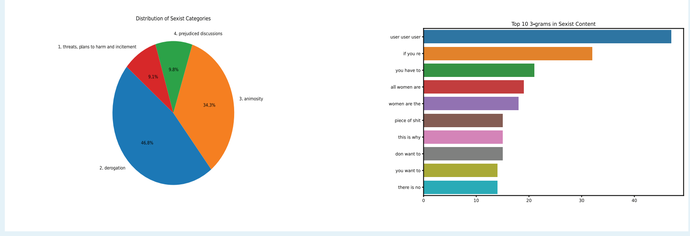
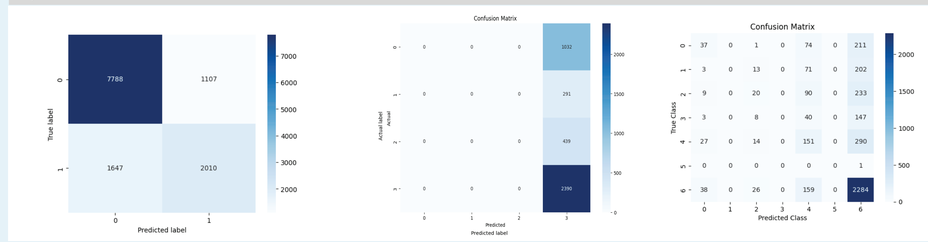


Figure 2: Exploratory Data Analysis for Different Sexist Class Content

A) Confusion Matrix for different tasks



B) Performance metrics during training

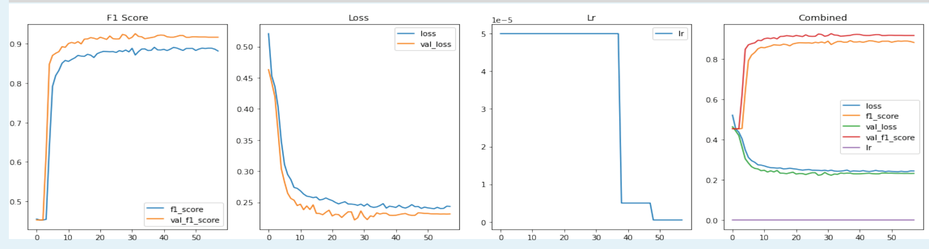


Figure 3: Performance metrics for the model on the different tasks. A) Confusion Matrix in different tasks B) Performance during training

A) Tokens with most impact on the class before applying Human Rationalization (left) and After (right) B) Examples of effects of each tokens on the class in two sentence

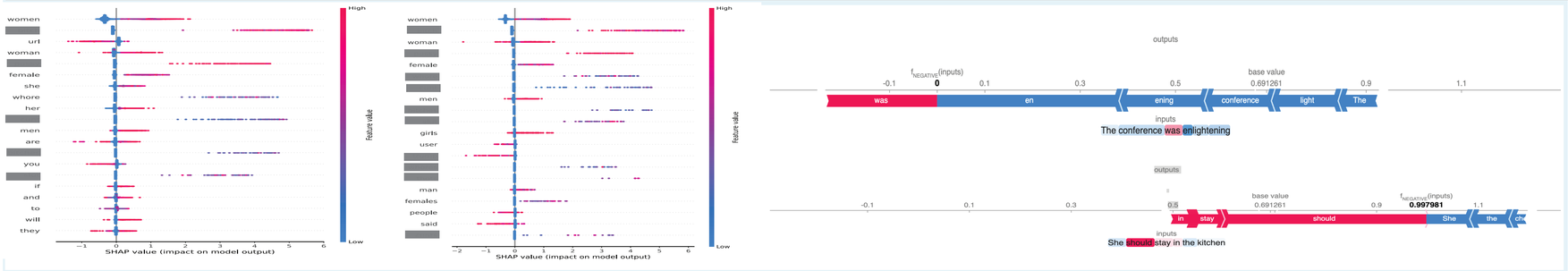


Figure 4: The Impact of Adding Human Rationalization to Explanations to Find out the Most Effective Factors (Left) and Two Examples (Right)

Conclusion

We combined pre-trained language models with a CNN to detect sexism in multilingual social media. While adding human insights slightly reduced accuracy, it made our model's decisions more trustworthy and human-like. Using the SHAP library, we pinpointed which parts of the text influenced the model's choices. Our goal moving forward is to create AI models that both perform well and think more like humans.

*Corresponding Author: h.mohammadi@uu.nl

Universiteit Utrecht

[1] the dataset utilized is specifically associated with SemEval 2023 - Task 10 - Explainable Detection of Online Sexism (EDOS) competition.

Rewire