"10 "11
"00 "01 "02 "03 "04 "05 "06 "07 "08 "09
"0A
"01 "0A

Hate Speech Detection

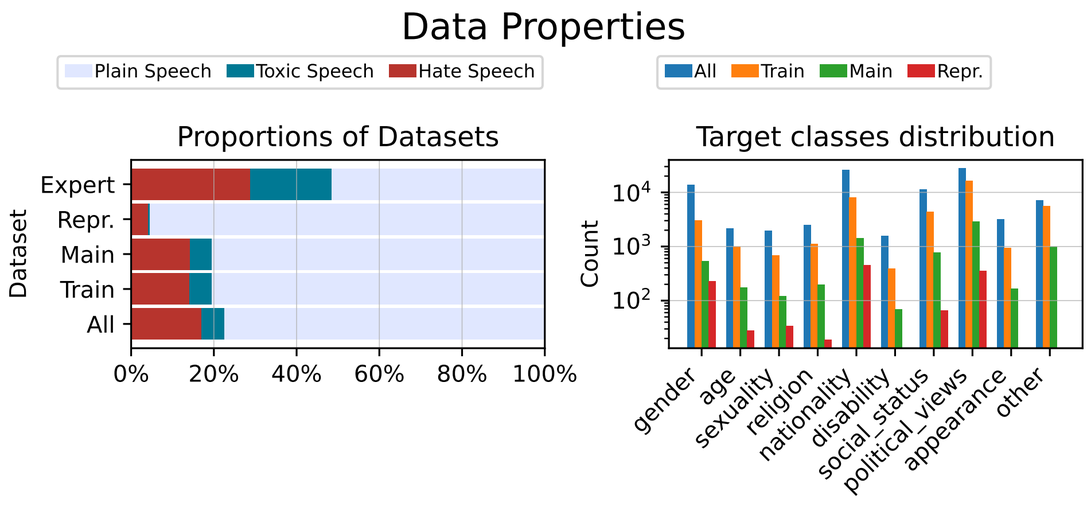
Olga Ovcharenko[1] Evžen Wybitul[1] [1]D-INFK, ETH Zürich

1 Introduction

The aim of our project is to automatically detect toxic speech and hate speech in comments from German-speaking forums, newspapers, and social media. We classify **Toxic Speech** and **Hate Speech**, and also detect the specific **Hate Speech Target Groups**.

2 Data

The dataset of 422,000 (mostly German and French) comments was collected by the Immigration Policy Lab and annotated by student research assistants. During the labeling, toxic comments were preferentially selected, which means the label distribution in the dataset **does not necessarily mirror the real world**. We only apply light data preprocessing to the comments, like removing mentions and URLs, and replacing emojis with their textual descriptions. We also remove comments with inconsistent labels and deduplicate the comments.



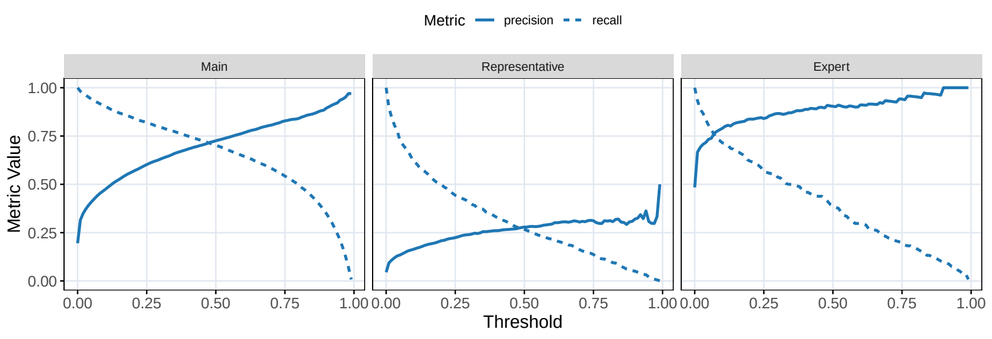
We evaluate on MAIN, a 15% split of the original dataset, REPRESENTATIVE, a set of 25,000 randomly sampled comments that is representative of the real-world distribution, and EXPERT, a set of 500 comments labeled by a three-expert consensus.

3 Toxic Speech Detection

Toxic Speech is any kind of offensive speech.
We set up two baselines and experiment with four different pre-trained models, which we fine-tune using **bottleneck adapters**.

Model	MAIN AUPRC	MAIN F1	REPRESENTATIVE AUPRC	REPRESENTATIVE F1	EXPERT AUPRC	EXPERT F1
Baseline: Constant	0.195	0.326	0.044	0.085	0.484	0.652
Baseline: FastText	0.295	0.523	0.058	0.086	0.533	0.153
mBERT	0.751	0.688	0.171	0.236	0.805	0.504
mDeBERTa-v3	0.766	0.701	0.184	0.254	0.824	0.544
XLM-RoBERTa Base	0.734	0.677	0.150	0.192	0.774	0.414
XLM-RoBERTa Large	**0.779**	**0.714**	**0.205**	**0.281**	**0.846**	0.588 **(0.777)**

We define an **indirect human baseline**, computed from comments labeled by multiple coders. We get an average **F1 score of 0.493 for non-experts, 0.918 for experts, and 0.714 for our best model**. The model is well-calibrated, especially on MAIN.



4 Hate Speech Detection

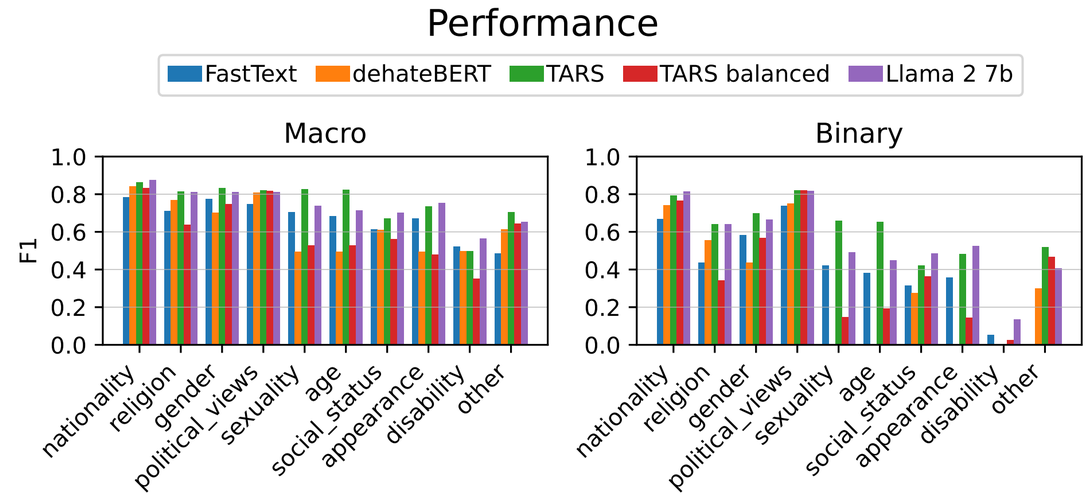
Hate Speech is a certain kind of toxic speech that targets a particular group of people. The task of hate speech detection is more complicated than that of pure toxic speech detection. Below, we compare the performance of our current best model with the model they use in the paper.

Model	MAIN AUPRC	MAIN F1	REPR. AUPRC	REPR. F1	EXPERT AUPRC	EXPERT F1
Baseline: Constant	0.249	0.249	0.041	0.079	0.288	0.447
Baseline: FastText	0.254	0.393 (0.836)	0.056	0.070	0.384	0.139
XLM-RoBERTa Large	0.650	0.612 (0.8891)	0.2083	0.2699	0.604	0.444

In parentheses for MAIN, we list the weighted F1, since that is the metric listed in the paper our project is based on. **The best model in the original paper achieves a weighted F1 of 0.805**, which means our best model outperforms it. The average positive-class expert F1 score is 0.860.

5 Target Class Detection

Hate Speech Target Groups detection aims to identify classes that hate speech is targeting. Our dataset contains ten targets with different distributions. One comment can have multiple targets, but there are mostly up to three targets per comment.
We set up FastText as a baseline and experiment with three different models: **DEHATEBERT**, **TARS FEW-SHOT**, **LLAMA 2 7B**.



To explore the performance differences between classes, we have tried different data settings for the **TARS FEW-SHOT**. Neither balancing classes nor treating them as multi-class improved performance.

6 Discussion & Conclusions

The quality of the dataset is the biggest limiting factor of the project. The data are noisy and heterogeneous, and there is data drift because of iterative data collection and multiple annotators.
Computational resources are another limiting factor. We tried to adhere to the requirements in our experiments, and therefore, our best approach for every task can be run within the constraints CPU and GPU constraints of the challenge givers.
Application of the models will be either direct or with fine-tuning to identify more fine-grained datasets for a field experiment. In the field experiment, individuals who post toxic comments will receive one of various treatments to reduce their inclination to post toxic speech. The end goal is a live-updating dashboard to expose the quality of the online discourse, the amount of hate speech, and its targets.