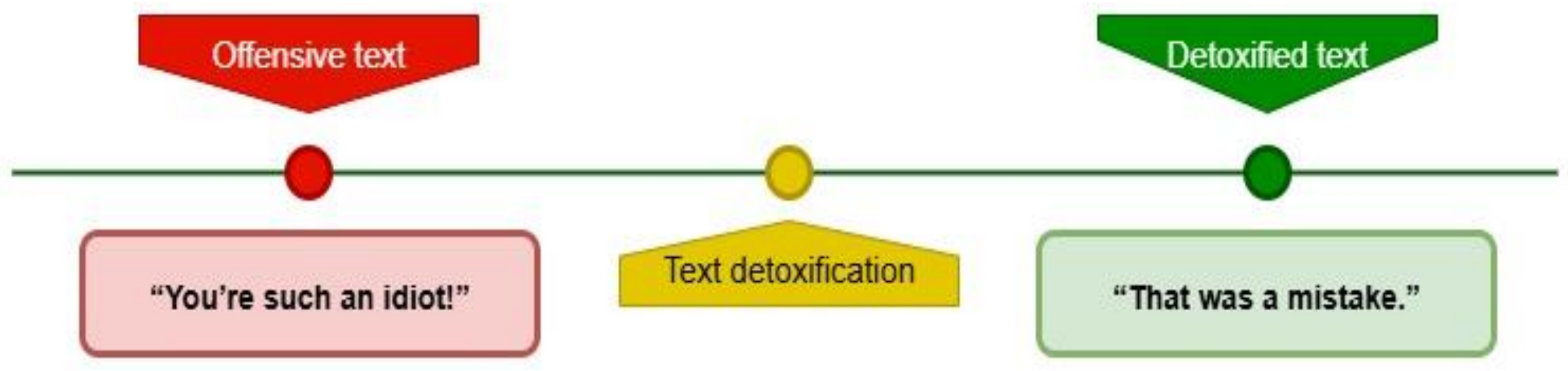
Towards Safer Hebrew Communica A Dataset for Offensive Language Detoxification

Natalia Vanetik, Lior Liberov, Marina Litvak, Chaya Liebeskind

JERUSALEM COLLEGE OF TECHNOLOGY

SHAMOON COLLEGE OF ENGINEERING

Text Detoxification



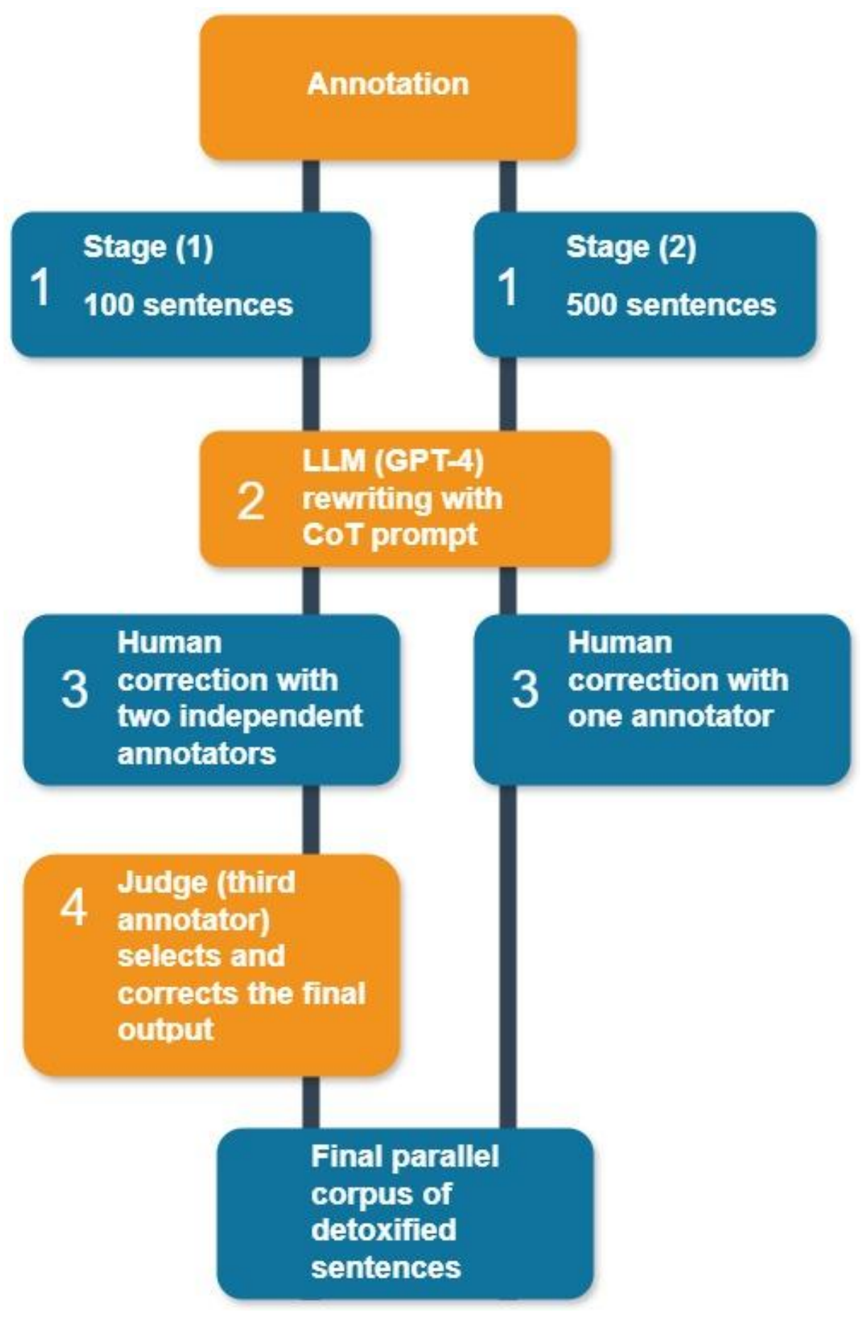
Challenges:

change of meaning, change of intent, excessive or non-sufficient text modification

Data Collection

- Emotionally charged user comments from Rotter.net news forum
- Used web crawler to scrape threads, metadata, and normalize text
- Applied full anonymization (removed usernames, links, PII)
- Few-shot CoT classification using Simplified Offensive Language (SOL) taxonomy
- Restricted dataset to explicit offensive samples to ensure high precision
- Oversampled by ~12% to remove borderline or ambiguous cases

Annotation


Annotation
1 Stage (1) 100 sentences
1 Stage (2) 500 sentences
2 LLM (GPT-4) rewriting with CoT prompt
3 Human correction with two independent annotators
3 Human correction with one annotator
4 Judge (third annotator) selects and corrects the final output
Final parallel corpus of detoxified sentences

Inter-annotator text similarity scores for Stage 1

representation	cosine similarity
heBERT SE	0.888
mlBERT SE	0.937
n-grams	0.649
tf-idf	0.685

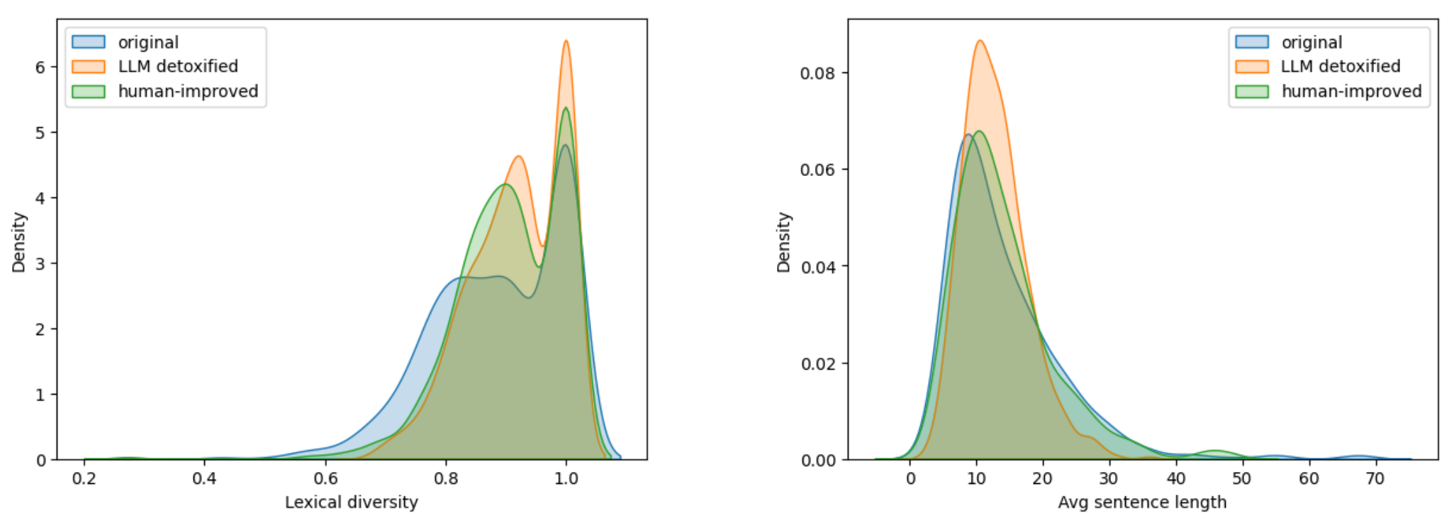
Lexical Diversity and word entropy for Stage 1 + Stage 2

text	MTLD (avg)	word entropy (avg)
original	0.714	3.490
LLM detoxified	0.027	3.523
human-improved	0.171	3.549

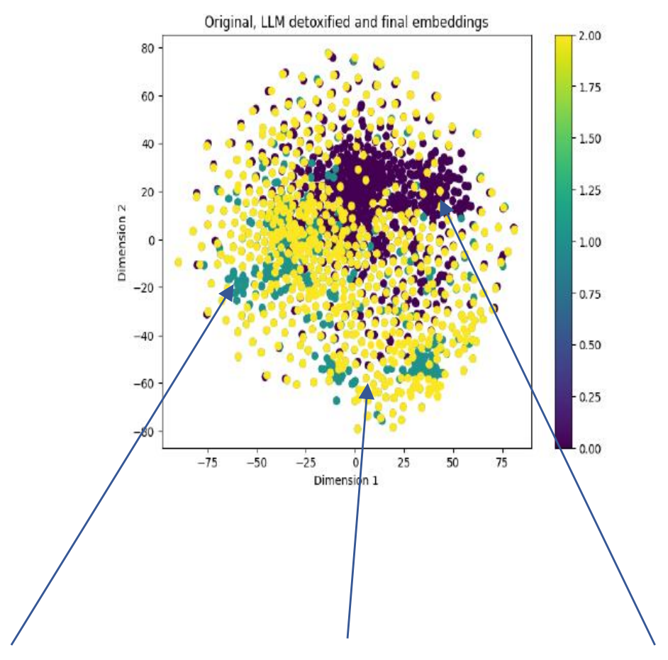
Syntactic and semantic similarity for texts in HeDedox

text comparison	BERTScore	BLEU	ROUGE-1	ROUGE-2	ROUGE-L
original vs. LLM detoxified	0.7373	0.0933	0.0330	0.0028	0.0330
original vs. human-improved	0.7655	0.1327	0.0547	0.0111	0.0547
LLM Detoxified vs. human-improved	0.8799	0.5520	0.0333	0.0033	0.0333

Data Analysis



Sentence embeddings

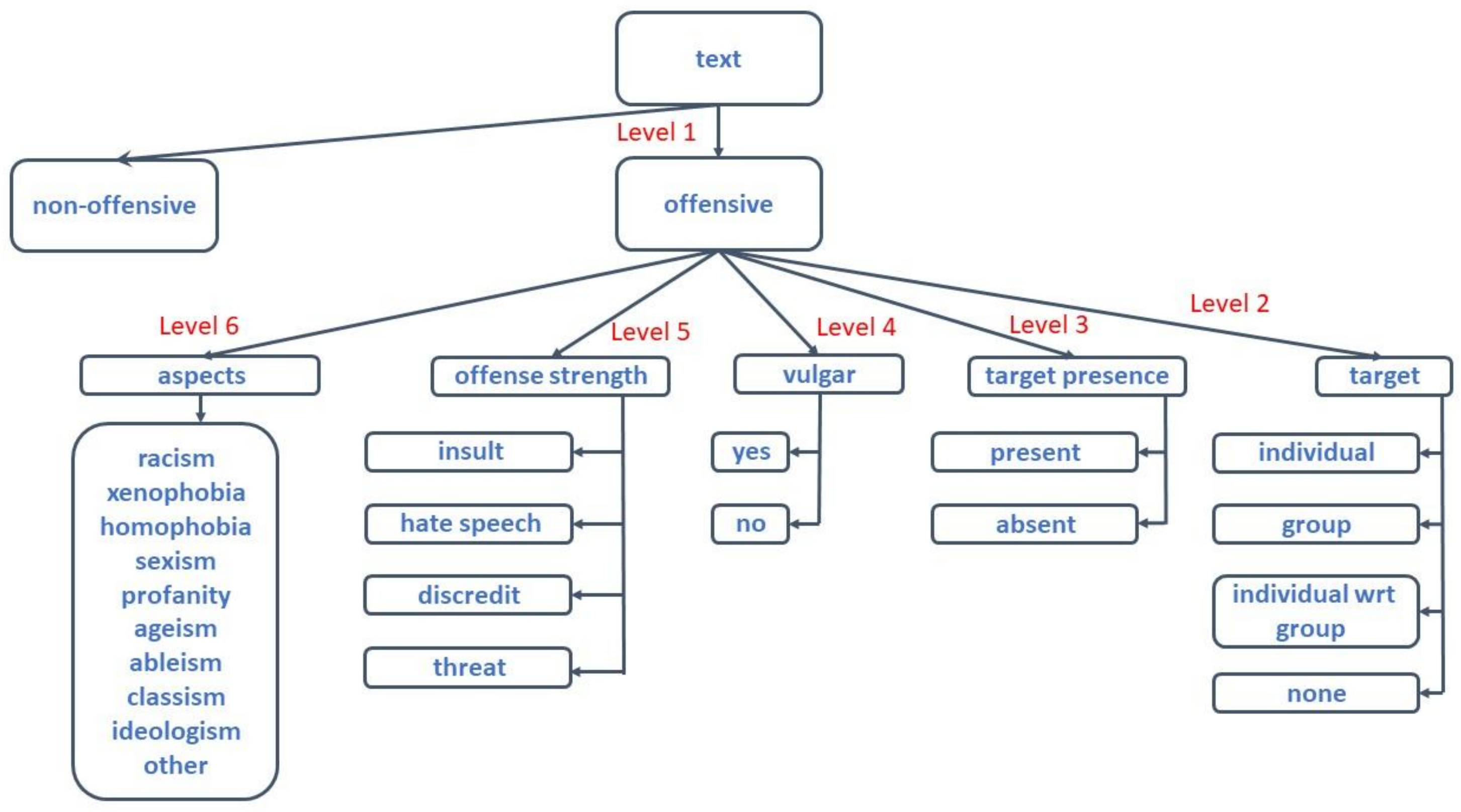


Human- corrected LLM-detoxified Original

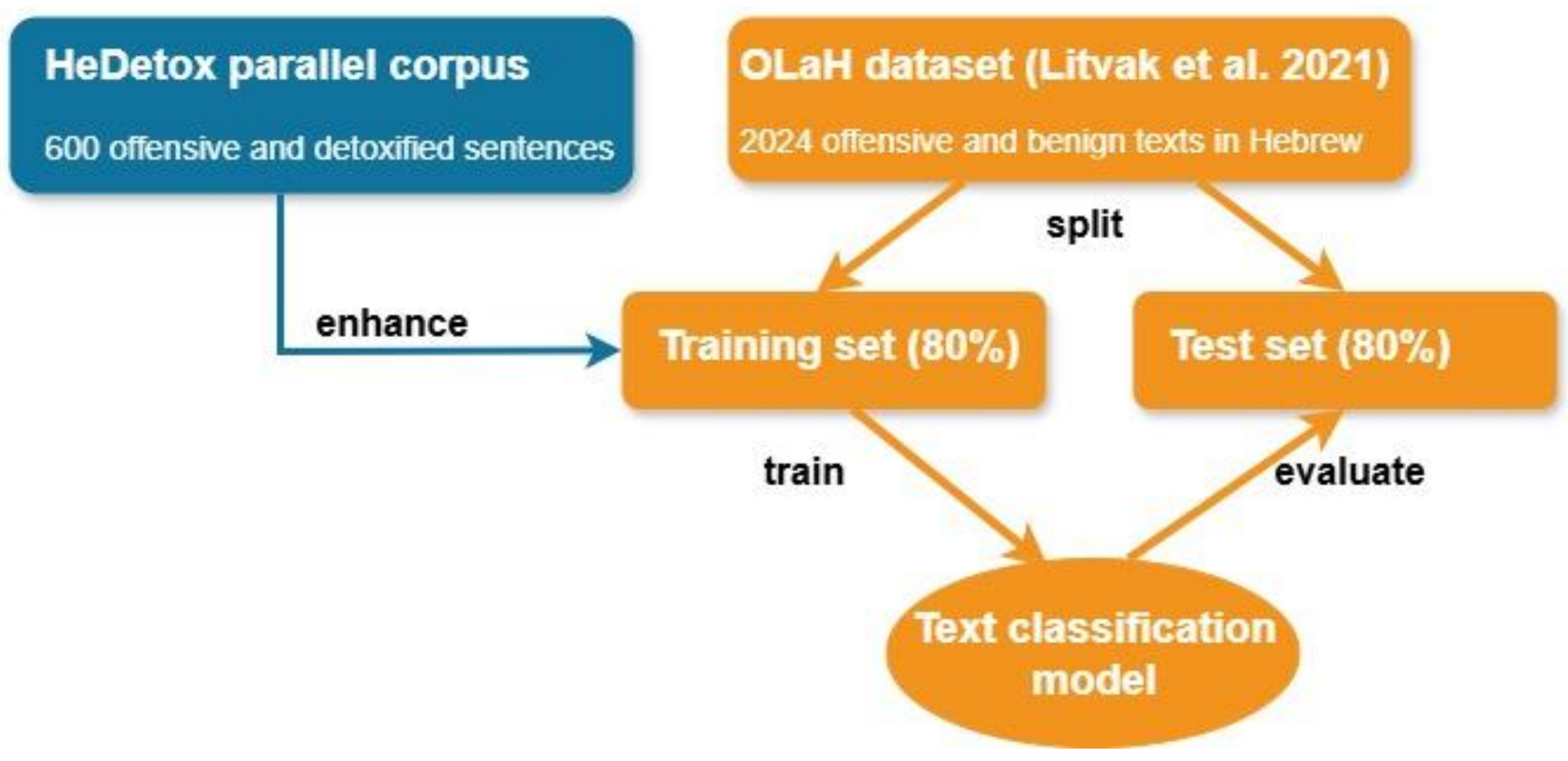
POS Tag	original	LLM detoxified	human-improved
ADJ	585	559	563
ADP	1555	1549	1704
ADV	664	790	842
AUX	136	191	190
CCONJ	337	291	275
DET	924	745	790
INTJ	3	–	–
NOUN	2337	1831	2101
NUM	104	37	67
PROPN	507	196	285
PRON	1122	1048	1116
PUNCT	1061	991	1102
SCONJ	426	475	546
SYM	2	–	–
VERB	1302	1502	1500
X	13	1	2

Part-of-Speech distribution

Simplified Offensive Language taxonomy



Evaluation


HeDetox parallel corpus
600 offensive and detoxified sentences
OLaH dataset (Litvak et al. 2021)
2024 offensive and benign texts in Hebrew
split
enhance
Training set (80%)
Test set (80%)
train
evaluate
Text classification model

Evaluation results

Model	Training Data	Accuracy	F1
mlBERT	OLaH	0.6897	0.5855
heBERT	OLaH	0.7660	0.7003
mlBERT	OLaH+HeDetox	0.7438	0.7029
heBERT	OLaH+HeDetox	0.7685	0.7202

Models

- Multilingual BERT (Devlin et al. 2019)
- Hebrew BERT (Shavit and Singer 2019)

Observation:

Exposure to detoxified rewrites enhances the classifier's ability to generalize beyond surface-level lexical cues

Conclusions

- HeDetox: first Hebrew detoxification dataset
- 600 offensive–detoxified sentence pairs created
- Hybrid LLM output and expert corrections
- Preserves original intent and conversational tone
- Improves lexical diversity and content structure
- Boosts offensive language classification performance
- Currently limited to explicit offenses, small size
- Ethical use ensured; research purposes only

RECENT ADVANCES IN NATURAL LANGUAGE PROCESSING 1989 2025 RANLP CONFERENCES AND EVENTS

HeDetox dataset is freely available at: https://github.com/NataliaVanetik/OffensiveLanguageResearchLab#hedetox

Download at:

Contact us: natalyav@sce.ac.il liorli1@sce.ac.il marinal@sce.ac.il liebchaya@gmail.com