Towards Safer Hebrew Communication: A Dataset for Offensive Language Detoxification

Natalia Vanetik [1] natalyav@sce.ac.il
Lior Liberov [1] liorli1@ac.sce.ac.il
Marina Litvak [1] marinal@sce.ac.il
Chaya Liebeskind [2] liebchaya@gmail.com
[1]Shamoon College of Engineering, Beer-Sheva, Israel
[2]Jerusalem College of Technology, Jerusalem, Israel

Abstract

Text detoxification is the task of transforming offensive or toxic content into a non-offensive form while preserving the original meaning. Despite increasing research interest in detoxification across various languages, no resources or benchmarks exist for Hebrew, a Semitic language with unique morphological, syntactic, and cultural characteristics. This paper introduces HeDetox, the first annotated dataset for text detoxification in Hebrew. HeDetox contains 600 sentence pairs, each consisting of an offensive source text and a non-offensive text rewritten with LLM and human intervention. We present a detailed dataset analysis and evaluation showing that the dataset benefits offensive language detection. HeDetox offers a foundational resource for Hebrew natural language processing, advancing research in offensive language mitigation and controllable text generation.

1 Introduction

Toxic and offensive language in online platforms presents significant challenges for content moderation, user safety, and inclusive communication (Fortuna and Nunes, 2018; Poletto et al., 2021). In Hebrew, detecting and mitigating offensive language is particularly complex, given the language's rich morphology, colloquial variations, and the frequent use of implicit or culturally embedded offensive expressions. Despite growing interest in offensive language detection across languages, Hebrew remains under-resourced in this domain, with only a few publicly available datasets of significant size (Litvak et al., 2021), annotated for offensive language detection only.

This study introduces a high-quality annotated dataset for Hebrew text detoxification, called HeDetox, aimed at supporting the development of systems capable of rewriting offensive or toxic content into non-offensive, semantically faithful alternatives. HeDetox contains 600 sentence pairs, including an original offensive sentence and its corresponding detoxified version.

The annotation process employed a hybrid approach combining LLM-guided rewriting with manual human verification and correction. In particular, we used a few-shot chain-of-thought (CoT) prompt (Wei et al., 2022; Kojima et al., 2022) with the GPT-4o model (OpenAI, 2024) to produce preliminary detoxified versions of offensive sentences, which were then examined, improved, and verified by skilled human annotators who adhered to strict annotation guidelines. To ensure clarity, grammatical accuracy, and cultural appropriateness in the revised language, these standards placed a strong emphasis on maintaining the original sentence's main meaning and intent while eliminating offending parts.

We thoroughly examined the dataset's linguistic and semantic characteristics and assessed its influence on offensive language identification performance to determine its usefulness for natural language processing (NLP) applications. Using baseline text classification models trained on offensive language detection, we demonstrate that integrating the detoxified dataset improves classification accuracy.

By providing the first publicly available dataset for Hebrew text detoxification, our work addresses a critical resource gap in Hebrew NLP. It contributes to broader efforts in offensive language detection, controlled text rewriting, and content moderation. The HeDetox dataset supports the development and testing of models that can both detect and reduce offensive language in Hebrew, helping to create a safer and more inclusive online environment (Dementieva et al., 2025, 2024b).

2 Related Work

Multiple studies have focused on automatic detection of offensive language, producing a range of annotated datasets and approaches (Fortuna and

Proceedings of Recent Advances in Natural Language Processing, pages 1289–1298
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-149

Nunes, 2018; Poletto et al., 2021).

Hate Speech Corpus and OLID (Zampieri et al., 2019a,b) were early standards for offensive language detection that only addressed the English language. Later datasets such as TRAC (Kumar et al., 2018) and HASOC (Mandl et al., 2019) extended coverage to several languages, including Hindi and German. Later, more language-specific datasets were created, including the Multilingual Hate Speech Corpus (Ousidhoum et al., 2019) and HaSpeeDe (Bosco et al., 2018) for Italian and GermEval (Wiegand et al., 2018) for German.

Parallel detoxification datasets have become essential for training and evaluating algorithms that transform offensive texts into neutral or non-offensive forms. The ParaDetox dataset, a crowd-sourced English corpus that includes non-toxic paraphrases for more than 10,000 English toxic statements, was introduced by Logacheva et al. (2022). Atwell et al. (2022) released APPADIA—the parallel corpus of offensive Reddit comments annotated by an expert sociolinguist, and the first discourse-aware style-transfer models that can effectively reduce offensiveness while preserving the meaning of the original text. However, both works explored approaches for parallel text detoxification corpora collection only in a monolingual setup.

Later, MultiParaDetox (Dementieva et al., 2024a, 2025) expanded the ParaDetox pipeline to multiple languages. The final dataset covers nine languages, containing 1000 samples per language, which are split into 400 training and 600 test instances, utilized for shared task evaluations (Dementieva et al., 2024b).

To address the scarcity of data for training and evaluation of the detoxification models, SynthDetoxM (Moskovskiy et al., 2025) introduced a synthetic parallel detoxification corpus containing 16,000 sentence pairs across German, French, Spanish, and Russian. These resources have significantly contributed to the advancement of detoxification models, particularly in multilingual contexts.

However, for Hebrew, these resources are very limited. The publicly available datasets for offensive language detection in Hebrew were introduced in very few works. Litvak et al. (2022) expanded OLaH (Litvak et al., 2021) and the Liebeskind (Liebeskind and Liebeskind, 2018) datasets. After merging both datasets and completing missing annotations, the final dataset contains 5,217 annotated comments. Hamad et al. (2023) collected 15,881 tweets, each labeled with one or more of five classes (abusive, hate, violence, pornographic, or non-offensive) by Arabic-Hebrew bilingual speakers. Liebeskind et al. (2023, 2024) introduced a taxonomy for categorizing offensive language in Hebrew, following (Lewandowska-Tomaszczyk et al., 2023b). They also collected a dataset that they used to annotate documents based on the proposed taxonomy and analyzed its usability for classifying offensive content using machine learning. Despite a large amount of collected tweets (around 8M) from all nine categories, only 450 samples (50 per category) were labeled by two independent annotators based on the introduced taxonomy.

To the best of our knowledge, no prior dataset contains paired offensive and detoxified texts in Hebrew. This dataset is to be adopted at PAN detoxification task (`https://pan.webis.de/clef25/pan25-web/text-detoxification.html`) and it will be made publically available for the community once the competition concludes. Our work addresses this gap by constructing a parallel corpus of 600 Hebrew sentences containing offensive content and their manually detoxified rewritings. We combine LLM-assisted annotation with human correction to ensure accuracy and consistency. In addition to dataset creation, we conduct an in-depth analysis of linguistic patterns in detoxified outputs and assess how fine-tuning baseline models on the corpus improves offensive language detection. The HeDetox dataset is a novel resource for both offensive language detection and detoxification in Hebrew, which is a low-resource language.

3 The HeDetox Dataset

3.1 Data Collection

We collected user comments from a highly active online news forum[1]. These emotionally charged responses to current events served as a rich source for detecting offensive and toxic language. We employed a standard web crawling pipeline to scrape entire discussion threads, extract metadata (e.g., timestamps, post IDs), and normalize the comment text. All collected data underwent a comprehensive anonymization process, whereby any personally identifiable information—including usernames, mentions, and embedded links—was removed or obfuscated to protect user privacy.

[1]`https://rotter.net/forum/listforum.php`

To identify content relevant for detoxification, we applied a few-shot classification approach based on chain-of-thought (CoT) prompting (Rouzegar and Makrehchi, 2024), using definitions derived from the *Simplified Offensive Language (SOL) Taxonomy* proposed by Lewandowska-Tomaszczyk et al. (2023a). This taxonomy offers a linguistically grounded yet computationally feasible framework for detecting offensive language. It introduces a stepwise structure that begins by assessing whether a comment is offensive, and proceeds to categorize its target (individual, group, or a group represented through an individual), as well as its level of vulgarity. Offenses are then classified into four primary types—*insult*, *hate speech*, *discredit*, and *threat*—each distinguished by the nature of the attack (personal vs. ideological), use of stereotypes, or intent to harm. In addition, the taxonomy encodes various *aspects* of offense, such as *racism*, *sexism*, *classism*, *ableism*, and *ideologism*, offering fine-grained interpretability of the offensive content. The model further accounts for implicit linguistic strategies—including *metaphor*, *irony*, *rhetorical questions*, and *exaggeration*—as vehicles for more covert or veiled expressions of hostility.

Despite this rich taxonomy, our manual inspection revealed that the implicit classifier tended to over-generate, frequently labeling figurative or emotionally expressive comments as implicitly offensive, even when no harmful intent or target was present. To maintain high precision and ensure the relevance of examples to detoxification tasks, we therefore restricted our dataset to samples classified as explicitly offensive, excluding implicitly offensive examples due to their lower reliability and semantic ambiguity.

The classifier annotated each comment as *explicitly offensive*, *implicitly offensive*, or *non-offensive*. To reduce uncertainty and improve overall dataset quality, we oversampled by approximately 12% beyond the desired dataset size. This allowed us to eliminate borderline or ambiguous cases, such as those with unclear targets, sarcastic tone without evident hostility, or marginal vulgarity, and retain only those samples that fit our criteria for explicit offensiveness.

The full few-shot prompt used for classification, including example annotations and taxonomy-based reasoning steps, can be seen on our GitHub account (Vanetik et al., 2025).

3.2 Detoxification

3.2.1 Few-Shot Chain-of-Thought Prompting

Dementieva et al. (2025) introduced Chain-of-Thought (CoT) prompting for detoxification, demonstrating that breaking down the detoxification process into intermediate reasoning steps improves the quality and fidelity of rewritten texts.

In our work, we adapted the A1 and A3 prompts from this work (available at (Dementieva et al., 2025; Vanetik et al., 2025) and successfully applied to other languages) for the detoxification of Hebrew-language texts with the GPT-4o model (OpenAI, 2024), which we selected for its strong performance in few-shot reasoning. We extended these prompts by adding more detailed instructions and multiple in-language examples tailored to the linguistic and cultural characteristics of Hebrew. Specifically, we designed a custom prompt that instructed the model to analyze provided Hebrew sentences for elements of toxicity using a predefined list of keywords and to output detoxified sentences in a structured format. The prompt emphasized preserving the original meaning, tone, and intent while removing toxic or offensive expressions without introducing unsolicited advice or commentary. We also added two negative examples in Hebrew where the modified sentences contain advice or interpretation not present in the original text. Our prompt is shown in Figure 1 (Hebrew sentences are accompanied by English translations for clarity).

We employed this prompt in a few-shot CoT setting, providing the LLM with two in-context examples before prompting it with new toxic sentences for detoxification. This structured approach ensured that the output remained faithful to the original message while systematically neutralizing offensive expressions. The combination of the CoT strategy with a domain-specific, language-tailored prompt allowed us to perform effective detoxification on Hebrew texts, despite limited parallel detoxification datasets for this language.

3.2.2 Manual Correction of LLM-Detoxified Sentences

To evaluate the outputs of large language models (LLMs) for offensive language detoxification, we conducted a manual analysis aimed at identifying systematic errors and improving model performance. Each detoxified sentence was reviewed and categorized by two interdependent judges accord-

Please analyze the provided sentences in Hebrew using the structure below to identify elements of toxicity and suggest improvements when I tell you. Use words from the keywords list (can be more than one word!):
keywords = [“Neutral”, “Informative”, “Casual”, “Assertive”, “Dismissive”, “Condescending”, “Friendly”, “Commanding”, “Instructive”,“Derogatory”, “Confrontational”, “Insulting”, “Vulgar”, “Formal”, “Informal”, “Offensive”, “Technical”, “Playful”, “Positive”, “Frustration”, “Analytical”, “Professional”, “Hostile”, “Hatred”, “Helpful”, “Angry”, “Friendly”, “Arrogant”]
Analysis Structure (do not use [] and “” in your answer, and do not suggest improvement!):
{
“Sentence”: “sentence”,
“Toxicity Level”: “Specify here (Low/Medium/High)”,
“Tone”: “the overall tone of the sentence – choose from keywords”,
“Language”: “Language style – choose from keywords”,
“Implied Sentiment”: “the overall sentiment- choose from keywords”,
“Context”: “Brief description of how context contributes to toxicity”,
“Negative Connotations”: “List specific negative words/phrases here”,
“Intent”: “Describe the perceived intent behind the sentence.”
}
Please detoxify the provided sentence using the structure below without changing the real meaning of the sentence.
Analysis structure (do not use ” and [] and ”” in your answer, and do not suggest improvement!):
{
“Sentence”: “sentence”,
“fixed sentence”: “the non-toxic sentence without changing the meaning”
},
Example 1: {
“Sentence”: נו לא ברור ! מתי הם מקבלים עתירה מימין !! ימח שמם שונאי יהודים !! משתין על בג”ץ ועליך !!
“Well, it’s not clear! When do they receive a petition from the right!! May the names of the haters of Jews be wiped out!! They are urinating on the Supreme Court and on you!!”
“fixed sentence”: נו, לא ברור מתי הם מקבלים עתירה מימין? אני ממש לא מרוצה מהבג”ץ ומהמצב.
“Well, it’s not clear when they receive a petition from the right? I am really not satisfied with the Supreme Court and the situation.”
}
Example 2: {
“Sentence”: בוגד ומניאק. מזל שלו שהוא לא גר ברוסיה פוטין היה מכניס את הבוגד לכלא במקרה התוב.
“Traitor and jerk. He’s lucky he doesn’t live in Russia; Putin would have put the traitor in prison in the best case.”
“fixed sentence”: אזה מזל שהו לא גר ברוסיה, שם פוטין היה יכול להכניס אתו לכלא במקרה התוב.
“What luck that he doesn’t live in Russia; there Putin could have put him in prison in the best case.”
}

When rephrasing offensive sentences into non-offensive versions, do not add advice, commentary, or suggestions that are not present in the original text. Only rephrase the text to remove offensive language while strictly preserving the original meaning, tone (as much as possible), and intent. Do not ”soften” the message into constructive criticism unless the original explicitly included advice.

Example 1: {
Original: אתה סתם רעש
“You are just noise.”
Incorrect Rephrase: לפעימים אתה לא תורם לדיונים בצורה חיובית, כדאי לשתף יותר במחשבות מועילות
“Sometimes you don’t contribute to discussions in a positive way; it’s worth sharing more helpful thoughts.”
Problem: Added advice not present in the original.
}
Example 2: {
Original: אפליו לצפות בצבע מתייבש יתר מעניין ממך
“Even watching paint dry is more interesting than you.”
Incorrect Rephrase: לפעמים אתה לא מצליח לעניין בשיח, כדאי לשתף בתוכן יותר מעניין ומועיל
“Sometimes you fail to be interesting in the conversation; it’s worth sharing more interesting and useful content.”
Problem: Added advice and interpretation not present in the original.
Advice: Always focus on neutralizing the offensive elements without introducing new ideas or interpretations.
}
Sentences to analyze: {`sentences`} }

Figure 1. Prompt used for detoxification of Hebrew texts.

ing to five predefined error types. The objective was to ensure that offensive content is removed while preserving the original meaning, tone, and communicative intent.

We defined five main error categories observed in the detoxification outputs and provided one ex-

1. Over-softening (22%): The system transforms the offensive sentence into an overly polite or mild statement, beyond what is necessary for toxicity removal, thus weakening the intended message.
Original: אתה אפס מאופס שלא מבין כלום בכלכלה ("You're a complete idiot who knows nothing about economics.")
Incorrect: "It's best to discuss economic topics in a positive and respectful manner." Preferred: "You don't understand economics."
2. Omission of key content (17%): The system removes not only toxic elements but also critical non-offensive content, resulting in semantic loss or excessive softening.
Original:
השרים האלה הם נוכלים והם רק דואגים לכיס שלהם ולא לעם
("These ministers are crooks; they only care about their own pockets, not the people.")
Incorrect: "The ministers care about their own pockets." Preferred: "These ministers care about their own pockets, not the people."
3. Addition of unoriginal content: (14%) The system introduces new information absent from the original sentence, thereby altering the intended meaning.
Original: האנשים האלה מטומטמים לחלוטין ("Those people are complete idiots.")
Incorrect: "Can we discuss the different aspects of this issue rationally?" Preferred: "Everyone who thinks that is completely wrong."
4. Unnecessary synonym replacement: (8%) The system substitutes words with less precise synonyms, unintentionally modifying tone or intensity.
Original: הוא היה גזען מגעיל שתמיד דיבר שטוית ("He was a disgusting racist who always talked nonsense.")
Incorrect: "He was an unpleasant person who always said incorrect things." Preferred: "He had racist views and always said incorrect things."
5. Retention of offensive words (2%): The system fails to remove toxic terms or retains them without contextual justification. In some cases, it may be acceptable to keep words with low toxicity if no insult is intended.
Original: האם כל השופטים הם מושחתים ונוכלימ ("Are all the judges corrupt crooks?")
Incorrect: "Are all the politicians terrible liars?" Preferred: "Are all the politicians untruthful?"

Figure 2. Error categories.

ample for each (see Figure 2). In the figure, the percentage of identified cases for each error type is shown in parentheses next to the error category name. Each output was manually assigned to one of these error categories or marked as correct. In total, 100 sentences were evaluated. For each erroneous case, a revised sentence was proposed. 38% of the sentences contained errors. This process aimed to document recurring patterns and identify system-level weaknesses to inform model refinement. Additionally, annotators were instructed to avoid heavy paraphrasing – substantial rewriting that alters sentence structure, vocabulary, or idea ordering – since such rewriting risks deviating from the speaker's authentic expression. The goal was to apply minimal edits that detoxify the sentence while preserving its semantic and pragmatic content. Our annotation process consisted of two phases to ensure high-quality corrections of the detoxified sentences produced by the LLM. All our annotators and the judge are native Hebrew speakers having at least a BSc academic degree. Two different annotators separately assessed the LLM-generated results during the initial annotation step. A revised version of the detoxified phrases was supplied by each annotator. A judge examined the adjustments after the annotators' assessments to make sure they were consistent and compliant with the rules. This stage was designed to gather different viewpoints on detoxifying offensive language and offer a more thorough examination of the possible modifications. We report the average cosine similarity between annotators' final corrections with various text representations. At this stage, 41 sentences out of 100 were identical, and 59 were different. We evaluated semantic similarity with sentence embeddings produced by heBERT (Chriqui and Yahav, 2022), a transformer-based language model pretrained on Hebrew corpora, and mlBERT (multilingual BERT) (Devlin et al., 2019) in Table 1. Additionally, we included bag-of-words models using n-grams and tf-idf features for comparison. The results show that both heBERT and mlBERT achieve high inter-annotator similarity, with mlBERT yielding the highest score (0.937), indicating strong semantic alignment despite syntactic variability. In contrast, the traditional vector-based representations (n-grams and tf-idf) exhibit lower similarity, reflecting lower syntactic similarity. This demonstrated the subjectivity of wording fixes and the flexibility of natural language, even if it did not pose an issue for maintaining semantic substance. In the second phase, we refined our annotation procedure to reduce un-

Table 1. Average inter-annotator cosine similarity for final sentence corrections.

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

predictability. Our methodology employed a two-phase human review: an initial annotation by an expert, followed by a thorough review and finalization by a dedicated judge/corrector. This sequential process ensured rigorous application of our five predefined error categories and precise formulation of revised sentences for erroneous cases. An annotator was responsible for correcting the LLM outputs while adhering to the established rules from the previous step. We used an LLM to assist with preliminary error identification and pre-annotation, which significantly shortened the overall process and allowed our human team to focus their expertise on the most challenging cases across 500 evaluated sentences. As a corrector, the judge examined the suggested adjustments, ensuring they maintained the sentence's original meaning and tone without excessive changes, ultimately providing the final validation for the dataset.

3.3 Data Analysis

To examine the linguistic characteristics of the HeDetox dataset, we computed lexical diversity (measured as the proportion of unique tokens relative to the total number of tokens), sentence length, and part-of-speech (POS) distributions across the original sentences, the LLM-detoxified texts, and their human-refined versions for all 600 texts in it.

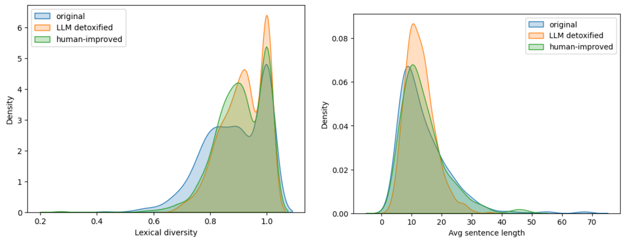
Figure 3 shows that both the LLM-detoxified and human-improved texts in HeDetox demonstrate increased lexical diversity compared to the original, with the LLM output exhibiting the highest mean value. This trend suggests that detoxification processes introduce more varied vocabulary, potentially as a result of rephrasing or paraphrasing strategies. Prior work has shown that LLM-generated Hebrew text is prone to morphological and syntactic errors due to the language's rich inflectional structure and ambiguity (Paz-Argaman et al., 2024; Gueta et al., 2023; Eyal et al., 2022). However, the average sentence length reveals a different dynamic. While human-improved texts maintain sentence lengths comparable to the original, the LLM-detoxified outputs are consistently shorter, with reduced variance. This phenomenon may reflect simplification strategies employed by the model, possibly to decrease the offensiveness of the text. Table 2 demonstrates notable shifts

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

Table 2. Part-of-speech (POS) tag distribution across all texts in HeDetox.

across the different text versions. Both the LLM-detoxified and human-improved texts exhibit increased use of content-bearing categories such as verbs, adverbs, and nouns, indicating a tendency toward more elaborated or descriptive constructions during detoxification. In contrast, a marked reduction in proper nouns is observed, most prominently in the LLM output, suggesting an implicit strategy of depersonalization, likely aimed at reducing the specificity or offensiveness of named references.

In addition to linguistic analysis, we evaluated the semantic similarity between the original sentences and their detoxified counterparts using BERTScore (Zhang et al., 2020), ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), and cosine similarity of sentence embeddings computed with mlBERT (Devlin et al., 2019) and heBERT (Chriqui and Yahav, 2022) models. We computed similarity scores for both the automatically detoxified and human-refined texts, allowing us to assess how closely each version preserved the meaning of the original. Table 3 shows that human-improved texts consistently score higher in BERTScore F1, BLEU, and ROUGE metrics when compared to the original versions, suggesting stronger semantic preservation and lexical cohesion. The similarity between LLM-detoxified and human-improved outputs is particularly notable. This pair achieves the highest BERTScore and BLEU scores among all comparisons, indicating a high degree of alignment in both meaning and surface structure. In contrast, ROUGE scores remain generally low across all text pairs,



(a) Lexical diversity distributions. (b) Sentence length distributions.

Figure 3. Lexical diversity and sentence length distributions for all texts in HeDetox.

likely reflecting substantial rephrasing and stylistic variation—a characteristic feature of detoxification tasks.

To further explore semantic patterns in the dataset, we computed sentence-level embeddings using the pre-trained heBERT model and visualized distribution via a t-SNE projection (Van der Maaten and Hinton, 2008). This two-dimensional representation (Figure 4) provides an intuitive view of clustering behaviors between the original and human-refined texts. While substantial clustering

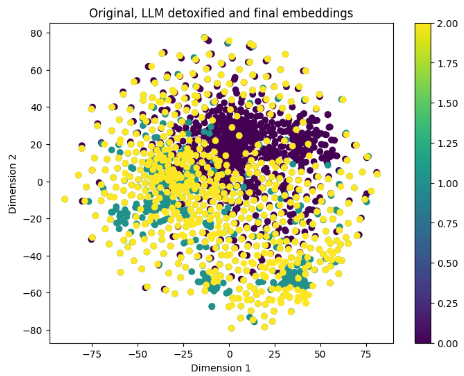


Figure 4. t-SNE visualization of original (blue), LLM detoxified (green), and final (yellow) texts.

suggests shared lexical cores among all three text versions, the broader spread of LLM and human-improved embedding indicates that both transformation processes introduce distinct semantic shifts. The greater overlap between the human-improved and original embeddings suggests that human edits preserve more of the original semantic space compared to LLM detoxification.

In addition, we computed the frequencies of top 10 words in the HeDetox dataset for all categories (presented in Table 4, caution – the table contains offensive words). We used the publicly available list of Hebrew stopwords (Mendels, 2015). We can see that LLM detoxification effectively removes explicit slurs and offensive language but also often eliminates politically charged terms and alters original meaning. In contrast, human edits retain more political and contextual content while rephrasing offensive expressions with constructive language. In this case, LLM tends to insert neutral or polite vocabulary, while humans prioritize meaning preservation.

We additionally evaluated lexical diversity and informational complexity across the three text versions by computing the Measure of Textual Lexical Diversity (MTLD) following the formulation by McCarthy (2005), with the default threshold of 0.72, and word entropy for each sentence. The results (Table 5) show average MTLD and entropy scores for the original, LLM-detoxified, and human-refined texts. The analysis of lexical diversity and word distribution reveals sharp contrasts between the text versions. The human-improved texts exhibit higher word entropy and moderately increased MTLD compared to the LLM output, suggesting richer vocabulary usage and more natural variation. In contrast, the extremely low MTLD observed in the LLM-detoxified texts points to a repetitive or overly constrained lexical style, highlighting potential limitations in generative diversity.

3.4 Evaluation

To evaluate whether exposure to detoxified variants can enhance offensive language classification, we conducted a fine-tuning experiment using the publicly available OLaH dataset (Litvak et al., 2021) that contains 2024 texts, 821 of them offensive. Our goal was not to increase the number of offensive examples, but to examine whether adding detoxified rewrites could improve the model's ability to detect offensive content. We fine-tuned two BERT-based models: a multilingual model, **ml-BERT** (Devlin et al., 2019), and a Hebrew-specific model, **heBERT** (Chriqui and Yahav, 2022), using a binary offensive/non-offensive classification objective. The OLaH dataset was split 80% for

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

Table 3. ROUGE, BLEU, and BERTScore metrics for different text comparisons (F1).

word (Heb)	En	original	LLM detoxified	human improved
יא	damn	93	0	0
זונה	whore	68	0	0
השמאל	the left	20	14	19
ביבי	Bibi (nickname for Netanyahu)	16	12	16
כלפי	towards	0	23	19
הכותב	the writer	0	39	0
מבין	understands	0	17	14
ההתנהלות	conduct	0	13	15
ההתנהגות	behavior	0	12	15
זבל	trash	24	0	0
חתיכת	piece of	24	0	0
מסכים	agrees	0	22	0
המצב	the situation	0	20	0
מזדיין	motherf***er	19	0	0
שרמוטה	slut	16	0	0
המדינה	the state	16	0	0
קוקסינל	tranny (slur)	16	0	0
להתמודד	to cope	0	14	0
הדרים	the things	0	0	11
הממשלה	the government	0	0	11
להבין	to understand	0	0	11
נתניהו	Netanyahu	0	0	10

Table 4. Counts of top words in HeDetox.

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

Table 5. MTLD and word entropy across texts.

training and 20% for validation.

To assess the effect of detoxified data, we repeated training after augmenting the original training set with paired original-detoxified sentences from our HeDetox dataset. Note that these additions did not simply increase the number of offensive examples but introduced alternative linguistic realizations of the same semantic content, aimed at improving the model's generalization. Table 6 shows that both models benefited from this augmentation. The F1 score of heBERT improved from 0.7003 to 0.7202, and mlBERT showed a more substantial gain from 0.5855 to 0.7029. These improvements suggest that exposure to detoxified rewrites enhances the classifier's ability to generalize beyond surface-level lexical cues.

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

Table 6. Classification results on the OLaH test set with and without HeDetox augmentation.

4 Conclusions and Future Work

This paper introduced HeDetox, the first parallel dataset for offensive language detoxification in Hebrew, addressing a major gap in Hebrew NLP resources. The dataset includes 600 pairs of offensive and detoxified sentences, created through a hybrid process that combines LLM outputs with expert human correction. This approach ensures that offensive content is neutralized while preserving the original intent and tone. Extensive linguistic and semantic analysis showed that both LLM and human interventions improve lexical diversity and content structure. Moreover, incorporating HeDetox into offensive language classification tasks enhanced model performance, demonstrating the practical value of detoxified data for downstream applications. Despite its contributions, HeDetox is currently limited to explicitly offensive texts and modest in size. Future work will focus on expanding the dataset to include implicit offenses, scaling its volume, addressing discourse-level detoxification, and incorporating active learning strategies for annotation (Rouzegar and Makrehchi, 2024; Li et al., 2024). We acknowledge the ethical concerns surrounding detoxification tasks and emphasize that our dataset is intended for research purposes, with full transparency and awareness of the potential risks of misuse.

References

Katherine Atwell, Sabit Hassan, and Malihe Alikhani. 2022. Appdia: A discourse-aware transformer-based style transfer model for offensive social media conversations. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6063–6074.

Cristina Bosco, Felice Dell'Orletta, Fabio Poletto, Manuela Sanguinetti, Maurizio Tesconi, et al. 2018. Overview of the EVALITA 2018 Hate Speech Detection Task. In *CEUR workshop proceedings*, volume 2263, pages 1–9. CEUR.

Avihay Chriqui and Inbal Yahav. 2022. Hebert & hebemo: a hebrew bert model and a tool for polarity analysis and emotion recognition. *INFORMS Journal on Data Science*.

Daryna Dementieva, Nikolay Babakov, and Alexander Panchenko. 2024a. Multiparadetox: Extending text detoxification with parallel data to new languages. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 124–140.

Daryna Dementieva, Nikolay Babakov, Amit Ronen, Abinew Ali Ayele, Naquee Rizwan, Florian Schneider, Xintong Wang, Seid Muhie Yimam, Daniil Alekhseevich Moskovskiy, Elisei Stakovskii, et al. 2025. Multilingual and explainable text detoxification with parallel corpora. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7998–8025.

Daryna Dementieva, Daniil Moskovskiy, Nikolay Babakov, Abinew Ali Ayele, Naquee Rizwan, Frolian Schneider, Xintog Wang, Seid Muhie Yimam, Dmitry Ustalov, Elisei Stakovskii, et al. 2024b. Overview of the multilingual text detoxification task at pan 2024. *Working Notes of CLEF*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.

Matan Eyal, Yoav Kantor, and Reut Tsarfaty. 2022. Multilingual sequence-to-sequence models for hebrew nlp.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys*, 51(4):1–30.

Eliyahu Gueta, Dafna Shahaf, and Reut Tsarfaty. 2023. Explicit morphological knowledge improves pre-training of language models for hebrew.

Nagham Hamad, Mustafa Jarrar, Mohammad Khalilia, and Nadim Nashif. 2023. Offensive hebrew corpus and detection using bert. In *2023 20th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA)*, pages 1–8. IEEE.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Ritesh Kumar, Atul Kr Ojha, Shervin Malmasi, and Marcos Zampieri. 2018. Benchmarking aggression identification in social media. In *Proceedings of the first workshop on trolling, aggression and cyberbullying (TRAC-2018)*, pages 1–11.

Barbara Lewandowska-Tomaszczyk, Anna Baczkowska, Olga Dontcheva-Navrátilová, Chaya Liebeskind, Giedrė Valūnaitė Oleškevičienė, Slavko Žitnik, Marcin Trojszczak, Renata Povolná, Linas Selmistraitis, Andrius Utka, and Dangis Gudelis. 2023a. Llod schema for simplified offensive language taxonomy in multilingual detection and applications. *Lodz Papers in Pragmatics*, 19(2):301–324.

Barbara Lewandowska-Tomaszczyk, Anna Baczkowska, Chaya Liebeskind, Giedre Valunaite Oleskeviciene, and Slavko Žitnik. 2023b. An integrated explicit and implicit offensive language taxonomy. *Lodz Papers in Pragmatics*, 19(1):7–48.

Dongyuan Li, Zhen Wang, Yankai Chen, Renhe Jiang, Weiping Ding, and Manabu Okumura. 2024. A survey on deep active learning: Recent advances and new frontiers. *IEEE Transactions on Neural Networks and Learning Systems*, 36(4):5879–5899.

Chaya Liebeskind and Shmuel Liebeskind. 2018. Identifying abusive comments in hebrew facebook. In *2018 IEEE International conference on the science of electrical engineering in Israel (ICSEE)*, pages 1–5. IEEE.

Chaya Liebeskind, Marina Litvak, and Natalia Vanetik. 2024. From linguistics to practice: a case study of offensive language taxonomy in hebrew. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 110–117.

Chaya Liebeskind, Natalia Vanetik, and Marina Litvak. 2023. Hebrew offensive language taxonomy and dataset. *Lodz Papers in Pragmatics*, 19(2):325–351.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain.

Marina Litvak, Natalia Vanetik, Chaya Liebeskind, Omar Hmdia, and Rizek Abu Madeghem. 2022. Offensive language detection in hebrew: can other languages help? In *Proceedings of the thirteenth language resources and evaluation conference*, pages 3715–3723.

Marina Litvak, Natalia Vanetik, Yaser Nimer, Abdulrhman Skout, and Israel Beer-Sheba. 2021. Offensive language detection in semitic languages. In *Multimodal Hate Speech Workshop 2021*, pages 7–12.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. Paradetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818.

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11).

Thomas Mandl, Sandip Modha, Prasenjit Majumder, Daksh Patel, Mohana Dave, Chintak Mandlia, and Aditya Patel. 2019. Overview of the HASOC track at FIRE 2019: Hate Speech and Offensive Content Identification in Indo-European Languages. In *Proceedings of the 11th annual meeting of the Forum for Information Retrieval Evaluation*, pages 14–17.

Philip M McCarthy. 2005. *An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD)*. Ph.D. thesis, The University of Memphis.

Gideon Mendels. 2015. Hebrewstopwords. https://github.com/gidim/HebrewStopWords. Accessed: 2025-07-27.

Daniil Moskovskiy, Nikita Sushko, Sergey Pletenev, Elena Tutubalina, and Alexander Panchenko. 2025. SynthDetoxM: Modern LLMs are Few-Shot Parallel Detoxification Data Annotators. *arXiv preprint arXiv:2502.06394*.

OpenAI. 2024. Gpt-4o technical report. https://openai.com/research/gpt-4o.

Nedjma Djouhra Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *EMNLP-IJCNLP 2019-2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA.

W. Tzuf Paz-Argaman, Guy Arviv, Yoav Kantor, and Reut Tsarfaty. 2024. Hesum: A novel dataset for abstractive text summarization in hebrew.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55:477–523.

Hamidreza Rouzegar and Masoud Makrehchi. 2024. Enhancing text classification through llm-driven active learning and human annotation. *arXiv preprint arXiv:2406.12114*.

Natalia Vanetik, Lior Liberov, Chaya Liebeskind, and Marina Litvak. 2025. Hedetox: Hebrew detoxification dataset and annotation guidelines. https://github.com/NataliaVanetik/OffensiveLanguageResearchLab#hedetox. Accessed: July 2025.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2018. Overview of the GermEval 2018 shared task on the identification of offensive language. In *Proceedings of GermEval*, pages 1–10.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019a. Predicting the type and target of offensive posts in social media. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1415–1420.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019b. Semeval-2019 task 6: Identifying and categorizing offensive language in social media (offenseval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.