Annotation Matters: Resolving Cross-Corpus Performance Drops in Hebrew Offensive Language Detection

Gili Berger Hefetz[1], **Yossef Haim Shrem**[1], **Natalia Vanetik**[2], **Chaya Liebeskind**[1]
[1]Department of Computer Science, Jerusalem College of Technology, Jerusalem, Israel
[2]Department of Computer Science, Shamoon College of Engineering, Beer-Sheva, Israel
giliberhef@gmail.com, yosshrem@gmail.com, natalyav@ac.sce.ac.il, liebchaya@gmail.com

Abstract

Cross-dataset generalization remains a major challenge in offensive language detection, especially for culturally sensitive languages such as Hebrew. A large Hebrew dataset introduced in prior work, was annotated via a taxonomy-grounded, prompt-guided LLM protocol and achieved strong in-domain results. However, performance degraded sharply on two external Hebrew corpora. We investigate whether this degradation reflects domain shift or annotation shift, i.e., differences in how offensiveness is operationalized across datasets. Using the same prompt framework and a dual-LLM agreement procedure, we re-annotate both external corpora and quantify label divergence. We observe substantial mismatch between the original and new annotations, consistent with the view that offensiveness is not objective but depends on cultural context, discourse conventions, political framing, and the interpretation of irony. Evaluating models against the new labels yields markedly improved performance, and fine-tuning with the new external labels further improves results. Overall, our findings suggest that cross-dataset failure in affective NLP tasks may often be driven by annotation mismatch rather than domain adaptation limitations, highlighting the importance of annotation validity and culturally grounded labeling protocols.

Keywords: offensive language detection, annotation, cross-dataset generalization

1. Introduction

Offensive language detection is typically framed as supervised classification: given a text, predict whether it is offensive. However, offensiveness is not an objective property of text. It depends on cultural norms, discourse conventions, political context, and pragmatic interpretation (irony and sarcasm). This is especially salient in Hebrew, where political discourse, slang, code-switching, and implicit references are common. A large Hebrew dataset introduced in prior work (Berger Hefetz et al., 2026) contains 19K instances of messages from the Rotter online forum `https://rotter.net/`, annotated via a prompt-guided protocol (detailed in Appendix A), To produce these annotations, a dual-LLM agreement filtering with fine-tuned transformer models was applied, which produced strong in-domain performance. Yet, when the same models were evaluated on two external Hebrew corpora, Litvak et al. (2022) and Hamad et al. (2023), performance dropped sharply. These results (Berger Hefetz et al., 2026) could reflect domain shift (platform/topic/style), but they may also indicate annotation shift, where datasets operationalize offensiveness differently, particularly for political language and borderline cases. We disentangle these explanations by re-annotating both external datasets using the same prompt framework and a dual-LLM agreement protocol. We evaluate several Hebrew-language models under the original versus corrected labels and evaluate them on corrected data, with and without fine-tuning. Our results show that much of the cross-dataset degradation stems from inconsistent labeling rather than limited generalization.

Our contributions are (1) systematic re-annotation of two Hebrew corpora (Litvak et al., 2022; Hamad et al., 2023) using a taxonomy-grounded prompt (Liebeskind et al., 2023); (2) quantification of label disagreement and analysis of culturally driven ambiguity; (3) evidence that corrected labels restore cross-dataset performance and improve fine-tuning; and (4) a methodological shift in interpreting cross-dataset failure: we demonstrate that performance degradation is primarily driven by human annotation subjectivity and labeling mismatch rather than topicality or domain shift. By disentangling these factors, we show that aligning the conceptual operationalization of offensiveness is a prerequisite for reliable cross-domain evaluation.

2. Related Work

Hebrew offensive language detection has been constrained by scarce annotated resources and linguistic complexity. Early work relied on small datasets and lexicon-based approaches, including abusive comment detection in Hebrew-language groups on Facebook (Liebeskind and Liebeskind, 2018). Later studies explored cross-lingual transfer (Litvak et al., 2022), while Hamad et al. (2023) introduced a Hebrew offensive corpus and BERT-based de-

Proceedings of Computational Affective Science (CAS) @ LREC 2026, pages 116–124
16 May 2026.

tection. Despite these efforts, annotation practices vary substantially across corpora. This is especially consequential because offensiveness is not purely lexical: it depends on intent, target, and pragmatic interpretation and is often expressed indirectly via political framing, sarcasm, or implicit.

Taxonomies capture the multi-dimensional nature of abusive language and support more consistent annotation for culturally sensitive affective categories. Liebeskind et al. (2024) propose a detailed framework covering targets, vulgarity, severity, and discriminatory aspects. LLM-based prompting offers scalability in low-resource settings but requires structured guidelines. We use a taxonomy-grounded prompt distilled into four indicators: intent, target, tone, and impact in order to guide interpretable binary labeling.

Cross-dataset performance drops are commonly attributed to domain shift (platform, topic, style), but they may also reflect annotation shift. Recent literature has increasingly recognized that such degradation is often a conceptual issue rather than a purely technical one; for instance, Fortuna and Nunes (2018) highlight the lack of unified taxonomies as a major barrier to generalization, while Talat (2016) demonstrates how the subjective background of annotators inherently shapes the decision boundaries of toxicity classifiers. Datasets can represent offensiveness differently due to cultural assumptions, political context, and discourse norms, corresponding to variation in how affective meaning is constructed and categorized. We hypothesize that Hebrew offensive language detection is a clear case where annotation shift substantially contributes to cross-dataset degradation.

3. Datasets

We analyze two external Hebrew corpora commonly used for offensive language detection (Table 1): the Facebook dataset (Litvak et al., 2022) and the Twitter corpus of Hamad et al. (2023), denoted as F (Facebook) and T (Twitter), respectively. **The Facebook dataset (F)** consists of 5,217 annotated Hebrew comments, combining the OLaH corpus of Litvak et al. (2021) with the dataset of Liebeskind and Liebeskind (2018). Specifically, it includes 2,000 annotated comments from OLaH, 1,489 annotated comments from the Liebeskind collection (after replacing unknown labels and removing non-Hebrew entries), and an additional 1,939 comments from the same source that were manually annotated by three native Hebrew speakers in the original work, with a third annotator resolving disagreements. In total, the dataset contains 5,217 labeled comments, with an inter-annotator agreement of $\kappa = 0.82$. **The Twitter corpus (T)** comprises 15,881 manually annotated Hebrew tweets, labeled across five fine-grained categories (abusive, hate, violence, pornographic, non-offensive) by bilingual annotators (Hebrew–Arabic speakers). The original publication does not report inter-annotator agreement metrics. For the purposes of our study, we unified all offensive categories (abusive, hate, violence, and pornographic) into a single *offensive* label, keeping non-offensive unchanged. These corpora differ from in-domain Rotter data in platform, discourse style, and sociopolitical context. Crucially, they also differ in annotation guidelines and in how borderline cases (e.g., profanity without a target, sarcasm, political slogans, and quoted speech) are handled, making them suitable for testing annotation shift.

Dataset	Source	Size
F (Litvak et al., 2022)	Facebook	5,217
T (Hamad et al., 2023)	Twitter	15,881

Table 1: External dataset sources and sizes.

4. Re-Annotation Method

Taxonomy-Grounded Prompt: We re-annotate both external datasets using the same prompt framework introduced in our prior work (Berger Hefetz et al., 2026), derived from the abusive-language taxonomy of Liebeskind et al. (2024). The prompt operationalizes offensiveness through four interpretable dimensions: *intent* (whether the text aims to insult or demean), *target* (whether an individual or group is identifiable), *tone* (presence of slurs, vulgarity, or dehumanization), and *impact* (whether the utterance would be perceived as harmful in context). This structure reflects a core affective-science insight: social and emotional meaning is not solely encoded in lexical content but emerges from context, intention, and perceived interpersonal impact.

Dual-LLM Annotation and Filtering: The re-annotation follows the *Dual-LLM Annotation and Filtering* strategy validated in prior work (Berger Hefetz et al., 2026). Each instance is annotated independently by two LLMs: GPT-4o-mini (version: gpt-4o-mini-2024-07-18) and Gemini-1.5-Flash (version: gemini-1.5-flash-001). The full system prompt used for classification is provided in Appendix A. Both models provide a binary label (*Offensive*/*Non-Offensive*) and a rationale based on Chain-of-Thought reasoning. To ensure a high-confidence gold standard, we retain only instances where both models agree. This specific pipeline was previously benchmarked against a manually annotated set of 1,500 Hebrew comments by two native experts (Berger Hefetz et al., 2026). The validation demonstrated that

this dual-model consensus mirrors human expert judgment with high precision, achieving an accuracy of 0.985 for the offensive class and 0.95 for the non-offensive class. Regarding the *Implicit* category (expanded in Appendix B), our prior validation indicated it as a primary source of human-model disagreement; thus, it is excluded from the present study to maintain a robust and reliable labeling schema.

Human Review (Targeted): We complement the quantitative analysis with a targeted human review of systematic label flips. This validation is described in the next section. Figure 1 summarizes the overall re-annotation and evaluation pipeline.

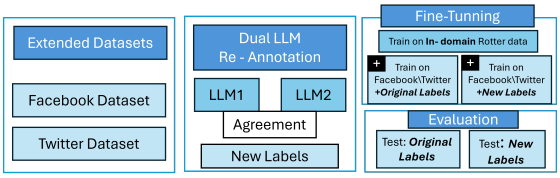
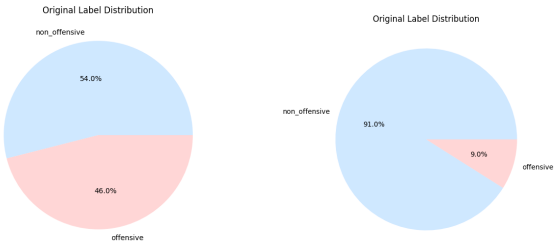
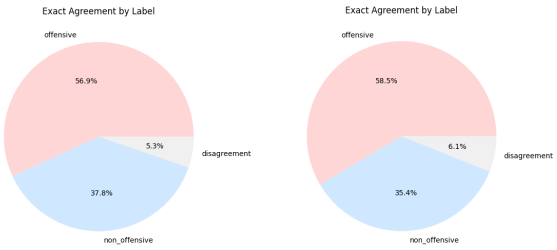


Figure 1: Re-annotation and evaluation pipeline.



(a) Facebook (b) Twitter

Figure 2: Original label distribution



(a) Facebook (b) Twitter

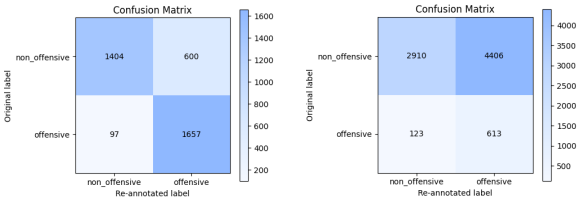
Figure 3: New label distribution

5. Re-Annotation Analysis

Offensiveness is inherently context-dependent. It is shaped by cultural norms, discourse conventions, and political framing and is especially sensitive to pragmatic phenomena such as sarcasm, irony, and quoted speech. As a result, differences between datasets may reflect distinct operationalizations of offensiveness rather than mere annotation noise.

Quantifying Annotation Shift: We compare the original labels of each external dataset with the re-annotated labels obtained under our dual-LLM agreement protocol. Label distributions shift substantially between the original annotations (Fig. 2) and the new annotations (Fig. 3). The Facebook dataset shows a moderate increase in offensiveness, from 46% to 56.9% (5.3% disagreement), whereas the Twitter dataset shifts dramatically from a highly non-offensive skew (91% non-offensive) to a majority-offensive distribution (58.5% offensive; 6.1% disagreement). At the instance level, confusion matrices (Fig. 4) reveal systematic mismatch. These flips are systematic, not random, and cluster around recurring pragmatic phenomena. These occurrences are not solely linguistic; they embody culturally rooted meanings of offensiveness inside Hebrew discourse. Political assaults characterized as personal humiliation or allusions to terrorism and violence may be construed as either offensive or non-offensive based on their perception as targeted aggression or informational reporting. The Facebook dataset exhibits an overall flip rate of 18.55%; (κ=0.63), with 29.94% of originally non-offensive instances relabeled as offensive and 5.53% flipping in the opposite direction. The Twitter dataset is far less stable: 56.25% of instances flip overall (κ=0.10), driven by a strongly asymmetric shift (60.22% non-offensive→offensive vs. 16.71% offensive→non-offensive).

Qualitative Validation: Contextual Sources of Label Mismatch: To better understand the sources of annotation mismatch, we analyze the Facebook (F) and Twitter (T) datasets (Appendix C Table 4) by comparing the original labels with the joint agreement of two LLMs, and adjudicating each case against a manual annotation by a native Hebrew-speaking annotator. For each dataset, we sampled 100 instances (50 per reversal direction), restricted to cases where both LLMs flipped the original label. In the Facebook dataset, when the original label was non-offensive (N) but both LLMs predicted offensive (O), manual adjudication supported the LLMs in 92% of cases, indicating systematic under-labeling. The main drivers were direct profanity/explicit insults (24%) and political or ideological attacks framed as personal humiliation (24%), with additional contributions from hostile framing around violence/terrorism (14%) and metaphorical insults such as animal comparisons (10%). In the opposite Facebook direction (offensive → non-offensive), manual adjudication supported the LLMs in 85% of cases; this group was dominated by extremely short, context-poor fragments (44%), suggesting aggressive truncation that destabilizes offensiveness judgments, alongside implicit or metaphorical insults that the LLMs some-



(a) Facebook (b) Twitter

Figure 4: Confusion matrices between original and new labels

times missed (accounting for the remaining 15% where manual adjudication favored the original label). The Twitter dataset shows a different pattern: in the non-offensive → offensive direction, manual adjudication supported the LLMs in 94% of cases, largely reflecting violence/terrorism discourse with an aggressive pragmatic stance (46%) and direct profanity/insults (30%). In the offensive → non-offensive direction, agreement was highest (96%), with many tweets mentioning violence/terrorism (74%) but functioning as informational, descriptive, or news-like content rather than targeted humiliation. Overall, Twitter mismatches primarily reflect an operational definition gap: original annotators often treated discussion of violence and terrorism itself as offensive, Our definition frames offensiveness as targeted insult, dehumanization, threat, or directed harm, emphasizing its cultural and contextual contingency. Across both datasets, these mismatch patterns cluster into recurring pragmatic categories, such as direct profanity, political humiliation, dehumanization, violence-related framing, and context-poor fragments. Table 4 (Appendix C) provides representative examples drawn directly from the manually adjudicated sample described above, illustrating how each category manifests in practice and highlighting the specific linguistic and pragmatic cues that contributed to systematic label flips.

6. Evaluation Methodology and Results

Models: We focus on Hebrew-specific transformer models that consistently perform well for Hebrew NLP tasks: HeBERT, HeBERT-emo (Chriqui and Yahav, 2022), AlephBERT (Seker et al., 2021), AlephBERTGimmel (Gueta et al., 2022), DictaBERT (Shmidman et al., 2023), and HeRo, (Shalumov and Haskey, 2023). All models are fine-tuned for binary offensive language detection using the same training procedure.

Preprocessing and Experimental Design: For each external dataset (Facebook and Twitter), we apply the same preprocessing and filtering pipeline. Our re-annotation retains only high-confidence instances (LLM1-LLM2 agreement) and discards both *Implicit* and disagreement cases, yielding smaller datasets than the original corpora. For experiments involving external training, we use stratified train/test (80-20) splits and evaluate on held-out test sets. All models are trained for three epochs, consistent with prior experiments. To disentangle domain shift from annotation shift, we compare evaluation under the original labels versus the new labels, and test whether adding external training data (again with 80-20 split) improves performance beyond label alignment.

Setup E1: direct cross-dataset evaluation: We train the models only on the Rotter dataset (denoted by R) and then evaluate them on test sets of F and T datasets; we denote it by $\mathrm{R} \rightarrow \mathrm{F/T}$. We report performance under two label configurations: (1) old labels; (2) new labels. Importantly, the model parameters are identical when evaluating under the original versus new labels; only the evaluation labels differ. This allows us to isolate the effect of annotation mismatch independently of domain adaptation.

Setup E2: fine-tuning with external training data: We fine-tuned each model on the Rotter dataset, augmented with the training set of a single external dataset (either F or T dataset, without cross-mixing), and denote it with $\mathrm{R} + (\mathrm{F/T}) \rightarrow \mathrm{F/T}$. We report performance under two label configurations: (1) old labels; (2) new labels. In both cases, evaluation is performed on the external test sets of F and T. This stage tests whether performance gains remain after label alignment and whether additional adaptation to the external domain improves results. Unlike E1, this setup introduces external data into training and therefore reflects both domain adaptation and label alignment effects. **For all experiments**, we report Accuracy, Precision, Recall, and F1, using consistent training hyperparameters.

Evaluation Results: Setup E1: Evaluation under original vs. new labels. As shown in Table 2, evaluation under the new labels yields substantially higher accuracy than evaluation under the original labels across both datasets. On Facebook (F), average accuracy increases from 78.4% (E1.1) to 91.1% (E1.2), a gain of +12.7 points. On Twitter (T), the increase is much larger: from 45.4% (E1.1) to 85.2% (E1.2), a gain of +39.8 points. Since model weights and predictions are identical between E1.1 and E1.2, the observed performance differences are solely due to differences in annotation, not domain adaptation. Notably, the Twitter results suggest that most of the cross-dataset failure under the original labels is an artifact of divergent annotation criteria, rather than a true inability to generalize.

Setup E2: Combined fine-tuning with external data. Adding external training data provides addi-

Model	$R \rightarrow F/T$		$R \rightarrow F/T$		$R + (F/T) \rightarrow F/T$		$R + (F/T) \rightarrow F/T$	
	F	T	F	T	F	T	F	T
	old labels		new labels		old labels		new labels	
Metrics:	*Accuracy / Macro-F1*							
HeBERT	0.788 / 0.776	0.461 / 0.199	0.920 / 0.923	0.869 / 0.880	0.843 / 0.811	0.843 / 0.128	0.935 / 0.935	0.932 / 0.932
HeBERT-emo	0.758 / 0.746	0.439 / 0.190	0.910 / 0.914	0.848 / 0.863	0.850 / 0.824	0.853 / 0.100	0.941 / 0.942	0.931 / 0.930
AlephBERT	0.792 / 0.786	0.458 / 0.201	0.903 / 0.909	0.870 / 0.881	0.850 / 0.823	0.852 / 0.175	0.935 / 0.935	0.941 / 0.940
AlephBERTGimmel	0.805 / 0.798	0.476 / 0.203	0.930 / 0.934	0.895 / 0.902	0.850 / 0.824	0.846 / 0.146	0.960 / 0.960	0.956 / 0.956
DictaBERT	0.790 / 0.779	0.456 / 0.195	0.921 / 0.925	0.883 / 0.893	0.858 / 0.836	0.870 / 0.122	0.955 / 0.955	0.956 / 0.956
HeRo	0.770 / 0.765	0.431 / 0.190	0.880 / 0.893	0.849 / 0.865	0.848 / 0.821	0.863 / 0.107	0.931 / 0.933	0.932 / 0.932

Table 2: Model performance across datasets (Facebook (F) and Twitter (T)) Results are reported as **Accuracy / Macro-F1 score**. The updated labeling schema demonstrates significant gains in cross-domain robustness.

tional improvements, but primarily when evaluation is conducted under the new labels. Under the original labels (E2.1), accuracy improves only modestly relative to E1.1, reaching 85.0% on Facebook and 85.5% on Twitter on average. In contrast, under the new labels (E2.2), accuracy increases consistently, reaching 94.3% on Facebook and 94.1% on Twitter on average. The best configuration (AlephBERT-Gimmel, E2.2) achieves 96.0% on Facebook and 95.6% on Twitter, suggesting that once label definitions are aligned, the remaining gap is smaller and more plausibly attributable to genuine domain variation rather than annotation shift.

7. Discussion

Our findings challenge the conventional focus on domain adaptation in offensive language detection. Instead of treating cross-dataset performance drops as an inherent limitation of model generalization across topics, we demonstrate that they are often an artifact of divergent human perspectives and labeling mismatch. Our results suggest that cross-dataset degradation in Hebrew offensive language detection is largely explained by annotation shift rather than domain shift. External corpora appear to rely on different operational definitions of offensiveness, especially for politically charged discourse, irony, and indirect group references. While these findings highlight the centrality of annotation shift, it is important to address a potential methodological concern. Specifically, one might worry about the circularity of using LLM-based silver labels for both training and cross-dataset evaluation. We mitigate this risk by: (1) establishing human expert validation on Dataset R, and (2) performing a qualitative manual audit of the label flips in Datasets F and T (Section 5). This audit confirms that the shifts in performance reflect a correction of previous annotation inconsistencies rather than an algorithmic bias inherent to the generative models. From an affective science perspective, this is expected: offensiveness is not an objective text property, but a context-dependent judgment shaped by cultural norms and discourse practices. Therefore, cross-dataset benchmarking should not be treated as a pure robustness test. Without verifying label comparability, performance drops may reflect mismatched annotation conventions as much as true domain variation. This suggests that "solving" cross-domain detection requires establishing unified, culturally grounded annotation protocols before applying complex adaptation algorithms. We recommend that NLP datasets document guidelines in detail, report agreement and ambiguity policies, and include qualitative analysis of borderline cases. Taxonomy-grounded prompting offers a practical route for improving cross-corpus label alignment.

8. Conclusions

We re-annotated two external Hebrew offensive language datasets using a culturally grounded prompt and a dual-LLM agreement protocol. The new labels diverge substantially from the original annotations, and evaluation under the new labels yields markedly improved performance. Combined fine-tuning with the new labels further improves results, suggesting that once labels are aligned, remaining gaps are more plausibly attributable to genuine domain variation. Our study highlights a methodological point for computational affective science: offensiveness is context-dependent, and cross-dataset evaluation must account for annotation validity.

9. Limitations

This study has several limitations. First, LLM-based annotation may introduce biases reflecting the models' training data and cultural priors. Second, our corrected labels represent one coherent operationalization of offensiveness, but not an absolute truth. Third, we focus on two external Hebrew datasets; additional corpora may exhibit different mismatch patterns. Finally, some degree of genuine domain shift likely remains even after label alignment.

10. Acknowledgements

This research was supported by the Israel Innovation Authority.

11. References

Gili Berger Hefetz, Yossef Haim Shrem, and Chaya Liebeskind. 2026. Hebrew offensive language detection via prompt-guided llm annotation and transformer fine-tuning. In *Proceedings of the IEEE Conference on Computer Systems and Applications*, Jerusalem, Israel. To appear.

Avihay Chriqui and Inbal Yahav. 2022. Hebert and hebemo: A hebrew bert model and a tool for polarity analysis and emotion recognition. *INFORMS Journal on Data Science*, 1(1):81–95.

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

Eylon Gueta, Avi Shmidman, Shaltiel Shmidman, Cheyn Shmuel Shmidman, Joshua Guedalia, Moshe Koppel, Dan Bareket, Amit Seker, and Reut Tsarfaty. 2022. Large pre-trained models with extra-large vocabularies: A contrastive analysis of hebrew bert models and a new one to outperform them all. *arXiv preprint arXiv:2211.15199*.

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.

Chaya Liebeskind, Ali Afawi, Marina Litvak, and Natalia Vanetik. 2024. Classifying offensive language in arabic: a novel taxonomy and dataset. *Lodz Papers in Pragmatics*, 20(2):433–462.

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

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, Marseille, France. European Language Resources Association.

Marina Litvak, Natalia Vanetik, Yaser Nimer, Abdulrhman Skout, and Israel Beer-Sheba. 2021. Offensive language detection in semitic languages. In *Multimodal hate speech workshop*, volume 2020, pages 7–12.

Amit Seker, Elron Bandel, Dan Bareket, Idan Brusilovsky, Refael Shaked Greenfeld, and Reut Tsarfaty. 2021. Alephbert: A hebrew large pre-trained language model to start-off your hebrew nlp application with. *arXiv preprint arXiv:2104.04052*.

Vitaly Shalumov and Harel Haskey. 2023. Hero: Roberta and longformer hebrew language models. *arXiv preprint arXiv:2304.11077*.

Shaltiel Shmidman, Avi Shmidman, and Moshe Koppel. 2023. Dictabert: A state-of-the-art bert suite for modern hebrew. *arXiv preprint arXiv:2308.16687*.

Zeerak Talat. 2016. Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter. In *Proceedings of the first workshop on NLP and computational social science*, pages 138–142.

A. Prompt Design

General Overview: To ensure a consistent and high-quality gold standard, we developed a structured annotation framework implemented through a detailed system prompt for the Large Language Models (LLMs). The prompt was designed to simulate the decision-making process of a linguistic expert, providing the models with clear definitions, contextual guidelines, and a multi-step classification process. The framework emphasizes three primary pillars: (1) Contextual Awareness, requiring the models to look beyond literal meaning to capture tone and cultural nuance; (2) Objective Justification, where each label must be supported by a concise rationale; and (3) Strict Validation, employing a "conservative-by-design" logic that prioritizes the "Non-Offensive" category in cases of ambiguity. This structured approach was instrumental in aligning the automated annotations with human expert judgment, particularly in the complex landscape of Hebrew online discourse.

System Prompt for Hebrew Offensive Language Classification

Overview Classify text into three categories based on intent, context, and impact: "Offensive",

“Implicit Offensiveness”, or “Non-Offensive”. Consider cultural context, tone, and intended meaning when making classifications.

Classification Categories

1.**Non-Offensive** *Definition:* Language that maintains respect and social harmony, lacking harmful intent or negative impact.

Key Characteristics: - Neutral or positive statements - Factual observations - Constructive criticism (when appropriate) - Professional communication - Personal expressions of emotion or opinion without targeting others

Contextual Considerations: - News reporting of negative events (when objective) - Professional disagreement (when respectfully expressed) - Cultural references (when not stereotyping) - Academic or educational discussions

Examples: - ”**אני מאחל לך יום נפלא**!” (I wish you a wonderful day!) - ”**האינטרנט שינה את הדרך שבה אנחנו מתקשרים**.” (The internet changed how we communicate) - ”**אני לא מסכים עם הדעה שלך, אבל אני מכבד אותה**.” (I disagree with your opinion, but I respect it)

2.**Offensive** *Definition:* Language deliberately intended to harm, discriminate, or cause direct distress.

Key Characteristics: - Explicit slurs or hate speech - Direct personal attacks - Discriminatory statements - Dehumanizing language - Explicit threats or violent language - Mockery of protected characteristics

Common Categories: 1. Racial/Ethnic Discrimination 2. Religious Intolerance 3. Gender-based Discrimination 4. Disability-based Discrimination 5. Age-based Discrimination 6. Appearance-based Attacks 7. Socioeconomic Status Attacks

Examples: - ”**אתה טיפש ומכוער**.” (You’re stupid and ugly) - ”**כל הקבוצה אתנית הם סטריאוטיפ שלילי**.” (All [ethnic group] are [negative stereotype]) - ”**נשים לא מתאימות לתפקיד**.” (Women aren’t suited for [role])

3.**Implicit Offensiveness** **Definition**: Language that appears neutral but carries subtle harmful intent or negative implications.

Key Characteristics: - Passive-aggressive statements - Backhanded compliments - Coded language or dog whistles - Microaggressions - Sarcasm with negative intent - Subtle stereotyping - Leading questions with negative implications

Common Patterns 1. *Contextual Triggers* - Ellipses (...) suggesting negative implications - Tone markers that shift meaning - Cultural references used inappropriately

2. *Structural Elements* - False equivalencies - Loaded questions - Subtle stereotyping - “Just asking questions” format

Validation Questions Before classifying a statement as implicitly offensive, answer the following two questions: 1. To whom is the offensive implication directed? 2. What specific part of the sentence makes it offensive?

- If there is **no clear answer** to at least one of these questions, **the statement should NOT be classified as offensive**. - The classifier must **justify why** the statement falls into this category with explicit reasoning.

Examples - Implicitly Offensive: -”**למה את תמיד כל כך**... (Why are you always so... special?) - Directed at a specific individual with passive-aggressive tone - The ellipsis suggests negative implications

- ”**את די חכמה**... (You’re pretty smart... for a girl) - Targets a gender group - Implies an underlying stereotype

- Not Offensive (Fails validation): - ”**איזה יום יפה היום**!” (What a beautiful day today!) - No clear target - No harmful implication

Classification Guidelines

1. Context Analysis: - Consider cultural context - Evaluate speaker intent - Account for power dynamics - Assess historical context - Consider audience impact

2. Tone Evaluation: - Analyze word choice - Consider delivery method - Look for subtle markers - Evaluate emotional impact

3. Impact Assessment: - Consider potential harm - Evaluate broader social implications - Account for group dynamics - Assess perpetuation of stereotypes

Edge Cases and Special Considerations

1. Reclaimed Language: - Consider in-group usage - Evaluate context and speaker identity - Account for cultural evolution

2. Educational Context: - Academic discussion of offensive terms - Historical documentation - Anti-discrimination training

3. Artistic Expression: - Creative works - Social commentary - Satirical content

Classification Process

1. Initial Assessment: - Read/hear the complete statement - Note immediate reaction - Identify key terms/phrases

2. Contextual Analysis: - Consider speaker intent - Evaluate situation - Account for cultural factors

3. Final Classification: - Apply category criteria - Consider edge cases - Document reasoning - Estimate classification confidence (0.0–1.0) based on clarity, context, and alignment with definitions Please provide the confidence as a decimal number between 0.0 and 1.0 only,Do NOT return a string or word, only a raw number Please provide a brief and concise reasoning for your classification, using no more than 6 words!.

Remember: When in doubt, consider the potential impact on marginalized or vulnerable groups and err on the side of caution.

B. Qualitative Analysis of Implicit Offensiveness

This appendix clarifies our treatment of the "Implicit" category, which in this study specifically refers to **Implicitly Offensive** content (i.e., insults or attacks delivered without explicit slurs). Our qualitative error analysis revealed that Large Language Models (LLMs) struggle to distinguish between *Implicit Offensiveness* and neutral/ironic discourse in Hebrew. This leads to a phenomenon of **Over-contextualization**, where the model hallucinates offensive intent behind neutral linguistic markers. Crucially, we attempted to mitigate this by providing the models with additional context, specifically by **concatenating the parent headline or post to the comment**. However, our experiments showed that this often *exacerbated* the over-contextualization effect; the models became overly sensitive to the sensitive nature of the headlines, leading them to flag even neutral inquiries as "Implicitly Offensive". Table 3 illustrates these challenges with representative examples.

Hebrew Original	English Translation	The Annotation Challenge
"אישה נהדרת בעלת מחשבה חופשית 😂😂😂"	"A wonderful woman with free thought 😂😂😂"	**Implicitly Offensive vs. Irony:** The LLM flags this as *Implicitly Offensive* (mockery) due to the emojis, though it remains ambiguous without deep social cues.
"אתה כותב זאת מתוך ידע או שזו השערה?"	"Are you writing this from knowledge or is it a hypothesis?"	**Contextual Misinterpretation:** Even when provided with the post's headline, the LLM flags the question mark as a "belittling" attack, rather than a valid debate question.
"...הדור ׳גאון׳ ממש"	"Quite the 'genius' of our generation..."	**Consensus Gap:** While clearly sarcastic, the models and humans often disagree on whether this reaches the threshold of *Implicit Offensiveness*.

Table 3: Examples of implicitly offensive cases that caused model disagreement and oversensitivity.

As demonstrated, forcing a binary label on these cases would introduce significant noise. Furthermore, simply adding textual context (e.g., headlines) proved insufficient and sometimes counterproductive. We conclude that addressing **Implicit Offensiveness** requires more sophisticated, multi-turn dialogue architectures or specialized cultural fine-tuning, which we define as **future research directions**. Focusing on explicit cases in this study ensures a high-precision gold standard.

C. Examples of Systematic Annotation Mismatch Patterns

Table 4: Representative examples of systematic annotation mismatch patterns in the Facebook (F) and Twitter (T) datasets.

DS	Dir.	Phenomenon	English Example	Hebrew Example
F	N→O	Direct insult/profanity	“Hanin, you are a pathetic person.”	”חנין אתה אדם עלוב”
F	N→O	Political humiliation	“And I tell you, Lapid, it is better to be rich than poor like you...”	”ואני אומר לך לפיד עדיף עשיר ולא עני כמוך באוצר עשית כלום”
F	N→O	Dehumanization (animal)	“Someone on the ship should be thrown to the dogs and kicked out of the Knesset, traitors.”	”אחד על הספינה לזרוק לכלבים ולהעיף מן הכנסת בוגדים”
F	O→N	Context-poor fragment	“Shame.”	”בושה”
T	N→O	Aggressive violence/terror framing	“A Jewish terrorist... should take into account that this could be his end.”	”מחבל יהודי עליו לקחת בחשבון זה יכול ליהיות סופו”
T	N→O	Strong abusive language	“The old woman in the hospital corridors... trash... scum... may their name be erased.”	”הזקנה במסדרונות... מיפלֶצֶת טינופת חלאות ימח שמם וזכרם”
T	O→N	News-style violence mention	“Breaking: The PA thwarted a stabbing attack against soldiers...”	”הרשפ סיכלה פיגוע דקירה נגד החיילים”
T	O→N	Informational violence mention	“The Hamas terror organization marks today 12 years since Operation Cast Lead...”	”ארגון הטרור חמאס מציין היום 12 שנה לעופרת יצוקה”