Explain the Flag: Contextualizing Hate Speech Beyond Censorship

Jason Liartis[1,2], Eirini Kaldeli[1,2], Lambrini Gyftokosta[3]
Eleftherios Chelioudakis[4,5], Orfeas Menis Mastromichalakis[1,6]

[1]National Technical University of Athens, [2]Datoptron, [3]Independent Researcher
[4]Homo Digitalis, [5]University of the Aegean, [6]Instituto de Telecomunicações

Abstract

Hate, derogatory, and offensive speech remains a persistent challenge in online platforms and public discourse. While automated detection systems are widely used, most focus on censorship or removal, raising concerns for transparency and freedom of expression, and limiting opportunities to explain why content is harmful. To address these issues, explanatory approaches have emerged as a promising solution, aiming to make hate speech detection more transparent, accountable, and informative. In this paper, we present a hybrid approach that combines Large Language Models (LLMs) with three newly created and curated vocabularies to detect and explain hate speech in English, French, and Greek. Our system captures both inherently derogatory expressions tied to identity characteristics and direct group-targeted content through two complementary pipelines: one that detects and disambiguates problematic terms using the curated vocabularies, and one that leverages LLMs as context-aware evaluators of group-targeting content. The outputs are fused into grounded explanations that clarify why content is flagged. Human evaluation shows that our hybrid approach is accurate, with high-quality explanations, outperforming LLM-only baselines. Our source code is available at https://github.com/ails-lab/detoex.

1 Introduction

Warning: This paper contains terms and expressions that may be offensive or disturbing to some readers. They are included solely for illustration and are not intended to endorse such language.

The spread of hate speech online has become a pressing concern with serious personal, social, and legal consequences. Beyond harming individuals, it can escalate social tensions and, in severe cases, contribute to discrimination, hostility, and violence. These risks highlight the need to monitor and address harmful language for the sake of well-being, social cohesion, and legal responsibility. Institutions have recognized this urgency: the European Union has launched initiatives to analyze, regulate, and counteract online hate speech (Commission et al., 2023)[1], and the United Nations has developed a dedicated strategy and plan of action to identify, prevent and confront hate speech[2].

In line with existing literature (Tonneau et al., 2024), this work adopts the United Nations' definition of hate speech. According to the UN, hate speech is *"any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor"*[3]. This interpretation is broader than purely legal definitions, which often restrict hate speech to explicit incitement to violence. It covers a wide spectrum of harmful communication, from derogatory expressions and stereotypes to insults and threats, as long as they are rooted in identity-based characteristics. Personal insults or generic offensive language are generally not included, unless they target individuals or groups based on such characteristics. Hate speech can thus emerge in two ways: through inherently derogatory expressions tied to identity, or through content targeting individuals or groups based on their identity, even without explicit profanity.

Current automated systems for moderating hate speech typically focus on flagging or removing harmful content, often without providing users with any justification. While such interventions may reduce exposure to offensive material, they also introduce two significant risks: users are not given the

[1] https://europeanonlinehatelab.com/
[2] https://www.un.org/en/hate-speech/un-strategy-and-plan-of-action-on-hate-speech
[3] https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech

Findings of the Association for Computational Linguistics: ACL 2026, pages 8305–8325

opportunity to understand how their language may perpetuate stereotypes or cause harm, and moderation actions may appear arbitrary or biased. As a result, many users encounter content removals or labels without explanation, leaving them unable to interpret or contest these decisions. In response to these shortcomings, recent research has begun advocating for explanatory approaches that contextualize offensive and derogatory language rather than simply removing it (Epstein et al., 2022; Menis Mastromichalakis et al., 2025).

In this work, we present a hybrid approach designed to capture the full scope of hate speech while providing grounded, contextualized explanations. A central contribution of our study is a set of curated vocabularies in English, French, and Greek containing terms that are inherently offensive or derogatory toward specific groups. These vocabularies were constructed by collecting and reviewing entries from Wiktionary[4] , including information about each term's meaning in contentious and non-contentious contexts, as well as the identity characteristic(s) it targets. This curated resource provides reliable grounding for the LLM, improving both detection accuracy and explanation quality while ensuring coverage of evolving or rare expressions that LLMs may fail to recognize. Our system integrates two complementary pipelines: a term-based pipeline that uses lemmatization and string matching to identify potentially problematic terms and leverages an LLM to disambiguate their meaning in context; and a term-free pipeline in which an LLM detects content explicitly targeting individuals or groups based on identity characteristics. The outputs of both pipelines are fused by an LLM to generate grounded explanations that clarify why specific content is flagged. We evaluate our approach through a human study on English, French, and Greek texts, assessing both detection performance and explanation quality, and show that our hybrid system outperforms LLM-only baselines.

2 Related Work

Online abusive and harmful language became a prominent challenge with the rise of early internet communities, where moderation was handled manually by platform administrators, forum moderators, and community managers. As online platforms and social media networks grew, the volume and velocity of user-generated content made manual moderation unsustainable, motivating the development of automated approaches. Early systems relied on traditional machine learning models such as SVMs (Warner and Hirschberg, 2012), often combined with curated dictionaries (Tulkens et al., 2016) or ensemble models (Burnap and Williams, 2014). With the emergence of deep learning, convolutional and recurrent neural networks (CNNs, LSTMs) were applied to classify abusive or hateful content (Del Vigna et al., 2017; Mathur et al., 2018; Meyer and Gambäck, 2019; Chakrabarty et al., 2019; Modha et al., 2018), followed more recently by transformer-based architectures and large language models (LLMs), which currently set the state of the art in detecting toxic language (Elmadany et al., 2020; Alonso et al., 2020; Davidson et al., 2020; Yao et al., 2024; Plaza-del Arco et al., 2023; Vargas et al., 2026).

Alongside methodological advances, many studies have focused on developing resources to support hate speech detection and contextualization. Structured frameworks for classifying offensive content provide a basis for systematic analysis (Banko et al., 2020; Kurrek et al., 2020), while lexicons, vocabularies, and annotated datasets offer collections of terms associated with abusive or derogatory speech for training and evaluation (Tonneau et al., 2024; ElSherief et al., 2021; Sap et al., 2020). Building on this line of work, we created curated vocabularies in English, French, and Greek containing terms inherently offensive or degrading toward identity-based groups. Unlike many existing lexicons that list only offensive terms, our vocabularies include the meaning and nuances of each term in both contentious and non-contentious contexts, along with the identity characteristic it targets (e.g., religion, sexual orientation), providing a foundation for grounded, context-aware analysis.

Despite these resources, automated detection remains challenging due to contextual and linguistic nuances. Many data-driven models exhibit biases (Wiegand et al., 2019; Davidson et al., 2019; Xia et al., 2020; Zhang et al., 2020; Sap et al., 2019) and robustness issues (Kaushik et al., 2020; Sen et al., 2022; Korre et al., 2023), which limit fairness and applicability in real-world settings. Incorporating context is essential to reduce false positives and improve detection quality (Kennedy et al., 2020; Bourgeade et al., 2024), an area where LLMs have shown particular promise.

Beyond detection accuracy, there is growing interest in approaches that justify and contextualize

[4] https://www.wiktionary.org/

hate speech moderation. Traditional flagging or removal of content, while reducing exposure to offensive material, often provides users with no rationale, limiting understanding and raising concerns about unfair censorship. Explanatory approaches aim to bridge this gap, helping users recognize and avoid offensive language. Recent works employ LLMs and curated rsources to generate rationales or highlight problematic segments (Epstein et al., 2022; Nirmal et al., 2024; Yang et al., 2023; Huang et al., 2023; Piot and Parapar, 2025; Menis Mastromichalakis et al., 2025). However, most such approaches focus on English and other high-resource languages, with lower-resource languages such as Greek receiving less attention. While automated approaches often perform well in high-resource languages, they may fall short in lower-resource contexts. Creating resources for these languages, such as the curated vocabulary presented here, can boost detection accuracy and explanation quality.

3 Methodology

3.1 Aspects of Hate Speech

The conceptualization of hate speech relies on the identity characteristics of the individuals or groups being targeted. For our analysis, we consider the following characteristics, based on prior research and online datasets (ElSherief et al., 2018; Ousidhoum et al., 2019): *gender, sexual orientation, race, ethnicity, religion, political affiliation, socioeconomic status, occupation, age, disability, addiction,* and *physical appearance*.

Drawing on these identity characteristics, hate speech manifests in two main ways. One relies on inherently derogatory terms that carry offensive meaning on their own, while the other involves hostile or demeaning expressions directed at individuals or groups specifically because of their identity, even when no explicit slur is present. Capturing both manifestations is essential to understanding the full spectrum of harmful language, including explicit and implicit expressions. We describe each in detail below, providing examples and highlighting the role of context in determining harmfulness.

Use of inherently derogatory terms: This refers to words or expressions whose meaning is offensive or degrading toward a specific group. Such terms include slurs and insults that are inherently tied to an identity characteristic, even when not directed at a member of that group. For example, the Greek term "αδελφή", when used with the meaning of "sissy" or "faggot", is derogatory toward gay people. It can be used as a direct insult against members of the queer community, but it can also appear in contexts where it is not aimed at a person because of their sexual orientation (e.g., saying "don't be such a sissy" to a friend). Even in these cases, the term perpetuates stereotypes about gay people and is therefore considered derogatory. Context is crucial because some words have neutral meanings (e.g., "αδελφή" also means "sister" or "nun") but also reclamation by targeted communities can change how they are perceived and used in different environments.

Language directed against groups or individuals based on their identity: This refers to expressions that explicitly attack a group or an individual because of their identity, even without inherently derogatory terms. For example, "He should be arrested and jailed; he's a Muslim, so he's a terrorist and a danger to public safety" targets an individual based on their religion, perpetuating a harmful stereotype that all Muslims are terrorists.

Often, both forms occur together. Calling a woman a "bitch" in "That bitch is always complaining, why can't women be more rational like us?" uses an inherently derogatory term while targeting her gender. Similarly, "That nigga always causes trouble, like the rest of them.", combines an inherently offensive term with a racial stereotype.

3.2 The Vocabularies

To address hate speech in the form of inherently derogatory terms, we constructed three dedicated vocabularies, one for each language in our study: English, French, and Greek. Each vocabulary contains terms and expressions that are, by themselves, derogatory or offensive toward a group or its members, providing a systematic basis for detection. Each entry includes: (1) the term; (2) a description covering all meanings (both offensive and non-offensive); (3) the identity characteristic(s) targeted; and (4) a link to the source repository. A detailed example is provided in Appendix A.

We selected Wiktionary as the source for our vocabularies, as it provides broad coverage across languages and has been shown to be "*a reliable and linguistically rich resource whose collaboratively constructed entries show a quality largely comparable to expert-made dictionaries*" (Meyer and Gurevych, 2012). Its open availability ensures transparency and reproducibility, while its active community continuously updates entries, keeping

the resource current with evolving language. Wiktionary also provides structured metadata, including usage labels such as "derogatory", "offensive", and "vulgarities", which can facilitate the systematic identification of inherently derogatory terms.

Our methodology for building the vocabularies involved five steps:

1. Initial collection: Retrieved Wiktionary terms tagged with relevant labels via the Wiktionary API, parsing and cleaning the returned HTML. This yielded 11,310 English, 3,749 French, and 965 Greek terms. See Appendix A for more details.

2. Filtering: Retrieved terms were filtered to remove common slurs and terms that are not inherently derogatory toward a group. For Greek, filtering was performed manually by human experts, while for English and French we used LLM-assisted filtering with Claude Sonnet 3.7[5], with human supervision and sampling validation.

3. Categorization: Tagged filtered terms with the identity characteristics they target. Manual categorization was performed for Greek, while English and French used LLM-assisted categorization with human supervision and sampling validation.

4. Generation of enriched description: All Wiktionary descriptions for each filtered term were fed to Claude Sonnet 3.7 with an appropriate prompt to generate continuous text that describes the meaning(s) of the term with an emphasis on why and under which circumstances the term is used in an offensive way (detailed prompts in Appendix C).

5. Human Validation: An expert team of 2 legal professionals[6] reviewed and corrected, where needed, the LLM outputs from steps 2–4. The Greek vocabulary was fully validated, while for English and French a representative sample (10% of entries) was reviewed. The experts found results satisfactory in over 90% of cases; in most of the remaining cases, the outputs were not wrong but required minor corrections.

The resulting vocabularies contain 3,904 English, 1,644 French, and 288 Greek entries. See Appendix B for the complete breakdown by category and language. The vocabularies are provided in the supplementary materials and will be made openly accessible upon publication.

[5] claude-3-7-sonnet-20250219-v1:0

[6] The experts hold a Master of Laws degree and work on matters related to hate speech, online abusive speech, anti-discrimination law, and AI-related regulatory and governance issues.

3.3 System Architecture

Our detection system implements two complementary objectives through parallel processing pipelines, as shown in Figure 1. The first pipeline, which we refer to as the *term-based* pipeline, detects inherently derogatory terms using the curated vocabularies. It combines string matching with LLM-powered semantic disambiguation. Terms are detected by matching lemmatized forms of the terms in the input text. Advanced matching logic handles multiple matches (retaining the longest) and grammatical variations sharing the same lemma (retaining the shortest Levenshtein distance). An LLM then resolves semantic ambiguity, determining whether the detected term is used contentiously. The LLM receives the source text and the vocabulary description for each term and outputs a decision along with a free-text explanation of why the term is offensive in context.

The second pipeline, which we refer to as the *term-free* pipeline, detects hate speech targeting individuals or groups without relying on specific vocabulary terms. It leverages an LLM grounded in curated identity characteristics to decide whether a text contains hate speech directed at individuals or groups based on their identity or beliefs, including slurs and other expressions not captured in the vocabularies. The LLM receives the source text, the identity characteristics, and dedicated instructions. We consider this pipeline as the system's baseline as it mirrors how many LLM detections systems operate, using a system prompt that describes the task without augmenting it with an external lexical resource.

A text is classified as free of hate speech only if both pipelines agree. If one pipeline detects hate speech, the text is flagged and the corresponding output is returned. If both pipelines flag it, an LLM fuses the outputs, removing redundancy and producing a coherent, unified explanation.

For comparative evaluation, we deployed two LLM configurations per language: Claude Sonnet 3.7 as the proprietary large model, and Llama-family variants as smaller open-weight alternatives. Language-specific LLMs included Llama-Krikri-8B-Instruct [7] for Greek and Hermes 3 Llama 3.1 8B [8] for English and French (selected after standard Llama-3.1-8B exhibited excessive guardrail

[7] https://huggingface.co/ilsp/Llama-Krikri-8B-Instruct

[8] https://huggingface.co/NousResearch/Hermes-3-Llama-3.1-8B

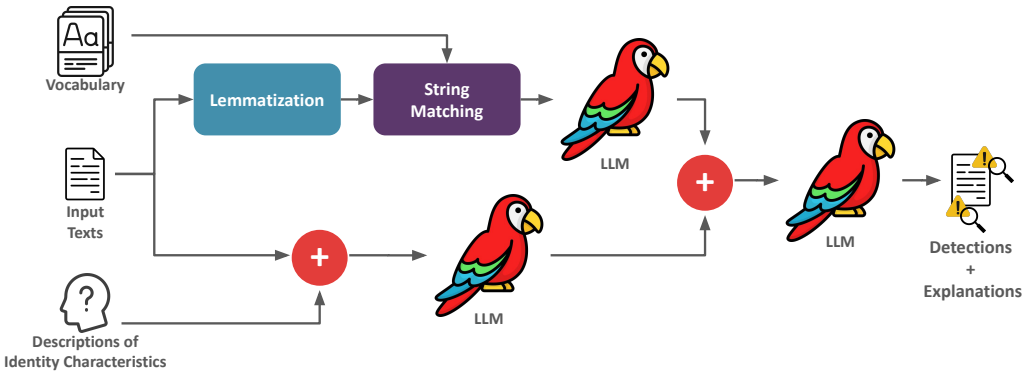

Figure 1: Overall system architecture illustrating the two pipelines for hate speech detection and explanation.

restrictions). The models from the Llama family were selected as a low-resource open-weight option as they can be deployed on a commercial GPU with at least 24GB of VRAM, while Claude Sonnet 3.7 was selected to test whether a sufficiently large model would implicitly internalize the kinds of contextual knowledge that our vocabulary explicitly provides.

Lemmatization employed Stanford Stanza (Qi et al., 2020) for English and French, and both Stanza and the ILSP lemmatizer (Prokopidis and Piperidis, 2020) for Greek, which has demonstrated superior performance for Greek terms.

Initial prompts were refined through qualitative evaluation of early outputs. Key enhancements included chain-of-thought formatting for term disambiguation, explicit instructions for reclaimed language, and guidelines for distinguishing direct from indirect speech and quotations (treated as non-hateful). While these refinements improved performance and explanation clarity, extensive prompt engineering or comparison of alternative prompting strategies was beyond the scope of this work. Cross-language evaluation showed no significant quality differences between prompts written in the target language and those with English instructions. See Appendix C for English prompts.

4 Experiments

To assess the effectiveness of our system, we conduct a series of experiments that evaluate both its detection performance and the quality of the explanations it produces. Beyond assessing overall system behaviour, we also examine how the different components of our approach contribute to these results. All experiments are conducted on a multilingual dataset that we constructed through manual annotation.

4.1 Evaluation Dataset

For our evaluation, we required real online texts that contain hate speech according to the definition adopted in this work, while also ensuring comparability across English, Greek, and French. To guide dataset selection, we established four criteria. First, the texts should be authentic user-generated content. Second, they should include cases of identity-targeting hate speech rather than only general toxicity or personal insults. Third, the datasets for the three languages should contain similar types of texts to maintain a consistent evaluation setup. Fourth, the selected texts should provide wide coverage of different identity characteristics, so that the evaluation reflects the diverse ways hate speech can manifest.

While several resources exist for English and French, Greek datasets are limited. Moreover, many available datasets across all three languages focus on offensive or toxic language more broadly, meaning they often include profanity or personal insults that do not meet our definition of hate speech. Some are also topic-specific (for example, restricted to sexism or racism), which limits coverage of the identity characteristics.

Taking these considerations into account, we selected the Greek subset of OffensEval2020 (Zampieri et al., 2020), the English Hate Speech Superset, and the French Hate Speech Superset (Tonneau et al., 2024) as our source datasets. Because the Greek dataset contains only tweets, we also restricted the English and French selections to tweets to maintain consistency across languages. These datasets provide suitable material containing both hateful and non-hateful content spanning a range of topics and target groups. It needs to be noted that these tweets do not contain user metadata such as gender, race, etc. which can influ-

ence the presence of hate speech, as in some edge cases slurs are present but used in a self-referential reclamatory way. Since these metadata are often not available in real world scenarios, we consider the pure-text setting to have wider applicability, and in practice contextual linguistic cues like tone, target, syntactic framing, and explicit references are often sufficient to distinguish reclaimed from derogatory use. Moreover, the original labels do not always align with our definition of hate speech (see Appendix E). Consequently, the original labels could not be used as ground truth.

To obtain reliable evaluation data, we selected and manually annotated 1,600 texts from the source datasets described above, 600 for English, 400 for French, and 600 for Greek. The selection ensured wide coverage of all identity characteristics and a roughly balanced distribution of hateful and non-hateful examples. This was achieved by considering the topic-related and toxicity labels in the original datasets to sample diverse and representative items across the three languages.

For the manual annotation, we recruited 18 annotators, all members or collaborators of organizations working on topics related to hate speech, digital rights, cyberbullying, and other online safety issues. Each text in the evaluation dataset was annotated independently by three annotators, who assigned one of three labels: *Yes* (contains hate speech), *Unsure*, or *No* (does not contain hate speech). Details on the inter-rater agreement between annotators are provided in Appendix D.

Based on the annotators' responses, we created our main evaluation dataset using majority voting. Since each text was annotated by multiple individuals, we assigned a final label when the annotators leaned in one direction (e.g. two *Yes* and one *No*, or one *Yes* and two *Unsure*), discarding only the fully divergent cases (e.g., one *Yes*, one *No*, and one *Unsure*). To gain further insights into the behavior of our system, such as whether it tends to be more "strict" or more "permissive", we also constructed three additional variants of the evaluation dataset. These variants differ in how they resolve borderline cases, defined as instances in which at least one annotator selected *Unsure* or where disagreement occurred. The three variants are:

– *Safe*: ignores all borderline cases.

– *Permissive*: if no evaluator has identified the text as positive, it is considered negative; otherwise it is considered positive. This favors treating borderline cases as non–hate speech.

– *Strict*: if no evaluator has identified the text as negative, it is considered positive; otherwise it is considered negative. This favors classifying borderline cases as hate speech.

See Appendix E for additional statistics on the constructed evaluation datasets, including class distributions and their (dis)agreement with the original dataset labels.

4.2 Accuracy of Detection

We evaluate the performance of our system with the 2 different LLM configurations across the three languages (English, French, and Greek) and the different variants of the annotated evaluation dataset.

4.2.1 Overall Evaluation

Table 1 reports precision, recall, and F1-score for each language, dataset variant, and LLM configuration. Overall, our results demonstrate the strong performance of the hybrid approach and reveal interesting patterns across models, languages, and dataset handling strategies.

Model Comparison Across English and French, Claude consistently achieves higher precision, recall, and F1-score than the Llama-based models, highlighting its effectiveness on higher-resource languages. For Greek, the situation is more nuanced: the Llama-based KriKri model outperforms Claude in terms of precision, while Claude exhibits higher recall. Despite these differences, the overall F1-scores for the two models are comparable in Greek, differing by only 2–3%, whereas in English and French the differences can reach up to 10%.

Language Comparison Focusing on overall performance (F1-score), the Llama-based models show comparable results across all three languages, indicating consistent behavior regardless of language. In contrast, Claude exhibits a notable drop in Greek, achieving F1-scores close to the much smaller Llama-based model. This difference is likely due to Claude's reduced capabilities in Greek, which is a lower-resource language compared to English and French. The drop in Claude's performance in Greek is mainly driven by lower precision. However, lower Precision in Greek is observed for both models compared to the other languages. This reduction in precision indicates that both models tend to over-flag texts as hateful in Greek, highlighting the difficulty of balancing sensitivity and specificity in a lower-resource language context.

		Safe			Majority			Permissive			Strict		
Lang	Model	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
EN	Claude	**0.97**	**0.90**	**0.93**	**0.92**	**0.89**	**0.90**	**0.47**	**0.90**	**0.62**	**0.98**	**0.78**	**0.87**
	Llama-based	0.84	0.80	0.82	0.82	0.82	0.82	0.42	0.80	0.55	0.92	0.74	0.82
FR	Claude	**0.97**	**0.90**	**0.94**	**0.96**	**0.91**	**0.93**	**0.78**	**0.91**	**0.84**	**0.98**	**0.81**	**0.89**
	Llama-based	0.93	0.74	0.83	0.92	0.73	0.81	0.77	0.74	0.76	0.95	0.65	0.77
EL	Claude	0.76	**0.95**	0.84	0.75	**0.92**	0.83	0.40	**0.95**	0.56	0.87	**0.82**	**0.84**
	Llama-based	**0.83**	0.90	**0.86**	**0.82**	0.86	**0.84**	**0.45**	0.90	**0.60**	**0.91**	0.71	0.80

Table 1: Precision, recall, and F1-score across dataset variations, models, and languages. Bold values indicate the best score between the two models for each metric.

Dataset Variant Comparison Comparing the different strategies for handling borderline cases, we observe that the system performs better under the *Strict* variant rather than *Permissive*, indicating that it is sensitive in flagging texts as hate speech. However, this sensitivity does not lead to over-flagging, as performance on the *Majority* variant, which is considered our primary gold standard, and the *Safe* variant, which is more "neutral", is equally strong, if not slightly higher. This demonstrates that our system balances accurate detection with careful handling of uncertain cases.

4.2.2 Pipeline Analysis

Table 2 presents the precision, recall, and F1-score of the two pipelines (term-based and term-free), along with the fused hybrid system, across model configurations, dataset variations and languages.

Across all languages and model configurations, the fused system consistently outperforms the individual pipelines in terms of overall performance (F1-score), showcasing that enhancing prompts with term descriptions from curated resources aids even large models, such as Claude Sonnet, despite their extensive implicit knowledge. The only exception is the English Llama-based model, where the fused system and the term-free pipeline obtain the same F1-score. This confirms the value of combining complementary signals from both pipelines.

As expected, the term-based pipeline shows high precision but substantially lower recall. This reflects its design: it captures only cases that involve inherently derogatory terms present in the curated vocabularies. Its contribution is therefore narrow but reliable. In contrast, the term-free pipeline covers a much broader space of hateful expressions and achieves higher recall.

We experimented with strengthening the term-free pipeline by explicitly prompting it to identify inherently derogatory terms in addition to broader hate speech. While this slightly improved its precision and recall, the term-free pipeline alone was still unable to reach the performance level of the fused system. This highlights the benefit of integrating the two complementary approaches rather than relying exclusively on a single LLM-based classifier.

Finally, it is important to note that the contribution of the term-based pipeline is not fully reflected in "macroscopic" evaluation metrics. Many commonly used slurs are already well covered by large language models, especially in high-resource languages, which limits the measurable impact of lexical resources when averaged across large datasets. However, the term-based pipeline plays a crucial role in capturing less frequent or newly emerging derogatory terms, which are underrepresented in existing datasets. It also strengthens the system's robustness against evolving or intentionally obfuscated hateful expressions, an area where prompting-based approaches alone are more vulnerable. The hybrid design therefore, provides both better performance and better long-term adaptability.

4.3 Quality of Explanations

Beyond detection accuracy, evaluating the quality of the explanations produced by our system is a key aspect of our study. To this end, the 18 annotators who contributed to the evaluation dataset assessed the explanations generated by our system for the detections on the texts of the evaluation dataset. Each annotator reviewed explanations produced by both LLM configurations (Claude and the Llama-based models), presented in a mixed order. The evaluators provided feedback along four dimensions:

- *Explanation quality (5 point Likert scale)*: How well the explanation reflects why the text constitutes hate speech, considering relevance, completeness, and correctness.

- *Explanation fluency (5 point Likert scale)*: How well-written the explanation is, in terms

Lang	Model	Type	Safe			Majority			Permissive			Strict		
			Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score	Precision	Recall	F1-Score
EN	Claude	TF	0.99	0.83	0.90	0.93	0.76	0.84	0.52	0.83	0.64	0.99	0.67	0.80
		TB	0.92	0.22	0.36	0.94	0.29	0.44	0.40	0.22	0.30	0.97	0.23	0.37
		F	0.97	0.90	0.93	0.92	0.89	0.90	0.47	0.90	0.62	0.98	0.78	0.87
	Llama-based	TF	0.90	0.77	0.83	0.87	0.78	0.82	0.46	0.77	0.57	0.95	0.68	0.79
		TB	0.80	0.30	0.44	0.81	0.36	0.50	0.36	0.30	0.33	0.91	0.32	0.48
		F	0.84	0.80	0.82	0.82	0.82	0.82	0.42	0.80	0.55	0.92	0.74	0.82
FR	Claude	TF	0.97	0.88	0.92	0.96	0.86	0.91	0.79	0.88	0.83	0.97	0.78	0.86
		TB	1.00	0.29	0.45	1.00	0.29	0.45	0.82	0.29	0.43	1.00	0.26	0.41
		F	0.97	0.91	0.94	0.96	0.91	0.93	0.78	0.91	0.84	0.98	0.81	0.89
	Llama-based	TF	0.94	0.55	0.70	0.93	0.53	0.67	0.79	0.55	0.65	0.95	0.48	0.63
		TB	0.94	0.36	0.52	0.94	0.36	0.52	0.79	0.36	0.50	0.95	0.31	0.47
		F	0.93	0.74	0.83	0.92	0.73	0.81	0.77	0.74	0.76	0.95	0.65	0.77
EL	Claude	TF	0.78	0.92	0.85	0.76	0.87	0.81	0.42	0.92	0.58	0.88	0.75	0.81
		TB	0.84	0.29	0.43	0.87	0.30	0.44	0.42	0.29	0.34	0.92	0.25	0.39
		F	0.76	0.95	0.84	0.75	0.92	0.83	0.40	0.95	0.56	0.87	0.82	0.84
	Llama-based	TF	0.88	0.80	0.84	0.85	0.75	0.80	0.50	0.80	0.62	0.93	0.59	0.72
		TB	0.78	0.30	0.43	0.81	0.32	0.46	0.37	0.30	0.33	0.89	0.28	0.43
		F	0.83	0.90	0.86	0.82	0.86	0.84	0.45	0.90	0.60	0.91	0.71	0.80

Table 2: Precision, recall, and F1-score of the Term Free (TF), Term Based (TB), and Fused (F) pipelines.

of grammar, syntax, and readability.

- *Feedback on explanation*: Categorical labels including "Irrelevant Information", "Too vague", "Duplication of Information", "Incorrect Details", "Too Verbose", and "Other".
- *Additional comments*: Optional free-text field.

Table 3 presents the average ratings and the most frequent issues (Issue 1 being the most frequent, and Issue 2 the 2nd most frequent) reported by annotators. Overall, explanations in English and French achieved high scores for both content and fluency (around 4 or higher), while Greek received lower scores. This reduction is likely due to a combination of factors: lower precision in Greek led to more false positives, which in turn caused the models to generate explanations that included irrelevant or incorrect information, lowering content quality.

The two LLM configurations produced explanations of comparable quality. Claude generally scored slightly higher in both content and fluency, though differences were modest. Fluency ratings were relatively high across all languages, with a minor decrease in Greek, reflecting the models' reduced effectiveness in this lower-resource language, which aligns with our findings in the quantitative evaluation of Section 4.2 where we saw Claude's performance dropping significantly in Greek. Content scores showed larger variation, largely reflecting misclassifications: explanations generated to justify incorrect classifications naturally received lower content ratings, even if their fluency remained acceptable.

Analysis of the categorical feedback reveals that

Lang	**Model**	Content	Fluency	Issue 1	Issue 2
EN	Claude	4.24 ± 1.27	4.66 ± 0.70	Irrelevant Info	Too Verbose
	Llama based	3.92 ± 1.43	4.59 ± 0.71	Irrelevant Info	Other
FR	Claude	4.45 ± 0.67	4.65 ± 0.61	Too Verbose	Irrelevant Info
	Llama based	4.40 ± 0.76	4.66 ± 0.59	Too Verbose	Irrelevant Info
EL	Claude	3.35 ± 1.13	4.12 ± 0.60	Irrelevant Info	Too Verbose
	Llama based	3.19 ± 1.14	4.04 ± 0.65	Irrelevant Info	Too Verbose

Table 3: Explanations evaluation

the most common issues across all languages were "Irrelevant Info" and "Too Verbose". These findings align with our observation that explanations for misclassified texts tended to include additional information to justify the model's decision, sometimes introducing irrelevant or excessive details. In some cases, this behavior reflects the model attempting to compensate for uncertainty, rather than outright hallucination. Free-text feedback also indicated that explanations occasionally exaggerated the level of toxicity in a text, which is consistent with our observation that the system is generally more eager to flag potential hate speech rather than under-flagging, while still maintaining good overall metrics. For additional details regarding the feedback, including the distribution of reported issues across languages and models, we refer the reader to Appendix F.

5 Conclusions

In this work, we presented a hybrid system that combines LLMs with curated domain knowledge and curated vocabularies for detecting hate speech

and generating contextual explanations across Greek, English, and French. Our evaluation demonstrates strong performance in both detection and explanation quality. The evaluation also offered valuable insights into the handling of borderline cases, revealing a tendency of our system to flag marginal expressions as positive (hate speech), without, however, over-flagging, as our system achieved good metrics in the majority variant of our evaluation dataset, which is used as a gold standard. We also analyzed the contribution of the different pipelines, showing the benefit of our hybrid approach over LLM-only approaches, and the complementarity of the 2 pipelines. A key contribution of our work is the development of three vocabularies of inherently offensive terms, along with information about their meanings and related identity categories, a valuable resource for domain adaptation and similar initiatives. These vocabularies, along with the annotated evaluation dataset and the source code of the system are available at https://github.com/ails-lab/detoex.

We see two main directions for the further development of this work. The first is expanding coverage to additional languages by creating new vocabularies. The openly available vocabularies, construction process, and detection system also provide a foundation for the broader community to build upon, whether by developing systems in other languages, enriching further the existing ones, or finding new applications. The second direction is improving the system itself, where we intend to explore techniques such as Retrieval-Augmented Generation, self-evaluation, and consistency-based methods to enhance robustness and reliability.

Limitations

The following limitations of the current work should be acknowledged:

- The preparation of the vocabularies of derogatory terms and corresponding categories and descriptions highly relied on an automatic process. In the case of Greek, all terms and associated information have been reviewed and improved, where necessary, by a human evaluator. For French and English, due to the high number of terms involved, human review was limited to a 10% sample of the vocabulary, with the results finding the automatically generated samples satisfactory in more than 90% of the cases. Thus, we expect the vocabularies to be of high quality; however, it's still possible that they may include inaccurate information, and some descriptions may suffer from inferior quality.

- The available ground truth data from previous initiatives does not fully align with the criteria following from the definition of hate speech we adopt. For Greek, the only dataset we were able to find includes annotations concerning toxic content, with only a subset of the positively labeled tweets actually being derogatory in relation to an identity characteristic. A similar remark holds for the French Hate Speech Superset, which also contains a significant number of aggressive language examples that do not constitute hate speech according to our definition. For English, we observe an opposite trend: a significant subset of the tweets drawn from the Hate Speech Superset is annotated as negative, even though they are clearly derogatory toward an identity characteristic according to our definition. These discrepancies may stem from annotation errors or other types of inconsistencies, such as those described in (Vidgen et al., 2020), or may reflect differences in how hate speech is interpreted. For these reasons, we were unable to use the labels of the datasets as ground truth for our evaluation, and we had to rely solely on human input to accurately calculate the performance of our system. However, the selection of annotators who have experience with the topic of hate speech gives us confidence in the quality of the annotations, giving as an high-quality evaluation dataset.

- Conducting a fair and reliable comparative evaluation with other hate speech detection systems is particularly challenging. As discussed above, issues related to ground truth data limit the availability of suitable benchmark datasets, especially for Greek and French. Moreover, existing systems vary widely in scope. As identified in previous work (Davidson et al., 2017), many systems do not distinguish between generally offensive language (e.g., personal insults, individual bullying), which our system treats as negative examples, and hate speech, which targets groups defined by identity characteristics. At the same time, several systems (Vidgen et al.,

2020; Chiril et al., 2020) focus only on a subset of identity characteristics (e.g., gender). Owing to these limitations, we rely on human validation for the evaluation presented in this work.

Ethical Considerations

Before conducting the human evaluation, all participants were thoroughly briefed on the task through an introductory workshop organized by the authors. During this session, we explained the evaluation procedure step-by-step, demonstrated example cases, and addressed questions to ensure that all evaluators clearly understood the objectives and evaluation criteria. The session was recorded and made available to participants for future reference, and a dedicated communication channel remained open throughout the evaluation period for any additional clarification or support.

Given the potentially disturbing nature of the content, participants were warned about the presence of offensive or harmful language before beginning the task. The study was designed and supervised by experts in abusive language and hate speech research, who have been actively involved in education and initiatives related to harmful language. These experts also facilitated the workshop and were available for ongoing consultation.

Participants were informed that their participation was voluntary and that they could withdraw from the study at any time without consequence. While the experts involved in evaluating the vocabularies, and designing and overseeing the human evaluation were properly compensated, the evaluators themselves participated on a voluntary basis, due to limited funding availability.

Acknowledgments

The work presented in this paper has been cofunded by the European Commission under the project "DEtection of TOxic and hateful speech with EXplanations -DETOEX" (UTTER - Unified Transcription and Translation for Extended Reality Agreement No. 101070631- HE; Grant Agreement No. 10039436 - UKRI; FSTP). We would also like to express our sincere gratitude to all the evaluators who participated in the DETOEX evaluation process. Their valuable contributions were instrumental to this work and greatly appreciated.

References

Pedro Alonso, Rajkumar Saini, and György Kovács. 2020. Hate speech detection using transformer ensembles on the hasoc dataset. In *International conference on speech and computer*, pages 13–21. Springer.

Michele Banko, Brendon MacKeen, and Laurie Ray. 2020. A unified taxonomy of harmful content. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 125–137.

Tom Bourgeade, Zongmin Li, Farah Benamara, Véronique Moriceau, Jian Su, and Aixin Sun. 2024. Humans need context, what about machines? investigating conversational context in abusive language detection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8438–8452, Torino, Italia. ELRA and ICCL.

Peter Burnap and Matthew Leighton Williams. 2014. Hate speech, machine classification and statistical modelling of information flows on twitter: Interpretation and communication for policy decision making.

Tuhin Chakrabarty, Kilol Gupta, and Smaranda Muresan. 2019. Pay "attention" to your context when classifying abusive language. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 70–79, Florence, Italy. Association for Computational Linguistics.

Patricia Chiril, Véronique Moriceau, Farah Benamara, Alda Mari, Gloria Origgi, and Marlène Coulomb-Gully. 2020. An annotated corpus for sexism detection in French tweets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*.

European Commission, Directorate-General for Justice, Consumers, and L. Kaati. 2023. *The European online hate lab*. Publications Office of the European Union.

Sam Davidson, Qiusi Sun, and Magdalena Wojcieszak. 2020. Developing a new classifier for automated identification of incivility in social media. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 95–101, Online. Association for Computational Linguistics.

Thomas Davidson, Debasmita Bhattacharya, and Ingmar Weber. 2019. Racial bias in hate speech and abusive language detection datasets. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 25–35, Florence, Italy. Association for Computational Linguistics.

Thomas Davidson, Dana Warmsley, Michael W. Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *International Conference on Web and Social Media*.

Fabio Del Vigna, Andrea Cimino, Felice Dell'Orletta, Marinella Petrocchi, and Maurizio Tesconi. 2017.

Hate me, hate me not: Hate speech detection on facebook. In *Proceedings of the first Italian conference on cybersecurity (ITASEC17)*, pages 86–95.

AbdelRahim Elmadany, Chiyu Zhang, Muhammad Abdul-Mageed, and Azadeh Hashemi. 2020. Leveraging affective bidirectional transformers for offensive language detection. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, with a Shared Task on Offensive Language Detection*, pages 102–108, Marseille, France. European Language Resource Association.

Mai ElSherief, Vivek Kulkarni, Dana Nguyen, William Yang Wang, and Elizabeth Belding. 2018. Hate lingo: A target-based linguistic analysis of hate speech in social media. In *Proc. of the 20th AAAI Conference on Web and Social Media (ICWSM)*.

Mai ElSherief, Caleb Ziems, David Muchlinski, Vaishnavi Anupindi, Jordyn Seybolt, Munmun De Choudhury, and Diyi Yang. 2021. Latent hatred: A benchmark for understanding implicit hate speech. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 345–363, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ziv Epstein, Nicolo Foppiani, Sophie Hilgard, Sanjana Sharma, Elena Glassman, and David Rand. 2022. Do explanations increase the effectiveness of ai-crowd generated fake news warnings? *Proceedings of the International AAAI Conference on Web and Social Media*, 16(1):183–193.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Chain of explanation: New prompting method to generate quality natural language explanation for implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023*, pages 90–93.

Divyansh Kaushik, Eduard Hovy, and Zachary Lipton. 2020. Learning the difference that makes a difference with counterfactually-augmented data. In *International Conference on Learning Representations*.

Brendan Kennedy, Xisen Jin, Aida Mostafazadeh Davani, Morteza Dehghani, and Xiang Ren. 2020. Contextualizing hate speech classifiers with post-hoc explanation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5435–5442, Online. Association for Computational Linguistics.

Katerina Korre, John Pavlopoulos, Jeffrey Sorensen, Léo Laugier, Ion Androutsopoulos, Lucas Dixon, and Alberto Barrón-Cedeño. 2023. Harmful language datasets: An assessment of robustness. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 221–230.

Klaus Krippendorff. 2018. *Content analysis: An introduction to its methodology*. Sage publications.

Jana Kurrek, Haji Mohammad Saleem, and Derek Ruths. 2020. Towards a comprehensive taxonomy and large-scale annotated corpus for online slur usage. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 138–149, Online. Association for Computational Linguistics.

Puneet Mathur, Rajiv Shah, Ramit Sawhney, and Debanjan Mahata. 2018. Detecting offensive tweets in Hindi-English code-switched language. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 18–26, Melbourne, Australia. Association for Computational Linguistics.

Orfeas Menis Mastromichalakis, Jason Liartis, Kristina Rose, Antoine Isaac, and Giorgos Stamou. 2025. Don't erase, inform! detecting and contextualizing harmful language in cultural heritage collections. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21836–21850, Vienna, Austria. Association for Computational Linguistics.

Christian M. Meyer and Iryna Gurevych. 2012. To exhibit is not to loiter: A multilingual, sense-disambiguated Wiktionary for measuring verb similarity. In *Proceedings of COLING 2012*, pages 1763–1780, Mumbai, India. The COLING 2012 Organizing Committee.

Johannes Skjeggestad Meyer and Björn Gambäck. 2019. A platform agnostic dual-strand hate speech detector. In *Proceedings of the Third Workshop on Abusive Language Online*, pages 146–156, Florence, Italy. Association for Computational Linguistics.

Sandip Modha, Prasenjit Majumder, and Thomas Mandl. 2018. Filtering aggression from the multilingual social media feed. In *Proceedings of the First Workshop on Trolling, Aggression and Cyberbullying (TRAC-2018)*, pages 199–207, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Ayushi Nirmal, Amrita Bhattacharjee, Paras Sheth, and Huan Liu. 2024. Towards interpretable hate speech detection using large language model-extracted rationales.

Nedjma Ousidhoum, Zizheng Lin, Hongming Zhang, Yangqiu Song, and Dit-Yan Yeung. 2019. Multilingual and multi-aspect hate speech analysis. In *Proc. of the 2019 Conf. on Empirical Methods in Natural Language Processing and the 9th Int. Joint Conf. on Natural Language Processing (EMNLP-IJCNLP)*, pages 4675–4684.

Paloma Piot and Javier Parapar. 2025. Towards efficient and explainable hate speech detection via model distillation. In *Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6–10, 2025, Proceedings, Part II*, page 376–392, Berlin, Heidelberg. Springer-Verlag.

Flor Miriam Plaza-del Arco, Debora Nozza, Dirk Hovy, and 1 others. 2023. Respectful or toxic? using zero-shot learning with language models to detect hate speech. In *The 7th Workshop on Online Abuse and Harms (WOAH)*. Association for Computational Linguistics.

Prokopis Prokopidis and Stelios Piperidis. 2020. A neural nlp toolkit for greek. In *11th Hellenic conference on artificial intelligence*, pages 125–128.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490.

Indira Sen, Mattia Samory, Claudia Wagner, and Isabelle Augenstein. 2022. Counterfactually augmented data and unintended bias: The case of sexism and hate speech detection. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4716–4726.

Manuel Tonneau, Diyi Liu, Samuel Fraiberger, Ralph Schroeder, Scott Hale, and Paul Röttger. 2024. From languages to geographies: Towards evaluating cultural bias in hate speech datasets. In *Proceedings of the 8th Workshop on Online Abuse and Harms (WOAH 2024)*, pages 283–311, Mexico City, Mexico. Association for Computational Linguistics.

Stéphan Tulkens, Lisa Hilte, Elise Lodewyckx, Ben Verhoeven, and Walter Daelemans. 2016. A dictionary-based approach to racism detection in dutch social media. *arXiv preprint arXiv:1608.08738*.

Francielle Vargas, Jackson Trager, Diego Alves, Surendrabikram Thapa, Matteo Guida, Berk Atil, Daryna Dementieva, Andrew Smart, and Ameeta Agrawal. 2026. Self-explaining hate speech detection with moral rationales. *Preprint*, arXiv:2601.03481.

Bertie Vidgen, Scott A. Hale, Ella Guest, Helen Margetts, David Broniatowski, Zeerak Waseem, Austin Botelho, Matthew Hall, and Rebekah Tromble. 2020. Detecting East Asian prejudice on social media. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 162–172.

William Warner and Julia Hirschberg. 2012. Detecting hate speech on the world wide web. In *Proceedings of the Second Workshop on Language in Social Media*, pages 19–26, Montréal, Canada. Association for Computational Linguistics.

Michael Wiegand, Josef Ruppenhofer, and Thomas Kleinbauer. 2019. Detection of abusive language: the problem of biased datasets. 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 602–608.

Mengzhou Xia, Anjalie Field, and Yulia Tsvetkov. 2020. Demoting racial bias in hate speech detection. In *Proceedings of the Eighth International Workshop on Natural Language Processing for Social Media*, pages 7–14, Online. Association for Computational Linguistics.

Yongjin Yang, Joonkee Kim, Yujin Kim, Namgyu Ho, James Thorne, and Se young Yun. 2023. Hare: Explainable hate speech detection with step-by-step reasoning.

Tsungcheng Yao, Ernest Foo, and Sebastian Binnewies. 2024. Personalised abusive language detection using llms and retrieval-augmented generation. In *Proceedings of the 7th International Conference on Natural Language and Speech Processing (ICNLSP 2024)*, pages 92–98.

Marcos Zampieri, Preslav Nakov, Sara Rosenthal, Pepa Atanasova, Georgi Karadzhov, Hamdy Mubarak, Leon Derczynski, Zeses Pitenis, and Çağrı Çöltekin. 2020. SemEval-2020 task 12: Multilingual offensive language identification in social media (OffensEval 2020). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 1425–1447, Barcelona (online). International Committee for Computational Linguistics.

Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and Tiejun Zhao. 2020. Demographics should not be the reason of toxicity: Mitigating discrimination in text classifications with instance weighting. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4134–4145, Online. Association for Computational Linguistics.

A Vocabulary Creation

We curated three vocabularies (one per the considered languages) that contain derogatory and offensive terms, which have at least one meaning that can be considered hate speech, i.e. terms that are per se derogatory and offensive towards a group and their members, based on their characteristics or beliefs. In this respect, the vocabularies were not meant to include common slurs, which may be directed against certain groups considered by our

categorization in certain contexts, but are not by themselves derogatory or offensive towards such a group (e.g., "asshole").

The vocabularies contain the following pieces of information:

- Term: the term that can be used as hate speech
- Description: Free-text description of all the meanings of the term (including both offensive and non-offensive ones). The description should mention why and under which circumstances the term is used in an offensive way.
- Category: The group or groups towards which the term is by itself offensive or derogatory, based on the groups defined in Section 3.1.
- Source: A link to the repository from which the term was sourced.

See Table 4 for an example.

We used the English, French and Greek Wiktionaries to create the respective vocabularies. This decision was primarily motivated by a lack of openly-available and extensive enough alternatives, especially in Greek.

In order to fetch an initial set of term candidates we made use of Wiktionary categories and tags that denoted a derogatory or insulting usage of a term. These categories and tags were:

- English
  1. Category:English derogatory terms
  2. Category:English vulgarities
  3. Category:English offensive terms
- French
  1. Catégorie:Termes péjoratifs en français
  2. Catégorie:Insultes en français
- Greek
  1. Μειωτικοί όροι (νέα ελληνικά) , μειωτικός , μειωτική , μειωτικό , μειωτικά
  2. Κατηγορία:Υβριστικοί όροι (νέα ελληνικά) , υβριστικός , υβριστική , υβριστικό , υβρισιτκά
  3. Κατηγορία: Χυδαιολογίες (νέα ελληνικά) , χυδαίος , χυδαία , χυδαίο
  4. βρισιά , βρισιές

The "query" action of the Wiktionary API was used to fetch pages. This is an example of an API call that fetches all terms under the category "Μειωτικοί όροι (νέα ελληνικά)" (derogatory terms in Greek):

```
GET https://el.wiktionary.org/w/api.php
Request Body:

{
    "action": "query",
    "list": "categorymembers",
    "cmtitle": "Κατηγορία:
        Μειωτικοί_όροι_(νέα_ελληνικά)",
    "cmprop": "title|ids",
    "format": "json",
    "cmlimit": "500"
}
```

This is an example of an API call that fetches all terms tagged as "υβριστικός" (insulting):

```
GET https://el.wiktionary.org/w/api.php
Request Body:

{
    "action": "query",
    "prop": "linkhere",
    "cmtitle": "υβριστικός",
    "cmprop": "title|ids",
    "format": "json",
    "lhlimit": "500"
}
```

This resulted in 11,310 English, 3,749 French and 965 Greek term candidates. Definitions were extracted from the term pages by parsing the HTML and detecting the sections containing definitions. Since the Wiktionary pages are only designed to be human-readable there is no consistent way to isolate the sections containing definitions, so some assumptions and work-arounds had to be made. Sections containing definitions almost universally start with a heading describing the part-of-speech of the word. A page may contain many such sections containing definitions (e.g. the term "αδερφή" contains definitions both under section "Ουσιαστικό" and "Κλιτικός τύπος επιθέτου"). All heading tags for all pages were collected and the ones referring to parts-of-speech were isolated. Other sections containing irrelevant information, such as the pronunciation of a term, were discarded. Within the sections kept, the term definitions are almost exclusively structured as a list (even if the section contains a single definition), so the HTML tags <ul> and <ol> were used to detect the lists of term definitions. For each definition, only plain text

Term	Description	Categories	Source
bitch	The term 'bitch' is primarily offensive when used to refer to women in a derogatory manner, implying they are aggressive, unpleasant, or overly assertive—traits that would often be viewed positively in men. It's also problematic when applied to men to suggest weakness or effeminacy, as this usage reinforces harmful gender stereotypes by equating femininity with inferiority. While the word has a neutral meaning when referring to female dogs, its use as a slur has overshadowed this definition in most contexts. In some LGBTQ+ communities and among close friends, the term has been reclaimed and may be used affectionately, but this usage is context-dependent and generally inappropriate for those outside these communities. The term's evolution from canine terminology to gendered insult reflects long-standing societal attitudes that devalue women and feminine characteristics.	Gender; Sexual Orientation	https://en.wiktionary.org/wiki/bitch

Table 4: A sample entry from the vocabulary.

was kept, discarding any links or tags contained in the HTML. For each term, all definitions collected were unified in a single numbered list and stored in a CSV file.

The next step was only keeping the terms that did actually have some derogatory usage and could be used as hate speech. Ideally, we would like to manually review all fetched terms. Although this proved feasible in Greek, with the help of some native speakers, it was not possible for English or French due to the sheer amound of the terms fetched. Instead, we opted for using Claude Sonnet 3.7 to filter the terms. Claude Sonnet 3.7 was also used to create the vocabulary description of each term by fusing the different Wiktionary definitions of each term into a single, coherent piece of text that describes all usages of the term, noting which ones are derogatory. See Appendix C for the prompt used for filtering and description creation. Claude's performance on filtering terms for English and French and constructing descriptions in all three languages was evaluated by humans. Evaluators agreed with over 90% of Claude's filtering and found the descriptions of sufficient quality.

A limitation that we inspected in some descriptions is that Wiktionary does not always sufficiently reflect probable recent shifts in the usage of terms in certain social environments, especially regarding the extent to which certain terms are used in a reclaimed manner (e.g. the reclaiming of "tranny" by trans people). As a result, the descriptions generated by Claude based on the Wiktionary definitions often treat such reclaimed usages as being derogatory. Where possible, this limitation has been addressed by adapting the automatic descriptions accordingly. Furthermore, the LLMs used for the tool's pipeline were also prompted to be cautious of such usages by the affected community itself, so as to avoid over-flagging (see Appendix C for the relevant prompts).

B Vocabulary Categories

In table 5 we present the number of terms with each category label in our vocabularies. Note that each term can have multiple category tags, hence the sum of each column is greater than the number of terms in the respective language.

C Prompts

Below we present the English prompts for all parts of the pipeline (term-based detection, term-free detection, and fusion of explanations), as well as vocabulary creation. The prompts for other languages can be found in the supplementary material and follow the exact same format, with the text having been translated by proficient speakers of the respective language.

Aspects of the prompts were adapted by manual evaluation of small samples, but without thorough benchmarking of prompting strategies and variations due to budgetary constraints and the different focus of this work. These are some examples of interest: The initial Term-Based prompt tended to

Category	English	French	Greek	Total
Addiction	38	20	8	66
Age	85	57	28	170
Disability	266	63	17	346
Ethnicity	1194	449	45	1688
Gender	1109	410	61	1580
Physical Appearance	201	80	26	307
Political Affiliation	774	382	28	1184
Public Institutions	261	148	13	422
Race	842	189	12	1043
Religion	400	177	10	587
Sexual Orientation	455	102	47	604
Socioeconomic	335	279	31	645
Other	11	7	23	41

Table 5: Counts of category tags for vocabulary terms by language

flag all occurrences of a term as harmful, even when it contextually carried a different meaning. Chain of thought reasoning was added to ensure that the term has been properly disambiguated before flagging the text, which seemed to improve results. Both the Term-Based and Term-Free pipelines were seen to flag quotations in several cases, one in particular being news headlines, so the explicit instruction to not flag quotations or indirect speech was added and it seemed to eliminate this error.

Term-based detection - System Prompt

You are an expert content moderator specializing in detecting hate speech in text. Your task is to analyze text and distinguish hateful from neutral uses of a specific term based on the following definition:
Hate speech refers to spoken or written communication that attacks or uses pejorative or discriminatory language with reference to a person or a group based on identity-related characteristics, including: Gender, Sexual orientation, Race, Ethnicity, Religion, Political affiliation, Socioeconomic status, Occupation, Age, Disability, Addiction, Physical appearance.
You will be given:

- A term
- A description of how that term can be used in hateful and non-hateful/neutral ways
- A piece of text containing this term
- One or more target characteristics the term may be offensive toward (e.g., "Sexual Orientation", "Ethnicity")

Your goal is to analyze the text and then decide if the term is used as hate speech. Use the following reasoning steps:

1. **Step 1**: If the description includes multiple possible meanings of the term, identify which meaning is used in the text. If disambiguation is particularly difficult, rely on non-hateful uses of the term. If it has only one clear meaning, write "Non ambiguous term". Do not evaluate the presence of hate speech yet.
2. **Step 2**: Based on the meaning you identified, consider whether the term corresponds to the hateful usage described earlier. Consider both the possibility of it being used in a hateful way and the possibility of it being used in a neutral/non hateful way.
3. **Step 3**: Decide whether the use of the term in the text is hateful or not and simply write "Hateful" or "Non hateful".
4. **Step 4**: Provide a clear, concise explanation (under 100 words) of your judgment. In your explanation use the phrasing provided in the term description you will be

given. Do not include, or refer to any previous Step.

Important considerations for analysis:

- **Indirect speech**: Any hate speech contained in the text as part of a quote or paraphrased from a different source should influence your decision significantly less or not at all.
- **Reclaimed language**: Some terms which are usually derogatory can be used in a reclaimed, empowering way by members of the same community they target. In these cases, the level of hate speech should be significantly lower or non-existent.
- **Self-identity versus targeting others**: Distinguish between someone who describes themself or their own community versus targeting others with the same language.

Format your output using XML tags as follows:
<STEP_1> [Step 1] </STEP_1>
<STEP_2> [Step 2] </STEP_2>
<STEP_3> [Step 3] </STEP_3>
<STEP_4> [Step 4] </STEP_4>

Term-based detection - User Prompt

Term: {}
Description: {}
Text: {}
Characteristics: {}

Non-term-based detection - System Prompt

You are an expert content moderator specializing in detecting hate speech in text. Your task is to analyze text and determine if it contains hate speech based on the following definition:
Hate speech refers to spoken or written communication that attacks or uses pejorative or discriminatory language with reference to a person or a group based on identity-related characteristics, including: Gender, Sexual orientation, Race, Ethnicity, Religion, Political affiliation, Socioeconomic status, Occupation, Age, Disability, Addiction, Physical appearance.
Important considerations for analysis:

- **Non-targeted speech**: Do not consider slurs as hate speech, unless they are directed towards an individual or group defined by their identity characteristics.
- **Indirect speech**: Any hate speech that the text contains as part of a quote or paraphrased from a different source should affect your decision significantly less or not at all.
- **Self-identity versus targeting others**: Distinguish between someone who describes themself or their own community versus targeting others with the same language.

Output Requirements: Provide a decision and concise explanation (under 100 words) covering:

- Which elements influenced your decision
- If hate speech was found, which identity characteristics are targeted
- Any ambiguities or nuances you considered
- Specific quotes from the text when necessary for your argument

Always use the exact XML format specified in the user prompt.

Non-term-based detection - User Prompt

Analyze the following text and evaluate whether it contains hate speech:
Text for analysis: {}
Please give your response exactly as specified in the following format:
<DECISION> ["Hate speech" or "Not hate speech"] </DECISION>
<EXPLANATION> [Explanation of your evaluation] </EXPLANATION>

Fusion - System Prompt

Task: Merge Hate Speech Analysis Texts
You will be given two or more separate texts describing hate speech content. Merge these analyses into a single, coherent description without redundancy.
Instructions:

1. Combine the information from all texts into a unified analysis
2. Reuse the existing text
3. Remove duplicate information
4. Reorganize for better flow
5. Maintain accuracy
6. Keep it focused
7. Keep it brief

Input Format: Text 1, Text 2, etc.
Output Format: Provide a single well-structured paragraph without opening/closing remarks.
Example:

- **Text 1:** The term "bitch" in this tweet is used as hate speech as it is part of a gender-based slur. The phrase aims to diminish and demean a woman through sexist language, linking her to derogatory references to sexual behavior and gendered stereotypes. The use of the term in this context violates basic principles of respect and gender equality.
- **Text 2:** The text contains hate speech that targets individuals based on their religion. Specifically, the term "diaperhead" is a derogatory and dehumanizing slur for Muslims, mocking traditional headwear like turbans or keffiyehs by comparing them to diapers. In addition, the text contains sexist language ("you stupid bitch") that targets the gender of the recipient and includes sexual insinuations of violence ("waiting to be fucked"). The language is directly offensive and targeted, without being a quotation or indirect speech.
- **Merged Output:** This particular text presents multiple levels of hate speech. The term "bitch" in this tweet is part of a gender-based slur. The phrase aims to diminish and humiliate a woman through sexist language and includes sexual innuendos of violence ("waiting to be fucked"). It is also a derogatory and dehumanizing characterization of Muslims, using the word "diaperhead", mocking traditional headwear like turbans or keffiyehs by comparing them to diapers.

Fusion - User Prompt

Please merge the following hate speech analyses: {}

Vocabulary Creation

You have an expert understanding of the English language and slang, and how it can be used in a derogatory manner to target individuals or groups through stereotypes, negative generalizations, or the use of identity-related markers (e.g., ethnicity, origin, profession) as a negative trait. This derogatory nature may be evident in the etymology or structure of the word (e.g., compound words using a component metaphorically to evoke a stereotype). Your task is to help the user create a vocabulary of **English terms that can constitute hate speech**. You will be given a term and one or more short descriptions (e.g., definitions or usage contexts) extracted from the English Wiktionary. Not all terms you will be given constitute hate speech. Your task is to determine:
1. Whether the term can constitute hate speech, based on the following definition: "Hate speech refers to spoken or written communication that attacks or uses pejorative or discriminatory language with reference to a person or a group based on identity-related characteristics. These characteristics include: gender, sexual orientation, race, ethnicity, religion, political affiliation, socioeconomic status, occupation, age, disability, addiction, and physical appearance." Words must be offensive due to their **meaning** and **harmful to a group** with shared identity characteris-

tics. Inherently degrading expressions qualify. Words that are merely offensive on an individual level, like common slurs, but not harmful to a social group, do **not** qualify. Some terms can have **neutral or non-offensive meanings** in other contexts. You should base your judgment on the provided description(s). If the description suggests the word is used in a derogatory sense targeting a group identity, it **can** constitute hate speech.
2. If the term **can** constitute hate speech, indicate which **category or categories** it targets, choosing from: Gender, Sexual orientation, Race, Ethnicity, Religion, Political affiliation, Socioeconomic status, Occupation, Age, Disability, Addiction, Physical appearance.
3. If the term **can** constitute hate speech, provide a vocabulary entry with clear, concise description for the term that explains:

- In which context(s) the term is considered offensive or inappropriate
- If and when the term can be used in a neutral or acceptable way
- Why or how the term came to acquire its derogatory meaning, if such information is available
- Any relevant linguistic, cultural, or historical background

Output your answer as JSON using the following fields:

- "reasoning": String with a short explanation of your decision in freeform text
- "hate_speech": Boolean (true or false)
- "categories": Array with one or more categories, or empty if false
- "vocabulary_entry": String, empty if false

Examples:
EXAMPLE 1
INPUT: TERM: welfare queen
DESCRIPTION(S): A woman collecting welfare, seen as doing so out of laziness, rather than genuine need.
OUTPUT: { "reasoning": "The term \"welfare queen\" is toxic because it's rooted in racist and sexist stereotypes that falsely portray people receiving government assistance as fraudulent and lazy, particularly targeting Black women.", "hate_speech": true, "categories": ["Race", "Socioeconomic Status"], "vocabulary_entry": "The term \"welfare queen\" refers to someone who allegedly exploits government welfare systems, but it's primarily considered derogatory and offensive. Popularized during Ronald Reagan's 1976 presidential campaign, the phrase has been widely criticized for perpetuating harmful stereotypes about welfare recipients, particularly women of color, and for reinforcing classist and racist assumptions about poverty. The term is inappropriate in respectful discourse about social policy, though it may appear in academic or analytical contexts when discussing political rhetoric or media representation, provided its problematic history is acknowledged." }
EXAMPLE 2
INPUT: TERM: nerd
DESCRIPTION(S):

- A person who is intellectual but generally introverted.
- One who has an intense, obsessive interest in something.
- A member of a subculture revolving around intellectualism, technology, video games, fantasy and science fiction, comic books and assorted media.
- One who is socially inept or unattractive, regardless of their intelligence; a social outcast.

OUTPUT: { "reasoning": "While potentially teasing, 'nerd' does not target an identity group based on protected characteristics.", "hate_speech": false, "categories": [], "vocabulary_entry": "" }
EXAMPLE 3
INPUT: TERM: slut
DESCRIPTION(S):

- A sexually promiscuous woman.
- A prostitute.
- Any sexually promiscuous person.

- Someone who seeks attention through inappropriate means or to an excessive degree.
- A disloyal individual; someone who does not commit to a particular thing.
- A slovenly, untidy person, usually a woman.
- A bold, outspoken woman.
- A female dog.
- A maidservant.
- A rag soaked in a flammable substance and lit for illumination.
- To wear slutty clothing or makeup, or otherwise behave in a slutty manner.
- To visit places frequented by men, with the intention of engaging in sexual intercourse by means of flirting.

OUTPUT: { "Reasoning": "The term slut has been used to stigmatized people for their sexual practices, especially women.", "hate_speech": true, "categories": ["Gender", "Sexual Orientation"], "vocabulary_entry": "The term \"slut\" is a derogatory word traditionally used to shame women for perceived sexual promiscuity or non-conformity to conventional sexual norms. Historically rooted in misogynistic attitudes that police women's sexuality while applying different standards to men, the term is considered offensive and inappropriate in most contexts because it perpetuates harmful double standards and slut-shaming. However, the word has undergone some reclamation efforts, particularly in feminist and LGBTQ+ communities, where individuals may use it self-referentially or positively to challenge sexual stigma—though this reclaimed usage should only be employed by those within these communities and with clear understanding of the context." }

D Inter-rater Agreement

Inter-rater agreement was measured using Krippendorff's alpha (Krippendorff, 2018), interpreting the three labels as interval values (0.0 for No, 0.5 for Unsure, 1.0 for Yes). Table 6 provides an overview of the agreement scores, along with the percentage of texts with strict disagreement (at least one '*Yes*' and one '*No*').

	Krippendorff's alpha	Percentage of tweets with disagreement between the raters
EN	0.54	0.29
FR	0.74	0.10
EL	0.65	0.22

Table 6: Inter-rater agreement

E Dataset Statistics

In Table 7 we present the percentage of positive (containing hate speech) samples in each variant of the evaluation dataset in each language. We see that the classes are relatively balanced, especially for the majority label.

	Safe	Majority	Permissive	Strict
EN	43%	50%	27%	64%
FR	52%	54%	43%	60%
EL	32%	38%	21%	55%

Table 7: Percentage of texts labeled as positive across dataset variations and languages

Table 8 shows the agreement between the labels that resulted from our annotation campaign and the labels of the original datasets. Agreement is generally low, especially for French. The English and French datasets seem to employ a more narrow definition of what constitutes hate speech, presenting higher agreement with the permissive labels. On the other hand, the Greek dataset flags more tweets as positive, which was expected as it is an offensive speech dataset rather than a hate speech one. This validates our choice of conducting annotation campaigns to acquire labels closer to our definition of hate speech, rather than rely on the original labels of the datasets.

	Safe	Majority	Permissive	Strict
EN	71%	66%	74%	53%
FR	58%	57%	58%	54%
EL	76%	72%	62%	72%

Table 8: Agreement rate between annotator labels and the original dataset labels

F Explanation Feedback

In Table 9, we report the percentages of issues noted by evaluators for explanations with lower content or fluency scores. Importantly, the feedback field was filled in only for these lower-scoring explanations, which occurred infrequently. The percentages in the table are calculated only over the explanations for which evaluators provided feedback, so each column sums to 100%, and they should not be interpreted as the proportion of all explanations in the dataset.

G Hardware

All inference using Claude was performed via Amazon Web Services. Inference using the Llama based models was done on a server with a NVIDIA GeForce RTX 4090 GPU. All other code was executed locally in a single-threaded setup on a laptop with an AMD Ryzen 7 5800H CPU and 16 GB of RAM.

	English		French		Greek	
Category	Claude	Llama	Claude	Llama	Claude	Llama
Duplication of Information	5.17%	2.78%	5.88%	12.50%	13.33%	6.56%
Incorrect Details	–	–	–	–	1.67%	1.64%
Irrelevant Information	**53.45**%	**69.44**%	17.65%	25.00%	**55.00**%	**57.38**%
Too Vague	–	–	–	–	1.67%	3.28%
Too Verbose	24.14%	12.50%	**76.47**%	**50.00**%	20.00%	24.59%
Other	17.24%	15.28%	–	12.50%	8.33%	6.56%

Table 9: Issues reported for the explanations accompanying hate speech detections per language and LLM