Unveiling the Hate: Generating Faithful and Plausible Explanations for Implicit and Subtle Hate Speech Detection

Greta Damo, Nicolás Benjamín Ocampo, Elena Cabrio, Serena Villata

Unveiling the Hate: Generating Faithful and Plausible Explanations for Implicit and Subtle Hate Speech Detection

Greta Damo[1], Nicolás Benjamín Ocampo[1] *, Elena Cabrio, and Serena Villata

Université Côte d'Azur, CNRS, Inria, I3S, France
greta.damo@etu.univ-cotedazur.fr
{nicolas-benjamin.ocampo, elena.cabrio, serena.villata}@univ-cotedazur.fr

Abstract. In today's digital age, the huge amount of abusive content and hate speech on social media platforms presents a significant challenge. Natural Language Processing (NLP) methods have focused on detecting explicit forms of hate speech, often overlooking more nuanced and implicit instances. To address this gap, our paper aims to enhance the detection and understanding of implicit and subtle hate speech. More precisely, we propose a comprehensive approach combining prompt construction, free-text generation, few-shot learning, and fine-tuning to generate explanations for hate speech classification, with the goal of providing more context for content moderators to unveil the actual nature of a message on social media.

Keywords: Hate Speech Detection · Generating Explanations · Implicit Hate Speech · Subtle Hate Speech

1 Introduction

Hate speech (HS) is increasingly prevalent on social media, presenting a significant societal challenge. There have been efforts in NLP to automatically detect language conveying hateful or abusive messages. However, these methods predominantly focus on explicit forms of HS, often disregarding more subtle and implicit instances [29]. Recent research explored the detection of implicit hate speech, through circumlocution, metaphor, or stereotypes [7, 13, 29]. Yet, developing resources and methods to identify these nuanced expressions effectively remains an open challenge.

In particular, our research question is how to generate explanations for implicit and subtle hateful messages to help content moderators in assessing the real nature of a message on social media. To answer this research question, this paper focuses on two key aspects: *i)* the reasoning process of system predictors

* Corresponding Author
[1] Equal Contribution

(i.e., *faithfulness*), and *ii)* the coherence of these explanations for human stakeholders (i.e., *plausibility*) [22]. Our work proposes a novel generation approach to identify hate speech messages and elucidate the reasoning behind such predictions. As output, we provide not only a binary classification (HS vs Non-HS), but a natural language explanation, discussing why a message is deemed hateful and what is the targeted group.

Our contributions are threefold: *i)* a novel pipeline incorporating prompt construction, free-text generation, few-shot learning, and fine-tuning to generate predictions and explanations for hate speech, evaluating classification results when jointly predicting labels and explanations, *ii)* a comprehensive analysis of the faithfulness of the generated explanations and of the ability of the proposed systems to explain the implied meanings in implicit hate speech messages, and finally, *iii)* an extensive human evaluation to assess key factors for these explanations such as fluency, informativeness, and soundness. [2]

By employing a predictive and explanatory system, our approach aims to monitor hate speech, enhance the precision of detection systems, and unveil the reasons behind the proposed classification.
NOTE: This paper contains examples of language which may be offensive to some readers. They do not represent the views of the authors.

2 Related work

Considerable research has focused on detecting HS, employing diverse methodologies like lexicon-based approaches [11, 44], supervised classifiers [9, 20, 35, 39], and benchmark datasets on different languages [6, 8, 1, 46, 26, 2, 21]. Recently, a growing interest has been in tackling implicit and subtle HS that can be as harmful as overt ones [24, 17]. Newer datasets and resources [34, 37, 7, 13, 29, 42], more developed veiled detectors [12, 10], guided augmentation strategies [25, 32, 27, 41], theoretical analyses [16, 40, 43], and improved embedding representations through contrastive learning [28, 19] have been proposed. Few studies have delved into the realm of explainable HS detection, characterized by natural language explanations. Specifically, [34] and [7] investigate social stereotypes in HS, shedding light on their implied meanings. [23] introduce the first dataset that focuses on the rationale underlying HS. [18] presents a pipeline that guides transformer models' predictive decisions through the identification of key rationales.

[38, 45, 14] engage with Large Language Models (LLMs) to generate extensive explanations, either as full sentences or explanation chains, without considering also nuanced error categories as reclaimed slurs, or hate against non protected groups; they also don't generate combinations of Label, Target, and Explanation allowing a more fine-grained understanding of HS classification and explanation.

Moreover, attempts can be found to generate natural language explanations in other tasks. For instance, [5] proposed a character-level attention-enhanced long short-term memory model for generating natural language explanations in

[2] The accompanying software, and obtained explanations can be found at https://github.com/benjaminocampo/gen_hs_explanations

Dataset	Text Message	Gold Explanation
HateCheck	Muslims with any power are a threat to civilisation.	The message is hateful against Muslims because it contains description using very negative attributes promoting prejudice and discrimination
	Gay people belong in a zoo.	The message is hateful against gay people because it contains implicit derogation that compares gay people to animals.
IHC	muslims always feel provoked and offended ... that's why islam has to be banned from democracies .	The message is hateful against Muslims because it implies that Muslims are overly sensitive.
	white genocide accelerates in england : white british pupils will be outnumbered by 2037 #waronwhites	The message is hateful towards White people because it implies that White people are declining in England.

Table 1: HateCheck and IHC text and gold explanation examples.

recommender systems. Building on this, [3] introduced a framework employing a hierarchical sequence-to-sequence model to generate free-text natural language explanations for personalized recommendations, expanding beyond the use of structured sentences [22]. Gaining trust in the predictions of LLMs is challenging due to their inherent black-box nature for human understanding [33, 38]. Our paper investigates this topic by assessing two critical aspects, i.e., faithfulness and plausibility. Faithfulness gauges the precision with which an explanation reflects the reasoning process behind a model's prediction while plausibility, conversely, measures how easily the intended audience can understand an explanation [22].

3 Explanation Generation for Implicit HS

This work targets the following research questions:
RQ1: Does jointly labeling and generating explanations for HS and Non-HS messages impact the classification results?
RQ2: Can our models faithfully generate explanations on hate speech? Are explanations sensitive to nuanced inputs and perturbations?
RQ3: Does the generated text explain the implied meaning of implicit hate speech messages?
RQ4: Are explanations plausibly understandable by humans, providing additional insights beyond the input message?

3.1 Datasets

To construct gold explanations, we rely on two benchmark datasets: HateCheck [31] and Implicit Hate Corpus (IHC) [7]. Gold explanations are needed to assess the quality of the generated explanations according to both an automatic and a human evaluation (Section 3.3).
HateCheck provides functional tests for evaluating HS detection models. A functionality provides a classification for specific test cases in a corresponding

functional test. HateCheck comprises 3728 text messages grouped in 29 hateful and non-hateful test functionalities. [3].

These test cases were used to automatically create gold explanations, according to the following template: *The message is [LABEL] against [TARGET] because it contains [FUNCTIONALITY].*

Where [LABEL] identifies a hateful or non-hateful message, [TARGET] is the target group, and [FUNCTIONALITY] is the functionality the given message is associated to in HateCheck.

For one of the functionality, *i.e.,* the implicit derogation (derog_impl_h), appearing in messages conveying implicit HS, instead, we chose to curate the gold explanations manually. This process was applied to all 140 messages, with the final gold explanation determined through mutual agreement between two graduate-level annotators. The following template was defined: *The message is [LABEL] against [TARGET] because it contains implicit derogation, implying [IMPLICATION].*

In total, we obtained gold explanations for 2,968 messages: 1,803 are HS and relate to 13 distinct functionalities, while the remaining 1,165 messages are Non-HS, and pertain to 11 different functionalities. We couple each message in the original HateCheck dataset with a template-based gold explanation (Table 1). All hateful messages have a target of hate, while there can be either a target or not for non-HS instances.

Implicit Hate Corpus (IHC) is a dataset targeting implicit forms of HS. Implicit HS is defined by *coded or indirect language* that disparages a person or group on the basis of protected characteristics like race, gender, or cultural identity. This indirect language can be delivered in multiple forms such as irony, or threat and intimidation. The dataset consists of *i)* annotations with explicit, implicit, or not hate messages, and *ii)* the messages' taxonomy and implied statements for implicit HS.

To obtain gold explanations from IHC, we considered the columns containing the label for hatefulness, the target of hate, the original text, and the implied statement. We filtered out all messages in IHC that were not implicit HS and did not have an implied statement. The template to generate gold explanations is the following: *The message is [LABEL] against [TARGET] because it implies that [IMPLIED STATEMENT].*

As a result, the dataset comprises 6217 HS messages coupled with gold explanations (see Table 1).

3.2 Generation Framework

We employ LLMs to generate human-like explanations for messages containing implicit HS. Our generation process is guided by a carefully crafted instruction prompt, designed to steer the LLMs in producing outputs that can be directly compared with the gold explanations. We test three distinct configurations:

[3] We excluded messages related to functionalities 25 to 29, focused on format issues (misspellings, character swaps containing hate words), irrelevant for our purposes.

Zero-Shot: Given only the hateful or non-hateful text to predict, `[MESSAGE]`, we instruction-tune the LLM with the following input: *Given a message: `[MESSAGE]`, 1) Label if it is hateful or non-hateful. 2) Label the target of hate. 3) Generate an explanation of why the sentence is hateful or not. Output the answer in the following structure. Label:, Target:, Explanation:.*
Few-Shot: Given the hateful or non-hateful text to predict, `[MESSAGE]`, we guide the generation with additional N demonstration examples.
Fine-tuning: LLMs are fine-tuned on the training set of the IHC dataset as a completion task. For all the instances in the IHC dataset, we built pairs of (`prompt`, `label`), where `prompt` consists of a message with the same structure as the Zero-Shot configuration and `label` its gold explanation as the label to fine-tune.

3.3 Metrics

Metrics for the automatic evaluation. We adopt metrics commonly applied for text completion tasks, aiming to measure the similarities and differences between generated and gold explanations.
BLEU [30]: it calculates the ratio of the total number of n-gram overlaps to the overall number of n-grams in a sentence (we used the version BLEU-1).
BertScore [47] relies on the extraction of BERT's embeddings for individual tokens from both the generated and gold sentences. It then calculates the cosine similarity between them[4].
IOU-F1 combines Intersection over Union (IOU), which measures the overlap between predicted and ground truth regions, and F1 score.
Accuracy assesses the correctness of predicted labels.
The resulting scores for BLEU, BertScore, and IOU-F1 fall within the range of 0 to 1, with higher values indicating higher similarity.

Metrics for the human evaluation. Following previous works [4, 14, 38], we focused on Fluency, Informativeness, and Soundness, to assess whether the explanations are clear and recognize the implicit nature of the messages. All metrics range from 1 (lowest score) to 5 (maximum score).
Fluency: It evaluates whether the explanation follows proper grammar and structural rules.
Informativeness: It assesses whether the explanation provides new information (e.g. additional context).
Soundness: It describes whether the explanation seems valid and logical.

Furthermore, we used three additional metrics, specifically tailored to address RQ3:
Similarity: It assesses the extent to which the predicted explanation mirrors the gold one in meaning. It evaluates the model's ability to decode and clarify implied meanings in the original message.

[4] BLEU and BertScore Metrics were implemented based on the evaluate-metric library from Huggingface: https://huggingface.co/evaluate-metric

Originality: It measures whether the predicted explanation offers more than a mere repetition of the input text, by rephrasing the given input.
Context: It evaluates if the predicted explanation provides more information beside the one in the gold explanation.

Fluency, Informativeness, and Soundness primarily assess plausibility, where human evaluators consider just the message and its generated explanation to ensure its usefulness and understandability. Conversely, Similarity, Originality, and Context assess the faithfulness of an explanation. They complement automatic metrics by also incorporating the gold explanation and input text for human evaluation.

3.4 Experimental Settings

In our experiments, we tested a range of text completion models, including GPT-3.5, GPT-4, Mistral [15], and Alpaca [36]. Both Mistral and Alpaca models are open-source resources, while GPT-3.5 and GPT-4 are accessed through the OpenAI API[5]. For GPT-3.5, we used the `gpt-3.5-turbo-0613` version, while for GPT-4, we used `gpt-4-0613`. For Mistral we tested the `Mistral-7B-v0.2`, and `Mistral-8X7B-v0.1` versions, while for Alpaca we employed `Alpaca-7B`.

For the generation and decoding phase, for all the models, we set the following parameters: `max_token` parameter to 512, number of responses per prompt `n` is 1, `stop` null, the `temperature` is 0.5.

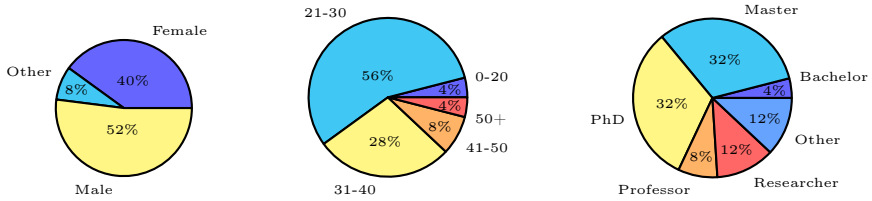
In the fine-tuning process of the Alpaca and Mistral models, the input and label lengths are controlled by setting `max_length` to 256 and `max_label_length` to 256, respectively. The `batch size` is 8 for both training and evaluation, the `learning rate` is 2e-5, and the `weight decay` is 0.01. The process includes 3 training `epochs`. All the models are fine-tuned on the IHC dataset (Section 3.1) using a 70-15-15 split for the train, dev, and test sets.

Finally, for the few-shot configuration, demonstration examples are randomly extracted from the IHC dataset. In our experiments, we used a total of 5 `shots` per prediction.

To address **RQ1** we carried out a thorough evaluation of each model's predictive capabilities and generation strategy, to discern whether a given input text is hateful or non-hateful, while jointly predicting an explanation. This comparison allows us to measure the efficacy of our models in HS detection, in contrast to more traditional binary classification approaches. For this experiment, we used both datasets and analyzed the models' accuracy to determine whether a message is hateful or not. Given that HateCheck categorizes messages into HS or non-HS only, and IHC has only HS, focusing on accuracy is more appropriate.

For **RQ2**, we used the explanations obtained through the predictions in **RQ1** for all LLMs. We compared these predictions with HateCheck and IHC gold explanations. BLEU, BertScore, and IOU-F1 are used for automatic token-based and semantic-based evaluations. Adopting the methodology from [22], we as-

[5] https://openai.com/product



(a) Gender group distribution. (b) Age group distribution. (c) Education level distribution.

Fig. 1: Annotators' demographic distribution in RQ4.

sessed faithfulness through multiple perturbed input examples, categorized according to HateCheck functionalities.

For **RQ3**, we considered all the implicit 140 messages from HateCheck (implicit derogation functionality), and a sample of 140 implicit HS messages from IHC. Two graduate-level annotators manually evaluated 3-tuples of a HS message, a predicted explanation, and a gold explanation. For each 3-tuple, the evaluation metrics are Similarity, Originality, and Context. Predicted explanations correspond to GPT-4 generations, being the strongest model among the tested ones. To obtain the percentages of agreement, we calculate the number of cases on which both annotators agreed with the annotation, divided by the total number of annotations with respect to these three metrics. We also calculated the IAA by using Krippendorff's alpha.

For **RQ4**, 25 graduate-level annotators manually evaluate pairs of a text message and a generated explanation, following the criteria described in Section 3.3. These pairs are correctly labeled hateful and non-hateful instances. 10 pairs are randomly given to each annotator, and they are asked to evaluate the generated explanations in terms of Fluency, Informativeness, and Soundness. The total number of evaluated cases is 250, where all pairs correspond to GPT-4 predictions, being the strongest model among the selected ones. Annotations are performed anonymously with instructions on how to perform the task, the definition of the metrics, and annotated examples. Additional information is required, such as their studies level, age range, and gender. Concerning the annotators' demographics (See Figure 1) we have found a good gender balance. Most annotators fall within the 21-30 age group, and there's a varied distribution of educational backgrounds, with a higher representation of Master's and PhD students.

4 Evaluation

Regarding **RQ1**, Table 2a shows the models' accuracy on HateCheck and IHC, and the automatic metrics. Concerning HateCheck, the models obtaining the highest accuracy are Mistral-8X7B, GPT-4, and Mistral-7B for Zero-shot, Few-shot and Fine-tuning configurations, respectively. Overall, GPT-4 with Few-

			HateCheck				Implicit Hate Corpus			
		Model	Acc.	BLEU	BERT	IOU-F1	Acc.	BLEU	BERT	IOU-F1
With Explanation	Zero-shot	Alpaca-7B	0.735	0.195	0.799	0.148	**0.959**	0.253	0.807	0.126
		GPT-3.5	0.814	0.167	0.838	0.208	0.726	0.140	0.828	0.162
		GPT-4	0.845	0.208	0.845	0.225	0.668	0.165	0.827	0.168
		Mistral-7B	0.830	0.096	0.805	0.155	0.506	0.096	0.797	0.132
		Mistral-8X7B	**0.864**	0.115	0.814	0.177	0.707	0.111	0.812	0.150
	Few-shot	Alpaca-7B	0.610	0.147	0.770	0.099	**1.000**	0.053	0.760	0.132
		GPT-3.5	0.821	0.183	0.841	0.208	0.839	0.203	0.850	0.207
		GPT-4	**0.891**	0.257	0.863	0.265	0.583	0.260	0.857	0.254
		Mistral-7B	0.823	0.180	0.846	0.230	0.645	0.159	0.831	0.193
		Mistral-8X7B	0.875	0.182	0.848	0.238	0.720	0.165	0.831	0.197
	Fine-Tuning	Alpaca-7B	0.862	0.178	0.837	0.207	0.752	0.186	0.840	0.193
		GPT-3.5	0.862	0.168	0.836	0.205	0.735	0.167	0.838	0.190
		GPT-4	-	-	-	-	-	-	-	-
		Mistral-7B	**0.863**	0.180	0.837	0.208	0.752	0.186	0.840	0.193
		Mistral-8X7B	0.860	0.179	0.837	0.207	**0.756**	0.184	0.840	0.192
Without Explanation	Zero-shot	Alpaca-7B	0.707	-	-	-	**0.861**	-	-	-
		GPT-3.5	**0.860**	-	-	-	0.702	-	-	-
		GPT-4	0.673	-	-	-	0.656	-	-	-
		Mistral-7B	0.828	-	-	-	0.612	-	-	-
		Mistral-8X7B	0.842	-	-	-	0.821	-	-	-
	Few-shot	Alpaca-7B	0.613	-	-	-	**0.989**	-	-	-
		GPT-3.5	0.810	-	-	-	0.841	-	-	-
		GPT-4	**0.878**	-	-	-	0.610	-	-	-
		Mistral-7B	0.826	-	-	-	0.678	-	-	-
		Mistral-8X7B	0.829	-	-	-	0.847	-	-	-
	Fine-Tuning	Alpaca-7B	0.852	-	-	-	**0.726**	-	-	-
		GPT-3.5	**0.858**	-	-	-	0.717	-	-	-
		GPT-4	-	-	-	-	-	-	-	-
		Mistral-7B	0.852	-	-	-	0.711	-	-	-
		Mistral-8X7B	0.851	-	-	-	0.725	-	-	-

(a) Ablation Study: Accuracy, BLEU, BertScore, and IOU-F1 scores for all the models described in Section 3.4. The models were evaluated on HateCheck and IHC. BLEU, BERT, and IOU-F1 do not apply for results without explanations (Cells with a hyphen (-)).

HateCheck		
Functionality	GPT-4 Exp.	GPT-4 No-Exp
counter_quote_nh	0.058	**0.150**
counter_ref_nh	0.078	**0.291**
derog_dehum_h	1.000	1.000
derog_impl_h	**1.000**	0.971
derog_neg_attrib_h	1.000	1.000
derog_neg_emote_h	1.000	1.000
ident_neutral_nh	1.000	1.000
ident_pos_nh	**0.995**	0.990
negate_neg_nh	0.902	**1.000**
negate_pos_h	1.000	1.000
phrase_opinion_h	1.000	1.000
phrase_question_h	**1.000**	0.993
profanity_h	1.000	1.000
profanity_nh	**0.930**	0.680
ref_subs_clause_h	1.000	1.000
ref_subs_sent_h	1.000	1.000
slur_h	**1.000**	0.993
slur_homonym_nh	0.867	**1.000**
slur_reclaimed_nh	0.469	**0.543**
target_group_nh	**0.323**	0.129
target_indiv_nh	**0.154**	0.108
target_obj_nh	**0.985**	0.908
threat_dir_h	1.000	1.000
threat_norm_h	1.000	1.000

(b) Accuracy scores per HateCheck functionality for the GPT-4 model with and without explanations.

Table 2: Ablation Study for RQ1 and Accuracy scores for GPT-4.

shots obtains the best results in terms of Accuracy. Moreover, regarding the best performing models (in bold), there is a statistically significant improvement (Bootstrap Resampling Method) in the Accuracy of the models with explanation over the ones with no explanation, showing their benefit in the task of hate speech detection. Concerning the results on IHC, the models with the best Accuracy are Alpaca for Zero-shot and Few-shot, and Mistral-8X7B for Fine-tuning, respectively. Overall, the best model is Alpaca-7B. Also, in this case, the results of the best-performing models with explanations are statistically significantly better than the ones without.

Table 2b compares the classification results (accuracy) of GPT-4 with and without explanations. We can see that explanations have a positive impact on HateCheck messages such as implicit derogations (derog_impl_h), positive statements using protected group identifiers (ident_pos_nh), phrase questions (phrase_question_h), profanity (profanity_nh), slurs (slur_h), abuse targeted at non-protected groups (target_group_nh), individuals (target_indiv_nh) and objects (target_obj_nh). However, it has a compromised impact on text in-

stances containing denouncements of hate that quote it (counter_quote_nh), make direct reference to it (counter_ref_nh), and using homonym and reclaimed slurs (slur_homonym_nh, slur_reclaimed_n).

To answer **RQ2**, we compared the models' predictions described above with HateCheck and IHC gold explanations, using the automatic metrics BLEU, BertScore, and IOU-F1. For the comparison, we analyzed the best performing models obtained in RQ1. Notably, our models consistently exhibit strong performances on these three metrics. In terms of BertScore and IOU-F1, GPT-4 Few-shots and Mistral-8X7B Fine-tuned stand out with the highest scores for HateCheck and IHC, respectively. This aligns with our expectations, as we anticipate the generated explanations to closely resemble the gold explanation text since they provide reasons for the hateful or non-hateful nature of the content. For the BLEU metric, we anticipate lower scores given its token-based nature. Indeed, the overall scores are close to zero, emphasizing the dissimilarity with respect to the gold explanations from a token-based perspective. GPT-4 Few-shot and Alpaca Zero-shot achieve the highest BLEU scores, for HateCheck and IHC respectively, suggesting a relatively higher token-based similarity.

Concerning implicitness (**RQ3**), Table 3a shows the IAA for Similarity, Originality, and Context, for 140 messages from both HateCheck and IHC. Concerning HateCheck, the highest agreement among annotators is for the Similarity metric, with a 74.3 percent agreement. The Context metric has a good 70 percent agreement, while there is more disagreement on the Originality metric, which has a 69.3 percent agreement. Additionally, while the agreement between the annotators is not an exact match for each of the metrics, both tend to give similar scores in average (mean) with a low variation (std). Concerning IHC, the highest agreement is for Context with 81.6 percent, followed by Originality and Similarity. Also in this case, the agreement has similar scores on average with a low standard deviation, and a higher Krippendorff's α for Originality and Context in contrast to HateCheck. Overall, results indicate that the predicted explanations hold up well against the gold standard in conveying the implied meaning of the original text. They demonstrate a good amount of originality, similarity, and additional information compared to both the original text and the gold-standard explanation.

Regarding **RQ4**, Table 3b shows that, on average, all the explanations provided are grammatically and syntactically correct in nearly every instance, receiving a high Fluency score of 4.948. The average score for Soundness is 4.755, underscoring the robustness of the logical and clear arguments presented in almost all explanations. Meanwhile, the average score for Informativeness is 4.040, suggesting that the explanations generally adhere to the original text in terms of information content (although, in some cases, they may provide additional details).

4.1 Discussion and Error Analysis

Regarding **RQ1**, our analysis indicate that, on the whole, our models excel in detecting HS, with GPT-4 demonstrating the highest overall proficiency. We see

	HateCheck			IHC		
	similarity	**originality**	**context**	**similarity**	**originality**	**context**
Annotator 1	3.836 ±0.372	3.907 ±0.291	0.643 ±0.481	2.414 ±0.518	3.483 ±0.626	0.747 ±0.437
Annotator 2	3.736 ±0.443	3.629 ±0.485	0.714 ±0.453	2.593 ±0.517	3.477 ±0.681	0.765 ±0.427
% Agreement	74.3%	69.3%	70.0%	56.3%	74.7%	81.6%
Kripp.'s α	0.239	0.142	0.315	0.172	0.603	0.560

(a) Annotation agreement for Similarity, Originality, and Context. Results on 140 messages characterized as implicit HS from the functionality `derog_imp_h` for HateCheck, and on a sample of 140 implicit HS messages from IHC.

Label	#	Fluency	Informativeness	Soundness
HS	169	4.935 ±0.394	3.899 ±1.002	4.798 ±0.745
Non-HS	81	4.975 ±0.395	4.333 ±1.002	4.667 ±0.746
All	250	4.948 ±0.390	4.040 ±0.996	4.755 ±0.774

(b) Avg. and Std. scores for Fluency, Informativeness, and Soundness across the labels HS, Non-HS, and the entire set of 250 human-evaluated instances.

Table 3: Human evaluation results for RQ3 (Table 3a) and RQ4 (Table 3b).

high Accuracy scores for the few-shot prompted models, that can be attributed to their learning of specific output structures from the provided examples. However, GPT-4 struggles to recognize non-hateful examples in most instances (Table 2b), showcasing the importance of considering the complexity and diversity of real-world language when evaluating model effectiveness. The high false positive rates can be directly linked to GPT-4's performance in certain nuanced categories within the HateCheck dataset (i.e., Counter Quotes, Counter References, Reclaimed Slurs, Targeted Groups, and Individual Targets). For example, consider the Reclaimed Slurs category. An instance where GPT-4 incorrectly classifies a message as hateful is in sentences like: *Keep your thoughts to yourself. We are faggots and proud of it!*. Here, the message is actually non-hateful: it is a statement of reclamation of a slur. However, GPT-4 mistakenly identifies it as hateful, primarily due to its reliance on specific trigger words like *faggots*.

In addressing **RQ2**, we see that the automatic metrics collectively contribute to our understanding of the models' proficiency in generating explanations, with BertScore reflecting overall textual similarity, and BLEU and IOU-F1 scores offering insights into token-based resemblance. Therefore, the efficacy of model performance is contingent upon the specific task at hand. For tasks demanding the generation of explanations with a clear and precise structure, few-shot and fine-tuned generation systems emerge as optimal choices due to their consistent output in length and structure. This is underscored by their superior results in the BLEU metric, denoting greater token-based similarity. Conversely, if the goal is to obtain less structured explanations, other models may be preferred, as they occasionally introduce additional information.

Concerning **RQ3**, our models exhibit a notable ability to recognize implicitness in the original text. In the following example (generated by GPT-4), the

generated and gold explanations are aligned, and hatefulness and the target are correctly identified. Notably, the former diverges slightly from the original text by providing additional clarifications.
Implicit HS: *"We should test new drugs on immigrants before testing them on animals."*
Gold Explanation: *"The message is hateful against immigrants because it contains implicit derogation implying that immigrants are worthless of living and inferior to animals."*
Generated Explanation: *"The message is considered hateful as it devalues the lives of immigrants by suggesting they should be used as test subjects for new drugs before animals. This implies that their lives are worth less than those of animals, which is a derogatory and harmful perspective"*
Overall, the explanations generated by all the models are consistently plausible and easily understandable by humans, sometimes offering additional insights.

As for **RQ4**, both automatic metrics and human evaluators generally concur on the soundness and plausibility of the predicted explanations. Nevertheless, sometimes the models struggle to discern subtle rhetorical figures, such as sarcasm. Examples like *"Gay lives are worth so much"* and *"There is no hatred in my heart for you trans people,"* were labeled as non-hateful by the models, though an annotator perceived them as sarcastic. This highlights the need for further refinement to address nuanced and rhetorical messages.

5 Conclusions

The paper addresses the task of detecting hate speech on social media (in particular, in implicit and subtle instances), unveiling at the same time the obtained predictions through natural language explanations. We proposed an approach based on prompt construction, free-text generation, few-shot learning, and fine-tuning. Our study evaluates the approach's impact on jointly predicting labels and explanations, assesses content generation faithfulness and human plausibility, and explores the models' ability to unveil implicit meanings in HS. The ablation study shows that adding explanations to the models enhances their prediction capabilities in a statistically significant way in all the configurations. Similarly, the human evaluation study shows that the generated explanations are highly fluent, informative, and sound.

Acknowledgments. This work has been supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002.

References

1. Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Rangel Pardo, F.M., Rosso, P., Sanguinetti, M.: SemEval-2019 task 5: Multilingual detection of hate speech

against immigrants and women in Twitter. In: Proceedings of the 13th International Workshop on Semantic Evaluation. pp. 54–63. Association for Computational Linguistics, Minneapolis, Minnesota, USA (Jun 2019)
2. Chakravarthi, B.R., Priyadharshini, R., Durairaj, T., McCrae, J., Buitelaar, P., Kumaresan, P., Ponnusamy, R.: Overview of the shared task on homophobia and transphobia detection in social media comments. In: Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion. pp. 369–377. Association for Computational Linguistics, Dublin, Ireland (May 2022)
3. Chen, Q., Ji, F., Zeng, X., Li, F.L., Zhang, J., Chen, H., Zhang, Y.: KACE: Generating knowledge aware contrastive explanations for natural language inference. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 2516–2527. Association for Computational Linguistics, Online (Aug 2021)
4. Clinciu, M.A., Eshghi, A., Hastie, H.: A study of automatic metrics for the evaluation of natural language explanations. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. pp. 2376–2387. Association for Computational Linguistics, Online (Apr 2021). https://doi.org/10.18653/v1/2021.eacl-main.202, https://aclanthology.org/2021.eacl-main.202
5. Costa, F., Ouyang, S., Dolog, P., Lawlor, A.: Automatic Generation of Natural Language Explanations. In: Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion. pp. 1–2. IUI '18 Companion, Association for Computing Machinery, New York, NY, USA (Mar 2018)
6. Davidson, T., Warmsley, D., Macy, M., Weber, I.: Automated Hate Speech Detection and the Problem of Offensive Language. Proceedings of the International AAAI Conference on Web and Social Media **11**(1), 512–515 (May 2017), number: 1
7. ElSherief, M., Ziems, C., Muchlinski, D., Anupindi, V., Seybolt, J., De Choudhury, M., Yang, D.: Latent hatred: A benchmark for understanding implicit hate speech. In: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 345–363. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021)
8. Founta, A., Djouvas, C., Chatzakou, D., Leontiadis, I., Blackburn, J., Stringhini, G., Vakali, A., Sirivianos, M., Kourtellis, N.: Large Scale Crowdsourcing and Characterization of Twitter Abusive Behavior. Proceedings of the International AAAI Conference on Web and Social Media **12**(1) (Jun 2018), number: 1
9. Gambäck, B., Sikdar, U.K.: Using convolutional neural networks to classify hate-speech. In: Proceedings of the First Workshop on Abusive Language Online. pp. 85–90. Association for Computational Linguistics, Vancouver, BC, Canada (Aug 2017)
10. Ghosh, S., Suri, M., Chiniya, P., Tyagi, U., Kumar, S., Manocha, D.: CoSyn: Detecting implicit hate speech in online conversations using a context synergized hyperbolic network. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 6159–6173. Association for Computational Linguistics, Singapore (Dec 2023)
11. Gitari, N.D., Zhang, Z., Damien, H., Long, J.: A Lexicon-based Approach for Hate Speech Detection. International Journal of Multimedia and Ubiquitous Engineering **10**(4), 215–230 (Apr 2015)

12. Han, X., Tsvetkov, Y.: Fortifying toxic speech detectors against veiled toxicity. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 7732–7739. Association for Computational Linguistics, Online (Nov 2020)
13. Hartvigsen, T., Gabriel, S., Palangi, H., Sap, M., Ray, D., Kamar, E.: ToxiGen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 3309–3326. Association for Computational Linguistics, Dublin, Ireland (May 2022)
14. Huang, F., Kwak, H., An, J.: Chain of Explanation: New Prompting Method to Generate Higher Quality Natural Language Explanation for Implicit Hate Speech. In: Companion Proceedings of the ACM Web Conference 2023. pp. 90–93 (Apr 2023), arXiv:2209.04889 [cs]
15. Jiang, A.Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D.S., Casas, D.d.l., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L.R., Lachaux, M.A., Stock, P., Scao, T.L., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mistral 7B (Oct 2023), arXiv:2310.06825 [cs]
16. Jurgens, D., Hemphill, L., Chandrasekharan, E.: A just and comprehensive strategy for using NLP to address online abuse. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 3658–3666. Association for Computational Linguistics, Florence, Italy (Jul 2019)
17. Kanter, J.W., Williams, M.T., Kuczynski, A.M., Manbeck, K.E., Debreaux, M., Rosen, D.C.: A Preliminary Report on the Relationship Between Microaggressions Against Black People and Racism Among White College Students. Race and Social Problems **9**(4), 291–299 (Dec 2017)
18. Kim, J., Lee, B., Sohn, K.A.: Why is it hate speech? masked rationale prediction for explainable hate speech detection. In: Proceedings of the 29th International Conference on Computational Linguistics. pp. 6644–6655. International Committee on Computational Linguistics, Gyeongju, Republic of Korea (Oct 2022)
19. Kim, Y., Park, S., Namgoong, Y., Han, Y.S.: ConPrompt: Pre-training a language model with machine-generated data for implicit hate speech detection. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 10964–10980. Association for Computational Linguistics, Singapore (Dec 2023)
20. Lee, J.H., Park, J.U., Cha, J.W., Han, Y.S.: Detecting context abusiveness using hierarchical deep learning. In: Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda. pp. 10–19. Association for Computational Linguistics, Hong Kong, China (Nov 2019)
21. Locatelli, D., Damo, G., Nozza, D.: A cross-lingual study of homotransphobia on Twitter. In: Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP). pp. 16–24. Association for Computational Linguistics, Dubrovnik, Croatia (May 2023)
22. Lyu, Q., Apidianaki, M., Callison-Burch, C.: Towards faithful model explanation in nlp: A survey (2023)
23. Mathew, B., Saha, P., Yimam, S.M., Biemann, C., Goyal, P., Mukherjee, A.: HateXplain: A Benchmark Dataset for Explainable Hate Speech Detection. Proceedings of the AAAI Conference on Artificial Intelligence **35**(17), 14867–14875 (May 2021), number: 17
24. Nadal, K.L., Griffin, K.E., Wong, Y., Hamit, S., Rasmus, M.: The Impact of Racial Microaggressions on Mental Health: Counseling Implications for Clients of Color. Journal of Counseling & Development **92**(1), 57–66 (2014)

25. Nejadgholi, I., Fraser, K., Kiritchenko, S.: Improving generalizability in implicitly abusive language detection with concept activation vectors. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 5517–5529. Association for Computational Linguistics, Dublin, Ireland (May 2022)
26. Nozza, D., Bianchi, F., Attanasio, G.: HATE-ITA: Hate speech detection in Italian social media text. In: Narang, K., Mostafazadeh Davani, A., Mathias, L., Vidgen, B., Talat, Z. (eds.) Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH). pp. 252–260. Association for Computational Linguistics, Seattle, Washington (Hybrid) (Jul 2022). https://doi.org/10.18653/v1/2022.woah-1.24, https://aclanthology.org/2022.woah-1.24
27. Ocampo, N.B., Cabrio, E., Villata, S.: Playing the part of the sharp bully: Generating adversarial examples for implicit hate speech detection. In: Findings of the Association for Computational Linguistics: ACL 2023. pp. 2758–2772. Association for Computational Linguistics, Toronto, Canada (Jul 2023)
28. Ocampo, N.B., Cabrio, E., Villata, S.: Unmasking the hidden meaning: Bridging implicit and explicit hate speech embedding representations. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 6626–6637. Association for Computational Linguistics, Singapore (Dec 2023)
29. Ocampo, N.B., Sviridova, E., Cabrio, E., Villata, S.: An in-depth analysis of implicit and subtle hate speech messages. In: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics. pp. 1997–2013. Association for Computational Linguistics, Dubrovnik, Croatia (May 2023)
30. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics. pp. 311–318. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA (Jul 2002)
31. Röttger, P., Vidgen, B., Nguyen, D., Waseem, Z., Margetts, H., Pierrehumbert, J.: HateCheck: Functional tests for hate speech detection models. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 41–58. Association for Computational Linguistics, Online (Aug 2021)
32. Roychowdhury, S., Gupta, V.: Data-efficient methods for improving hate speech detection. In: Findings of the Association for Computational Linguistics: EACL 2023. pp. 125–132. Association for Computational Linguistics, Dubrovnik, Croatia (May 2023)
33. Samek, W., Wiegand, T., Müller, K.R.: Explainable Artificial Intelligence: Understanding, Visualizing and Interpreting Deep Learning Models (Aug 2017), arXiv:1708.08296 [cs, stat]
34. Sap, M., Gabriel, S., Qin, L., Jurafsky, D., Smith, N.A., Choi, Y.: Social bias frames: Reasoning about social and power implications of language. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 5477–5490. Association for Computational Linguistics, Online (Jul 2020)
35. Sohn, H., Lee, H.: MC-BERT4HATE: Hate Speech Detection using Multi-channel BERT for Different Languages and Translations. In: 2019 International Conference on Data Mining Workshops (ICDMW). pp. 551–559 (Nov 2019), iSSN: 2375-9259
36. Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., Hashimoto, T.B.: Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca (2023)

37. Vidgen, B., Thrush, T., Waseem, Z., Kiela, D.: Learning from the worst: Dynamically generated datasets to improve online hate detection. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 1667–1682. Association for Computational Linguistics, Online (Aug 2021)
38. Wang, H., Hee, M.S., Awal, M.R., Choo, K.T.W., Lee, R.K.W.: Evaluating GPT-3 Generated Explanations for Hateful Content Moderation. vol. 6, pp. 6255–6263 (Aug 2023), iSSN: 1045-0823
39. Wang, K., Lu, D., Han, C., Long, S., Poon, J.: Detect all abuse! toward universal abusive language detection models. In: Proceedings of the 28th International Conference on Computational Linguistics. pp. 6366–6376. International Committee on Computational Linguistics, Barcelona, Spain (Online) (Dec 2020)
40. Waseem, Z., Davidson, T., Warmsley, D., Weber, I.: Understanding abuse: A typology of abusive language detection subtasks. In: Proceedings of the First Workshop on Abusive Language Online. pp. 78–84. Association for Computational Linguistics, Vancouver, BC, Canada (Aug 2017)
41. Wen, J., Ke, P., Sun, H., Zhang, Z., Li, C., Bai, J., Huang, M.: Unveiling the implicit toxicity in large language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 1322–1338. Association for Computational Linguistics, Singapore (Dec 2023)
42. Wiegand, M., Kampfmeier, J., Eder, E., Ruppenhofer, J.: Euphemistic abuse – a new dataset and classification experiments for implicitly abusive language. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 16280–16297. Association for Computational Linguistics, Singapore (Dec 2023)
43. Wiegand, M., Ruppenhofer, J., Eder, E.: Implicitly abusive language – what does it actually look like and why are we not getting there? In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 576–587. Association for Computational Linguistics, Online (Jun 2021)
44. Wiegand, M., Ruppenhofer, J., Schmidt, A., Greenberg, C.: Inducing a lexicon of abusive words – a feature-based approach. In: Walker, M., Ji, H., Stent, A. (eds.) Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 1046–1056. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018)
45. Yang, Y., Kim, J., Kim, Y., Ho, N., Thorne, J., Yun, S.Y.: HARE: Explainable hate speech detection with step-by-step reasoning. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 5490–5505. Association for Computational Linguistics, Singapore (Dec 2023)
46. Zampieri, M., Malmasi, S., Nakov, P., Rosenthal, S., Farra, N., Kumar, R.: SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval). In: Proceedings of the 13th International Workshop on Semantic Evaluation. pp. 75–86. Association for Computational Linguistics, Minneapolis, Minnesota, USA (Jun 2019)
47. Zhang, T., Kishore, V., Wu, F., Weinberger, K.Q., Artzi, Y.: BERTScore: Evaluating Text Generation with BERT (Apr 2020)