Data & Knowledge Engineering 162 (2026) 102535



Contents lists available at ScienceDirect

Data & Knowledge Engineering

journal homepage: www.elsevier.com/locate/datak





"Detectors Lead, LLMs Follow": Integrating LLMs and traditional models on implicit hate speech detection to generate faithful and plausible explanations☆

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

Université Côte d'Azur, Inria, CNRS, I3S, France

ARTICLE INFO

Dataset link: https://github.com/benjaminocampo/dke_gen_hs_explanations

Keywords:
Hate speech detection
Generating explanations
Implicit hate speech
Subtle hate speech

ABSTRACT

Social media platforms face a growing challenge in addressing abusive content and hate speech, particularly as traditional natural language processing methods often struggle with detecting nuanced and implicit instances. To tackle this issue, our study enhances Large Language Models (LLMs) in the detection and explanation of implicit hate speech, outperforming classical approaches. We focus on two key objectives: (1) determining whether jointly predicting and generating explanations for why a message is hateful improves LLMs' accuracy, especially for implicit cases, and (2) evaluating whether incorporating information from BERT-based models can further boost detection and explanation performance. Our method evaluates and enhances LLMs' ability to detect hate speech and explain their predictions. By combining binary classification (Hate Speech vs. Non-Hate Speech) with natural language explanations, our approach provides clearer insights into why a message is considered hateful, advancing the accuracy and interpretability of hate speech detection.

1. Introduction

Hate speech (HS) is increasingly prevalent on social media, presenting a significant societal challenge. There have been efforts in natural language processing 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 [1]. Recent research explored the detection of implicit hate speech, through circumlocution, metaphor, or stereotypes [1–3]. Yet, developing resources and methods to identify these nuanced expressions effectively remains an open challenge [4].

Social media companies use these automated systems to enforce content policies, determining what is allowed on their platforms. High-confidence flagged content is removed automatically, while uncertain cases require human moderation. As human evaluation is expensive, a significant challenge is managing the sheer volume of potential hate speech requiring review. One solution is enhancing moderation by not only detecting hateful messages but also automatically generating explanations for why the content is deemed hateful [5].

In these directions, our aim is to evaluate and improve the detection and explanation capabilities of hateful messages of Large Language Models (LLMs). Our work proposes a novel generation approach to identify hate speech messages and elucidate the

☆ This article is part of a Special issue entitled: 'NLDB 2024 - Storey' published in Data & Knowledge Engineering.

* Corresponding author.
E-mail addresses: greta.damo@univ-cotedazur.fr (G. Damo), nicolas-benjamin.ocampo@univ-cotedazur.fr (N.B. Ocampo), elena.cabrio@univ-cotedazur.fr (E. Cabrio), serena.villata@univ-cotedazur.fr (S. Villata).

[1] Equal contribution

https://doi.org/10.1016/j.datak.2025.102535
Received 22 November 2024; Received in revised form 29 July 2025; Accepted 22 November 2025
Available online 24 November 2025
0169-023X/© 2025 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

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. More specifically, we focus on two aspects, to investigate: (i) whether jointly predicting and generating explanations for why a message is deemed hateful or not could improve LLMs prediction accuracy, especially for implicit and subtle forms, and (ii) whether adding information from Bidirectional Encoder Representations from Transformers (BERT) to LLMs can further improve detection and explanation results. We addressed the first objective in Damo et al. [6], tacking the following research questions:

RQ1: Does jointly labeling and generating explanations for HS and Non-HS messages impact the classification results?

For this, we developed 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.

RQ2: Can our models faithfully generate explanations on hate speech? Are explanations sensitive to nuanced inputs and perturbations?

We performed 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.

RQ3: Does the generated text explain the implied meaning of implicit hate speech messages?

We performed a human evaluation to assess key factors for these explanations on 3 defined dimensions, Similarity, Originality, and Context.

RQ4: Are explanations plausibly understandable by humans, providing additional insights beyond the input message?

We performed an extensive human evaluation with 25 participants in order to assess the plausibility of the explanations by evaluating Fluency, Informativeness, and Soundness.

This work extends the previous contribution targeting the second objective, and addressing the following research questions:

RQ5: Are the LLMs coupled with explanations better or worse than commonly used HS detectors, such as BERT, and HateBERT (trained on HS datasets) in classification?

We obtain hateful or non-hateful labels for the best performing LLM and two BERT models (BERT, and HateBERT), and we calculate the Accuracy and F1 scores. We do this on HateCheck and IHC datasets, and we compare the results.

RQ6: Can LLMs coupled with explanations improve detection performance if they are guided with HS detectors models' information?

We develop three methods for trying to improve LLMs' detection performance, by adding the following information to the prompt: *(1)* `[LABEL] + [PROBABILITY]` from a HS detector model; *(2)* if the message is hateful, we provide the `[IMPLIED STATEMENT]`, i.e., the underlying negative stereotype; *(3)* relevant tokens that impact the decision of an HS detector model.

RQ7: Does HS detectors models' information also impact the quality of LLM-generated explanations?

We compute the scores of two automatic metrics, Repetition Rate and ArgumentJudge, to measure the quality of the generated explanations.

We believe our findings on detecting and explaining hate speech can benefit both the NLP community and companies developing solutions to address issues like mental health disorders triggered by online hate speech [7–9]. Our research lies at the intersection of Generative AI via LLMs and the faithful, plausible evaluation of LLMs in the context of hate speech detection-key challenges within NLP to Information and Communication Systems [10–13].

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. An illustration of the workflow overview is depicted in Fig. 1. [2]

In Section 2, we review existing methods for hate speech detection, along with recent advancements in explainability and the generation of natural language explanations. Section 3 outlines our methodology, including the datasets used, evaluation metrics, and experimental settings for each of our seven RQs. Section 4 presents a detailed evaluation and analysis of each RQ, accompanied by a discussion and error analysis. Finally, in Section 5, we conclude by summarizing our findings, discussing the limitations of our work, and highlighting open challenges and future research directions.

NOTE: This paper contains examples of language which may be offensive to some readers. They do not represent the views of the authors.

[2] The data and accompanying software can be found at https://github.com/benjaminocampo/dke_gen_hs_explanations.

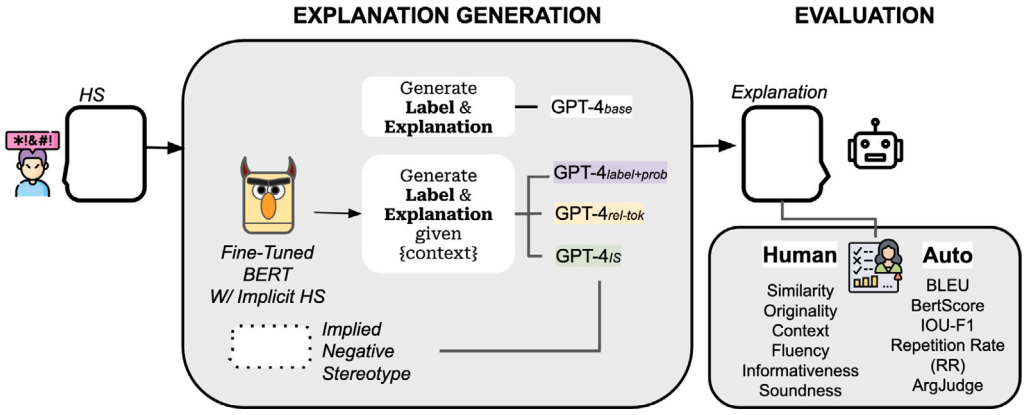


Fig. 1. Pipeline overview for HS detection and explanation. An example HS message is processed by a BERT-based detector fine-tuned on implicit HS data. From the model, we extract relevant tokens contributing to the HS classification, along with the predicted label, probability, and an inferred implied statement. These outputs are then used to prompt a LLM to generate explanations. The resulting outputs are evaluated using both automatic and human-based metrics.

2. Related work

This section is divided into two parts: Section 2.1 reviews advancements, performance, and results of hate speech detection models, while Section 2.2 examines the state-of-the-art in explainability for NLP tasks.

2.1. Hate speech detection

Hate speech is generally defined as any form of communication that demeans an individual or group based on characteristics such as race, color, ethnicity, gender, sexual orientation, nationality, religion, or similar attributes [14]. With the rapid increase in user-generated content, particularly on social media platforms, the prevalence of hate speech has also been steadily rising. In recent years, there has been growing interest in detecting online hate speech, especially through automated methods, as its societal impact continues to expand. Addressing this issue effectively requires NLP techniques specifically tailored to this challenge. Considerable research has focused especially on detecting explicit forms of HS, employing diverse methodologies like lexicon-based approaches [15,16], supervised classifiers [17–20], and benchmark datasets on different languages [21–27]. More generally, different methods have been developed for text classification, that can be applied also to hate speech detection, like the use of a hybrid model that combines BERT-based transformers, Many To Many Long Short-Term Memory (MTM LSTM), and Decision Templates (DT) for text classification [28]. Recently, a growing interest has been in tackling implicit and subtle HS that can be as harmful as overt ones [29,30]. Newer datasets and resources [1–3,31–33], more developed veiled detectors [34,35], guided augmentation strategies [36–39], theoretical analyses [4,40,41], and improved embedding representations through contrastive learning [42,43] have been proposed. While these studies provide a foundation for exploring and modeling the issue of implicit hate speech, there remains significant room for improvement. This includes developing more suitable resources to study these aspects and enhancing computational methods to address them effectively.

2.2. Explainability

Few studies have delved into the realm of explainable HS detection, characterized by natural language explanations. Specifically, Sap et al. [31] and Elsherief et al. [2] investigate social stereotypes in HS, shedding light on their implied meanings. Mathew et al. [44] introduce the first dataset that focuses on the rationale underlying HS. Kim et al. [45] presents a pipeline that guides transformer models' predictive decisions through the identification of key rationales. Damo et al. [46], Wang et al. [47], Yang et al. [48], and Huang et al. [49], engage with LLMs to generate extensive explanations, either as full sentences or explanation chains, highlighting the importance of careful application in content moderation. Our work goes beyond these contributions by (i) identifying issues in generations not only involving positive terms or the absence of overtly derogatory words in HS messages, but also nuanced error categories as reclaimed slurs, or hate against non protected groups; (ii) jointly predicting the label and explanation of a text message, highlighting the fact that both parts of the prediction are inherently tied; (iii) comparing with a wider range of systems, providing a broader perspective on the capabilities and limitations of different language models in HS detection; (iv) specifically targeting implicit hate, on HateCheck [50] and Implicit Hate Corpus [2], providing novel insights into the subtleties of identifying and explaining implicit HS; and (v) building gold explanations for such messages on both datasets.

Moreover, attempts can be found to generate natural language explanations in other tasks. For instance, Costa et al. [51] proposed a character-level attention-enhanced long short-term memory model for generating natural language explanations in recommender systems. Building on this, Chen et al. [52] 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 [53]. Furthermore, model explainability has been addressed in studies for real world applications, as in Abbas et al. [54], where the authors addresses the challenges of interpretability and safety in applying deep reinforcement learning to industrial decision-making, and in the study of Štrumbelj et al. [55]. Here, the authors present a method for explaining any classifier's decisions by breaking down predictions into attribute contributions, and they do tests on artificial datasets and a real-world breast cancer recurrence prediction, showing that explanations align with the model's decision-making and improve with prediction accuracy. However, gaining trust in the predictions of LLMs is still a challenging task, due to their inherent black-box nature for human understanding [47,56]. Our paper investigates this topic, starting from the work of Lyu et al. [57]. In the paper, the authors do a thoroughly survey on the importance of considering explainability in NLP tasks, defined as *the extent to which the internal mechanics of a model can be presented in understandable terms to a human*. They particularly focus on two key aspects, namely Faithfulness and Plausibility.

Faithfulness: It gauges the precision with which an explanation reflects the reasoning process behind a model's prediction.

Plausibility: Conversely to faithfulness, it measures how easily the intended audience can understand the explanation of the model's prediction.

As highlighted in previous research [58–60], faithfulness is one of the most essential principles of explainability. The purpose of a model explanation is to convey the mechanism behind its predictions in a way that humans can understand. Faithfulness ensures that the explanation accurately reflects the model's underlying reasoning process. Put simply, if an explanation lacks faithfulness, it fails to meet the very definition of an explanation. On the other hand, there exists a tension between faithfulness and plausibility. In practice, faithfulness is frequently confused with plausibility. For instance, explanations that align more closely with human perception are often assumed to be more faithful, when in reality, they are merely more plausible. For this reason, we decided to assess both the Faithfulness and the Plausibility of our generated explanations, in order to assess whether they are correct in terms of explainability.

3. Explanation generation for implicit HS

In this section, we first introduce the datasets used (Section 3.1), we provide a definition of the chosen metrics (Section 3.2), and, finally we describe the experimental setting (Section 3.3) of our research.

3.1. Datasets

In this work, we use two benchmark datasets for Hate Speech detection tasks, namely HateCheck [50] and Implicit Hate Corpus [2], in order to assess the quality of the generated explanations according to both automatic and human evaluation (Section 3.2).

HateCheck provides functional tests for evaluating HS detection models. A functionality provides a classification for specific test cases in a corresponding functional test. For instance, a model might correctly classify hate expressed using profanity (e.g., "F*ck all black people") but misclassify non-hateful uses of profanity (e.g., "F*cking hell, what a day"). HateCheck comprises 3728 text messages grouped in 29 hateful and non-hateful test functionalities.[3]

These test cases were utilized to automatically generate *gold explanations*, i.e., ground-truth explanations containing the minimum essential information required for a hate speech explanation. For this dataset, gold explanations are based on the following template:

> *The message is [LABEL] against [TARGET] because it contains [FUNCTIONALITY].*

`[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 functionalities, *i.e.,* the implicit derogation `(derog_impl_h)`, appearing in messages conveying implicit HS, 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 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 [2]. 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 hateful messages, and (ii) the messages' taxonomy and implied statements for implicit instances.

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

Table 1
HateCheck and IHC text and gold explanation examples.

Dataset	Text Message	Gold Explanation
HateCheck	Muslims with any power are a threat to civilization.	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 is 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.

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).

We selected the IHC dataset because it is specifically designed for implicit hate speech data, whereas HateCheck is structured as a series of functional tests. In HateCheck, messages within each functional category are systematically perturbed to examine how LLMs respond to input variations, which we expect the models can effectively explain. Moreover, both datasets offer valuable elements for creating gold-standard explanations: IHC provides implicit negative stereotypes, while HateCheck includes manual descriptions of functionality. These features make both datasets suitable for our study over other ones.

3.2. 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 [61] 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 [62] 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]

Intersection Over Union F1 (IOU-F1) combines intersection over union, which measures the overlap between predicted and ground truth regions, with F1 score.

Accuracy assesses the correctness of predicted labels, by computing the proportion of correct predictions over all the total number of predictions.

The resulting scores for BLEU, BertScore, and IOU-F1 fall within the range of 0 to 1, with higher values indicating higher similarity.

Repetition Rate (RR) quantifies the lexical repetitiveness of a text by calculating the average ratio of n-grams that appear more than once [63,64].

ArgJudge (ArgJ) is a BERT-based model trained on human-assessed scores of counter-argument quality [65]. It evaluates the strength of a counter-argument in effectively rebutting a given argument.

Metrics for the human evaluation. Following previous works [47,49,66], 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 evaluates whether the explanation follows proper grammar and structural rules.

Informativeness assesses whether the explanation provides new information (e.g. additional context).

Soundness describes whether the explanation seems valid and logical.

The following example illustrates an explanation with high scores for Fluency (5), Informativeness (5), and Soundness (5):

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

HS: *It is time we make a serious push for the "ethnostate". if we could get 30–40 million whites to wake up it could happen.*

Explanation: *The message is promoting the idea of an "ethnostate", which is a political unit that is populated by and run in the interest of a particular ethnic group. This idea inherently excludes and marginalizes non-members of that ethnic group, in this case, non-whites. The phrase "wake up" suggests that those who do not support this idea are ignorant or misguided, further promoting division and hostility.*

This explanation is highly fluent because it is well-structured. It begins by defining the concept of an "ethnostate" and systematically breaks down its implications. The explanation is also informative, as it provides context for why the message is hateful, not present in the hateful message itself, offering a concise definition of an "ethnostate" and explaining how it marginalizes non-members. Finally, the explanation is highly sound because it is logically consistent, well-reasoned, and supported by clear evidence from the text.

The following example illustrates an explanation with low scores for Fluency (1), Informativeness (1), and Soundness (1):

Explanation: *Uh, this say like, we need make place only for some people. If lot people, woke up, then maybe it happen.*

On the other hand, this explanation is not fluent because it has grammatical and syntactical errors, and informal language that make the explanation hard to follow. It is also non informative, since it misses context, as it does not analyze or consider key aspects of the HS message, like "ethnostate". Finally, it is not sound because the explanation does not build a logical argument or provide evidence from the text to support its points. It merely paraphrases the message in a vague and incomplete way.

Furthermore, we use three additional metrics, specifically tailored to address RQ3:

Similarity 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.

Originality measures whether the predicted explanation offers more than a mere repetition of the input text, by rephrasing the given input.

Context evaluates if the predicted explanation provides more information beside the one present in the gold explanation.

Both Similarity and Originality scores range from 1 (lowest score) to 5 (highest score), while Context can take two values, i.e. 0 (no additional context) and 1 (more context).

The following example illustrates an explanation with high scores for Similarity (5), Originality (5), and Context (1):

HS: *funny how tolerating black crime is key to keep black politicians in power ; safe streets would bring whites back and they had want good govt.*

Gold Explanation: *The message is hateful against black people because it implies that white people are the only ones capable of keeping the streets safe.*

Gen. Explanation: *The message is hateful as it generalizes black crime and black politicians, implying that they are inherently linked. It also suggests that black politicians are in power only because of crime, and that whites inherently provide better governance and are the only ones capable of keeping the streets safe.*

The following example illustrates an explanation with low scores for Similarity (1), Originality (1), and Context (0):

Gen. Explanation: *The message is hateful because it insults black people.*

The differences between the two explanations are the following. The explanation with high similarity, originality, and context excels because it closely aligns with the message's meaning, accurately capturing its nuances and intent (similarity), it provides a detailed reasoning rather than simple rephrasing (originality), and situates the hate speech within broader societal implications, with respect to the gold explanation (context). Conversely, the explanation with low scores fails to match the message's specific nuances, relies on generic and shallow language, and lacks connection to the broader implications of the hate speech, making it less effective in conveying why the content is harmful.

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. The use of these metrics is motivated by the fact that the gold explanation and input messages are implicit: human evaluators verify that the generated explanations are aligned to the gold one regarding also implicit aspects.

3.3. Experimental settings

To jointly detecting and generating explanations for HS messages (our first objective,[6]), we carried out a thorough evaluation of different LLMs and generation strategies (**RQ1**), to discern whether a given input text is hateful or non-hateful, while jointly predicting an explanation. Therefore, we compare the classification performance of LLMs under two configurations: one where the models generate both labels and the corresponding explanation (`with-explanations` configuration) and another where the

models are tasked solely with generating labels (`without -explanations` configuration). This comparison allows us to measure the efficacy of LLMs in HS detection, in contrast to more traditional binary classification approaches. For this experiment, we used HateCheck and IHC datasets and analyzed the models' accuracy to determine whether a message is hateful or not.

Generations `with-explanations` are 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:

Zero-Shot: Given only the hateful or non-hateful text to predict, `[MESSAGE]`, we instruction-tune the LLM with the following input:

> *Given the text [MESSAGE], label if it is "hateful" or "non-hateful", and generate an explanation of why the text is "hateful" or "non-hateful" in no more than 3 sentences. Output the answer in the following structure. Label:, 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 dataset, we built pairs of (`prompt`, `label`), where `prompt` consists of a message with the same structure as the Zero-Shot configuration, and `label` is the gold explanation of the message.

Generations in the `without-explanations` configuration are guided by a prompt similar to that used in the `with-explanations` configuration, but with the request for explanations removed:

> *Given the text [MESSAGE], label if it is "hateful" or "non-hateful". Output the answer in the following structure. Label:, Explanation:.*

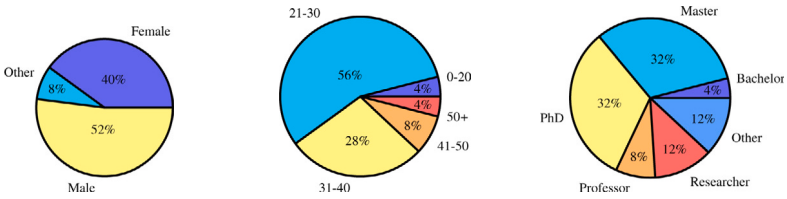
Concerning the models, we tested a range of text completion models, including GPT-3.5, GPT-4, Mistral [67], and Alpaca [68]. Both Mistral and Alpaca models are open-source resources, while GPT-3.5 and GPT-4 are accessed through the OpenAI API.[5] For GPT-3.5, we used the `gpt-3.5-turbo -0613` version, while for GPT-4, we used `gpt-4-0613`. For Mistral we tested the `Mistral-7B-v0.2`, and `Mixtral-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, 5 demonstration examples per prediction are randomly extracted from the IHC dataset.

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 Lyu et al. [53], we assessed 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 evaluated pairs of a text message and a generated explanation, following the criteria described in Section 3.2. These pairs are correctly labeled as hateful or 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 for RQ1. 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 Fig. 2) we have found a good gender balance. Most annotators fall within the 21–30 age group, and there is a varied distribution of educational backgrounds, with a higher representation of Master's and Ph.D. students. To extend Damo et al. [6], we aim at further improving detection and explanation results. We build on the findings from the previous research questions, which indicate that GPT-4, when used with joint explanations, outperforms other LLMs in hate speech detection tasks. However, we hypothesize that fine-tuned BERT-based models, trained on both explicit and implicit hate speech datasets, may still achieve superior performance (**RQ5**). To test this hypothesis, we evaluate HS versus Non-HS classification performance using Accuracy and F1 scores. The evaluation is conducted on HateCheck and IHC comparing GPT-4 with explanations against BERT [69] and HateBERT [70] models. The hate speech detectors used are BERT and HateBERT, fine-tuned on two benchmark datasets for implicit and subtle hate speech, ISHate, and IHC, described in [42,71].

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



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

Fig. 2. Annotators' demographic distribution in RQ4.

For **RQ6**, we explore the assumption that BERT-based models outperform LLMs in hate speech detection, particularly in implicit scenarios. Building on this, we aim to improve the performance of the best-performing LLM in our study—GPT-4 `with-explanations`. To achieve this, we assume that incorporating insights from BERT models could guide GPT-4 to enhance its detection capabilities. Based on this approach, we develop the following methods:

Label + Probability: The method employs a BERT-based model to generate the prediction of hatefulness for each message in HateCheck and IHC, and it also provides the probability of being hateful or not. This information is inserted into the base prompt, and we generate the new final label with an explanation of why the message is hateful or not. The prompt for this configuration is as follows:

> *Given the text `[MESSAGE]`, and the label `[Y_PRED]` and probability `[Y_PROB]` from a detector, check if the label `[Y_PRED]` is true, and generate an explanation of why the text is "hateful" or "non-hateful" in no more than 3 sentences. Output the answer in the following structure. Label:, Explanation:.*

It is important to note that the final label produced by the LLM may differ from the initial label provided by the detector, as the LLM has the ability to reassess and potentially correct the detector's prediction.

Implied Statement: We adapt the method described in Bonaldi et al. [72] about the use of the Implied Statement to create good quality counter speech to our framework. The idea is similar to the base generation of GPT-4 with explanations. However, this method first detects with a BERT model if a message is actually hateful or not. If it is hateful, its associated implied statement (i.e. the implicit negative stereotype present in a hateful message) is extracted by using the best-performing configuration $HS \rightarrow IS$ of a fine-tuned BART model described in Akazawa et al. [73]. All the extracted implied statements have a predefined structure: `subject--predicate--object`, e.g. "Muslims are terrorists". We used the following prompt:

If the message is considered "hateful" by the detector, we use the prompt:

> *Given the text `[MESSAGE]`, the label "hateful", probability `[Y_PROB]`, and implied statement `[IMPLIED_STATEMENT]`, check if the label "hateful" is true, and generate an explanation of why the text is "hateful" or "non-hateful" in no more than 3 sentences. Output the answer in the following structure. Label:, Explanation:.*

If the message is considered "non-hateful" by the detector, we use the Label + Probability prompt configuration.

Relevance Tokens: We also explore the extraction of the most relevant tokens within a text message that contribute to make its prediction as "hateful". Specifically, we utilize a detector to calculate these scores using SHAP values [74] for the positive (hateful) class. Each token is assigned a relevance score, and tokens with scores above the 90th percentile were identified as key contributors to the decision. These tokens are then used as contextual information to guide the prediction.

To validate the accuracy of the relevant tokens identified, we test them against the HateXplain dataset [44], achieving an F1 score of 0.30 on the task of rationale HS detection. While this result might appear modest compared to scores typically seen within the dataset, it is important to note that the model is not fine-tuned on HateXplain data, making the results significant in this context. The prompt for this configuration is as follows:

> *Given the text `[MESSAGE]`, and the relevance scores for each word in the text `[RELEVANCE SCORES]`. Label if it is "hateful" or "non-hateful", and generate an explanation of why the text is "hateful" or "non-hateful" in no more than 3 sentences. Output the answer in the following structure. Label:, Explanation:.*

Therefore, for this research question, we evaluate the following methods: GPT-4 `with--explanations` as baseline (GPT-4_{base}), generations guided by the detector's prediction and confidence scores (GPT-$4_{label+prob}$), generations based on implied statements (GPT-4_{IS}), and generations utilizing relevant tokens identified by the detector (GPT-4_{rel_tokens}).

Finally, for **RQ7**, we take the generations obtained by the systems described in **RQ6**, and evaluate them using Repetition Rate and ArgJudge scores (described in Section 3.2). We choose these scores as we would like to analyze the quality of generated explanations through the perspective of repetitivity with respect to other explanations and cogency. We use the definition of cogency extracted from [72], which measures the amount and logical correctness of the supporting arguments provided by the explanation.

Table 2
Ablation Study: Accuracy, BLEU, BertScore, and IOU-F1 scores for all the models described in Section 3.3. The models were evaluated on HateCheck and IHC. BLEU, BERT, and IOU-F1 do not apply for results without explanations (Cells with a hyphen).

		Model	HateCheck				Implicit Hate Corpus			
			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
		Mixtral-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
		Mixtral-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
		Mixtral-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	–	–	–
		Mixtral-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	–	–	–
		Mixtral-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	–	–	–
		Mixtral-8X7B	0.851	–	–	–	0.725	–	–	–

Table 3
Accuracy scores per HateCheck functionality for the GPT-4 model with and without explanations.

	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

4. Evaluation

Regarding **RQ1**, Table 2 shows the models' accuracy on HateCheck and IHC, and the automatic metrics. Concerning HateCheck, the models obtaining the highest accuracy are Mixtral-8X7B, GPT-4, and Mistral-7B for Zero-shot, Few-shot and Fine-tuning configurations, respectively. Overall, GPT-4 with Few-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 [75]) 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 Mixtral-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 4
Human evaluation results for RQ3 (Table 4(a)) and RQ4 (Table 4(b)).

	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 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 instances 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 Mixtral-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 4(a) 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 4(b) 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).

Regarding **RQ5**, Table 5(a) presents the Accuracy and F1 scores for hate speech detection on both HateCheck and IHC. Overall, GPT-4 with explanations demonstrates strong performance, surpassing BERT and HateBERT on HateCheck. However, when dealing with implicit hate speech from the IHC, GPT-4 lags behind these more traditional models, consistent with findings from similar studies [71]. Analyzing HateCheck per functionality (Table 5(b)), as noted in RQ1, GPT-4 underperforms in non-hateful categories such as Counter Quote, Counter References, Profanity, Reclaimed Slur, References to non-hateful target groups, and target individuals, where BERT and HateBERT achieve higher scores. For the hateful functionalities, GPT-4 maintains good performance, with no significant declines compared to the other models.

Concerning **RQ6**, Table 5(a) presents the classification results of GPT-4$_{IS}$, GPT-4$_{label+pred}$, and GPT-4$_{rel_tokens}$. These methods demonstrate the ability to maintain or even surpass baseline performance on HateCheck, while also improving detection of implicit messages in IHC. Among them, GPT-4$_{IS}$ stands out as the most balanced approach. Additionally, it performs better on non-hateful

Table 5
Comparison between GPT-4 `with-explanations`, and BERT/HateBERT models fine-tuned on implicit and subtle data. BERT_{ISHate} and HateBERT_{ISHate} correspond to models fine-tuned on ISHate, while BERT_{IHC} and HateBERT_{IHC} correspond to models fine-tuned on IHC.

Model	HateCheck		IHC	
	Accuracy	F1	Accuracy	F1
BERT_{ISHate}	0.7449	0.8129	0.7556	0.8608
HateBERT_{ISHate}	0.7685	0.8252	0.6977	0.8220
BERT_{IHC}	0.7086	0.7775	**0.8521**	**0.9201**
HateBERT_{IHC}	0.7156	0.7775	0.8435	0.9151
GPT-4_{base}	**0.8454**	**0.8871**	0.6677	0.8008
GPT-4_{IS}	0.8538	0.8918	0.8264	0.9049
$\text{GPT-4}_{label+prob}$	0.8689	0.9023	0.7760	0.8739
$\text{GPT-4}_{rel-tok}$	0.8521	0.8914	0.7738	0.8725

(a) Overall Accuracy and F1 results on HateCheck and Implicit Hate Corpus.

Functionality	GPT-4_{base}	BERT_{ISHate}	HateBERT_{ISHate}	BERT_{IHC}	HateBERT_{IHC}	GPT-4_{IS}	$\text{GPT-4}_{label+prob}$	$\text{GPT-4}_{rel-tok}$
counter_quote_nh	0.0578	0.2948	**0.4855**	0.3410	0.4566	0.075	0.127	0.098
counter_ref_nh	0.0780	0.4113	0.4539	0.3333	**0.5248**	0.184	0.298	0.191
profanity_nh	0.9300	**1.0000**	0.9900	0.8400	0.8400	0.950	0.950	**1.000**
slur_reclaimed_nh	0.4691	0.6296	**0.8148**	0.2222	0.2222	0.617	0.728	0.605
target_group_nh	0.3226	0.4194	0.4516	**0.7258**	**0.7258**	0.323	0.258	0.226
target_indiv_nh	0.1538	0.7231	**0.8308**	0.6308	0.4769	0.262	0.308	0.077

(b) Accuracy and F1 results per functionalities on HateCheck.

Table 6
Repetition rate and ArgumentJudge results on HateCheck and Implicit Hate Corpus.

Model	HateCheck		IHC	
	RR	ArgJudge	RR	ArgJudge
GPT-4_{base}	21.7720	4.6071	11.9750	4.5659
GPT-4_{IS}	8.5560	4.8548	6.3700	4.6163
$\text{GPT-4}_{label+prob}$	**8.4530**	4.9040	**5.8510**	**4.8060**
$\text{GPT-4}_{rel-tokens}$	8.6930	**4.9107**	6.0060	4.8017

categories when evaluated per functionality, as shown in Table 5(b). This suggests that incorporating detector information can effectively guide LLM predictions, leading to improved results.

Finally, concerning **RQ7**, from Table 6, we see that adding BERT-based model information to the base GPT-4 also improves the quality of the generated explanations. More specifically, regarding the *Repetition Rate*, all our methods obtain a lower score with respect to the one of GPT-4 base. This means that the new explanations are more diverse in content and more specific to the hateful message they explain. Regarding *ArgJudge* results, we obtain a similar effect obtaining an increment on the cogency over the baseline for each of the methods.

4.1. Discussion and error analysis

Regarding **RQ1**, our analysis indicate that, on the whole, our models have a sufficient performance in detecting HS, with GPT-4 demonstrating the highest overall results. We see 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 3), 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 misclassifications 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 and IoU 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. However, it can rise to the problem of hallucinations, text that is nonsensical or unfaithful to the provided source input [76], often containing factually incorrect content.

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 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 correctly labeled as non-hateful by the models, though an annotator perceived them as sarcastic.

Regarding **RQ5**, the results indicate that BERT-based detectors, fine-tuned on implicit and subtle messages, outperform GPT-4 in terms of accuracy for non-hateful instances in the HateCheck dataset and for implicit messages in the IHC dataset. This highlights the continued effectiveness of specialized fine-tuned models for these specific tasks, particularly when implicitness are key challenges.

Moving to **RQ6**, the integration of BERT-based models information into LLMs shows promising results. Specifically, GPT-4_{IS} emerges as the most balanced model configuration. By incorporating implicit negative stereotypes and relevant token for prediction into the generation of explanations, this configuration successfully maintains the overall performance of the LLM on HateCheck while improving it on IHC. This suggests that jointly using fine-tuned BERT detectors with LLMs can improve overall performance on the HS detection task.

Finally, concerning **RQ7**, each of the methods leads to generated explanations with a lower repetition rate, being more relevant and specific to the hateful message being explained, while also improving the cogency, that is, the number of logical and sound arguments used for describing each explanation. Therefore, jointly using BERT detectors with LLMs can also lead to gains on producing better quality explanations when detecting HS and Non-HS messages.

4.2. Challenges and considerations

We identify a set of challenges in generating natural language explanations, encompassing both limitations of current models and open challenges:

Language and Culture. Hate speech is culturally dependent, so generated responses must be culturally informed [77]. Terms can carry different meanings depending on the country [78,79], language [80], or historical context [81,82].

Explicit vs. Implicit Language and Connotation. Explanations should be appropriate to the group targeted and the specific terms used [72,83]. Some expressions carry negative connotations either explicitly or implicitly. For example, the phrase "taking jobs" may be flagged by detectors (especially traditional models) as implicit hate speech because of its association with anti-immigrant sentiment. LLMs might generate an explanation like: "The phrase 'taking jobs from hardworking citizens' is a common dog whistle, implying immigrants are a threat to the economy. While not explicitly hateful, it reinforces harmful stereotypes".

On the other hand, similar phrases may be used with a neutral or positive intent, such as: "I am taking on jobs to meet my financial obligations". Understanding connotation—whether positive, neutral, or negative—and whether a term is used explicitly or implicitly is crucial for both detection and explanation.

Type of Haters. Hate speech detection often focuses on the target but overlooks the type and motivation of the hater—such as true believers, trolls, the unaware (ignorant of harm), or conformists (seeking social approval). These distinctions, along with demographic factors like age and location, are important to consider [77].

Hallucinations. Highlighting the absurdity of hate speech can be effective for tasks such as explanation generation [48], counter-speech [77], or misinformation verdicts [84]. In this work, we explore LLMs using zero-shot, few-shot, and fine-tuning approaches, as well as models guided by detector context. However, open-ended generation—especially in zero- or few-shot settings—often leads to unfaithful or nonsensical outputs, frequently containing factually incorrect content (hallucinations) [76,77,85].

Evaluation. There is currently no standardized way to evaluate explanations for hate speech. To address this gap, we leverage the implied statements from IHC and the test functionality explanations from HateCheck to construct gold-standard explanations. These serve as references against which generated explanations can be compared. However, evaluation datasets remain limited—Hatexplain [44] is a useful resource, but it only provides rationales rather than full explanations. This stands in contrast to other domains such as counterspeech generation [86] or misinformation verdict generation [87], where the tasks are better defined and supported by dedicated resources. Moreover, there is a lack of datasets aimed at assessing or classifying explanation quality, such as those used to evaluate the effectiveness of counterspeech [88]. In human evaluations, dimensions are often not properly standardized, and what constitutes a high-quality explanation depends not only on the authors' subjectivity but also on the specific aspects being measured and the research questions being addressed.

5. Conclusions

Our previous study focused on **RQ1** to **RQ4**—establishing the benefits of generating explanations alongside labels, evaluating their faithfulness, and confirming their human interpretability. Our results confirmed that explanations significantly enhance HS classification accuracy (**RQ1**), with models like GPT-4 achieving strong performance, particularly when paired with Few-Shot learning. Explanations were also shown to be faithful (**RQ2**), effective in addressing implied meanings in implicit HS (**RQ3**), and highly fluent, informative, and sound according to human evaluations (**RQ4**).

This work extends our prior research on hate speech detection and explanation [6] by addressing new questions about the performance and enhancement of LLMs. This paper adds a deeper analysis by exploring **RQ5**, **RQ6**, and **RQ7**. Specifically, we investigate how LLMs compare to traditional HS detectors and whether incorporating external information improves their detection and explanation quality.

This study further demonstrates that while LLMs perform well in explicit HS detection, traditional models such as BERT and HateBERT are better in non-hateful and implicit categories (**RQ5**). Importantly, supplementing LLMs with external information—such as labels, probabilities, and implied stereotypes—improves their performance on implicitness and non-hs detection (**RQ6**) and enhances the clarity and quality of their explanations (**RQ7**).

In summary, coupling LLMs with explanation-driven approaches and external guidance improves both detection accuracy and interpretability.

This work is significant for both NLP research and practical applications. It advances explainable AI by demonstrating how explanations can improve model transparency, faithfulness, and user trust. By tackling implicit and subtle hate speech, it addresses a critical challenge in NLP, paving the way for more nuanced and context-aware models. Moreover, the integration of traditional models and LLMs showcases a hybrid approach that enhances both accuracy and interpretability. Future work will refine these methods and explore their application across broader datasets and diverse languages.

CRediT authorship contribution statement

Greta Damo: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Nicolás Benjamín Ocampo:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Elena Cabrio:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization. **Serena Villata:** Writing – review & editing, Supervision, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been supported by the French government, through the 3IA Cote d'Azur investments in the project managed by the National Research Agency (ANR) with the reference number ANR-23-IACL-0001.

Data availability

The data and accompanying software can be found at https://github.com/benjaminocampo/dke_gen_hs_explanations.

References

[1] N.B. Ocampo, E. Sviridova, E. Cabrio, S. Villata, 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, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 1997–2013.

[2] M. ElSherief, C. Ziems, D. Muchlinski, V. Anupindi, J. Seybolt, M. De Choudhury, D. Yang, Latent hatred: A benchmark for understanding implicit hate speech, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 345–363.

[3] T. Hartvigsen, S. Gabriel, H. Palangi, M. Sap, D. Ray, E. Kamar, 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), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 3309–3326.

[4] M. Wiegand, J. Ruppenhofer, E. Eder, 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, Association for Computational Linguistics, 2021, pp. 576–587, Online.

[5] A. Calabrese, L. Neves, N. Shah, M. Bos, B. Ross, M. Lapata, F. Barbieri, Explainability and hate speech: Structured explanations make social media moderators faster, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 398–408, http://dx.doi.org/10.18653/v1/2024.acl-short.38, URL https://aclanthology.org/2024.acl-short.38.

[6] G. Damo, N.B. Ocampo, E. Cabrio, S. Villata, Unveiling the hate: Generating faithful and plausible explanations for implicit and subtle hate speech detection, in: A. Rapp, L. Di Caro, F. Meziane, V. Sugumaran (Eds.), Natural Language Processing and Information Systems, Springer Nature Switzerland, Cham, 2024, pp. 211–225.

[7] Council of Europe, Dealing with propaganda, misinformation and fake news, 2025, URL https://www.coe.int/en/web/campaign-free-to-speak-safe-to-learn/dealing-with-propaganda-misinformation-and-fake-news. (Accessed: 02 April 2025).

[8] K. Saha, E. Chandrasekharan, M.D. Choudhury, Prevalence and psychological effects of hateful speech in online college communities, Proc. the ACM Web Sci. Conf. (2019) 255–264, http://dx.doi.org/10.1145/3292522.3326032, URL https://pmc.ncbi.nlm.nih.gov/articles/PMC7500692/. (Accessed: 02 April 2025).

[9] A. Dreißigacker, P. Müller, A. Isenhardt, J. Schemmel, Online hate speech victimization: consequences for victims' feelings of insecurity, Crime Sci. 13 (4) (2024) http://dx.doi.org/10.1186/s40163-024-00204-y, URL https://crimesciencejournal.biomedcentral.com/articles/10.1186/s40163-024-00204-y. (Accessed: 02 April 2025).

[10] H. Horacek, E. Kapetanios, E. Metais, F. Meziane, Editorial for VSI:NLDB-saarbruecken-2021, Data Knowl. Eng. 149 (2024) 102259, http://dx.doi.org/10.1016/j.datak.2023.102259, URL https://www.sciencedirect.com/science/article/pii/S0169023X23001192.

[11] R. Sepúlveda-Torres, M. Vicente, E. Saquete, E. Lloret, M. Palomar, Leveraging relevant summarized information and multi-layer classification to generalize the detection of misleading headlines, Data Knowl. Eng. 145 (2023) 102176, http://dx.doi.org/10.1016/j.datak.2023.102176, URL https://www.sciencedirect.com/science/article/pii/S0169023X23000368.

[12] C. Veres, J. Sampson, Self supervised learning and the poverty of the stimulus, Data Knowl. Eng. 147 (2023) 102208, http://dx.doi.org/10.1016/j.datak.2023.102208, URL https://www.sciencedirect.com/science/article/pii/S0169023X2300068X.

[13] J. Kersting, F. Maoro, M. Geierhos, Towards comparable ratings: Exploring bias in german physician reviews, Data Knowl. Eng. 148 (2023) 102235, http://dx.doi.org/10.1016/j.datak.2023.102235, URL https://www.sciencedirect.com/science/article/pii/S0169023X23000952.

[14] J.T. Nockleby, Hate speech, in: L.W. Levy, K.L. Karst, D.J. Mahoney (Eds.), Encyclopedia of the American Constitution, second ed., Macmillan, 2000, pp. 1277–1279.

[15] N.D. Gitari, Z. Zhang, H. Damien, J. Long, A lexicon-based approach for hate speech detection, Int. J. Multimed. Ubiquitous Eng. 10 (4) (2015) 215–230.

[16] M. Wiegand, J. Ruppenhofer, A. Schmidt, C. Greenberg, Inducing a lexicon of abusive words – a feature-based approach, in: M. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 1046–1056.

[17] B. Gambäck, U.K. Sikdar, Using convolutional neural networks to classify hate-speech, in: Proceedings of the First Workshop on Abusive Language Online, Association for Computational Linguistics, Vancouver, BC, Canada, 2017, pp. 85–90.

[18] J.-H. Lee, J.-U. Park, J.-W. Cha, Y.-S. Han, Detecting context abusiveness using hierarchical deep learning, in: Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 10–19.

[19] H. Sohn, H. Lee, MC-BERT4hate: Hate speech detection using multi-channel BERT for different languages and translations, in: 2019 International Conference on Data Mining Workshops, ICDMW, (ISSN: 2375-9259) 2019, pp. 551–559.

[20] K. Wang, D. Lu, C. Han, S. Long, J. Poon, Detect all abuse! toward universal abusive language detection models, in: Proceedings of the 28th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 6366–6376.

[21] T. Davidson, D. Warmsley, M. Macy, I. Weber, Automated hate speech detection and the problem of offensive language, Proc. Int. AAAI Conf. Web Soc. Media 11 (1) (2017) 512–515, Number: 1.

[22] A. Founta, C. Djouvas, D. Chatzakou, I. Leontiadis, J. Blackburn, G. Stringhini, A. Vakali, M. Sirivianos, N. Kourtellis, Large scale crowdsourcing and characterization of Twitter abusive behavior, Proc. Int. AAAI Conf. Web Soc. Media 12 (1) (2018) Number: 1.

[23] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F.M. Rangel Pardo, P. Rosso, M. Sanguinetti, 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, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 54–63.

[24] M. Zampieri, S. Malmasi, P. Nakov, S. Rosenthal, N. Farra, R. Kumar, SemEval-2019 task 6: Identifying and categorizing offensive language in social media (OffensEval), in: Proceedings of the 13th International Workshop on Semantic Evaluation, Association for Computational Linguistics, Minneapolis, Minnesota, USA, 2019, pp. 75–86.

[25] D. Nozza, F. Bianchi, G. Attanasio, HATE-ITA: Hate speech detection in Italian social media text, in: K. Narang, A. Mostafazadeh Davani, L. Mathias, B. Vidgen, Z. Talat (Eds.), Proceedings of the Sixth Workshop on Online Abuse and Harms, WOAH, Association for Computational Linguistics, Seattle, Washington (Hybrid), 2022, pp. 252–260, http://dx.doi.org/10.18653/v1/2022.woah-1.24, URL https://aclanthology.org/2022.woah-1.24.

[26] B.R. Chakravarthi, R. Priyadharshini, T. Durairaj, J. McCrae, P. Buitelaar, P. Kumaresan, R. Ponnusamy, 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, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 369–377.

[27] D. Locatelli, G. Damo, D. Nozza, A cross-lingual study of homotransphobia on Twitter, in: Proceedings of the First Workshop on Cross-Cultural Considerations in NLP, C3NLP, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 16–24.

[28] S. Jamshidi, M. Mohammadi, S. Bagheri, H.E. Najafabadi, A. Rezvanian, M. Gheisari, M. Ghaderzadeh, A.S. Shahabi, Z. Wu, Effective text classification using BERT, MTM LSTM, and DT, Data Knowl. Eng. 151 (2024) 102306.

[29] K.L. Nadal, K.E. Griffin, Y. Wong, S. Hamit, M. Rasmus, The impact of racial microaggressions on mental health: Counseling implications for clients of color, J. Couns. Dev. 92 (1) (2014) 57–66.

[30] J.W. Kanter, M.T. Williams, A.M. Kuczynski, K.E. Manbeck, M. Debreaux, D.C. Rosen, A preliminary report on the relationship between microaggressions against black people and racism among white college students, Race Soc. Probl. 9 (4) (2017) 291–299.

[31] M. Sap, S. Gabriel, L. Qin, D. Jurafsky, N.A. Smith, Y. Choi, Social bias frames: Reasoning about social and power implications of language, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 5477–5490, Online.

[32] B. Vidgen, T. Thrush, Z. Waseem, D. Kiela, 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), Association for Computational Linguistics, 2021, pp. 1667–1682, Online.

[33] M. Wiegand, J. Kampfmeier, E. Eder, J. Ruppenhofer, 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, Association for Computational Linguistics, Singapore, 2023, pp. 16280–16297.

[34] X. Han, Y. Tsvetkov, Fortifying toxic speech detectors against veiled toxicity, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP, Association for Computational Linguistics, 2020, pp. 7732–7739, Online.

[35] S. Ghosh, M. Suri, P. Chiniya, U. Tyagi, S. Kumar, D. Manocha, 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, Association for Computational Linguistics, Singapore, 2023, pp. 6159–6173.

[36] I. Nejadgholi, K. Fraser, S. Kiritchenko, 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), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 5517–5529.

[37] S. Roychowdhury, V. Gupta, Data-efficient methods for improving hate speech detection, in: Findings of the Association for Computational Linguistics: EACL 2023, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 125–132.
[38] N.B. Ocampo, E. Cabrio, S. Villata, 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, Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2758–2772.
[39] J. Wen, P. Ke, H. Sun, Z. Zhang, C. Li, J. Bai, M. Huang, Unveiling the implicit toxicity in large language models, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 1322–1338.
[40] D. Jurgens, L. Hemphill, E. Chandrasekharan, 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, Association for Computational Linguistics, Florence, Italy, 2019, pp. 3658–3666.
[41] Z. Waseem, T. Davidson, D. Warmsley, I. Weber, Understanding abuse: A typology of abusive language detection subtasks, in: Proceedings of the First Workshop on Abusive Language Online, Association for Computational Linguistics, Vancouver, BC, Canada, 2017, pp. 78–84.
[42] N.B. Ocampo, E. Cabrio, S. Villata, Unmasking the hidden meaning: Bridging implicit and explicit hate speech embedding representations, in: Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 6626–6637.
[43] Y. Kim, S. Park, Y. Namgoong, Y.-S. Han, 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, Association for Computational Linguistics, Singapore, 2023, pp. 10964–10980.
[44] B. Mathew, P. Saha, S.M. Yimam, C. Biemann, P. Goyal, A. Mukherjee, HateXplain: A benchmark dataset for explainable hate speech detection, Proc. the AAAI Conf. Artif. Intell. 35 (17) (2021) 14867–14875, Number: 17.
[45] J. Kim, B. Lee, K.-A. Sohn, Why is it hate speech? Masked rationale prediction for explainable hate speech detection, in: Proceedings of the 29th International Conference on Computational Linguistics, International Committee on Computational Linguistics, Gyeongju, Republic of Korea, 2022, pp. 6644–6655.
[46] G. Damo, N.B. Ocampo, E. Cabrio, S. Villata, PEACE: Providing explanations and analysis for combating hate expressions, in: ECAI 2024, IOS Press, 2024, http://dx.doi.org/10.3233/faia241029.
[47] H. Wang, M.S. Hee, M.R. Awal, K.T.W. Choo, R.K.-W. Lee, Evaluating GPT-3 generated explanations for hateful content moderation, 6, (ISSN: 1045-0823) 2023, pp. 6255–6263.
[48] Y. Yang, J. Kim, Y. Kim, N. Ho, J. Thorne, S.-Y. Yun, HARE: Explainable hate speech detection with step-by-step reasoning, in: Findings of the Association for Computational Linguistics: EMNLP 2023, Association for Computational Linguistics, Singapore, 2023, pp. 5490–5505.
[49] F. Huang, H. Kwak, J. An, 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, 2023, pp. 90–93, arXiv:2209.04889 [cs].
[50] P. Röttger, B. Vidgen, D. Nguyen, Z. Waseem, H. Margetts, J. Pierrehumbert, 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), Association for Computational Linguistics, 2021, pp. 41–58, Online.
[51] F. Costa, S. Ouyang, P. Dolog, A. Lawlor, Automatic generation of natural language explanations, in: Proceedings of the 23rd International Conference on Intelligent User Interfaces Companion, in: IUI '18 Companion, Association for Computing Machinery, New York, NY, USA, 2018, pp. 1–2.
[52] Q. Chen, F. Ji, X. Zeng, F.-L. Li, J. Zhang, H. Chen, Y. Zhang, 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), Association for Computational Linguistics, 2021, pp. 2516–2527, Online.
[53] Q. Lyu, M. Apidianaki, C. Callison-Burch, Towards faithful model explanation in NLP: A survey, 2023, arXiv:2209.11326.
[54] A.N. Abbas, G.C. Chasparis, J.D. Kelleher, Hierarchical framework for interpretable and specialized deep reinforcement learning-based predictive maintenance, Data Knowl. Eng. 149 (2024) 102240.
[55] E. Štrumbelj, I. Kononenko, M.R. Šikonja, Explaining instance classifications with interactions of subsets of feature values, Data Knowl. Eng. 68 (10) (2009) 886–904.
[56] W. Samek, T. Wiegand, K.-R. Müller, Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models, 2017, arXiv:1708.08296 [cs, stat].
[57] Q. Lyu, S. Havaldar, A. Stein, L. Zhang, D. Rao, E. Wong, M. Apidianaki, C. Callison-Burch, Faithful chain-of-thought reasoning, in: J.C. Park, Y. Arase, B. Hu, W. Lu, D. Wijaya, A. Purwarianti, A.A. Krisnadhi (Eds.), Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Nusa Dua, Bali, 2023, pp. 305–329, http://dx.doi.org/10.18653/v1/2023.ijcnlp-main.20, URL https://aclanthology.org/2023.ijcnlp-main.20.
[58] S. Jain, B.C. Wallace, Attention is not explanation, 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 3543–3556.
[59] J. Bastings, K. Filippova, The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? in: Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP, Association for Computational Linguistics, 2020, pp. 149–155, Online.
[60] A. Jacovi, Y. Goldberg, Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020, pp. 4198–4205, Online.
[61] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 2002, pp. 311–318.
[62] T. Zhang, V. Kishore, F. Wu, K.Q. Weinberger, Y. Artzi, BERTScore: Evaluating text generation with BERT, 2020.
[63] N. Bertoldi, M. Cettolo, M. Federico, Cache-based online adaptation for machine translation enhanced computer assisted translation, in: Proceedings of MT-Summit, 2013, pp. 35–42.
[64] M. Cettolo, N. Bertoldi, M. Federico, The repetition rate of text as a predictor of the effectiveness of machine translation adaptation, in: Proceedings of the 11th Biennial Conference of the Association for Machine Translation in the Americas, AMTA 2014, 2014, pp. 166–179.
[65] J. Lin, R. Ye, M. Han, Q. Zhang, R. Lai, X. Zhang, Z. Cao, X.-J. Huang, Z. Wei, Argue with me tersely: Towards sentence-level counter-argument generation, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 16705–16720.
[66] M.-A. Clinciu, A. Eshghi, H. Hastie, 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, Association for Computational Linguistics, 2021, pp. 2376–2387, http://dx.doi.org/10.18653/v1/2021.eacl-main.202, Online, URL https://aclanthology.org/2021.eacl-main.202.
[67] A.Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D.S. Chaplot, D.d.l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L.R. Lavaud, M.-A. Lachaux, P. Stock, T.L. Scao, T. Lavril, T. Wang, T. Lacroix, W.E. Sayed, Mistral 7B, 2023, arXiv:2310.06825 [cs].
[68] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T.B. Hashimoto, Stanford alpaca: An instruction-following llama model, 2023.
[69] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), 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), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186, http://dx.doi.org/10.18653/v1/N19-1423, URL https://aclanthology.org/N19-1423.
[70] T. Caselli, V. Basile, J. Mitrović, M. Granitzer, HateBERT: Retraining BERT for abusive language detection in english, in: A. Mostafazadeh Davani, D. Kiela, M. Lambert, B. Vidgen, V. Prabhakaran, Z. Waseem (Eds.), Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021), Association for Computational Linguistics, 2021, pp. 17–25, http://dx.doi.org/10.18653/v1/2021.woah-1.3, Online, URL https://aclanthology.org/2021.woah-1.3.

[71] G. Damo, N.B. Ocampo, E. Cabrio, S. Villata, PEACE: Providing explanations and analysis for combating hate expressions, in: Proceedings of the 27th European Conference on Artificial Intelligence, ECAI 2024, Santiago de Compostela, Spain, 2024, HAL-04684950.
[72] H. Bonaldi, G. Damo, N.B. Ocampo, E. Cabrio, S. Villata, M. Guerini, Is safer better? The impact of guardrails on the argumentative strength of LLMs in hate speech countering, in: Y. Al-Onaizan, M. Bansal, Y.-N. Chen (Eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Miami, Florida, USA, 2024, pp. 3446–3463, URL https://aclanthology.org/2024.emnlp-main.201.
[73] N. Akazawa, S.S. Tekiroğlu, M. Guerini, Distilling implied bias from hate speech for counter narrative selection, in: Y.-L. Chung, H. Bonaldi, G. Abercrombie, M. Guerini (Eds.), Proceedings of the 1st Workshop on CounterSpeech for Online Abuse, CS4OA, Association for Computational Linguistics, Prague, Czechia, 2023, pp. 29–43, URL https://aclanthology.org/2023.cs4oa-1.3.
[74] S.M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, in: I. Guyon, U.V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, R. Garnett (Eds.), Advances in Neural Information Processing Systems 30, Curran Associates, Inc., 2017, pp. 4765–4774, URL http://papers.nips.cc/paper/7062-a-unified-approach-to-interpreting-model-predictions.pdf.
[75] B. Efron, Bootstrap methods: Another look at the jackknife, Ann. Statist. 7 (1) (1979) 1–26, http://dx.doi.org/10.1214/aos/1176344552.
[76] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y.J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Comput. Surv. 55 (12) (2023) http://dx.doi.org/10.1145/3571730.
[77] H. Bonaldi, Y.-L. Chung, G. Abercrombie, M. Guerini, NLP for counterspeech against hate: A survey and how-to guide, in: K. Duh, H. Gomez, S. Bethard (Eds.), Findings of the Association for Computational Linguistics: NAACL 2024, Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 3480–3499, http://dx.doi.org/10.18653/v1/2024.findings-naacl.221, URL https://aclanthology.org/2024.findings-naacl.221/.
[78] M. Tonneau, D. Liu, S. Fraiberger, R. Schroeder, S.A. Hale, P. Röttger, From languages to geographies: Towards evaluating cultural bias in hate speech datasets, in: Y.-L. Chung, Z. Talat, D. Nozza, F.M. Plaza-del Arco, P. Röttger, A. Mostafazadeh Davani, A. Calabrese (Eds.), Proceedings of the 8th Workshop on Online Abuse and Harms, WOAH 2024, Association for Computational Linguistics, Mexico City, Mexico, 2024, pp. 283–311, http://dx.doi.org/10.18653/v1/2024.woah-1.23, URL https://aclanthology.org/2024.woah-1.23/.
[79] G. Castillo-lópez, A. Riabi, D. Seddah, Analyzing zero-shot transfer scenarios across spanish variants for hate speech detection, in: Y. Scherrer, T. Jauhiainen, N. Ljubeˇ sić, P. Nakov, J. Tiedemann, M. Zampieri (Eds.), Tenth Workshop on NLP for Similar Languages, Varieties and Dialects, VarDial 2023, Association for Computational Linguistics, Dubrovnik, Croatia, 2023, pp. 1–13, http://dx.doi.org/10.18653/v1/2023.vardial-1.1, URL https://aclanthology.org/2023.vardial-1.1.
[80] K. Korre, A. Muti, A. Barrón-Cedeño, The challenges of creating a parallel multilingual hate speech corpus: An exploration, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC-COLING 2024, ELRA and ICCL, Torino, Italia, 2024, pp. 15842–15853, URL https://aclanthology.org/2024.lrec-main.1376/.
[81] M. Laurent, Project hatemeter: helping NGOs and social science researchers to analyze and prevent anti-muslim hate speech on social media, Procedia Comput. Sci. 176 (2020) 2143–2153, http://dx.doi.org/10.1016/j.procs.2020.09.251, Knowledge-Based and Intelligent Information & Engineering Systems: Proceedings of the 24th International Conference KES2020. URL https://www.sciencedirect.com/science/article/pii/S1877050920321554.
[82] S. Hoeken, S. Spliethoff, S. Schwandt, S. Zarrieß, Ö. Alacam, Towards detecting lexical change of hate speech in historical data, in: N. Tahmasebi, S. Montariol, H. Dubossarsky, A. Kutuzov, S. Hengchen, D. Alfter, F. Periti, P. Cassotti (Eds.), Proceedings of the 4th Workshop on Computational Approaches To Historical Language Change, Association for Computational Linguistics, Singapore, 2023, pp. 100–111, http://dx.doi.org/10.18653/v1/2023.lchange-1.11, URL https://aclanthology.org/2023.lchange-1.11.
[83] J. Bengoetxea, Y.-L. Chung, M. Guerini, R. Agerri, Basque and spanish counter narrative generation: Data creation and evaluation, in: N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, N. Xue (Eds.), Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC-COLING 2024, ELRA and ICCL, Torino, Italia, 2024, pp. 2132–2141, URL https://aclanthology.org/2024.lrec-main.192/.
[84] D. Russo, S. Kaszefski-Yaschuk, J. Staiano, M. Guerini, Countering misinformation via emotional response generation, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 11476–11492, http://dx.doi.org/10.18653/v1/2023.emnlp-main.703, URL https://aclanthology.org/2023.emnlp-main.703.
[85] G. Damo, E. Cabrio, S. Villata, Beating harmful stereotypes through facts: rag-based counter-speech generation, 2025, URL https://arxiv.org/abs/2510.12316. arXiv:2510.12316.
[86] H. Bonaldi, M.E. Vallecillo-Rodríguez, I. Zubiaga, A. Montejo-Raez, A. Soroa, M.-T. Martín-Valdivia, M. Guerini, R. Agerri, The first workshop on multilingual counterspeech generation at COLING 2025: Overview of the shared task, in: H. Bonaldi, M.E. Vallecillo-Rodríguez, I. Zubiaga, A. Montejo-Ráez, A. Soroa, M.T. Martín-Valdivia, M. Guerini, R. Agerri (Eds.), Proceedings of the First Workshop on Multilingual Counterspeech Generation, Association for Computational Linguistics, Abu Dhabi, UAE, 2025, pp. 92–107, URL https://aclanthology.org/2025.mcg-1.10/.
[87] D. Russo, S.S. Tekiroğlu, M. Guerini, Benchmarking the generation of fact checking explanations, Trans. Assoc. Comput. Linguist. 11 (2023) 1250–1264, http://dx.doi.org/10.1162/tacl_a_00601, URL https://aclanthology.org/2023.tacl-1.71.
[88] G. Damo, E. Cabrio, S. Villata, Effectiveness of counter-speech against abusive content: A multidimensional annotation and classification study, 2025, URL https://arxiv.org/abs/2506.11919. arXiv:2506.11919.

Greta Damo is a Ph.D. student at Université Côte d'Azur, at the Interdisciplinary Institute for Artificial Intelligence 3IA Côte d'Azur. She is a member of the MARIANNE Research Team and is currently working with Natural Language Processing in the domain of Hate Speech, and Counter Speech (how to automatically generate it to combat hate expressions). She also teaches Computer Science related courses at the University Côte d'Azur.

She obtained her Bachelor's and Master's degrees in Data Science at Bocconi University.

Nicolás Benjamín Ocampo is a Postdoc researcher at Centrum Wiskunde & Informatica (CWI). He obtained his Ph.D. in Computer Science (CS) at Université Côte d'Azur, 3IA, Inria, CNRS, I3S, as a member of the MARIANNE Research Team working on Implicit and Subtle Abusive Language using Machine Learning and Natural Language Processing (NLP).

His current research interests lie in the intersection of linguistics, misinformation, hate speech, and NLP. His thesis topic focused on properly representing implicit and subtle (instances with irony, sentiment, exaggeration, circumlocution, etc.) for abusive language detection and downstream tasks. After properly characterizing these notions, he studied how to automatically counter hate through natural language explanations and counter-narratives.

Before his Ph.D., he got his Bachelor's and Master's degrees in CS at Facultad de Matematica, Astronomia, Fisica, y Computacion (FAMAFyC), Universidad Nacional de Cordoba (Argentina). There he graduated with the second-best GPA in 2021 and Special Mention Diploma.

Elena Cabrio is a Full Professor at Université Côte d'Azur, Inria, CNRS, I3S and member of the MARIANNE Research Team. She has been recently awarded with a Chair in Artificial Intelligence at the Interdisciplinary Institute for Artificial Intelligence 3IA Côte d'Azur on "AI and Natural Language".

She defended her HDR (habilitation) in October 2020.

Goal of her research is to design debating technologies for advanced decision support systems, to support the exchange of information and opinions in different domains (as healthcare and politics), leveraging interdisciplinarity and advances in machine learning for Natural Language Processing. She is currently coordinating the ANTIDOTE (ArgumeNtaTIon-Driven explainable artificial intelligence fOr digiTal mEdicine) project (CHIST-ERA XAI 2019).

Serena Villata is a senior researcher (Directrice de recherche) in computer science at the CNRS and she pursues her research at the I3S laboratory. Since February 2025, she is the head of the new joint InriaI3S MARIANNE team, focusing on natural language argumentation.

Her research area is Artificial Intelligence (AI), and her current work focuses on computational argumentation, considering both reasoning on computational models of argument and mining natural arguments from text (argument mining). She is particularly interested in multidisciplinary applications of argumentation to the legal domain, to medical texts, to political debates, and to social networks.

In November 2021, she was awarded the Prix Inria Académie des sciences jeunes chercheurs et jeunes chercheuses (Award website).

Since July 2019, she has held a Chair in Artificial Intelligence at the Interdisciplinary Institute for Artificial Intelligence 3IA Côte d'Azur on "Artificial Argumentation for Humans". Since January 2021, she has been the Deputy Scientific Director of the 3IA Côte d'Azur Institute.

She is a nominated member of the French Council for Artificial Intelligence and Digital Technology and of the French Advisory Committee on Digital Ethics, since September 2025. From December 2019 to December 2024, she was a member of the National Pilot Committee for Digital Ethics (CNPEN), and from October 2020 to June 2025 she was a member of the Scientific and Ethical Committee of the Cancer Data Platform of the Institut National du Cancer.

Since January 2025, she is Vice-President of the Steering Committee of COMMA (Computational Models of Arguments). She is part of the Steering Committee of ECA (European Conference on Argumentation). She was a member of the IAAIL executive committee from 2021 to 2025.