Negotiating LLMs for Enhanced Hate Speech Classification and Interpretability

João Assaoka , Lilian Berton

[1] Universidade Federal de São Paulo (UNIFESP)
São José dos Campos, SP – Brazil

{joao.assaoka , lberton}@unifesp.br

Abstract. *Traditional hate speech classification frameworks rely on a single Large Language Model (LLM) operating in an isolated, single-turn decision-making process. However, this approach can suffer from limitations in handling nuanced linguistic phenomena such as sarcasm, ambiguity, and contextual shifts. To address these challenges, we introduce a multi-LLM negotiation framework, where two specialized models engage in iterative exchanges to refine the classification decision. The generator proposes a classification alongside a rationale, while the discriminator evaluates its credibility and requests adjustments until a consensus is reached. Experiments conducted on Twitter in Portuguese hate speech detection dataset, demonstrate that the negotiation-based approach using Sabizinho 3 and GPT 4.1 Nano with zero-shot, achieve competitive precision and recall. Furthermore, results indicate that this framework allow interpretability of the final classification.*

1. Introduction

Automated hate speech detection is one of the greatest challenges for content moderation systems [Kumar et al. 2025, Rawat et al. 2024]. The subjective and often implicit nature of hostile speech makes accurate classification difficult, requiring deep reasoning to differentiate between legitimate criticism, sarcasm, and offensive language. Seemingly harmless phrases can be deeply offensive when used with sarcasm or irony. For example: “What a joke, [minority group] is so intelligent!” without context and intonation, it is difficult for a model (and often for humans) to identify intent. Hate speech often manifests through slang, codes, acronyms, or euphemisms that evolve rapidly to evade detection. Hate speech can be subtle and implicit, depending on cultural, historical, or social references that are not explicitly stated in the text.

Traditionally, machine learning models handle this task using a single LLM [Albladi et al. 2025, Guo et al. 2023], which makes decisions in a single turn based on prompting and in-context learning. However, this approach has significant limitations, as a single model may fail to capture important linguistic nuances. Few papers have employed more than one LLM to overcome these problems, [Park et al. 2024] explored multiple agents, each representing the labeling criteria of a specific dataset. Each agent generates its positions and justifications on whether a text constitutes hate speech or not.

In this work, we propose an LLM negotiation framework, in which two models, a generator and a discriminator, engage in iterative interactions to refine the final decision. The generator proposes a classification along with a rationale, while the discriminator evaluates its credibility and requests adjustments until an optimized consensus is reached.

This approach allows models to leverage their complementary capabilities, refining their interpretations through explanations and counterarguments.

The main contributions of this work include:

- Proposes an interactive zero-shot classification approach where two LLMs (Sabizinho 3 and GPT 4.1 Nano) collaborate iteratively to refine hate speech classification.
- Provide an interpretable approach that allows tracking the LLM's reasoning process.
- Conducts experiments on hate speech detection benchmarks in Portuguese, as far as we know being the first work to perform multi-label in a benchmark dataset.
- The code and prompts used in this work are available on GitHub.[1]

The remainder of this paper is organized as follows. Section 2 presents the works related to our approach. Section 3 describes the methodology followed to set up the experiments, whereas Section 4 presents the performance analysis of our approach. Finally, Section 5 draws conclusions and points to future work.

2. Related work

Traditionally, hate speech detection has been approached with supervised machine learning models based on lexical and syntactic features, as well as with deep learning models such as recurrent neural networks (RNNs) and pre-trained transformers (e.g., BERT, RoBERTa). However, the rise of Large Language Models (LLMs), like GPT-3, GPT-4, Llama, and Gemini, has introduced new perspectives and challenges for hate speech classification.

Recent literature demonstrates a growing exploration of the potential of LLMs for this complex task. Approaches can be categorized into three main strands: zero-shot/few-shot learning, fine-tuning, and multi-agent negotiation frameworks.

2.1. Zero-shot and Few-shot Learning

One of the greatest advantages of LLMs is their ability to perform tasks with little to no exemplification. This is particularly appealing for hate speech classification, where annotating large volumes of data can be costly and ethically sensitive. For instance, [Oliveira et al. 2023] evaluated ChatGPT's zero-shot capability for detecting hate speech in Portuguese, achieving a competitive F1-score of 0.73 and demonstrating robustness in a cross-dataset evaluation. Other works, such as those by [Chiu et al. 2021] and [Ghorbanpour et al. 2025], further explore the capacity of LLMs to classify hate speech using only textual instructions (prompts), showing performance comparable to fine-tuned models. Effectiveness, however, depends on prompt engineering and the clarity of hate speech definitions provided.

2.2. Fine-tuning of LLMs

Although pre-trained LLMs possess vast linguistic knowledge, fine-tuning them on specific hate speech datasets can optimize their performance. This approach

[1] https://github.com/Assaoka/Negotiating_LLMs_for_Enhanced_Hate_Speech_Classification_and_Interpretability

typically involves transformer models like BERT. For the Portuguese language, [Silva and Freitas 2022] demonstrated the effectiveness of this approach by fine-tuning BERTimbau on the dataset from [Fortuna et al. 2019a], achieving a state-of-the-art F1-score of 0.86 for binary classification. Their result surpassed the baseline established by [Fortuna et al. 2019a] themselves, who used an LSTM model and obtained an F1-score of 0.78. A persistent challenge is the sensitivity to class imbalance, often addressed with oversampling techniques, and the interpretability of model decisions. The ethics of fine-tuning on potentially toxic data is also a concern [Vidgen and Derczynski 2020].

2.3. Multi-Agent and Negotiation Frameworks

While single-LLM approaches are prevalent, some research has begun to explore multi-agent systems to enhance reasoning and resolve ambiguity. For example, [Sun et al. 2023] introduced a negotiation framework for *sentiment analysis*, where a generator LLM proposes a decision and a discriminator LLM evaluates it. The models interact iteratively until a consensus is reached, leveraging their complementary abilities to correct imperfect responses and improve accuracy over single-model baselines. While their work focused on sentiment, our paper adapts a similar negotiation philosophy to the more complex and nuanced task of hate speech classification, aiming to improve both performance and interpretability.

3. Methodology

3.1. Resources

Our framework was developed in Python and uses Langchain [Chase 2022] to orchestrate calls to the LLM providers' APIs. We employed two LLMs, *Sabiá* [Almeida et al. 2024] is optimized for the nuances, regionalisms, slang, and cultural context of Brazil. Sabiá models are built on the Transformer architecture, undergoing extensive pre-training and instruction-tuning on a vast corpus of Brazilian Portuguese texts. The *GPT-4.1 Nano* [OpenAI 2025] is the lightest, fastest, and most cost-effective model in the GPT-4.1 family, launched by OpenAI in mid-2025. Designed to optimize speed and cost, this model positions itself as the ideal solution for tasks requiring real-time responses and high efficiency, although with a more limited reasoning capacity compared to its larger counterparts, GPT-4.1.

3.2. Dataset

The dataset used was proposed by [Fortuna et al. 2019a]. Authors collected tweets focused on hate speech detection on Twitter in Portuguese, employing a dual strategy of keyword and profile-based searches to gather messages. The final dataset comprises 5,668 tweets from 1,156 different users. Over 95% of these tweets were from January, February, and March 2017. The dataset underwent a two-phase annotation process:

- Binary annotation: Three annotators (Portuguese native speaker volunteers from Information Science) independently classified each message as either 'hate speech' or 'not hate speech' following provided guidelines. Each annotator received an equal number of messages. Fleiss's Kappa was used to check agreement among the three classifications, yielding a low value of K = 0.17. A majority vote was applied to finalize the annotations, resulting in 31.5% of the messages being labeled as 'hate speech' in the final dataset.

- Hierarchical hate speech classification: Recognizing that hate speech can be categorized into subtypes (e.g., racism, sexism) and exhibit intersectionality (e.g., targeting Black women), the researchers proposed a hierarchical annotation schema using a Rooted Directed Acyclic Graph (rooted DAG).

After the annotation phase, they ended up with a multilabeled dataset (Table 1) where 22% of the instances were identified as hate speech.

Table 1. Occurrences and Definitions of Hate Speech Categories

Class	Occurrences	Definition
Sexism	672	Hate speech based on gender, e.g. against women.
Homophobia	322	Hate speech based on sexual orientation.
Body	164	Hate speech based on body characteristics, e.g. fat, thin, etc.
Racism	94	Hate speech based on ethnicity.
Ideology	92	Hate speech based on ideas, e.g. feminist or left-wing ideologies.
Migrants	82	Hate speech directed at immigrants or refugees.
Religion	30	Hate speech based on religion.
Origin	26	Hate speech based on place of origin.
OtherLifestyle	20	Hate speech based on lifestyle choices, e.g. vegetarianism.
Health	6	Hate speech based on health conditions, e.g. PWD.
Ageing	4	Hate speech based on age, such as against elderly people.

3.3. Negotiation Framework

In our multi-stage hate speech classification pipeline, two distinct LLMs, a generator and a discriminator, engage in a three-turn "negotiation" over each input text, refining their judgments until a final consensus is reached. Unlike prior work that uses sentiment analysis, we employ a zero-shot approach tailored specifically for hate speech detection and multi-label categorization.

Generator: The generator receives only a brief task description in natural language, a definition of the possible classes (no demonstrations), and the raw input text. Its task is to produce a chain-of-thought reasoning ("Thought: ...") explaining whether the text is hate speech, followed by "Classes: ..." listing one or more of the predefined labels (or "None").

Discriminator: The discriminator is given the same input text plus the generator's "Thought" and "Classes." It must evaluate that classification, outputting "Analysis: ..." to point out any errors or omissions and "Classes: ..." to propose its own label assignment.

Generator (Final Verdict): The generator then re-examines its original "Thought" and "Classes," the discriminator's "Analysis" and "Classes," and the input text. It issues "Result: ..." with a reconciled reasoning that addresses points of agreement or disagreement, and "Classes: ..." as its definitive multi-label assignment.

3.4. Evaluation

To evaluate the performance of the models on the hate speech detection task, the following metrics were used to quantify the classification quality:

Accuracy (Acc): Accuracy is a measure of how often the classifier is correct, defined by the proportion of correctly predicted instances (both positive and negative) to the total number of instances. It is given by Equation 1:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{1}$$

where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives.

Precision (P): Precision measures how many of the instances predicted as positive are actually positive. High precision indicates a low false positive rate. It is defined by Equation 2:

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

Recall (R): Recall measures the ability of the model to correctly identify all positive instances. High recall indicates that the model is capable of detecting most of the positive samples. It is defined by Equation 3:

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

F1-Score (F1): The F1-score is the harmonic mean of precision and recall, providing a balance between these two metrics, which is particularly useful when the dataset has imbalanced classes. The F1-score ranges from 0 to 1, with higher values indicating better performance. It is defined by Equation 4:

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{4}$$

Cohen's Kappa (k): The Cohen's Kappa coefficient is a statistical measure that assesses the agreement between two raters (in this case, the models or a model and the ground truth) on classification tasks. It takes into account the possibility of the agreement occurring by chance. It is calculated by Equation 5:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{5}$$

where p_o is the observed agreement (proportion of correct predictions) and p_e is the expected agreement by chance. A k value close to 1 indicates perfect agreement, while 0 indicates agreement equivalent to chance.

These metrics will provide a comprehensive evaluation of the models, considering overall correctness (Accuracy), the balance between precision and recall (F1-Score), and the level of agreement between classifications (Kappa), which is fundamental for validating the robustness of the hate speech detection models.

4. Results

This chapter details the main experimental results from the three iterations of our classification framework based on negotiation between LLMs, addresses the computational cost analysis, and compares our performance with existing methods on the dataset from [Fortuna et al. 2019a].

4.1. First Iteraction

At this stage, each model, GPT-4 Nano (referred to as GPT) and Sabiazinho-3 (referred to as Sabia), individually generated an initial classification without any interaction between them.

As presented in Table 2, GPT achieved higher precision, indicating fewer false positives, while Sabia obtained superior recall, detecting more instances of hate speech. Despite similar accuracy, the F1-Score and Kappa coefficient reveal distinct trade-offs: GPT favors precision, whereas Sabia prioritizes coverage. This divergence reflects differences in adaptation to colloquial Portuguese versus the multilingual nature of GPT. Regarding the F1-Score, Sabia led in five categories, whereas GPT stood out in three rare classes.

Table 2. Evaluation metrics per class for GPT, Sabia, with best values in bold and macro-average in the final row. Presented metrics: Accuracy (Acc), Precision (P), Recall (R), F1-score (F1), Kappa (k).

	GPT					Sabia					k
	Acc	**P**	**R**	**F1**	**k**	**Acc**	**P**	**R**	**F1**	**k**	**k**
Sexism	0.89	**0.85**	0.13	0.23	0.16	**0.90**	0.61	**0.51**	**0.55**	**0.50**	0.21
Homophobia	**0.96**	**0.80**	0.41	0.54	0.41	**0.96**	0.69	**0.63**	**0.66**	**0.64**	0.45
Body	**0.97**	0.54	0.41	0.47	0.31	**0.98**	0.56	**0.90**	**0.69**	**0.67**	0.32
Racism	**0.98**	**0.42**	0.41	**0.42**	0.24	0.96	0.29	**0.82**	0.43	0.41	0.28
Ideology	**0.98**	**0.36**	0.15	**0.21**	**0.10**	0.96	0.14	**0.33**	0.19	0.17	0.10
Migrants	**0.99**	**0.48**	0.20	0.28	0.13	0.98	0.46	**0.72**	**0.56**	**0.55**	0.17
Religion	**0.99**	**0.31**	0.67	**0.43**	0.18	0.98	0.17	**0.93**	0.29	0.28	0.32
Origin	**0.99**	**0.21**	0.23	**0.22**	0.06	0.98	0.07	**0.35**	0.12	0.11	0.09
OtherLifestyle	**0.98**	0.02	0.10	0.04	0.01	0.97	**0.06**	**0.45**	**0.10**	**0.10**	0.09
Health	**1.00**	**0.22**	**0.33**	**0.27**	**0.03**	**1.00**	0.06	0.17	0.09	0.08	0.04
Ageing	**1.00**	0.00	0.00	0.00	-0.00	**1.00**	0.00	0.00	0.00	-0.00	-0.00
Average	**0.97**	**0.41**	0.30	0.32	0.18	**0.97**	0.37	**0.53**	**0.41**	**0.39**	0.22

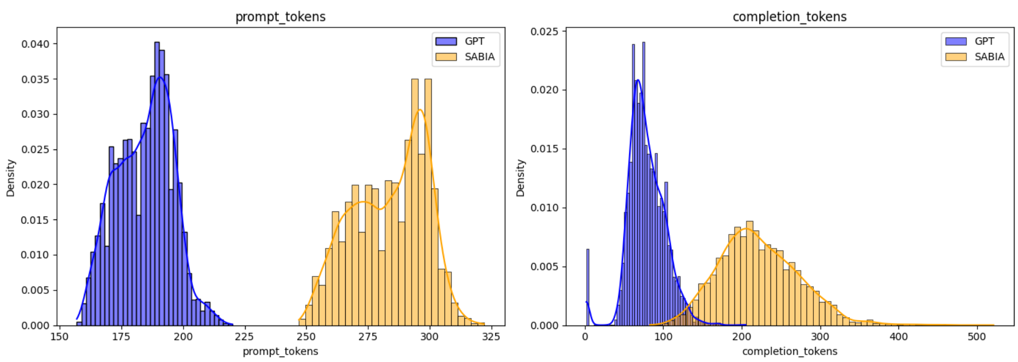

Figure 1. Cost histogram – 1st Iteration

Figure 1 presents the token cost for both models in the first iteration. Sabia consumed more tokens than GPT in this iteration (1.61M vs. 1.05M input tokens; 1.26M vs. 0.45M output tokens). Additionally, it is important to consider token pricing. Sabiazinho-3 costs R$ 0.70 per million input tokens and R$ 2.10 per million output tokens. GPT-

4.1-nano costs \$0.10 (R\$ 0.55) per million input tokens and \$0.40 (R\$ 2.21) per million output tokens.

4.2. Second Iteration

In the second iteration, each model reviewed the partner's analyses, acting as a discriminator.

The interaction rebalanced precision and recall: the GPT discriminator increased Sabia's precision, while the Sabia discriminator improved GPT's recall. This led to higher overall F1 scores and an improved Kappa coefficient.

The interaction gains persisted, as presented in Table 3, reinforcing the value of inter-model dialogue.

Table 3. Evaluation metrics per class for GPT-Sabia and Sabia-GPT, with best values in bold and macro-average in the final row. Presented metrics: Accuracy (Acc), Precision (P), Recall (R), F1-score (F1), Kappa (k).

	GPT-Sabia					**Sabia-GPT**					**k**
	Acc	**P**	**R**	**F1**	**k**	**Acc**	**P**	**R**	**F1**	**k**	**k**
Sexism	0.90	**0.66**	0.37	0.48	0.37	**0.90**	0.62	**0.48**	**0.54**	**0.48**	0.53
Homophobia	0.96	0.69	0.52	0.60	0.46	**0.97**	**0.74**	**0.62**	**0.68**	**0.65**	0.53
Body	0.97	0.54	0.73	0.62	0.45	**0.98**	**0.57**	**0.81**	**0.67**	**0.63**	0.50
Racism	**0.97**	0.33	0.59	0.42	0.28	**0.97**	**0.33**	**0.72**	**0.45**	**0.42**	0.45
Ideology	**0.97**	**0.19**	0.26	0.22	0.13	0.96	0.16	**0.33**	0.21	0.19	0.27
Migrants	0.99	0.49	0.40	0.44	0.23	**0.99**	**0.51**	**0.59**	**0.55**	**0.50**	0.31
Religion	**0.98**	**0.19**	0.70	0.30	0.15	**0.98**	0.18	**0.90**	**0.30**	**0.28**	0.47
Origin	**0.99**	**0.14**	0.31	**0.19**	**0.07**	0.98	0.09	**0.35**	0.14	0.12	0.19
OtherLifestyle	**0.98**	0.04	0.25	0.07	0.03	**0.98**	**0.06**	**0.30**	**0.09**	**0.08**	0.17
Health	**1.00**	**0.14**	0.33	**0.20**	**0.03**	**1.00**	0.13	**0.33**	0.18	0.11	0.07
Ageing	**1.00**	**0.20**	**0.25**	**0.22**	**0.01**	**1.00**	0.00	0.00	0.00	-0.00	0.02
Average	**0.97**	0.37	0.43	0.34	0.22	**0.97**	**0.39**	**0.49**	**0.36**	**0.34**	0.32

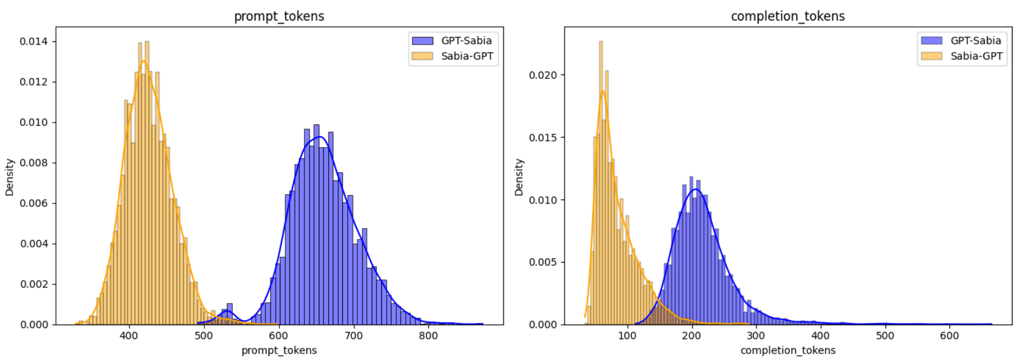

Figure 2. Cost histogram – 2nd Iteration

At this stage (as shown in Figure 2), Sabia used 3.74M input tokens and 1.23M output tokens; GPT used 2.42M and 0.48M, respectively. The increased input token costs refer to the input of tokens generated in the first iteration.

4.3. Third Iteration

Finally, the generator re-evaluates the discordant responses to refine its classification. In this study, this will be our last iteration, but the process could continue until consensus is reached or a higher iteration limit is met.

As expected, in the third iteration the models somewhat revert to their original responses (after all, they can choose to accept changes or maintain their position). GPT regained part of its precision, and Sabia recovered part of its recall, but the changes were minimal since only the disagreements were reanalyzed, as presented in Table 4.

Table 4. Evaluation metrics per class for GPT-Sabia-GPT and Sabia-GPT-Sabia, with best values in bold and macro-average in the final row. Presented metrics: Accuracy (Acc), Precision (P), Recall (R), F1-score (F1), Kappa (k).

	GPT-Sabia-GPT					**Sabia-GPT-Sabia**					**k**
	Acc	**P**	**R**	**F1**	**k**	**Acc**	**P**	**R**	**F1**	**k**	**k**
Sexism	**0.90**	**0.66**	0.36	0.46	0.36	**0.90**	0.62	**0.49**	**0.55**	**0.49**	0.53
Homophobia	0.96	0.72	0.53	0.61	0.48	**0.97**	**0.73**	**0.62**	**0.67**	**0.65**	0.55
Body	0.97	0.52	0.68	0.59	0.43	**0.98**	**0.56**	**0.85**	**0.68**	**0.66**	0.49
Racism	**0.98**	**0.37**	0.60	**0.45**	**0.29**	0.97	0.33	**0.72**	**0.45**	**0.44**	0.44
Ideology	**0.97**	**0.19**	0.23	**0.21**	0.12	0.96	0.15	**0.33**	**0.21**	**0.19**	0.21
Migrants	0.99	0.52	0.40	0.45	0.24	**0.99**	**0.54**	**0.65**	**0.59**	**0.57**	0.34
Religion	**0.98**	**0.20**	0.77	**0.32**	**0.17**	**0.98**	0.17	**0.90**	0.29	0.28	0.50
Origin	**0.99**	**0.13**	0.31	**0.19**	**0.07**	0.98	0.09	**0.38**	0.15	0.14	0.22
OtherLifestyle	**0.98**	0.04	0.25	0.07	0.03	**0.98**	**0.06**	**0.35**	**0.10**	**0.09**	0.17
Health	**1.00**	0.13	0.33	0.19	0.03	**1.00**	**0.16**	**0.50**	**0.24**	**0.22**	0.06
Ageing	**1.00**	**0.17**	**0.25**	**0.20**	**0.01**	**1.00**	0.00	0.00	0.00	-0.00	0.01
Average	**0.97**	0.33	0.44	0.33	0.22	**0.97**	**0.36**	**0.55**	**0.36**	**0.35**	0.34

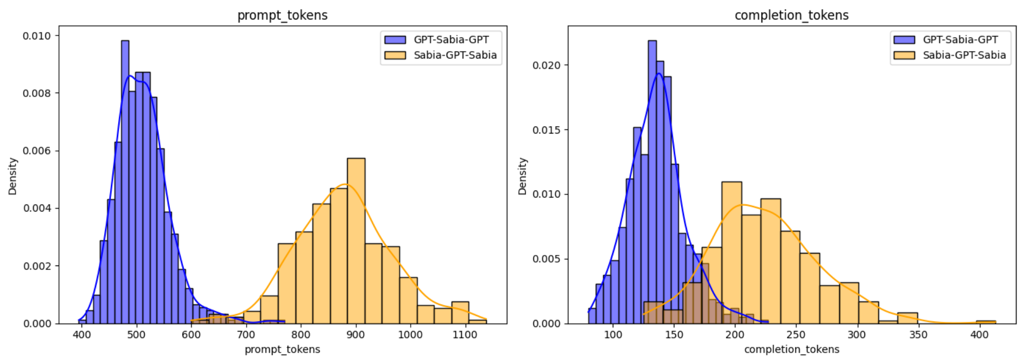

Figure 3. Histograma de Custo – 3ª Iteração

Regarding costs, Figure 3 shows that token usage in this iteration was reduced. Sabia used 260,896 input tokens and 66,609 output tokens, while GPT used 358,137 input tokens and 70,626 output tokens.

4.4. Disagreement Over the Course of the Iterations

We can observe in Table 5 that most agreements occur by the second iteration. Therefore, continuing beyond the third iteration is unlikely to bring significant improvements

Table 5. Disagreements per hate-speech category across 2ª and 3ª iterations. Bold indicates the higher disagreement per row.

	GPT-Sabia-GPT		Sabia-GPT-Sabia	
	2ª It.	**3ª It.**	**2ª It.**	**3ª It.**
Sexism	**283**	34	70	32
Homophobia	**91**	15	61	**30**
Body	**114**	23	64	**42**
Racism	89	**24**	**98**	29
Ideology	**90**	**27**	53	28
Migrants	36	5	**57**	**29**
Religion	**60**	12	42	**32**
Origin	36	10	**54**	**27**
OtherLifestyle	87	31	**87**	**39**
Health	**9**	1	26	**25**
Ageing	3	1	**25**	**24**
Total	700	145	297	84

in performance metrics. On the other hand, with each iteration after the second, costs are substantially reduced due to the smaller number of samples, which may offer a favorable cost-benefit trade-off.

Another key observation is that the initiating model plays a significant role, starting with the ”stronger” model consistently yielded better outcomes in both scenarios.

4.5. Literature Comparison

Table 6 presents the classification results from three literature paper. All of them used only one LLM for binary classification. As far as we know, we are the only paper that performed multi-label classification. Authors from [Fortuna et al. 2019b] and [Silva and Freitas 2022] employed traditional classification approaches using both labeled and unlabeled data. In addition, [Silva and Freitas 2022] balanced the dataset, although the criteria for data removal were not clearly specified. Notably, only [Oliveira et al. 2023] implemented zero-shot classification, where no labeled examples are provided to the language model. As such, their work is the most closely aligned with our approach.

The multi-label problem involves 11 categories, presenting a significant challenge. As shown in the previous results, some classes are heavily imbalanced, for instance, ’sexism’ contains 672 examples, while ’health’ has only 6. Moreover, some examples were highly ambiguous, resulting in low agreement among the annotators. This way, the model struggled to correctly classify the underrepresented classes, which negatively impacted the overall F1 score.

4.6. Interpretability

This study demonstrates that negotiation strategies between LLMs can iteratively improve hate speech classification, balancing precision and recall as the models interact. The generator–discriminator interaction brings clear benefits in terms of self-correction and consistency, even in the presence of class imbalance.

Table 6. Comparison of F_1 scores for binary and multi-label classification.

Study	Approach	Task	F1 (bin)	F1-macro	F1-micro
[Fortuna et al. 2019b]	LSTM + embeddings	Binary	0.78	—	—
[Silva and Freitas 2022]	BERTimbau	Binary	**0.86**	—	—
[Oliveira et al. 2023]	ChatGPT 3.5 (zero-shot)	Binary	0.73	—	—
This work	GPT	multi-label	-	0.28	0.34
This work	Sabia	multi-label	-	0.33	0.49
This work	GPT-Sabia	multi-label	-	0.34	0.45
This work	Sabia-GPT	multi-label	-	0.35	**0.50**
This work	GPT-Sabia-GPT	multi-label	-	0.34	0.45
This work	Sabia-GPT-Sabia	multi-label	-	**0.36**	**0.50**

Moreover, this technique has an advantage: it is interpretable. Consider the example below:

Input: "@realmen don't have girlfriends, they have temporary females."

GPT: "Thought: The comment makes a derogatory and pejorative statement about a man, using terms like 'temporary females' to refer to women, which shows a disrespectful and dehumanizing view. However, there is no mention of specific groups based on race, religion, sexual orientation, origin, or other aspects protected under hate speech categories. It is an expression of disrespect or disdain, but it does not constitute hate speech according to the established categories. Classes: None"

Sabia: "Analysis: The previous model's reasoning seems to have failed in not identifying the hate speech present in the comment. The phrase 'don't have girlfriends, they have temporary females' is clearly sexist, as it devalues women and reinforces gender stereotypes, treating them as temporary objects rather than human beings with feelings and dignity. The correct classification should include the category of sexism, as the comment promotes gender inequality and the objectification of women. Classes: Sexism"

GPT: "Result: I agree with the other model in identifying that the comment has a sexist connotation, as it devalues women and reinforces gender stereotypes by calling them 'temporary females'. However, my analysis also highlights that the comment is pejorative and dehumanizing, which reinforces the classification of sexism. Therefore, my final evaluation is that the correct category is indeed 'Sexism', not 'None'. I agree that the statement should be classified as sexist, as it promotes the objectification and disrespect of women. Classes: Sexism"

5. Conclusion

This study highlights the potential of leveraging in-context learning (ICL) and negotiation-based prompting strategies for hate speech classification using large language models (LLMs). The integration of dual LLMs in a generator–discriminator framework introduces a compelling avenue for refining decision-making through self-reflection and critique.

Future work could examine the impact of incorporating more than two LLMs, engaging in multi-agent collaborative or adversarial settings to simulate richer negotiation dynamics. Additionally, investigating a broader variety of LLMs, including those fine-tuned for specific languages or social contexts, may enhance performance in multilingual and culturally nuanced scenarios. Adaptive prompt engineering, confidence calibration, and real-time moderation applications also represent promising directions to expand the impact of this approach.

6. Ackowledgement

We thanks Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP) grant 2024/17511-2 and Serasa Experian.

References

Albladi, A., Islam, M., Das, A., Bigonah, M., Zhang, Z., Jamshidi, F., Rahgouy, M., Raychawdhary, N., Marghitu, D., and Seals, C. (2025). Hate speech detection using large language models: A comprehensive review. *IEEE Access*.

Almeida, T. S., Abonizio, H., Nogueira, R., and Pires, R. (2024). Sabi\'a-2: A new generation of portuguese large language models. *arXiv preprint arXiv:2403.09887*.

Chase, H. (2022). Langchain. `https://www.langchain.com/`. Accessed: 2025-06-18.

Chiu, K.-L., Collins, A., and Alexander, R. (2021). Detecting hate speech with gpt-3. *arXiv preprint arXiv:2103.12407*.

Fortuna, P., da Silva, J. R., Wanner, L., Nunes, S., et al. (2019a). A hierarchically-labeled portuguese hate speech dataset. In *Proceedings of the third workshop on abusive language online*, pages 94–104.

Fortuna, P., Rocha da Silva, J., Soler-Company, J., Wanner, L., and Nunes, S. (2019b). A hierarchically-labeled Portuguese hate speech dataset. In Roberts, S. T., Tetreault, J., Prabhakaran, V., and Waseem, Z., editors, *Proceedings of the Third Workshop on Abusive Language Online*, pages 94–104, Florence, Italy. Association for Computational Linguistics.

Ghorbanpour, F., Dementieva, D., and Fraser, A. (2025). Can prompting llms unlock hate speech detection across languages? a zero-shot and few-shot study. *arXiv preprint arXiv:2505.06149*.

Guo, K., Hu, A., Mu, J., Shi, Z., Zhao, Z., Vishwamitra, N., and Hu, H. (2023). An investigation of large language models for real-world hate speech detection. In *2023 International Conference on Machine Learning and Applications (ICMLA)*, pages 1568–1573. IEEE.

Kumar, M. et al. (2025). Exploring hate speech detection: challenges, resources, current research and future directions. *Multimedia Tools and Applications*, pages 1–37.

Oliveira, A., Cecote, T., Silva, P., Gertrudes, J., Freitas, V., and Luz, E. (2023). How good is chatgpt for detecting hate speech in portuguese? In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 94–103, Porto Alegre, RS, Brasil. SBC.

OpenAI (2025). Models Overview.

Park, S., Kim, J., Jin, S., Park, S., and Han, K. (2024). Predict: Multi-agent-based debate simulation for generalized hate speech detection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20963–20987.

Rawat, A., Kumar, S., and Samant, S. S. (2024). Hate speech detection in social media: Techniques, recent trends, and future challenges. *Wiley Interdisciplinary Reviews: Computational Statistics*, 16(2):e1648.

Silva, F. and Freitas, L. (2022). Brazilian portuguese hate speech classification using bertimbau. *The International FLAIRS Conference Proceedings*, 35.

Sun, X., Li, X., Zhang, S., Wang, S., Wu, F., Li, J., Zhang, T., and Wang, G. (2023). Sentiment analysis through llm negotiations.

Vidgen, B. and Derczynski, L. (2020). Directions in abusive language training data, a systematic review: Garbage in, garbage out. *PLOS ONE*, 15(12):e0243300.