ISCTE IUL
REPOSITÓRIO
INSTITUTO UNIVERSITÁRIO DE LISBOA

Repositório ISCTE-IUL

Deposited in *Repositório ISCTE-IUL*:
2026-11-17

Deposited version:
Accepted Version

Peer-review status of attached file:
Peer-reviewed

Citation for published item:
Ramos, G., Batista, F., Ribeiro, R., Fialho, P., Moro, S., Fonseca, A....Silva, C. (2025). Bypassing the nuances of Portuguese covert hate speech through contextual analysis. In Manuel Filipe Santos, José Machado, Paulo Novais, Paulo Cortez, Pedro Miguel Moreira (Ed.), Progress in artificial intelligence: EPIA 2024 (pp. 241?253). Springer. https://doi.org/10.1007/978-3-031-73503-5_20

Further information on publisher's website:
10.1007/978-3-031-73503-5_20

Publisher's copyright statement:
This is the peer reviewed version of the following article: Ramos, G., Batista, F., Ribeiro, R., Fialho, P., Moro, S., Fonseca, A....Silva, C. (2025). Bypassing the nuances of Portuguese covert hate speech through contextual analysis. In Manuel Filipe Santos, José Machado, Paulo Novais, Paulo Cortez, Pedro Miguel Moreira (Ed.), Progress in artificial intelligence: EPIA 2024 (pp. 241?253). Springer. https://doi.org/10.1007/978-3-031-73503-5_20, which has been published in final form at https://dx.doi.org/10.1007/978-3-031-73503-5_20. This article may be used for non-commercial purposes in accordance with the Publisher's Terms and Conditions for self-archiving.

Use policy

Creative Commons CC BY 4.0
The full-text may be used and/or reproduced, and given to third parties in any format or medium, without prior permission or charge, for personal research or study, educational, or not-for-profit purposes provided that:

- a full bibliographic reference is made to the original source
- a link is made to the metadata record in the Repository
- the full-text is not changed in any way

The full-text must not be sold in any format or medium without the formal permission of the copyright holders.

Serviços de Informação e Documentação, Instituto Universitário de Lisboa (ISCTE-IUL)
Av. das Forças Armadas, Edifício II, 1649-026 Lisboa Portugal
Phone: +(351) 217 903 024 | e-mail: administrador.repositorio@iscte-iul.pt
https://repositorio.iscte-iul.pt

Bypassing the Nuances of Portuguese Covert Hate Speech through Contextual Analysis

Gil Ramos[1], Fernando Batista[2,3], Ricardo Ribeiro[2,3], Pedro Fialho[1], Sérgio Moro[1,4], António Fonseca[1], Rita Guerra[5], Paula Carvalho[3], Catarina Marques[6], and Cláudia Silva[7]

[1] Instituto Universitário de Lisboa (ISCTE-IUL), ISTAR, Lisbon, Portugal
[2] Instituto Universitário de Lisboa (ISCTE-IUL), Lisbon, Portugal
[3] INESC-ID, Lisbon, Portugal
[4] University of Jordan, Amman, Jordan
[5] ISCTE-Instituto Universitário de Lisboa and Center for Psychological Research and Social Intervention (CIS-ISCTE), Lisbon, Portugal
[6] ISCTE-Instituto Universitário de Lisboa and Business Research Unit (BRU-ISCTE), Lisbon, Portugal
[7] ITI-LARSyS and IST, Lisbon, Portugal

Abstract. Detecting and addressing covert hate speech poses significant challenges in online platforms where discriminatory messages are often disguised within seemingly innocuous content. In this study, we investigate the effectiveness of contextual analysis in bypassing the nuances of covert hate speech. Our research explores the impact of prompt engineering and context addition on the classification of overt and covert hate speech across diverse target groups, including Roma, migrants, LGBTQ+, and individuals of African descent. Through experimental trials using generative models like GPT-3.5 and GPT-4, our findings reveal that the addition of context, not only improves the overall performance of the models in general hate speech (from 75.0% to 79.6% F1 score, for the positive class), but also significantly improves the classification of covert hate speech, increasing True Positives by 21.64% (absolute) compared to the 6.5% in overt hate speech. Despite these improvements, the addition of context also increased the number of False Positives, indicating that a further refinement is needed for this contextual analysis. Moreover, target group analysis demonstrates a correlation between the prevalence of covert hate speech and model performance.

Keywords: Hate speech · GPT · natural language processing

1 Introduction

In the modern era, the widespread adoption of social media has transformed digital communication, reshaping how individuals express themselves and connect with others [22]. With smartphones and internet access becoming universal, social media platforms are readily available, allowing people worldwide to share their thoughts and viewpoints effortlessly. This democratization of expression,

while promoting empowerment and dialogue, has also highlighted a pressing issue: the alarming increase in Hate Speech (HS) and related offenses [25].

There are no universally accepted and precise definitions of HS [19], however, for this work, Online HS refers to bias-motivated, derogatory language that spreads, incites, promotes, or justifies hatred, exclusion, and/or violence/aggression against a person/group because of their group membership, and can be operationalized through the following coexisting conditions [6]:

- HS is an intergroup phenomenon that targets social groups or individuals because of their perceived membership in certain social groups. Therefore, in this work, HS does not encompass hateful expressions that occur at the individual/interindividual level of analysis.
- The target of HS is attacked primarily because of perceived membership in a given social group, and not necessarily because of a specific behavior or action.

HS can be both overt (explicit or direct) and covert (implicit or indirect). In overt HS (OHS) the message often conveys overtly biased, inflammatory speech, insults, abusive and derogatory language [2]:

- Racismo o c@ralho! se não fossem esses parasitas da sociedade que não querem fazer nada, Portugal era um paraíso.. [*Racism my ass! If it weren't for these parasites of society who don't want to do anything, Portugal would be a paradise..*].

Whereas in covert HS (CHS) the message does not usually contain derogatory or insulting terms, and its meaning is implicit and needs to be inferred [2]:

- Coitadinhos dos "feirantes", vão ficar sem os benefícios. [*Poor 'market vendors', they're going to lose their benefits.*].

Such formulations may use a diversity of rhetorical tactics, including humour, irony, analogies and metaphors [1,3], euphemisms [13], and rhetorical questions [14,12].

HS often precipitates psychological and emotional distress in recipients, potentially culminating in stress, anxiety, and even depression [23]. Moreover, prolonged exposure to HS can corrode societal bonds, fostering an environment of mistrust and polarization. This fragmentation exacerbates the cycle of animosity, as individuals retreat into their respective echo chambers, reinforcing preexisting biases and prejudices [16]. This problem gains even more significance when we take into account the fact that most hateful comments in social media are covert or implicit, and their interpretation requires information on the sociopolitical and pragmatic context [2], making it even harder to detect and combat.

Many organizations have tackled the issue of HS on social media by implementing guidelines and policies. However, the sheer volume of data generated by these platforms makes manual classification impractical. Therefore, there is a growing necessity to employ Machine Learning (ML) techniques to automate the classification process, yielding more efficient and reliable outcomes [20]. This

technological shift has spurred extensive research and development focused on utilizing ML for HS detection, encompassing various techniques ranging from traditional ML to Ensemble Models and Deep Learning (DL) models, all showing promising results, with the recent emergence of Transformer-based models like BERT [24] having significantly expanded the landscape of HS detection methods.

In this work, to tackle the problem of detecting CHS effectively we used a YouTube corpus created in the scope of the project *kNOwHATE: kNOwing online HATE speech*[1] and applied transfer learning through two generative models, GPT-3.5-turbo and GPT-4-turbo, with context, to compare the results of CHS and OHS and the effect that this context has in them.

2 Related Work

Generative large language models, such as GPT, are being used in several text classification tasks, and hate speech is no exception: Chiu *et al.* [9] employed GPT-3 for the detection of sexist and racist comments on YouTube and Reddit through zero-, one-, and few-shot classification where example sexist/racist comments were given as context within the prompt. The few-shot method proved to lead to better performance, indicating that the inclusion of more information in the prompt can be beneficial. Also exploring ChatGPT as a zero-shot classifier, Oliveira *et al.* [17] concluded that it is a feasible alternative for HS detection in Brazilian Portuguese, in part due to its ability to perform well on cross-domain testing.

Besides generative models, several researchers have implemented solutions based on contextual analysis, where some information was fed to the model in addition to the target message, with success. Pérez *et al.* [18] experimented with classifying a tweet as HS by adding the news article title that the tweet was replying to, and obtained better performance with this addition of context. Casavantes *et al.* [7] used the information of the metadata of the tweet and the author to feed the model and managed to obtain good results.

For Portuguese, the literature on HS detection is relatively limited, with a small number of studies focusing specifically on this language. The studies have primarily focused on Brazilian Portuguese, with few addressing European Portuguese.

For European Portuguese, one study focused primarily on constructing a hierarchically labeled dataset for HS detection, and also developed an initial baseline classification for the dataset, where they used pre-trained word embeddings and LSTM and achieved a 71% micro F1 Score [10]. Other studies focused solely on the task of detecting HS, with one study using a combination of Generative Adversarial Networks and a BERT-based model to obtain a 66.4% positive class F1 Score [21]. Also using a BERT-based approach, Matos *et al.* [15] explored a BERT-CNN architecture for this classification task and managed to obtain

[1] https://knowhate.eu/pt-pt/

72.1% F1 Score on the positive class, using the best combination of annotators. Both of these studies used newly developed annotated datasets for European Portuguese with HS instances from YouTube and Twitter [5,4].

For CHS the literature is still limited, with a single study found that made a preliminary attempt to classify it, achieving a 48.87% F1 Score [15]. Besides this study, no other studies were found that tried to detect CHS, and no studies explored the use of previous comments or messages from the same thread or conversation to provide contextual clues to the model as a means to improve the performance of the detection of CHS.

After exploring the literature on the use of generative models and context for HS detection, particularly in the context of European Portuguese, a gap was identified, with the research done in this area still in the preliminary stages. Therefore, there is a need for further research that aims to fill this gaps. By developing tools for a language that, as it does not yet have many resources, may be more vulnerable to the risk of discrimination and abuse, we have the potential to make a significant difference in mitigating the harmful effects of HS on the Portuguese-speaking community.

3 Data

For this study, we utilized a dataset of annotated online HS messages comprised of 23912 comments extracted from 88 unique YouTube videos centered around topics and occurrences that directly or indirectly target four specific target groups: individuals of African descent, Roma, migrants, and members of the LGBTQ+ communities. Initially, videos containing instances of HS were identified. To expand this dataset, we included additional videos recommended in the related videos section. This strategy was based on the hypothesis that videos frequently suggested when viewing a previously flagged HS video were more likely to harbor HS content. To quantify the frequency of video recommendations and identify those most likely to contain HS, we compiled a ranked list of suggested videos, and videos appearing more than 85 times in this list were incorporated into our dataset. Subsequently, videos were excluded if they did not meet a minimum threshold of 1000 views and 100 comments, resulting in a final selection of 88 videos distributed across the target groups as follows: Roma - 16, migrants - 19, LGBTQ+ - 24, and African descent - 29.

The corpus underwent a rigorous manual annotation process by a team of trained annotators comprising social psychologists and linguists. These annotators were also actively involved in the creation of the annotation framework. Each annotator was tasked with annotating a subset of 825 comments that were assigned to all annotators to assess inter-annotator agreement (IAA) and annotation reliability using Krippendorff's alpha [11], with the IAA being moderate at 0.546. This IAA subset also served as the test set that was used for our experiments in this work. The described training set of 23912 comments was not used in this work seeing that no training was done with the models employed. Given the task subjectivity, only the messages that were labeled as conveying

HS by at least two annotators were considered hatred content in the test set. The distribution of the test set used for our experiments is presented in Table 1. Please note that a message can be associated with zero, one, or more than one target group, and can be simultaneously OHS and CHS, that being the reason for the totals to not match with the sums of the several categories.

Table 1. YouTube corpus' test set distribution.

Group	Total messages	HS messages	OHS messages	CHS messages
Roma	127	123	74	70
LGBTQ+	153	133	50	109
Migrants	231	209	140	113
African descent	190	158	63	120
Total	825	596	292	402

4 Experimental Setup

In this section, we present an overview of the different models used for HS detection, along with the experimental settings used to run these models and the metrics used to evaluate the performance of each model.

We employed two Transformer-based models for text generation: GPT-3.5-turbo and GPT-4-turbo. The following prompts were used to obtain the labels for runs with and without context:

- No context: "*Below is the description of a text classification task, and a YouTube comment. Your task is to classify the comment into one of the following categories: hate speech, non hate speech. Return only the number 1 if you classify the comment as hate speech and 0 if not, do not return any additional text besides 1 or 0.*"
- With context: "*Below is the description of a text classification task, and a YouTube comment. Your task is to classify the comment into one of the following categories: hate speech, non hate speech. Return only the number 1 if you classify the comment as hate speech and 0 if not, do not return any additional text besides 1 or 0. You'll be provided with the target comment to classify, delimited by '<target>comment</target>', and the preceding comments in the conversation, marked as '<context>preceding comments</context>', which will serve as context for your classification. If the context is missing, classify the target comment based solely on its own, without considering context.*"

All prompts were composed in English, although the messages fed for classification were written in Portuguese, since in our prompt engineering experiments, detailed in Section 5.1 proved that this configuration led to better results. For

the context runs, all previous comments, from the same conversation, were provided in the prompt, to serve as context for the models. All runs were conducted between April 7th and June 28th, 2024. All experiments were conducted using Google Colab, a cloud-based platform providing access to an Nvidia Tesla K80 GPU with 12 GB of memory. We employed the OpenAI API to use the GPT models for our experiments – the temperature parameter was set to 0 to assure consistent results.

The performance of the models was evaluated using three standard metrics, namely Precision, Recall, and F1 score. When we assess the models we give more importance to the positive class F1 score since it evaluates the class we want to detect and is a more balanced measure, when taking into account both False Positives and False Negatives. For the comparisons between both types of HS, we compared them based on the number of True Positives and False Positives that each model achieved, since we employed a binary classification where the models were prompted to classify text as either HS or non-HS. This approach does not allow us to determine whether the model distinguishes between OHS or CHS. As a result, it is only possible to obtain the above-mentioned metrics for the general HS, and not for these finer types of HS.

5 Results

This section is divided into the various experiments we made. First, we present the results of the Prompt Engineering attempts to obtain the prompt that had the best performance for our task, then we provide the results of the addition of context to both CHS and OHS and how this affects both types and finally, we present the results for the different target groups.

5.1 Prompt Engineering

To identify the most effective prompt for our experiments, we conducted several trials using different prompt formulations and evaluated their performance based on the Positive F1 Score metric. For resource optimization, these experiments were primarily conducted on GPT-3.5-turbo. However, we also tested a smaller sample of data on GPT-4-turbo to confirm the results. The prompts tested and their corresponding scores are summarized in Table 2.

The results from our prompt engineering experiments indicate that the "Task" prompt, which includes only the description of the task, achieved the highest Positive F1 Score of 70.6%. This prompt formulation demonstrated better performance when compared to other variations that included additional components such as HS definitions or example messages of HS. It is worth noting that we also experimented with different prompt formulations and languages (Portuguese and English), to select the prompt, ultimately arriving at the best version, as presented in Section 4.

Based on these findings, we selected the "Task" prompt as the optimal formulation to be used for subsequent experiments, leveraging its effectiveness and simplicity in eliciting high-quality responses from our generative models.

Table 2. Prompt engineering attempts

Prompt	Description	Positive F1 Score
Task	Prompt with only the description of the task	70.6%
Definition	Prompt with the description of the task + our definition of HS	69.2%
Example	Prompt with the description of the task + an example message of HS	57.4%
Definition + Example	Prompt with the description of the task + our definition of HS + an example message of HS	59.5%

5.2 Context in Overt vs Covert Hate Speech

Following the determination of the optimal prompt design, we conducted experiments using both the GPT-3.5-turbo and GPT-4-turbo models to assess the impact of context on the classification of OHS and CHS. Looking first at general HS, the GPT-3.5-turbo model achieved Positive F1 Scores of 75.0% and 79.6% without and with context, respectively. In comparison, the GPT-4-turbo model exhibited lower performance, with Positive F1 Scores of 72.8% and 74.2%. We can see that both models showed improvement with the addition of context, increasing 4.6% and 1.4%, respectively, in general HS.

Turning our attention specifically to OHS and CHS, Table 3 presents the results for both types of HS classifications with and without context. Overall, the addition of context led to improvements in True Positives, particularly for CHS. This increase was more noticeable in GPT-3.5-turbo, which increased in 21.64% the number of True Positives in CHS compared to 6.5% for OHS. GPT-4-turbo, although less noticeably, also increased the number of True Positives for CHS by 3.23% compared to 0% in OHS. Despite these improvements, it is noteworthy that the inclusion of context also resulted in a substantial increase in False Negatives, with an additional 84 and 11 messages misclassified by GPT-3.5-turbo and GPT-4-turbo, respectively.

5.3 Target Group

In examining the performance across different target groups, Table 4 presents the Positive F1 Score obtained in the context run, as well as HS distribution and IAA for each group. Notably, both Roma and Migrants groups exhibit the best results, each achieving a Positive F1 Score exceeding 90%, in contrast to 73.7% for LGBTQ+ and 84.5% for African descent. Furthermore, our analysis reveals a correlation between performance and IAA with the proportion of CHS messages within each group. Groups with lower performance and IAA, such as LGBTQ+ and African descent, exhibit a higher percentage increase in CHS messages (more than 20%) compared to groups with higher performance.

Table 3. Results of OHS and CHS detection without and with addition of context to the prompt.

Metric	Type of HS	No context	Context	Gain
GPT-3.5-turbo				
True Positives	OHS	262	281	+19 (6.5%)
	CHS	212	299	**+88 (21.64%)**
False Positives	Both	57	141	+84 (10.18%)
GPT-4-turbo				
True Positives	OHS	255	255	+0 (0.0%)
	CHS	198	211	**+13 (3.23%)**
False Positives	Both	52	63	+11 (1.33%)

Table 4. Results of HS detection by Target Group, with HS distributions and IAA.

Target Group	GPT-3.5 F1	GPT-4 F1	%HS (%CHS)	IAA
Roma	91.9%	85.6%	98.42% (48.6%)	0.691
LGBTQ+	73.7%	57.0%	86.93% (68.6%)	0.494
Migrants	90.8%	83.6%	90.48% (44.7%)	0.627
African descent	84.5%	80.7%	83.16% (65.6%)	0.464

6 Discussion

The observed impact of prompt design on model performance raises questions about the role of information inclusion and rule specificity. Our findings suggest that the addition of explicit definitions or specific examples of HS within the prompt may inadvertently constrain the model's ability to generalize across diverse instances of HS. This phenomenon could result from the model becoming overly fixated on the provided definitions or examples, potentially missing out on nuanced or less typical instances of HS. Surprisingly, our results contrast with existing literature, which often indicates that the provision of examples and information enhances model performance [9]. This disagreement confirms the complexity of prompt engineering and the need for investigation into optimal prompt strategies for each case study.

Regarding the impact of the addition of context to the prompts on OHS and CHS classification, CHS, which relies heavily on contextual cues and subtleties, appears to benefit significantly more from the addition of context to the prompt, as is visible in Fig. 1. This observation aligns with expectations, as context plays a pivotal role in deciphering the hidden or implicit nature of CHS messages.

Observing Table 5 it is possible to find examples where the model without context misclassified a CHS instance as non-HS, and then the same model with the addition of context managed to classify the comment as HS, confirming that the addition of context is necessary in cases where the target comment is

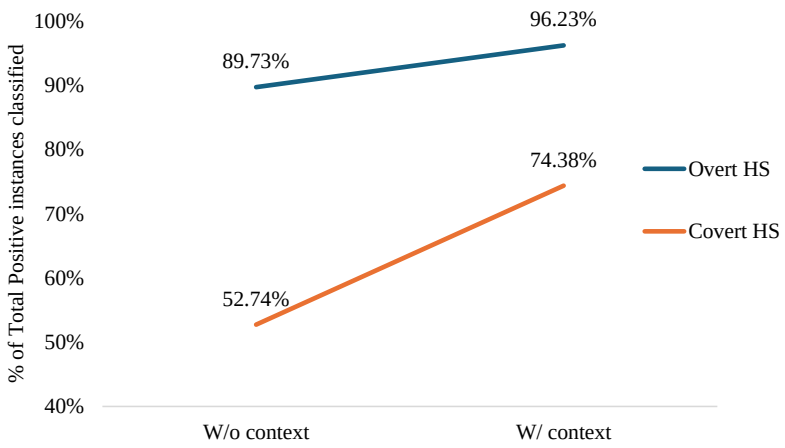


Fig. 1. Impact of context in the prediction of Overt and Covert HS in GPT-3.5-turbo

being implicitly hateful. In these examples, we can see that the comment on its own does not contain any of the characteristics of a HS comment, but when we integrate it in a thread of messages its intent becomes apparent.

However, the notable increase in False Positives with context suggests a potential drawback. The inclusion of contextual information may inadvertently lead the model to misclassify non-HS messages as HS, emphasizing the need for careful consideration when integrating contextual cues into HS detection systems. It is worth mentioning that the False Positives misclassified with the addition of context were not cases of counter speech, which can pose potential issues.

Finally, our analysis reveals a correlation between the percentage of CHS and IAA within target groups and model performance. Target groups characterized by a higher proportion of CHS, such as LGBTQ+ and African descent, exhibit lower model performance and IAA compared to groups with a lower prevalence of CHS. This observation further highlights the challenges posed by CHS and the need for tailored approaches to address the specific dynamics of different target groups.

It is important to note that the results obtained with GPT models exhibited substantial variability since subsequent runs with the same configuration led to differences in performance, where variations of approximately 25% were observed. This suggests that the outcomes from these models are not static and fluctuate across different runs and across time, as confirmed by the literature [8].

Table 5. Effect of prompt context in predicted label.

Context	Prompt	Label
no	**Target:** Isso sim uma limpeza sem dor nem remorsos [*Yes a cleansing without pain or remorse*]	0
yes	**Context:** Está na hora de deixarmos de ser um povo de brandos costumes e limparmos a nossa casa. [*It's time for us to stop being a people of soft customs and clean our house.*] **Target:** Isso sim uma limpeza sem dor nem remorsos	1
no	**Target:** existem brasileiros do bem?????????	0
yes	**Context:** Esses bastardos só causam desgosto e vergonha aos brasileiros de bem. **Target:** existem brasileiros do bem?????????	1
no	**Target:** Tal e qual.	0
yes	**Context:** Infelizmente quem imigrava respeitava o país que o recebia ficava grato por trabalho e estudo. Hoje dia eles levam o inferno com eles. **Target:** Tal e qual.	1

7 Conclusion

In this study, we investigated the impact of context on the classification of OHS and CHS, as well as the differences between diverse target groups. First, our exploration of prompt engineering revealed that while the inclusion of definitions and examples of HS may initially seem beneficial, our results suggest that overly prescriptive prompts could constrain model generalization, leading to potential performance limitations. This observation highlights the need for adaptive prompt strategies that balance specificity with flexibility. Second, the addition of context had a discernible impact on the classification of OHS and CHS. Notably, CHS, which relies heavily on contextual cues, exhibited significant performance improvements with the inclusion of context. However, this enhancement was accompanied by an increase in False Positives. Furthermore, our analysis of target groups revealed distinct patterns of HS prevalence and model performance. Groups characterized by a higher proportion of CHS demonstrated lower model performance, highlighting the challenges associated with detecting subtle forms of HS within vulnerable communities. In conclusion, our study contributes valuable insights into the complex dynamics of HS detection using generative models. Moving forward, future research should focus on refining prompt strategies, enhancing contextual modeling techniques, and developing tailored approaches by fine-tuning these generative models on HS corpora to further specialize them for this task.

Acknowledgments. This work was funded in part by the European Union under Grant CERV-2021-EQUAL (101049306). However, the views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European

Union or Knowhate Project. Neither the European Union nor the Knowhate Project can be held responsible.

References

1. Baider, F.: Pragmatics lost?: Overview, synthesis and proposition in defining online hate speech. Pragmatics and Society **11**(2), 196–218 (2020). https://doi.org/10.1075/ps.20004.bai
2. Baider, F.: Covert hate speech, conspiracy theory and anti-semitism: Linguistic analysis versus legal judgement. International Journal for the Semiotics of Law - Revue internationale de Sémiotique juridique **35**, 1–25 (04 2022). https://doi.org/10.1007/s11196-022-09882-w
3. Baider, F., Constantinou, M.: Covert hate speech: A contrastive study of greek and greek cypriot online discussions with an emphasis on irony. Journal of Language Aggression and Conflict **8**(2), 262–287 (2020). https://doi.org/10.1075/jlac.00040.bai
4. Carvalho, P., Matos, B., Santos, R., Batista, F., Ribeiro, R.: Hate speech dynamics against African descent, Roma and LGBTQ+ communities in Portugal. In: Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022). European Language Resources Association (ELRA (2022)
5. Carvalho, P., Caled, D., Silva, C., Batista, F., Ribeiro, R.: The expression of hate speech against afro-descendant, roma, and lgbtq+ communities in youtube comments. Journal of Language Aggression and Conflict (2023). https://doi.org/10.1075/jlac.00085.car
6. Carvalho, P., Guerra, R.: D3.2/D3.3 Annotation Guidelines OHS & OCS. Tech. rep., Iscte-Instituto Universitário de Lisboa (05 2023)
7. Casavantes, M., Aragón, M., González, L., Montes-y Gómez, M.: Leveraging posts' and authors' metadata to spot several forms of abusive comments in Twitter. Journal of Intelligent Information Systems **61**(2), 519–539 (2023). https://doi.org/10.1007/s10844-023-00779-z
8. Chen, L., Zaharia, M., Zou, J.: How is chatgpt's behavior changing over time? (2023)
9. Chiu, K.L., Collins, A., Alexander, R.: Detecting hate speech with gpt-3 (2022)
10. Fortuna, P., Rocha da Silva, J., Soler-Company, J., Wanner, L., Nunes, S.: A hierarchically-labeled Portuguese hate speech dataset. In: Proceedings of the Third Workshop on Abusive Language Online. pp. 94–104. Association for Computational Linguistics, Florence, Italy (Aug 2019). https://doi.org/10.18653/v1/W19-3510, https://aclanthology.org/W19-3510
11. Giacomo Marzi, M.B., MArchiori, D.: K-alpha calculator–krippendorff's alpha calculator: A user-friendly tool for computing krippendorff's alpha inter-rater reliability coefficient. Methods X **12** (2024)
12. Krobová, T., Zàpotocký, J.: "i am not racist, but ...": Rhetorical fallacies in arguments about the refugee crisis on czech facebook. Journal of Intercultural Communication **21**(2), 58–69 (Jul 2021). https://doi.org/10.36923/jicc.v21i2.14, https://immi.se/index.php/intercultural/article/view/14
13. Magu, R., Luo, J.: Determining code words in euphemistic hate speech using word embedding networks. In: Proceedings of the 2nd Workshop on Abusive Language Online (ALW2). pp. 93–100. Association for Computational Linguistics, Brussels, Belgium (Oct 2018). https://doi.org/10.18653/v1/W18-5112, https://aclanthology.org/W18-5112

14. Marco, M.A.: Rhetorical questions as reproaching devices. Journal of Language Aggression and Conflict **11**(2), 176–199 (2023). https://doi.org/10.1075/jlac.00077.alb
15. Matos, B.C., Santos, R.B., Carvalho, P., Ribeiro, R., Batista, F.: Comparing Different Approaches for Detecting Hate Speech in Online Portuguese Comments. In: 11th Symposium on Languages, Applications and Technologies (SLATE 2022). Open Access Series in Informatics (OASIcs), vol. 104, pp. 10:1–10:12. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany (2022). https://doi.org/10.4230/OASIcs.SLATE.2022.10
16. Media Smarts: Impact of Online Hate (2021), https://mediasmarts.ca/online-hate/impact-online-hate
17. Oliveira, A., Cecote, T., Silva, P., Castro Gertrudes, J., Freitas, V., Luz, E.: How good is chatgpt for detecting hate speech in portuguese? In: Anais do Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL). pp. 94–103 (09 2023). https://doi.org/10.5753/stil.2023.233943
18. Perez, J., Luque, F., Zayat, D., Kondratzky, M., Moro, A., Serrati, P., Zajac, J., Miguel, P., Debandi, N., Gravano, A., Cotik, V.: Assessing the Impact of Contextual Information in Hate Speech Detection. IEEE Access **11**, 30575–30590 (2023). https://doi.org/10.1109/ACCESS.2023.3258973
19. Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., Patti, V.: Resources and benchmark corpora for hate speech detection: a systematic review. Language Resources and Evaluation **55**(2), 477–523 (Jun 2021). https://doi.org/10.1007/s10579-020-09502-8
20. Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S. Yu: A Survey on Text Classification: From Traditional to Deep Learning | ACM Transactions on Intelligent Systems and Technology. ACM Transactions on Intelligent Systems and Technology **13**(2), 1–41 (Apr 2022). https://doi.org/10.1145/3495162
21. Santos, R.B., Matos, B.C., Carvalho, P., Batista, F., Ribeiro, R.: Semi-Supervised Annotation of Portuguese Hate Speech Across Social Media Domains. In: 11th Symposium on Languages, Applications and Technologies (SLATE 2022). Open Access Series in Informatics (OASIcs), vol. 104, pp. 11:1–11:14. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, Dagstuhl, Germany (2022). https://doi.org/10.4230/OASIcs.SLATE.2022.11
22. Statista: Number of social media users worldwide from 2017 to 2027 (Aug 2023), https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/
23. Tynes, B.M., Giang, M.T., Williams, D.R., Thompson, G.N.: Online racial discrimination and psychological adjustment among adolescents. The Journal of Adolescent Health: Official Publication of the Society for Adolescent Medicine **43**(6), 565–569 (Dec 2008). https://doi.org/10.1016/j.jadohealth.2008.08.021
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is All you Need. In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc., Long Beach, CA (2017). https://doi.org/10.48550/arXiv.1706.03762
25. Watanabe, H., Bouazizi, M., Ohtsuki, T.: Hate Speech on Twitter: A Pragmatic Approach to Collect Hateful and Offensive Expressions and Perform Hate Speech Detection. IEEE Access **6**, 13825–13835 (2018). https://doi.org/10.1109/ACCESS.2018.2806394