Exploring ChatGPT for Hate Speech Detection in *Rioplatense Spanish*

Anonymous ACL submission

Abstract

The irruption of Large Language Models (LLMs) has revolutionized the field of Natural Language Processing (NLP). In particular, the GPT-3.5 model (also known as ChatGPT) has been shown to be effective in a wide range of NLP tasks. In this work we present a brief analysis of the performance of ChatGPT in the detection of Hate Speech for Rioplatense Spanish. We performed classification experiments leveraging chain-of-thought (CoT) reasoning with ChatGPT, and compare their results against a state-of-the-art BERT classifier. Our experiments show that ChatGPT show a lower precision compared to the fine-tuned BERT classifier, but a higher recall for highly nuanced cases (particularly, homophobic/transphobic hate speech). In some cases, we observe that ChatGPT is not getting slurs or colloquialisms. We make our code and models publicly available for future research.

1 Introduction

In recent years, an increasingly unfolding of violent, discriminatory and hateful speeches can be observed on digital platforms, media and networks (Berecz and Devinat, 2017). Along with the rising of the so-called "alternative right" movements, which have a strong presence on social networks (Woods and Hahner, 2019; Hodge and Hallgrímsdóttir, 2021), discriminatory and hateful discourses surface in different enunciation areas and modalities, specially in public spaces such as social media.

Social media, such as Twitter, offers valuable data access to a relatively natural environment for the study of hate speech. Most of the studies for this pervasive phenomenon have been conducted in English. Spanish takes second place as first language worldwide (after Chinese), with more than 450 million native speakers, primarily in Spain, Latin America and also parts of the US (Tellez et al., 2023). This aggregation includes many varieties and dialects. Among them, Rioplatense Spanish, mainly spoken both in Argentina and Uruguay, accounts for a tied second place with Colombia, Spain and US, and surpassed in speakers only by Mexico (Lipski, 2012; Coloma, 2018).

Large-language models (LLMs) have been shown to be effective in a wide range of NLP tasks (Brown et al., 2020; Wei et al., 2021; Ouyang et al., 2022). Being GPT-3.5 (also known as ChatGPT) one of the most popular and growing LLMs (Wu et al., 2023; Deng and Lin, 2022) it arises the question of how well it could detect hateful messages in a particular dialectal variant of Spanish.

In this work we present a brief analysis of the performance of ChatGPT in the detection of Hate Speech for Rioplatense Spanish. We performed classification experiments leveraging chain-of-thought (CoT) reasoning with ChatGPT, and compare their results against a fine-tuned BERT classifier. Our experiments show that ChatGPT show a lower precision compared to the fine-tuned BERT classifier, but a higher recall for highly nuanced cases (particularly, homophobic/transphobic hate speech). However, explanations given by ChatGPT are —while not equal to human annotators— convincing in most cases.

We make our code and models publicly available.[1]

2 Related work

A broad amount of literature has been written in the past years about the automatic detection and treatment of hate speech. We refer the readers to Poletto et al. (2021); Schmidt and Wiegand (2017); Fortuna and Nunes (2018) for extensive reviews of work in the field. In this section, we focus on the most recent work on hate speech detection, explanation and treatment using LLMs.

With the recent advent of LLMs (Brown et al.,

[1] TBD

2020; Wei et al., 2021; Ouyang et al., 2022; **?**), some studies have been conducted to evaluate their performance in hate speech detection, explanation and treatment. Sap et al. (2020) used GPT-2 to detect and generate hate speech explanations. Wang et al. (2023); Huang et al. (2023) evaluated the performance of GPT-3/GPT-3.5 to detect and explain hate speech messages, finding that LLM-generated explanations are equally good (and even preferred to) human-written explanations. Some of these explanations are inducted by chain-of-thought reasoning (Wei et al., 2022), also known as the "let's think step by step" technique. Oliveira et al. (2023) tested ChatGPT for hate speech detection in Portuguese, particularly on its Brazilian dialect, achieving almost state-of-the-art results in a zero-shot setting. Çam and Özgür (2023) performed experiments for Turkish, with similar results.

3 Data

We used the dataset from Pérez et al. (2023), which consists of Twitter replies to posts from news outlets from Argentina. These comments were annotated for hate speech detection and categorized into eight possible types: misogyny, homophobia/transphobia, racism/xenophobia, class hatred, appearance, against criminals and disabled people. All the instances have then a context (the post from the news outlet) and text being analyzed (the comment from an user). Contextual information situates the comment and has been shown quite relevant to detect hate speech (Sheth et al., 2022; Xenos et al., 2021; Pérez et al., 2023).

For this work, we only considered the first four categories, from now on dubbed WOMEN, LGBTI, RACISM, CLASS. We selected these categories based on their prevalence and societal impact. The remaining categories (appearance, against criminals, and disabled people) were excluded as they are not usually considered in the literature, and we want to check ChatGPT abilities on a more standard ground.

Table 1 shows some examples of the dataset.

4 Classification experiments

To test ChatGPT performance on hate speech detection, we prompted the model with the following text:

> Determine if the following message contains hate speech. We understand that there is hate speech if it has statements of an intense and irrational nature of rejection, enmity and abhorrence against an individual or against a group, being the targets of these expressions for possessing a protected characteristic. The protected characteristics that we consider are:
>
> - women: women or feminist movement
> - lgbti: against gays, lesbians, transsexuals and other gender identities
> - racism: immigrants, xenophobia, or against aboriginal peoples
> - class: low-income people or class reasons
>
> Answer one or more of the characteristics separated by a comma, or "nothing" if there is no hate speech. Think step by step before answering.

We leveraged chain-of-thought reasoning (CoT) (Wei et al., 2022) to both enhance the model's performance and to provide an explanation for the prediction. Two settings were provided for the model: one-shot and few-shot. In the one-shot setting, the model was prompted with a single example of hate speech, particularly for *racism*. In the few-shot setting, the model was prompted with 12 examples of hate speech for the different characteristics. The examples were selected from the training set, and we tested different configurations against the validation set. Each example consisted of three lines, such as this:

> **context:** Wuhan celebrates the end of the coronavirus quarantine with a message for the rest of the world: "Learn from our mistakes"
>
> **text:** Motherfuckers! I wish you all chinese people die
>
> **output:** The text wishes that Chinese people would die, blaming them for the COVID-19 pandemic. The final answer is "racism".

Both the context and the text tweet were preprocessed using the *pysentimiento* library (Pérez et al., 2023).

We relied on GPT-3.5 turbo[2], accessed through the API via its *python* API. We compare Chat-

[2] gpt-3.5-turbo-0613

Category	Context	Comment
WOMEN	Mia Khalifa: acted in porn videos for a few months, became world famous and now fights to erase her past	HAHAHA KEEP SUCKING....
LGBTI	The story of the Colombian trans model kissing the belly of her eight-month pregnant husband	A male kissing another male
RACISM	Yanzhong Huang: "It is quite likely that a Covid-21 is already brewing"	Urgent bombs to that damned race
CLASS	Social movements cut off 9 de Julio Av.: they demand a minimum wage of $45,000	get to work, mfs

Table 1: Hateful examples from the analyzed dataset.

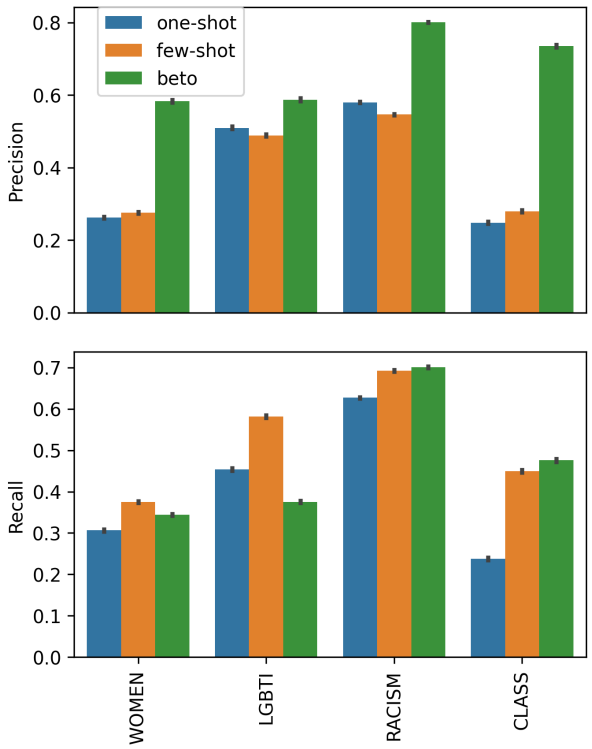

Figure 1: Precision and recall of the classifiers: ChatGPT in one-shot and few-shot learning settings, and a fine-tuned *BETO* classifier.

GPT's performance with a fine-tuned *BETO* classifier (Cañete et al., 2020) trained on the same corpus, following the guidelines of Pérez et al. (2023).

To evaluate the performance of the classifiers, we assessed the precision, recall, and F1-score in two modalities: multi-label classification, and binary classification (that is, not taking into account if a racist message was labeled as class-hatred). We get bootstrap 95%-ci intervals using the *confidence-intervals* library (Ferrer and Riera, 2023).

5 Results

Figure 1 shows the results for the multi-label experiments, disaggregated by characteristic. Table 2 shows the binary classification results. We can observe that the fine-tuned *BETO* classifier has the highest precision in all the analysis, but the few-shot classifier matches it in recall, and even outperforms it for WOMEN and LGBTI. This last category is particularly interesting, as this attacked group features culturally-nuanced, and metaphorical language, which are quite difficult to detect for BERT-based classifiers (Pérez et al., 2023). We sampled a few false positives for the *BETO* classifier that were correctly classified by ChatGPT, and found that the explanations generated by ChatGPT were correct, capturing subtle, non-slur related language attacking transgender people, one of the most difficult characteristics to detect. Table 3 shows some examples of these cases.

	Precision	Recall	F1
BETO	72.9 ± 2.3	56.2 ± 2.1	63.5 ± 1.9
One shot	45.5 ± 2.1	50.4 ± 2.2	47.8 ± 1.9
Few shot	43.5 ± 1.9	61.0 ± 2.2	50.8 ± 1.8

Table 2: Binary classification results. One shot and few shot refer to ChatGPT's one-shot and few-shot settings, respectively.

5.1 Error Analysis

To address the cases in which the ChatGPT prediction differed from the aggregated labels from the corpus, we sampled 126 examples and two authors reviewed them. As the CoT explanation gives a hint of the model's reasoning, we decided to score the **soundness** (Wang et al., 2023) in the logical reasoning behind the decision. A Likert scale was used, ranging from 1 to 5, where 1 and 2 were considered *errors*, 3 was considered *doubtful*, and 4 and 5 were considered *acceptable*.

As a result, we found that in 77% cases the answers were acceptable for both scorers (5-4 points), while 15.5% were definitely errors (1-2 points)

News Context	Reaction Tweet	ChatGPT CoT output
BETO false negatives correctly classified by ChatGPT		
"Why does a kiss between two men make them so violent?": life after being a victim of homophobic attacks	Could be because it goes against the nature of the race.	The text alludes to the fact that a kiss between two men goes against the "nature of the race", insinuating that homosexuality is not natural. The final answer is LGBTI, racism.
Why don't we see trans doctors?: the claim of a prestigious cardiologist for America to be more inclusive	It is difficult to be a doctor with a burned head	The comment alludes to the condition of being transgender (trans doctors) and discriminates against them, saying that it is difficult be a doctor with that characteristic. The final answer is LGBTI.
Incorrectly classified by ChatGPT		
Lázaro Báez's partner told how they were waiting for him inside the country: "What they did to him was outrageous."	The paskin with the ***trola*** (whore).	The text does not mention any protected characteristics nor does it contain statements of rejection or enmity towards an individual or group. The final answer is 'nothing'.
Moria Casán's anger against Rocío Oliva: 'A lot of oxygenated water, she had half a neuron left to play ball'	patoruzú [3] trava	The text is an insult towards the person mentioned, but without making any mention of the protected characteristics. The final answer is 'nothing'.

Table 3: Some examples of true positives and false negatives.

and 7.5% were doubtful cases (3 points), being the agreement between scorers of 79%.

As aforementioned most of ChatGPT's CoT answers were found acceptable, whether it finds hate speech or not, the argument is plausible. In cases where hate speech is detected, that was not annotated in the corpora, a broader sensitivity towards certain topics can be noticed. In cases where ChatGPT fails to detect hate speech, even if the CoT is convincing, some details escape, particularly when detecting discrimination against, for example, the Mapuche community and in other cases of comments that were detected as racist speech.

Regarding errors and doubtful cases in labeling, we can find cases in which the mistake is linked to incomprehension of nuances in the lexicon, such as shown in the second part of Table 3. First row shows a case in which ChatGPT fails to detect a hateful message against the woman, identified as a prostitute with the Rioplatense slang "trola". In the second case ChatGPT's fails to understand the slur "trava", which refers to a transgender woman.

6 Conclusions

In this work we presented a brief analysis of the performance of ChatGPT in the detection of Hate Speech for Rioplatense Spanish. In the comparison with a state-of-the-art fine-tuned BETO classifier, ChatGPT showed a lower precision but a higher recall in some categories, particularly in difficult cases that the supervised classifier could not detect. A deeper analysis of the chain-of-thought explanations given by the LLM showed that, while not agreeing with human annotations, its reasoning was sound in most cases but showing a higher bias towards flagging hate speech.

Regarding cultural and linguistic nuances, we found that ChatGPT was able to detect some of them, but not all, missing some slurs, expressions and insults typical of the Rioplatense dialect. Future work could focus on improving the prompting to have a better handling of dialectal variants.

While ChatGPT shows as a powerful tool for hate speech detection, supervised classifiers still outperform it in precision, and are more suitable for detecting hate speech at large scale. This highlights the importance and value of producing corpora on specific topics and linguistic variants.

7 Limitations

One of the main limitations of this work is the dataset we worked with, and the task itself (hate speech detection) which tries to capture a complex social phenomenon. The original dataset does not have natural language explanations for the annotations.

The analysis of explanations was performed in a very limited way, only assessing its soundness and by two of the authors. A more thorough analysis of the explanations is needed, including a larger sample, more annotators and using better metrics (such as informativeness).

Also, the analysis of the whether ChatGPT detects culturally nuanced cases was mostly restricted to the LGBTI category. Future work should explore other categories and also the use of slang and colloquialisms in depth.

Acknowledgements

To be defined in the final version of the paper.

References

Tamás Berecz and Charlotte Devinat. 2017. Relevance of cyber hate in europe and current topics that shape online hate speech. *INACH, EU*, 7:2020.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Nur Bengisu Çam and Arzucan Özgür. 2023. Evaluation of chatgpt and bert-based models for turkish hate speech detection. In *2023 8th International Conference on Computer Science and Engineering (UBMK)*, pages 229–233. IEEE.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish Pre-Trained BERT Model and Evaluation Data. *PML4DC at ICLR*.

Germán Coloma. 2018. Argentine spanish. *Journal of the International Phonetic Association*, 48(2):243–250.

Jianyang Deng and Yijia Lin. 2022. The benefits and challenges of chatgpt: An overview. *Frontiers in Computing and Intelligent Systems*, 2(2):81–83.

Luciana Ferrer and Pablo Riera. 2023. Confidence Intervals for evaluation in machine learning. Original-date: 2023-12-06T12:26:21Z.

Paula Fortuna and Sérgio Nunes. 2018. A survey on automatic detection of hate speech in text. *ACM Computing Surveys (CSUR)*, 51(4):1–30.

Edwin Hodge and Helga Kristín Hallgrímsdóttir. 2021. Networks of hate: the alt-right,"troll culture", and the cultural geography of social movement spaces online. In *British Columbia's Borders in Globalization*, pages 102–119. Routledge.

Fan Huang, Haewoon Kwak, and Jisun An. 2023. Is chatgpt better than human annotators? potential and limitations of chatgpt in explaining implicit hate speech. In *Companion Proceedings of the ACM Web Conference 2023*, pages 294–297.

John M Lipski. 2012. Geographical and social varieties of spanish: An overview. *The handbook of Hispanic linguistics*, pages 1–26.

Amanda S Oliveira, Thiago C Cecote, Pedro HL Silva, Jadson C Gertrudes, Vander LS Freitas, and Eduardo JS Luz. 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. SBC.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Juan Manuel Pérez, Franco M Luque, Demian Zayat, Martín Kondratzky, Agustín Moro, Pablo Santiago Serrati, Joaquín Zajac, Paula Miguel, Natalia Debandi, Agustín Gravano, et al. 2023. Assessing the impact of contextual information in hate speech detection. *IEEE Access*, 11:30575–30590.

Fabio Poletto, Valerio Basile, Manuela Sanguinetti, Cristina Bosco, and Viviana Patti. 2021. Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55(2):477–523.

Juan Manuel Pérez, Mariela Rajngewerc, Juan Carlos Giudici, Damián A. Furman, Franco Luque, Laura Alonso Alemany, and María Vanina Martínez. 2023. pysentimiento: A python toolkit for opinion mining and social nlp tasks.

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

Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In *Proceedings of the fifth international workshop on natural language processing for social media*, pages 1–10.

Amit Sheth, Valerie L Shalin, and Ugur Kursuncu. 2022. Defining and detecting toxicity on social media: context and knowledge are key. *Neurocomputing*, 490:312–318.

Eric S Tellez, Daniela Moctezuma, Sabino Miranda, Mario Graff, and Guillermo Ruiz. 2023. Regionalized models for spanish language variations based on twitter. *Language Resources and Evaluation*, pages 1–31.

Han Wang, Ming Shan Hee, Md Rabiul Awal, Kenny Tsu Wei Choo, and Roy Ka-Wei Lee. 2023. Evaluating gpt-3 generated explanations for hateful content moderation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 6255–6263. International Joint Conferences on Artificial Intelligence Organization. AI for Good.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Heather Suzanne Woods and Leslie A Hahner. 2019. *Make America meme again: The rhetoric of the Alt-Right*. Peter Lang New York.

Tianyu Wu, Shizhu He, Jingping Liu, Siqi Sun, Kang Liu, Qing-Long Han, and Yang Tang. 2023. A brief overview of chatgpt: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica*, 10(5):1122–1136.

Alexandros Xenos, John Pavlopoulos, and Ion Androutsopoulos. 2021. Context sensitivity estimation in toxicity detection. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 140–145, Online. Association for Computational Linguistics.