Exploring Large Language Models for Hate Speech Detection in *Rioplatense* Spanish

Anonymous ACL submission

Abstract

Hate speech detection deals with many variants, slang, slurs, specific lexicons, expression modalities, and cultural nuances. This outlines the importance of working with specific corpora, when addressing hate speech within the scope of Natural Language Processing, recently revolutionized by the irruption of Large Language Models. This work presents a brief analysis of the performance of large language models in the detection of Hate Speech for *Rioplatense* Spanish. We performed classification experiments leveraging chain-of-thought reasoning with *ChatGPT 3.5*, *Mixtral*, and *Aya*, comparing their results with those of a state-of-the-art BERT classifier. These experiments outline that, even if large language models show a lower precision compared to the fine-tuned BERT classifier and, in some cases, they find hard-to-get slurs or colloquialisms, they still are sensitive to highly nuanced cases (particularly, homophobic/transphobic hate speech). 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, especially in public spaces such as social media. Social media, as Twitter, offers valuable data access to a relatively natural environment for the study of hate speech, being particularly interesting the activation of hate speech regarding public topics, such as news (Zannettou et al., 2020; Erjavec and Kovačič, 2012).

From the Natural Language Processing (NLP) perspective, hate speech detection has to deal with languages crossed by variants, slang, slurs, and other specific modalities found (Nunberg, 2018; Diaz-Legaspe, 2020). This is why it is important to be aware of cultural nuances and specific contexts of use. There is a plethora of resources for automatic detection of hate speech. Nevertheless, when it comes to Spanish, corpora are scarce, despite being one of the main languages in the number of worldwide native speakers (after Chinese and Hindi). With over 450 million native speakers, primarily in Spain, Latin America, and also parts of the US (Tellez et al., 2023; Eberhard et al., 2023), Spanish includes many varieties and dialects. Each variety and dialect represents a common cultural background and semantic field, expressing different uses for some words or, contrary wise, the use of specific words or phrases addressing the same purpose. Among them, *Rioplatense* Spanish, mainly spoken both in Argentina and Uruguay, is thought to be spoken by more than one-tenth of Spanish native speakers. This variant accounts for a tied second place with Colombia, Spain, and US Spanish, and is surpassed in speakers only by Mexico (Lipski, 2012; Coloma, 2018).

Rioplatense Spanish also includes argot and slang, especially *lunfardo*, an integrated lexical repertoire, which has around 6,000 voices, of which only about 300 are recognized by the Dictionary of the Royal Spanish Academy (Conde, 2013).[1] While almost all languages have repertoires of expression outside of general use, the case of *lunfardo* constitutes a linguistic phenomenon in which words and expressions of very diverse origin converge (Italian, popular Spanish, French, Portuguese, Guaraní, Quechua, among others), as result of the migratory processes in Argentina, with its epicenter

[1] The Royal Spanish Academy (RAE) is a cultural institution dedicated to linguistic regularization among the Spanish-speaking world.

in Buenos Aires, especially during the 19th century and the first half of the 20th century (Conde, 2013).

This underlines the relevance of developing dialectal corpora and analysis that allows to automatically detect specific hateful expressions in different lexical contexts. Addressing this issue emerges, at first, a particular interest into the performance of Large-language models (LLMs) by analyzing hate speech regarding local expression nuances. LLMs have 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 raising 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, focusing in *Rioplatense* variety.

This paper aims to develop an exploratory approach to the effectiveness of LLMs in detecting specific texts and tagging corpora. We take as benchmark a fine-tuned BERT classifier trained with a corpus written in Rioplatense Spanish, annotated to detect hate speech. In this case an specific analysis of the performance of LLMs in the detection of Hate Speech for *Rioplatense* Spanish is showcased.[2]

This work focuses on expressions of hate based on gender (against women or LGBTI), racism and classism, based on their prevalence among hateful speech and their societal impact. These topics are widely covered and considered in the available literature, meaning the results could be a useful contribution to standard ground and state of the art (Paz et al., 2020; Tontodimamma et al., 2021). We performed classification experiments leveraging the chain-of-thought (CoT) reasoning within the LLMs ChatGPT, Mixtral and Aya, and compare their results against a fine-tuned BERT classifier. Our experiments point out that LLMs 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.

2 Related work

In order to identify hate speech, the first step was to define an operational definition of "hate speech" and "discrimination". At the same time, it was necessary to achieve a restricted definition of these problematic speeches to simultaneously sustain freedom of expression. Here, the conceptual framework relies upon the human rights paradigm and international instruments linked to freedom of expression and non-discrimination.[3] Departing from this general basis, we follow the Argentinean case, where *Rioplatense* is mainly spoken. There, the Argentinian National Plan against Discrimination, developed by the Presidency of the Nation, proposes a classification of characteristics and groups subject to discrimination in the country. They report different areas that showcase the activation of discriminatory discourses and could be identified, such as nationality or ethnicity, migration, religion, gender, and sexual identity, among others (Presidencia de la Nación, 2006).

Recently, a broad amount of literature has been written 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 of NLP. 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., 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. Plaza-del arco et al. (2023) assessed the performance of several language models (such as the instruction-finetuned *mT0* (Muennighoff et al., 2023) and *FLAN-T5* (Chung et al., 2022) in zero-shot setting over several hate speech and toxicity datasets. 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

[2]We make our code and models publicly available. TBD

[3]We start from a broad definition of freedom of expression, such as that proposed by the American Convention on Human Rights (ACHR) and the International Covenant on Civil and Political Rights (ICCPR), which indicate that everyone has the right to freedom of thought and of unrestricted expression. Likewise, these same international treaties establish that hate speech is not protected by freedom of expression.

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

For our experiments we use a dataset in *Rioplatense* Spanish, specifically annotated for hate speech detection. The dataset consists of Twitter replies to posts from Argentinean news outlets[4]. In this dataset, comments to news posted by regional users were annotated for the presence of hate speech and categorized into one or more of four possible types: misogyny, homophobia/transphobia, racism/xenophobia, and class hatred according to the attacked characteristics, from now on dubbed WOMEN, LGBTI, RACISM, and CLASS. All annotated instances have a context (the tweet posted by the news outlet, plus the whole content of the news) and the text being analyzed and annotated (each Twitter user's comment). Contextual information situates the comment and has been shown relevant to detect hate speech (Sheth et al., 2022; Xenos et al., 2021).

We worked with 5670 comments to news, half of the original dataset. Of them, 479 comments contain messages of discrimination or hate against at least one of the targeted categories. Accounting for 230 comments for RACISM, 131 for WOMEN, 88 for LGBTI, and 76 for CLASS. Some messages express attacks to more than one category. In 44 comments, hate speech addresses two categories, finding two relevant combinations: RACISM associated with CLASS (21 cases), followed by the association of WOMEN and LGBTI (10 cases). Only one comment targeted 3 categories.

Table 1 shows some examples of the dataset.

3.1 Regionalism identification

The tweets corresponding to the test dataset were also categorized according to their inclusion of regional terms. Therefore, we used a dictionary of regional terms[5].

The three authors of the paper also manually annotated a set of 51 Tweets to determine whether they contained regionalisms or not. Therefore, we referred to regionalisms as idiomatic phrases, words exclusively used in Rioplatense Spanish, or with a meaning that is differentiallyhhh used in Argentina or only in certain Spanish-speaking countries (eg. "pelotudo" for idiot).

The inter-annotator agreement among the dictionary categorization and each of the human annotations using Krippendorff's Alpha (Krippendorff, 2011) were 0.32, 0.39, and 0.47. The inter-annotator agreement using Krippendorff's Alpha among the three annotations was 0.6.

4 Classification experiments

We compared two kinds of classification algorithms:

- Pre-trained language models based on *BERT*: fine-tuned on supervised data from the corpus.
- Large Language Models (LLMs) using few-shot learning and chain-of-thought reasoning (CoT) (Wei et al., 2022).

For the first group of classifiers, we tested pre-trained models in Spanish, namely *BETO* (Cañete et al., 2020), RoBERTa (Gutiérrez-Fandiño et al., 2022) and RoBERTuito (Pérez et al., 2022). For each model, we performed a small hyperparameter search following the guidelines of Godbole et al. (2023), searching for the best-performing values for the number of epochs, the learning rate and warm-up ratio. To track our experiments, we used the *wandb* library (Biewald, 2020). For each of the pre-trained models, we previously fine-tuned them on an unsupervised corpus provided in the used dataset, as it has been shown to improve the performance in domain-specific tasks (Gururangan et al., 2020). More details on the fine-tuning process of the supervised models can be found in the Appendix A.1.

As for the LLMs, we resorted to few-shot or in-context learning with the following prompt, which was translated from Spanish to English for the sake of clarity:

> Determine if the following text, corresponding to a tweet, presented with a context, 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 we consider are:

[4] Published and publicly available dataset. Not mentioned for anonymity reasons.

[5] The reference will be included in the final version of this paper.

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.

> - women: refers to women or the feminist movement
> - LGBTI: refers to gays, lesbians, transgender individuals, and other gender identities
> - racism: refers to immigrants, xenophobia, or against indigenous peoples
> - class: refers to low-income people or class-related issues
>
> The tweets are written in Rioplatense Spanish, and within the cultural context of Argentina. Respond with one or more of the characteristics separated by commas, or "nothing" if there is no hate speech. Think and justify the response step by step before answering.

We leveraged chain-of-thought reasoning (Wei et al., 2022) to both enhance the model's performance and to provide an explanation for the prediction. The model was prompted with a total of 12 examples of hate speech considering the different characteristics. The examples were selected from the training set, and 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. answer is "racism".

The output consists of a natural language explanation. The full list of examples and the original prompt in Spanish can be found in Appendix A.2.

Regarding the large language models, we selected three models that show good performance in Spanish:

- *GPT-3.5* turbo[6] (Ouyang et al., 2022): a closed-source large language model provided by OpenAI, that has an outstanding performance in several tasks.
- *Mixtral* (Jiang et al., 2024): a mixture-of-experts open-source language model pre-trained in English, French, Italian, German and Spanish.
- *Aya* (Üstün et al., 2024): a massively-multilingual sequence-to-sequence language model, that follows the architecture of *T5* (Raffel et al., 2020), pre-trained in 101 languages.

Mixtral and Aya were run in two NVIDIA A30, using the Transformers library. The same prompt was used for the three LLMs.

4.1 Evaluation

To evaluate the performance of the classifiers, we assessed the precision, recall, and F1-score in two modalities: multi-label classification (we consider a true positive if at least one category matches), and binary classification (that is, if the message contains hateful speech or not). We get bootstrap 95%-CI intervals using the *confidence-intervals* library (Ferrer and Riera, 2023). We also evaluated a subset of the dataset, that specifically contains regional terms.

5 Results

Table 2 shows the results for the binary classification task. It shows that fine-tuned BETO classifier outperforms in terms of precision and F1, but GPT-3.5, Mixtral, and Aya have higher recall. As Aya

[6] gpt-3.5-turbo-0125

Model	F1	Precision	Recall
Aya	21.2 ± 0.8	11.9 ± 0.5	93.0 ± 1.2
Mixtral	38.6 ± 1.3	25.1 ± 1.0	83.8 ± 1.7
GPT-3.5	47.8 ± 1.8	39.2 ± 1.8	61.2 ± 2.2
FT *BETO*	63.5 ± 1.8	72.9 ± 2.4	56.3 ± 2.1

Table 2: Binary classification results of Aya, Mixtral, fine-tuned (FT) BETO, GPT 3.5, and Aya.

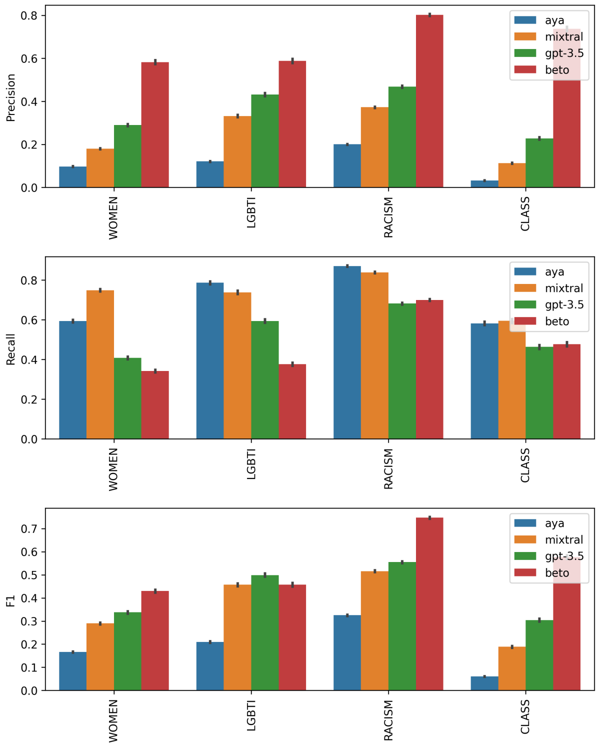

Figure 1: Precision, recall and F1 of the classifiers: ChatGPT 3.5, Aya, Mixtral and the fine-tuned *BETO* classifier.

model performs poorly in terms of precision, qualitative analysis is focused on the LLMs GPT-3.5 and Mixtral.

A closer inspection of each of the considered characteristics is presented in the multi-label classification results, shown in Figure 1. It is shown that *Mixtral* obtains a better recall for all of the characteristics but at the cost of low precision, while *GPT-3.5* has a better trade-off between them. The case of the LGBTI characteristic, is particularly interesting given that is the case where *GPT-3.5* outperforms the fine-tuned classifier ($F1 = 49.9 \pm 4.4$ for *GPT-3.5* and $F1 = 45.7 \pm 5.1$ for *BETO*, Mann-Whitney U=$16386.5, p \leq 1 * 10^{-6}$). This is particularly relevant as this characteristic is difficult to detect, as it often involves culturally nuanced language,

MODEL	BETO		GPT-3.5		Mixtral	
Regionalism	reg.	wo. reg.	reg.	wo. reg.	reg.	wo. reg.
HATEFUL	0.65	0.61	0.54	0.44	0.45	0.35
CLASS	0.67	0.30	0.46	0.11	0.28	0.09
LGBTI	0.47	0.44	0.50	0.48	0.44	0.47
RACISM	0.70	0.76	0.60	0.54	0.56	0.50
WOMEN	0.51	0.27	0.42	0.25	0.39	0.20
Macro	0.60	0.48	0.50	0.37	0.42	0.32

Table 3: F1 by category for messages containing regionalism (reg.) and those not containing it (wo. reg). Hateful represents the binary classification.

irony and metaphors and where *BERT*-based classifiers show a significant gap compared to humans (Yigezu et al., 2023).[7]

Focusing in those messages that contain regionalisms –that means 1547 comments, being 245 hateful ones, detected as described in Subsection 3.1-, when the test subset with regionalisms is evaluated, no conclusive differences are detected. As seen in 3, in general, the performance of all the classifiers follows similar patterns, they stay equal or get better when regionalisms are present in the text. This is due to the fact that regionalisms are likely to be matched with slang slurs which both the *BETO* and the LLMs (particularly, *GPT-3.5*) profit as leverage. Hate speech messages represented 9% of the dataset. That rises to a 16% of hateful messages within those that have regionalisms. So, hate speech has a prevalence when regionalisms are present.

Following up, for example, in CLASS category, words such as *planero, villero*, constituted part of the repertoire of classist denigrational speech; and regionalisms in form of slurs against WOMEN (*trola, yegua, abortera*) are also very common. These two classes are, in fact, those with more hateful comments containing regionalisms: 71% of CLASS hateful messages contains regionalisms and the same account for 62% of those against women. It is thought that allows an improvement in the precision of *BETO* in the presence of regionalisms. This helps this classifier addressing the task of flagging hateful speech, compared to comments without any slur slang, which makes hate speech more difficult to be detected.

In the case of LGBTI hateful instances, when hate speech detection task faces subtle, non-slur, slang related language attacking transgender people, most explanations generated by LLMs were correct, getting to detect difficult-to-tag messages that were overlooked by the *BETO* classifier. Table

[7]Another reference will be added upon publication.

News Context	Reaction Tweet	GPT-3.5 output	Mixtral output
BETO false negatives correctly classified by LLMs			
After a barrage of more than 20 shots, a trans woman was murdered in Rosario	The prostate should not appear in the autopsy by law!	The message refers to the gender identity of the trans woman murdered in a discriminatory manner, using an emoji to mock the situation. The final answer is "LGBTI".	The text refers to the victim as "a man" and denies her gender identity. The final answer is "LGBTI".
¿More Rial found love in a personal trainer? The mediatic shared a live broadcast with a handsome personal trainer where they looked very intimate	That one sees a *pingo* (Arg. for "horse", but also "dick") and jumps on.	The text refers to More Rial's sexuality in a derogatory way, using the term "pingo" in a pejorative way to refer to a penis. The final answer is "woman".	The text refers to More Rial as "that one" and accuses her of getting on a man because of her physical attractiveness. This is sexist and derogatory. The final answer is "woman".
Mara Gómez: the story of the first trans soccer player in the Argentine championship	*Ponga huevos, Mara ponga huevos...* (Eggs on [winning], Mara, Eggs on! Here "eggs" stands for testicles)	The text cheers Mara Gómez as a soccer player, but she uses the expression "*ponga huevos*" in a discriminatory way, with the aim of questioning her masculinity and her gender identity as a trans woman. The final answer is "LGBTI".	The text alludes to Mara Gómez's lack of "huevos", a term used to refer to the lack of courage or bravery. This is offensive and derogatory. The final answer is "woman, LGBTI."
Incorrectly classified or incorrect explanations by LLMs			
British arrived in Peru for 10 days and stayed for love: she lives without safe water and among insects	The Peruvian's "*gaver*". Mother. ("gaver" stands for "verga", meaning dick. It is a sexist comment, not racist)	The text contains a derogatory comment towards Peruvians, using a vulgar and discriminatory term. The final answer is "racism."	The text refers to Peruvians as "gaver", a derogatory term. The final answer is "racism."
Mara Gómez: the story of the first trans soccer player in the Argentine championship	It comes to bring what the team lacks: goals and a lot of "eggs" (in the sense of courage or bravery, refering to testicles. The comment should be tagged as LGBTI.)	The text praises Mara Gómez's soccer skills and mentions that she plans to contribute goals and effort to the team. It does not contain hate speech or mention any protected characteristics. The final answer is "nothing."	The text alludes to Mara Gómez's ability as a footballer, without making any allusion to a protected characteristic. The final answer is 'nothing'.
Tension in Villa Mascardi: Mapuches resist the eviction of a usurped property.	How strange, [Joe] Lewis and the millionaire squatters are not going to be evicted.	The text mentions the Mapuches in a derogatory manner and suggests preferential treatment towards non-indigenous occupants. Although protected characteristics are not directly mentioned, the comment denotes prejudice towards the Mapuche people. The final answer is "racism."	The text compares the Mapuches with the "millionaire squatters", suggesting that they are equal. This shows contempt for indigenous peoples and their fight for their lands. The final answer is "racism."
Regional slangs and some cultural nuances detected by LLMs			
Hebe de Bonafini[8] criticized supermarkets and called for "replacing meat with rice, chickpeas and lentils"	@usuario Old "chupa pi." ("pi." stands for "pija" ("cock"), meaning "cocksucker"), you like "morcilla" (literaly "blood sausage", but also "dick").	The text contains an insult towards Hebe de Bonafini, calling her an "vieja chupa pi." (old cocksucker), but also makes derogatory reference to her diet, using a sexist and vulgar comment. The final answer is "woman."	The text insults Hebe de Bonafini by calling her a "vieja chupa pi." (old cocksucker), a derogatory expression. Additionally, she makes fun of her dietary preferences. The final answer is "woman."
China: another virus identified "with the potential to become pandemic" #NewVirus #China	First they *morfan* (Arg. lunf. "eat") bats, now they must have *garchado* (Arg. "fucked") a pig.	The text contains a derogatory comment towards Chinese, alluding to their eating habits and suggesting a repulsive sexual intercourse with a pig. The final answer is "racism".	The text shows contempt for Chinese culture and its relation to animals. The final answer is "racism."

Table 4: Examples of LLMs results

4 shows some examples of those cases.[9]

These results suggest that despite possible preconceptions about the lack of cultural grounding of LLMs, they are somehow able to grasp cultural nuances from *Rioplatense* Spanish and Argentinian culture, even compared to fine-tuned BERT modeled to address that specific dialect. Some of the examples in Table 4 show that LLMs, sometimes are able to detect, explain, and provide information about regional slang not detected by the fine-tuned *BETO* classifier.

However, LLMs show a higher rate of false positives than the fine-tuned *BETO* classifier, when it comes to the reference of majority-vote labels in the dataset.[10] This might indicate, first, that these models are more sensitive to the presence of hate speech and toxicity (probably due to preference tuning or some other safety mechanisms and second, that the comparison against one single binary label might not be the best way to evaluate these models. Evaluating some of their explanations with other metrics, such as human evaluation of soundness, informativeness, among others (Wang et al., 2023), or also by using a perspectivist framework taking into account the disagreement of the annotations (Sachdeva et al., 2022; Basile et al., 2021) may provide a better comparison between these models.

[9]The analysis is shown for GPT 3.5 and Mixtral (with the benchmark of fine-tuned Beto), as Aya underperformed at this task.

[10]In the original dataset, each comment was annotated by three annotators. Therefore, it was used a majority-vote label.

6 Conclusions

This brief analysis attempts to showcase the performance of LLMs addressing the task of hate speech detection in *Rioplatense* Spanish tweets. In the comparison with a state-of-the-art fine-tuned BETO classifier, *ChatGPT* and *Mixtral* showed a lower precision but a higher recall in some categories, particularly in difficult cases that the supervised classifier was not able to detect. A deeper analysis of the chain-of-thought explanations given by LLMs reveals that, while not agreeing with human annotations, their reasoning showed soundness in most cases but expressed a higher bias towards classifying texts as containing hate speech.

While LLMs have proven to be a powerful tool for hate speech detection, supervised classifiers still outperform it in F1 and 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. Regarding cultural and linguistic nuances, we found that LLMs were able to detect some of them, but not all, missing some slurs, expressions and insults typical of the *Rioplatense* dialect. The culture and communication of Latin America is diverse. Full of different expressions, idioms, slang, specific uses and adaptations of the Spanish language which offers subtle differences that cannot be captured outside of their context of use. Future work could focus on improving the prompting to have a better handling of dialectal variants. Also, it could be of interest to conduct similar experiments with other Spanish variants, such as Iberian, where there are more available corpora and/or Mexican Spanish, which represents the majority of spoken Spanish.

7 Limitations

One of the main challenges that face this work is the task itself: hate speech detection, which tries to capture a complex social phenomenon. And, regarding the dataset, it has to be noted that the original dataset does not have natural language explanations for the annotations.

The analysis of LLMs explanations was performed in a very limited way, being their soundness assessed by the authors only. A deeper analysis of those explanations could be of interest, by including larger samples, more annotators, and the use of other metrics (such as informativeness).

The task of regionalism detection could be enhanced, whether by human annotation or by dictionary enrichment, based on human annotations. It also could be worthwhile to consider regional specificity and/or contextual information, to distinguish text containing challenging elements, such as wordplays, metaphors related to regional knowledge, idiomatic expressions, and instances of irony. Taking that into account, would lead to better identification of regional terms, and future work could be enhanced by exploring in depth different categories and the specific use of slang and colloquialisms tied to them.

Acknowledgements

To be disclosed in the final version of the paper.

References

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*, pages 15–21, Online. Association for Computational Linguistics.

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

Lukas Biewald. 2020. Experiment tracking with weights and biases. Software available from wandb.com.

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*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

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

Oscar Conde. 2013. Lunfardo rioplatense: delimitación, descripción y evolución. *De parces y troncos. Nuevos enfoques sobre los argots hispánicos*, pages 77–106.

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

Justina Diaz-Legaspe. 2020. What is a slur? *Philosophical Studies*, 177(5):1399–1422.

David M Eberhard, Gary Francis Simons, and Charles D Fenning. 2023. Ethnologue: Languages of the world.

Karmen Erjavec and Melita Poler Kovačič. 2012. "you don't understand, this is a new war!" analysis of hate speech in news web sites' comments. *Mass Communication and Society*, 15(6):899–920.

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.

Varun Godbole, George E. Dahl, Justin Gilmer, Christopher J. Shallue, and Zachary Nado. 2023. Deep learning tuning playbook. Version 1.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Asier Gutiérrez-Fandiño, Jordi Armengol-Estapé, Marc Pàmies, Joan Llop-Palao, Joaquin Silveira-Ocampo, Casimiro Pio Carrino, Carme Armentano-Oller, Carlos Rodriguez-Penagos, Aitor Gonzalez-Agirre, and Marta Villegas. 2022. MarIA: Spanish Language Models. *Procesamiento del Lenguaje Natural*, 68(0):39–60.

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.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of Experts.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

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

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15991–16111, Toronto, Canada. Association for Computational Linguistics.

Geoffrey Nunberg. 2018. The social life of slurs. *New work on speech acts*, pages 237–295.

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.

María Antonia Paz, Julio Montero-Díaz, and Alicia Moreno-Delgado. 2020. Hate speech: A systematized review. *Sage Open*, 10(4):2158244020973022.

Juan Manuel Pérez, Damián Ariel Furman, Laura Alonso Alemany, and Franco M Luque. 2022. Robertuito: a pre-trained language model for social media text in spanish. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7235–7243.

Flor Miriam Plaza-del arco, Debora Nozza, and Dirk Hovy. 2023. Respectful or toxic? using zero-shot learning with language models to detect hate speech. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 60–68, Toronto, Canada. Association for Computational Linguistics.

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.

Presidencia de la Nación. 2006. Hacia un plan nacional contra la discriminación “. *Boletín Oficial*, 30.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Pratik Sachdeva, Renata Barreto, Geoff Bacon, Alexander Sahn, Claudia von Vacano, and Chris Kennedy. 2022. The measuring hate speech corpus: Leveraging rasch measurement theory for data perspectivism. In *Proceedings of the 1st Workshop on Perspectivist Approaches to NLP @LREC2022*, pages 83–94, Marseille, France. European Language Resources Association.

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.

Alice Tontodimamma, Eugenia Nissi, Annalina Sarra, and Lara Fontanella. 2021. Thirty years of research into hate speech: topics of interest and their evolution. *Scientometrics*, 126:157–179.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D’souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model.

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.

Mesay Gemeda Yigezu, Olga Kolesnikova, Grigori Sidorov, and Alexander Gelbukh. 2023. Transformer-based hate speech detection for multi-class and multi-label classification.

Savvas Zannettou, Mai ElSherief, Elizabeth Belding, Shirin Nilizadeh, and Gianluca Stringhini. 2020. Measuring and characterizing hate speech on news websites. In *Proceedings of the 12th ACM conference on web science*, pages 125–134.

A Appendix

In this appendix, we describe details of the fine-tuning process of our BERT classifiers (Section A.1) and the original prompt and instruction provided to the LLMs (Section A.2).

A.1 Fine-tunning

In this subsection, we provide details on the fine-tuning process of the supervised models.

The classifiers were trained with Adam (**?**) as the optimizer and a triangular learning rate schedule.

Hyperparameter	Values
Epochs	3, 4, 5
Batch Size	32
Learning Rate	2e-5, 3e-5, 5e-5, 6e-5, 7e-5, 8e-5, 1e-4
Weight Decay	0.1
Warmup Ratio	0.06, 0.08, 0.10

Table 5: Hyperparameter search space considered for each model.

We optimized hyperparameters for six versions of BERT: BETO, RoBERTa, and RoBERTuito, each one in its original version and fine-tuned over the non-annotated data.[11]

To determine the best hyperparameters for each model, we performed a random search using the *wandb* library (Biewald, 2020). BETO achieved the best results.

Table 5 outlines the spectrum of values applied to each hyperparameter. For every model, task, and language, we conducted between 30 and 60 runs, choosing the optimal model based on the Macro F1 score from the validation set. We adopted a batch size of 32, tailored to accommodate our GPU memory limitations (either a GTX 1080Ti or Tesla T4, with memory ranging from 11 to 14GB).

A.2 Original prompt and examples

In this subsection, we present the original prompt and the provided examples for the few-shot scenario, both in Spanish. *Instrucción* stands for Instruction, *Ejemplos* for Example, *Contexto*, *Texto* and *Salida* for Context, Text, and Output.

Instrucción: "Determinar si el siguiente texto, correspondiente a un tweet y presentado con un contexto, contiene discurso de odio. Entendemos que hay discurso de odio si tiene declaraciones de carácter intenso y/o irracional de rechazo, enemistad o aborrecimiento contra un individuo o contra un grupo, siendo el individuo o el grupo objetivo de las expresiones por poseer una característica protegida. Las características protegidas que contemplamos son:

- mujer: hace referencia a mujeres o al movimiento feminista.
- lgbti: hace referencia a gays, lesbianas, transexuales y otras identidades de género.
- racismo: hace referencia a inmigrantes, xenofobia, o a pueblos aborígenes.
- clase: hace referencia a personas de bajos recursos o motivos de clase.

Los tweets están escritos en español rioplatense, y dentro del contexto cultural de Argentina. Responder una o varias de las características separadas por coma, o bien "nada" si no hay discurso de odio. Pensar y justificar paso a paso la respuesta antes de responder."

Ejemplos:

contexto: 'Wuhan celebra el fin de la cuarentena por el coronavirus con un mensaje para el resto del mundo: "Aprendan de nuestros errores"'

texto: "Chinos culiados de mierda ojalá se mueran todos"

salida: 'El texto insulta a los chinos ("culiados") y desea que se mueran, implicando que son los culpables del COVID-19. La respuesta final es "racismo".'

contexto:"Al borde del llanto, Nati Jota pidió disculpas por sus repudiables tuits pero los justificó: 'Quería sumar seguidores'"

texto: "Show o nada mi amorrr!! Sos una imbesil!"

salida: 'El texto contiene un insulto pero no menciona nada relacionado a las características protegidas, pese a mencionar a una mujer. La respuesta final es "nada".'

contexto:'Funcionario público y drag queen: la "doble vida" del político que representa a Salta en tacos, peluca y barba con brillos',

texto: "Que asco..!'",

salida:'El texto expresa desagrado hacia la identidad de género de la persona. La respuesta final es "lgbti"

contexto: "Violento motín en la cárcel de Devoto: tomaron e incendiaron pabellones",

texto: "Hermanito lo que son esos grones. Ahí adentro le robaron la billetera al coronavirus"

salida: 'El texto utiliza la palabra "grones", que significa negros (se obtiene "grones" a partir de "negros" cambiando el orden de las sílabas). Negro o grone es una manera despectiva de refererirse a alguien por su color de

[11] More details to be provided upon publication of the article.

piel. Además, da a entender que los presos son ladrones. La respuesta final es "racismo".'

context:'Mayra Mendoza criticó el banderazo contra el Gobierno: "No tiene sentido y es antidemocrático"'

text: "Que negrita villera....ni con la plata q roba puede arreglarse"

output: 'El texto alude al color de piel de Mayra Mendoza ("negrita") y a su condición social ("villera") de manera discriminatoria. La respuesta final es "racismo, clase".'

context: "No da tregua: sin ropa y desafiante Flor Peña volvió a mostrar su lado más hot"

text: "K lado hot tiene más varices k mi abuela. Tiene mas puesta de espalda k karagian trola"

output: 'El texto alude al aspecto de Flor Peña (sus várices) y la desprecia por "trola", un insulto que significa "prostituta". La respuesta final es "mujer".'

context: "Otro ataque de furia contra un colectivero: una mujer trans le pateó la unidad y le tiró un ladrillazo"

text: "Un tipo operado. Con la fuerza de un hombre y no la de una mujer"

output: "salida": 'El texto asegura que la mujer trans es "un tipo operado", implicando que en los hechos es un hombre y negando su identidad de género. La respuesta final es "lgbti"'

context:'Elisa Carrió denunció que el Gobierno usa la pandemia para "establecer un estado de sitio"'

text: "Gorda psiquiátrica"

output: 'El texto habla de su aspecto (gorda) y la acusa de tener problemas psiquiátricos. Sin embargo, no hace ninguna alusión a una característica protegida. La respuesta final es "nada".

context: 'Loly Antoniale mostró su impresionante casa en Miami: "Soy la reina de mi castillo"'

text: "No pudo enganchar al viejo famoso..se busco otro..y este le puso su castillo en Miami..."

output: 'El texto alude a la vida amorosa de Loly Antoniale, sugiriendo que busca hombres por interés económico. Esta apreciación es sexista y despectiva. La respuesta final es "mujer".'

context: "Les darán DNI provisorio a personas en situación vulnerable, para que puedan empezar a cobrar planes sociales"

text: "Seguimos alimentando vagos"

output: 'El texto se refiere a quienes cobran planes sociales como vagos a quienes hay que alimentar. Esto muestra un desprecio hacia las clases bajas que requieren de la asistencia estatal. La respuesta final es "clase".'

context: "Los dos presos heridos de bala en el motín de Devoto tienen Covid-19 y uno quedó hemipléjico"

text: "justicia divina!"

output: "salida": "El comentario contiene una celebración frente a un ataque a presos. A pesar de ser sumamente ofensivo, no hace ninguna alusión a una característica protegida. La respuesta final es 'nada'."