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

Anonymous ACL submission

Abstract

Hate speech detection deals with many language variants, slang, slurs, 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 found on digital platforms, media and networks (Berecz and Devinat, 2017; Woods and Hahner, 2019; Hodge and Hallgrímsdóttir, 2021). While discriminatory discourses emerge in different enunciation areas and modalities, we are particularly interested in the way this kind of problematic speech is spreed along with certain themes in the public arena and the circulation of news (Assimakopoulos et al., 2017). The news published on social networks encourages the debate or discussion of public issues and problems among users. And many times the tweets displaying hate speech are reactions to that news tweets, meaning the context of those messages is extremely important (Cotik et al., 2020). Social media, such 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, for example topics triggered by news that are subject of discussion (Zannettou et al., 2020; Erjavec and Kovačič, 2012).

Hate speech detection, from the Natural Language Processing (NLP) perspective, has to deal with languages crossed by variants, slang, slurs, and other specific modalities found (Nunberg, 2018; Diaz-Legaspe, 2020). Being aware of cultural nuances and specific contexts of use is crucial to address this task and shows the relevance of developing dialectal corpora and analysis that allows to automatically detect specific hateful expressions in different lexical contexts.

Regarding this issue, a particular interest arises in the performance of large language models (LLMs) by analyzing hate speech in relation to 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). Given that GPT-3.5 (also known as ChatGPT) is one of the most popular and emerging large language models (Wu et al., 2023; Deng and Lin, 2022), we have the question of how well it can detect hateful messages in a specific dialectal variant of Spanish, focusing on the *Rioplatense* variety.

With over 450 million native speakers, primarily in Latin America, Spain, and also parts of the US (Eberhard et al., 2015), Spanish includes many varieties and dialects. Each one 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 (Lipski, 2012; Coloma, 2018).

One of the most interesting features of *Rioplatense* Spanish is that includes ruralisms, indigenisms, argot and slang, especially *lunfardo*, among other vocabularies. 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 in Buenos Aires, especially during the 19th century and the first half of the 20th century (de Letras, 2003; Conde, 2013). *Lunfardo* shows as 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,[1] such as the term "*laburar*", from Italian "*lavorare*", standing as the verb "to work" or "*trabajar*" in Spanish. Many words have other meanings, for example "*trola*" means a lie o deceit in Iberian Spanish, but in *Rioplatense* Spanish it refers to "*whore*", and it also has a masculine form "*trolo*" that means "*faggot*". These terms are commonly used to derogatorily address both women and gays. These examples point out that the understanding of many expressions constitutes a challenge even to Spanish native speakers unfamiliar with the *Rioplatense* variant. Would that be the case of LLMs?

Following that question, this paper develops an exploratory approach to the effectiveness of LLMs addressing the task of detecting specific texts and tagging corpora linked to hate speech. We take as a benchmark a fine-tuned BERT classifier trained with a corpus of news articles and Twitter's comments in *Rioplatense* Spanish, annotated to detect hate speech (Pérez et al., 2023). 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.

The results of this work, focused in *Rioplatense* Spanish, contribute to understanding possible forms of bias and to evaluate the specificity of *Rioplatense* variant that is not part of the typical corpus used for training LLMs. This highlights its unique characteristics and the importance of developing and working with dialect-specific corpora for the automated detection of hate speech. Focusing exclusively on *Rioplatense* Spanish is a first step in addressing cultural nuances and specific local expressions inherent in this dialect, particularly in relation to the performance of LLMs. Future work could be pursued on this basis, aiming to compare Rioplatense Spanish with other variants, such as Mexican or Iberian Spanish.

2 Related work

There is a plethora of resources for automatic detection of hate speech in English. Nevertheless, when it comes to Spanish, despite being one of the main languages in terms of the number of native speakers worldwide (Eberhard et al., 2024), corpora are scarce, with only a few datasets publicly available. Some of the main references of published and available resources in Spanish are found in datasets like *IberEval*, which presents a ~4k Twitter dataset for the Automatic Misogyny Identification (AMI) (Fersini et al., 2018); those released by the MEX-A3T task, that included a dataset of ĩ1k Mexican Spanish tweets annotated for aggressiveness (Álvarez-Carmona et al., 2018); and the dataset launched in the context of the HatEval challenge, a ~6.6k tweets dataset annotated for misogyny and xenophobia (Basile et al., 2019). To the best of our knowledge, only one dataset annotated for hate speech detection is available in *Rioplatense* (Pérez et al., 2023). This highlights the positive impact of contributions tending to build resources in Spanish and its variants.

Regarding the automatic detection and treatment of hate speech, a broad amount of literature has been published. 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.

Upon the blazing development 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

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

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 in 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

Our experiments are based on an anonymized dataset available in *Rioplatense* Spanish, consisting of Twitter replies to posts from Argentinean news outlets; with models, partitions and class distributions arealso available (Pérez et al., 2023). 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 eight possible types: politics, misogyny, homophobia/transphobia, racism/xenophobia, class hatred, appearance, hate against criminals and disabled people. All annotated instances have a context (the tweet posted by the news outlet, plus the news title and whole text content of the news) and the text being analyzed and annotated for hate speech (each Twitter user's comment). This is important as contextual information situates the comment and that it has been shown to be relevant to detect hate speech (Sheth et al., 2022; Xenos et al., 2021). Along with the annotated dataset, another unannotated dataset suitable for unsupervised training is provided. The unsupervised corpus used to continue pre-training includes around 5,000,000 *Rioplatense* tweets comments, with a contextual information of 288,000 news tweets and full articles.

In this work, we address the presence of hate speech linked to four possible types: misogyny, homophobia/transphobia, racism/xenophobia, and class hatred/classism, according to the attacked characteristics, from now on dubbed WOMEN, LGBTI, RACISM, and CLASS (see Appendix A.1 for a broader definition). The selection of those categories was made based on the prevalence of each category among hateful speech and their societal impact. Also, 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 used the same train and development samples as Pérez et al. (2023) ($36,420$ and 9120 examples respectively), but, for budget reasons, we subsampled the test set to 5670 examples (50% of the original test set). Among these, 479 comments contain hate speech, distributed as shown in Table 1. Some comments express attacks to more than one category. In those cases, we find 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 at the same time. Table 2 shows some examples of the dataset.

In order to assess whether the large language models were able to capture the meaning of regional terms and expressions or not, we marked the test dataset for the presence of regionalisms. We referred to regionalisms as idiomatic phrases, words exclusively used in *Rioplatense* Spanish, or with a meaning that is differentially used in Argentina or only in *Rioplatense*-speaking countries (eg. "pelotudo" for "asshole"). As aforementioned, there is a vacancy area regarding *Rioplatense*: resources are scarce and, when available, they are not the best to work with. Therefore, we generated a very basic and not exhaustive list of regional terms, scraped from an crowd-sourced dictionary of regionalisms in *Rioplatense* Spanish.[2]

[2] https://www.diccionarioargentino.com/

Category	Number	Percentage
RACISM	230	4.06 %
WOMEN	131	2.31 %
LGBTI	88	1.55 %
CLASS	76	1.34 %
TOTAL	479	8.45 %

Table 1: Number and percentage of messages containing hate speech in each category.

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 2: Hateful examples from the analyzed dataset.

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 two pre-trained models in Spanish: *BETO* (Cañete et al., 2020), 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 the unsupervised corpus provided along with the used dataset (Pérez et al., 2023), as it has been shown to improve the performance in domain-specific tasks (Gururangan et al., 2020).

The fine-tuning process of the supervised models followed standard practices in fine-tuning BERT-like models. The classifiers were trained with Adam (Kingma and Ba, 2014) as the optimizer with weight decay, and a triangular learning rate schedule.

Table 3 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). The best hyperparameters found in the tuning process were roughly the same for all the models: 0.1 for warm-up ratio, 3 and 4 (only RoBERTuito on its non-finetuned version) the number of training epochs, and finally $6e^{-5}$ for the learning rate.

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 3: Hyperparameter search space considered for each model.

4.1 Prompting strategies

Regarding the large language models, we selected three models that show good performance in Spanish to run the experiments and prompts:

- *GPT-3.5* turbo-0125 (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.

To build the prompt, we resorted to few-shot or in-context learning. Early experiments using zero-shot, one-shot settings and few-shot learning were conducted. Then, different prompts were evaluated in the development split until no improvement in performance was observed.

That prompt engineering process led to the following prompt:[3]

> 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:
>
> - 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".

[3]Originally in *Rioplatense* Spanish (see in Appendix), translated to English for the purpose of this paper.

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

Table 4: Classification results for LLMs and fine-tuned BETO, expressed as macro averages of F1, Precision and Recall for all the considered labels.

The output consisted of a natural language explanation.

The 12 examples considered the different characteristics and target labels and were balanced by their labels. They also included "nothing" examples, that pointed out cases where there was no hate speech towards the targeted categories (women, LGBTI, racism or class). The selection of examples provided in the prompt had the following distribution of categories: 2 examples for racism, 2 for LGBTI, 2 for women and 2 for "nothing"; 1 example for classism; and 1 multi-class example for racism and class. The full list of examples and the original prompt in Spanish can be found in Appendix A.2.

4.2 Evaluation

To evaluate the performance of the classifiers, we assessed the precision, recall, and F1-score for each class, in a multi-label classification schema. 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 4 shows the results for the multi-label classification task, represented as the macro averages of F1, Precision and Recall. We report only the results of the best BERT classifier, the fine-tuned version of BETO.

Table 5 shows the F1 scores for all the considered BERT classifiers. As mentioned in Section 4, we considered BETO (Cañete et al., 2020) and RoBERTuito (Pérez et al., 2022), and also their fine-tuned versions in the unsupervised corpus provided by the considered dataset (Pérez et al., 2023). We can observe that, although RoBERTuito performs

	WOMEN	RACISM	LGBTI	CLASS	Macro
BETO	0.366	0.698	0.414	0.473	0.570
RoBERTuito	0.414	0.675	0.435	0.451	0.582
BETO (FT)	0.422	0.736	0.468	0.511	0.614
RoBERTuito (FT)	0.405	0.694	0.474	0.471	0.598

Table 5: F1 scores for the considered BERT classifiers. FT stands for fine-tuned, marking those pre-trained models that were further fine-tuned in the unsupervised corpus.

Model	BETO		GPT-3.5		Mixtral	
Regionalism	reg.	wo. reg.	reg.	wo. reg.	reg.	wo. reg.
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 6: F1 by category for messages containing regionalism (reg.) and those not containing it (wo. reg.).

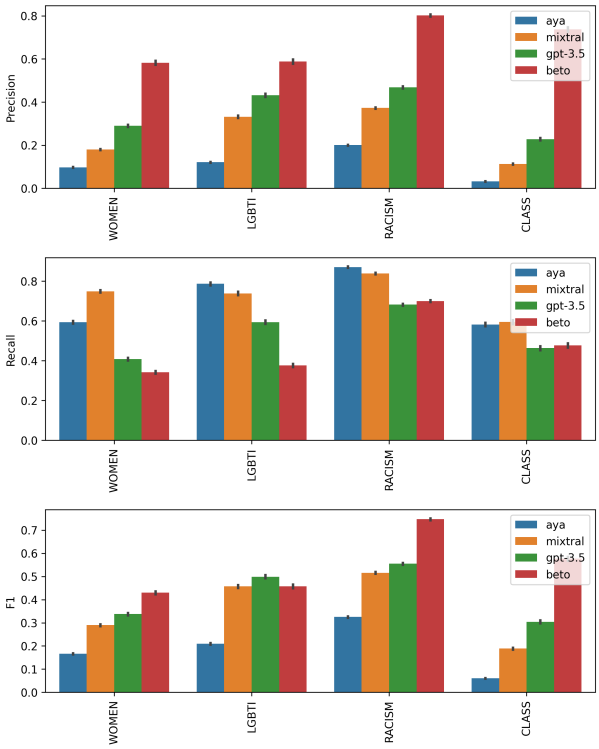

Figure 1: Precision, recall and F1 of the classifiers: ChatGPT 3.5, Aya, Mixtral and the fine-tuned *BETO* classifier.

better in the non-finetuned version, BETO achieves the best results after the pre-training process.

While the BETO classifier outperforms their counterparts in terms of precision and F1, the LLMs have higher recall. As Aya model performed poorly, 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 complex language, irony and metaphors and where *BERT*-based classifiers show a significant gap compared to humans (Yigezu et al., 2023; Pérez et al., 2023).

Focusing in those messages that contain regionalisms, no conclusive differences are detected. As seen in 6, 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 might be 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*) leverage for this detection task. Indeed, hate speech messages represented 9% of the original dataset, and this rises to a 16% of hateful messages within those that have regionalisms.

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. These slur, regional expressions may make the task of flagging hateful speech easier compared to comments lacking it.

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 shows some examples of those cases.[5]

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 mod-

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

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[4] 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 7: Examples of LLMs results

eled to address that specific dialect. Some of the examples in Table 7 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.[6] 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.

6 Conclusions

The culture and communication of Latin America is diverse and full of different expressions, idioms, slang, specific uses, and adaptations of, among others, the Spanish language, which embodies subtle differences that are often context-dependent and cannot be captured outside of their context of use. We believe this work is a first step toward filling a gap in resources regarding *Rioplatense*, a particular variety of Spanish, and aims to contribute to the

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

understanding of how large language models perform in these specific, "under-resourced" language varieties and their cultural contexts, which is particularly important when addressing the phenomenon of hate speech.

This study shows the effectiveness of *ChatGPT* and *Mixtral* classifiers in the task of hate speech detection, specifically in the context of cultural and linguistic complexities present in *Rioplatense* Spanish tweets. The findings highlight that while *Mixtral* excels in recall across various characteristics, it does so at the expense of precision. Conversely, *ChatGPT* offers a more balanced performance, particularly in detecting the LGBTI characteristic, where it outperforms traditional fine-tuned classifiers, such as BETO. Comparing LLMs 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. These results underline the potential for LLMs to capture cultural nuances and regional expressions effectively and to interpret culturally sensitive language rich with ironies and metaphors. The results of *ChatGPT* to identify hate speech laden with regionalisms show ability in processing colloquial language. Notably, the prevalence of regional slurs in classist and misogynistic discourse appears to indicate the importance of incorporating regional context to improve hate speech detection. However, the observed higher rate of false positives in LLM classifications relative to BETO emphasizes the need for careful evaluation metrics when assessing the efficacy of these models.

While LLMs have proven to be a powerful tool for hate speech detection, supervised classifiers still outperform them in the general case and are more suitable for detecting hate speech at large scale. 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 insights from this study emphasize the importance of building models that are sensitive to cultural and linguistic diversity, while also highlighting the value of producing corpora on specific topics and linguistic variants that serve as benchmarks. In that sense, this paper contributes to variant-focused research in NLP, being in this case *Rioplatense* Spanish, underlining the importance of a comprehensive understanding of variants, its cultural nuances and language uses as a vital step in addressing hate speech in a linguistically diverse context.

7 Limitations

One of the main challenges faced in this work is related to the task itself: hate speech detection, which tries to capture a complex social phenomenon. Additionally, it is important to note that the original dataset lacked natural language explanations for the annotations.

Also, the task of regionalism detection could be enhanced, whether by human annotation or by dictionary enrichment, based on human annotations.

Finally, the analysis of LLMs explanations was performed in a very limited way, being their soundness assessed solely by the authors. 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).

8 Future work

As a future work, it would be interesting to develop a multi-variant of Spanish evaluation, as this kind of analytic effort and experiments were out of the scope of this work at this stage. In particular, 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 major variant of spoken Spanish.

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 of hate speech and the specific use of slang and colloquialisms tied to them. Along with this, future work could focus on improving the prompting engineering to have a better handling of dialectal variants.

Acknowledgements

References

Miguel Ángel Álvarez-Carmona, Estefanía Guzmán-Falcón, Manuel Montes-y Gómez, Hugo Jair Escalante, Luis Villasenor Pineda, Verónica Reyes-Meza, and Antonio Rico Sulayes. 2018. Overview

of mex-a3t at ibereval 2018: Authorship and aggressiveness analysis in mexican spanish tweets. In *IberEval@ SEPLN*, pages 74–96.

Stavros Assimakopoulos, Fabienne H Baider, and Sharon Millar. 2017. *Online hate speech in the European Union: a discourse-analytic perspective*. Springer Nature.

Valerio Basile, Cristina Bosco, Elisabetta Fersini, Debora Nozza, Viviana Patti, Francisco Manuel Rangel Pardo, Paolo Rosso, and Manuela Sanguinetti. 2019. Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In *Proceedings of the 13th international workshop on semantic evaluation*, pages 54–63.

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.

Viviana Cotik, Natalia Debandi, Franco M. Luque, Paula Miguel, Agustín Moro, Juan Manuel Pérez, Pablo Serrati, Joaquin Zajac, and Demián Zayat. 2020. A study of hate speech in social media during the COVID-19 outbreak.

Academia Argentina de Letras. 2003. Diccionario del habla de los argentinos.

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 F. Simons, and Charles D. Fennig. 2024. *Ethnologue: Languages of the World*. "Dallas, Texas: SIL International.".

David M Eberhard, Gary Francis Simons, and Charles D Fenning. 2015. 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.

Elisabetta Fersini, Paolo Rosso, Maria Anzovino, et al. 2018. Overview of the task on automatic misogyny identification at ibereval 2018. *Ibereval@ sepln*, 2150:214–228.

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.

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.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

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.

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.

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.

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 F Gelbukh. 2023. Transformer-based hate speech detection for multi-class and multi-label classification. In *IberLEF@ SEPLN*.

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 present the definition of each targeted class of hate speech and we describe details of the original prompt and instruction provided to the LLMs

A.1 Classes Definition

Definition of each class addressed in hate speech detection according to the *Plan Nacional contra la Discriminación* (Argentinean National Plan Against Discrimination) guidelines (Presidencia de la Nación, 2006).

A.1.1 RACISM

Racism can be defined as a set of beliefs, attitudes and practices that discriminate against individuals or groups based on their race, ethnicity, or physical characteristics, denying them equal rights and dignity. This discrimination can be both individual and institutional and can occur in various ways: Individual racism; Institutional racism; Structural racism; Cultural racism; Xenophobia (rejection or fear of people from other countries or cultures); and also includes Antisemitism, Arabophobia, Islamophobia and Afrophobia.

A.1.2 WOMEN

Discrimination against women is a social phenomenon that manifests itself as a set of actions and beliefs that devalue women compared to men, through attitudes, practices and structures that perpetuate gender inequality and limit women's access to resources, rights and opportunities in various areas of life. Discrimination against women can be expressed in different areas, such as Economic Inequality; Gender Violence; Access to Education; Political Representation; Cultural Norms and Stereotypes; among others. Added to this is intersectional discrimination (women who belong to minority groups may experience additional forms of discrimination, which further complicates their situation).

A.1.3 LGBTI

Gender-based discrimination refers to attitudes, behaviors, and policies that marginalize and devalue people based on their sexual orientation or gender identity. This form of discrimination manifests itself in various dimensions and contexts, affecting the daily lives and fundamental rights of LGBTQ+ people. Gender-based discrimination can be expressed as violence and harassment; legal inequality; social stigmatization; unequal access to services; non-positive media representation; among others. Likewise, gender-based discrimination can intersect with other forms of discrimination, such as racism or poverty, exacerbating the difficulties faced by individuals belonging to multiple marginalized groups.

A.1.4 CLASS

Class-based discrimination refers to attitudes, practices, and social structures that marginalize people based on their socioeconomic status/poverty. This form of discrimination manifests itself in various

dimensions, affecting access to resources, rights, and opportunities, and perpetuating social exclusion. Expressing itself, for example, as Limited Access to Resources; Social Stigmatization; Labor Inequality; Violence and Crime; Lack of Political Representation, among others. Furthermore, class-based discrimination can intersect with other forms of discrimination, such as racism or sexism, exacerbating the difficulties faced by individuals belonging to multiple marginalized groups.

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 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'."