Procesamiento del Lenguaje Natural, Revista nº 71, septiembre de 2023, pp. 227-245 recibido 01-04-2023 revisado 10-05-2023 aceptado 11-05-2023

# Automatic counter-narrative generation for hate speech in Spanish

## *Generación automática de contranarrativas para discursos de odio en español*

**M. Estrella Vallecillo-Rodríguez, Arturo Montejo Ráez, M. Teresa Martín-Valdivia**

Department of Computer Science, Advanced Studies Center in ICT (CEATIC)
Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, Spain
{mevallec, amontejo, maite}@ujaen.es

**Abstract:** This paper analyzes the use of language models to automatically generate counter-narratives for hate speech in Spanish. Despite the existence of a few studies in English and other languages, no previous work has explored this topic focused on Spanish. The article shows that the use of GPT-3 outperforms other models in generating non-offensive and informative counter-narratives, which sometimes present compelling arguments. We have used few-shot learning algorithms applying different prompt strategies and analyzing the results for each of them. Additionally, a new corpus called CONAN-SP, which consists of 238 pairs of hate speech and counter-narratives in Spanish, has been made available to the research community to facilitate further investigations in this area. These findings highlight the potential of language models to combat hate speech in Spanish by counter-narrative generation.
**Keywords:** Spanish counter-narrative generation, Hate speech, Natural language Generation, Few-shot learning.

**Resumen:** Este trabajo analiza el uso de modelos lingüísticos para generar automáticamente contranarrativas al discurso del odio en español. A pesar de la existencia de algunos estudios en inglés y otros idiomas, ningún trabajo previo ha explorado este tema centrado en el español. El artículo muestra que el uso de GPT-3 supera a otros modelos en la generación de contranarrativas no ofensivas e informativas incluyendo en ocasiones argumentos convincentes. Hemos utilizado diferentes algoritmos de *few-shot learning* aplicando varias estrategias de *prompting* y analizando los resultados para cada una de ellas. Además, se ha puesto a disposición de la comunidad investigadora un nuevo corpus llamado CONAN-SP, que consta de 238 pares de discursos de odio y contranarrativas en español, para facilitar nuevas investigaciones en este ámbito. Estos resultados ponen de relieve el potencial de los modelos del lenguaje para combatir el discurso de odio en español mediante la generación de contranarrativas.
**Palabras clave:** Generación de contranarrativas en español, Discurso del odio, Generación de lenguaje natural, Aprendizaje con pocos ejemplos.

## *1 Introduction*

Hate Speech (HS)[1] refers to any form of communication that promotes hostility, discrimination, or violence towards a group or individual based on their race, gender, sexual orientation, religion, or any other characteristic. This is a problem that has been enhanced by the massive use of social networks that allow the easy and rapid propagation of hate messages (Mathew et al., 2019). For this reason, it is important to implement computational systems capable of detecting and combating online hate speech. There are already some approaches based on Machine Learning (ML) and Natural Language Processing

[1] *Warning: This paper discusses and contains content that may be deemed offensive or upsetting.*

ISSN 1135-5948 DOI 10.26342/2023-71-18 

(NLP) that try to identify linguistic patterns to detect both hate speech in general (Djuric et al., 2015) and messages targeting certain groups like sexism (Frenda et al., 2019) or xenophobia (Plaza-Del-Arco et al., 2020). It is important to note that for such a system to work properly, it is necessary to adapt it to the language in which the messages to be detected are written.

On the other hand, in order to protect people and promote a civilized and respectful online dialogue, it is not only important to detect hate speech, but it is also increasingly necessary to use strategies that allow us to mitigate its consequences.

Some strategies employ blocking or suspending tools to counter online hate speech on social media. However, despite the efforts, social media sites that have implemented strict policies against hate speech have not been particularly successful. Some argue that selectively blocking or suspending free speech sets a dangerous precedent, and therefore it should not be done. While blocking hateful speech may decrease its impact on society, it also risks infringing upon free speech. Richards and Calvert (2000) argue that a preferable solution to hate speech is to promote more speech using for example Counter-Narrative (CN). Counter-narrative is the strategy to combat hate speech that consists of providing an alternative narrative to the one that promotes hate and violence (Hangartner et al., 2021). CN can include messages that promote tolerance, respect and inclusion, and can be a powerful tool to challenge and dismantle hate speech.

In this sense, using NLP algorithms to automatically generate counter-narratives may be an option to explore in the fight against hate speech. However, there are still not many approaches to apply NLP tools in CN generation and, in addition, most of the work focuses on combating messages written in English. Thus, the main objective of our paper is to study the use of NLP techniques to generate counter-narratives in Spanish to combat hate speech online. Our research is focused on generating Spanish texts through an approach based on one of the most referenced works on the automatic generation of counter-narratives, which is described by Chung, Tekiroglu, and Guerini (2021). The authors of this work used the CONAN corpus (Chung et al., 2019) to develop a variation called CONAN-KN, which includes 195 pairs of hate speech and counter-narratives along with the background knowledge used to generate the counter-narrative. The main difference between CONAN and CONAN-KN is the inclusion of background information in the latter. Although our experiments are based on this corpus, we only focused on the hate speech and counter-narrative pairs, leaving the incorporation of background knowledge for future research.

We propose an approach for automatically generating counter-narratives to hate speech in Spanish, using a few-shot learning strategy. Since datasets for this task are scarce and expensive to create, we translated the CONAN-KN dataset into Spanish and manually verified its accuracy. We use a subset of this corpus as examples in a prompt and evaluate various language generation models to determine the most accurate one. Our contributions can be summarized as follows:

1. To study the automatic generation of CN for hate-speech in Spanish.
2. To compare different models for Spanish CN generation.
3. To generate a new Spanish corpus with pairs of HS and CN using a prompting approach (CONAN-SP).
4. To evaluate different prompt strategies for automatic CN generation.

The paper is organized as follows: first, we provide a Background section that outlines the hate speech and generation of CN problem and previous research in the area. Next, we describe our proposal including the description of the dataset and preprocessing, the selection of algorithms and models, and the evaluation methodology. Then, we present the details of our experiments and the results obtained, continuing with an Error analysis section that discusses potential limitations and sources of error in our approach, as well as ways to address them. Finally, we conclude the paper with a concise summary of our findings, a discussion of their implications and possible future directions for research.

## 2 Background

Over the past few years, the NLP community has dedicated significant research efforts

to studying the issue of hate speech (Fortuna and Nunes, 2018; Plaza-del Arco et al., 2021; Fortuna et al., 2022). Despite this ongoing research, the amount of research on computational systems designed to combat the dissemination and propagation of HS messages is very limited. Nevertheless, mitigating the consequences of HS is crucial not only through its detection but also via effective strategies to combat it. Counter-narrative is one such strategy to combat hate speech that consists of providing an alternative narrative to the one that promotes hate and violence. Counter-narrative can include messages that promote tolerance, respect, and inclusion, and can be a powerful tool to challenge and dismantle hate speech by mitigating its negative impacts. Thus, the use of NLP technologies to automatically generate counter-narratives is an option to explore in the fight against HS and promote the creation of a safer and more inclusive online environment for everyone.

Actually, some works have proved that counter-narrative are effective in hate countering (Benesch, 2014; Mathew et al., 2019). This is why some Non-Governmental Organizations (NGOs) train human operators to intervene in online conversations by writing counter-narratives. However, manual human action against hate speech is not a scalable solution. Consequently, data-driven Natural Language Generation (NLG) methods are now being explored to aid NGO operators in creating CNs.

In this line, a range of CN collection strategies have been put forward, each with its own set of benefits and drawbacks (Mathew et al., 2018; Qian et al., 2019). Chung et al. (2019) created the multilingual CONAN corpus of HS and CN pairs for three different languages: English, French, and Italian. CONAN is a high-quality dataset manually generated by NGO human operators trained to combat online hate speech and who can be considered experts in counter-narrative generation. Chung, Tekiroglu, and Guerini (2021) used this dataset to fine-tune GPT-2 to automatically generate counter-narratives. They also adopted the same methodology of data collection on hate speech targeting other religions, races, and gender to fine-tune GPT-2 for automated generation (Fanton et al., 2021).

Our work relies on a Spanish adaptation of the CONAN-KN corpus, which is widely recognized as a high-quality dataset. To adapt it to Spanish, we used DeepL as an automatic translation tool and ensured that the quality of the HS-CN pairs was maintained. The strategy used to ensure the quality of HS-CN pairs consists of a manual revision of some texts and comparing them with the original texts. We use this corpus as a basis for generating various prompts using selected LNG models. To generate Spanish CN, we employed Pre-trained Language Models (LMs) known for achieving impressive results on challenging generation tasks after fine-tuning. In this line, some works have investigated the performance of some architectures generating CN such as GPT-2 or XNLG (Tekiroglu, Chung, and Guerini, 2020). However, the popularity and the impressive results that GPT-3 is showing have led us to include this model in our experimentation. As we can see in Section 4.2, the results obtained with GPT-3 are by far the best and truly simulate a human operator. For this reason, we have focused on GPT-3 and we have applied different prompt strategies in the experiments (Brown et al., 2020).

After performing the experiments with GPT-3 we have created a corpus including all generated HS-CN pairs which we call CONAN-SP (CONAN-SPanish).

The main problem we faced was to undertake the evaluation of the results. For this purpose, we have chosen to carry out a manual evaluation following the same methodology as Ashida and Komachi (2022), in which three perspectives are taken into account: offensiveness, stance and informativeness. More details are given in the next section.

## *3 Data and Methodology*

### 3.1 Dataset

To run our experiments, we use a Knowledge grounded hate-countering dataset (CONAN-KN). This dataset contains 195 pairs of hate speech texts and the respective counter-narrative of the offensive texts, both written in English. Figure 1 shows an example of a CONAN-KN HS-CN pair translated to Spanish. The offensive texts are labeled with five different classes that represent the target of hate speech: Antisemitism, Islamophobia, Racism, Homophobia, and Misogyny. In Table 1 we can see the number of instances of

each class in the CONAN-KN dataset.

| Class | # Instances |
|---|---|
| Antisemitism | 50 |
| Islamophobia | 51 |
| Racism | 29 |
| Homophobia | 39 |
| Misogyny | 26 |
| **Total** | **195** |

Table 1: HS-CN pairs class distribution in the CONAN-KN corpus.

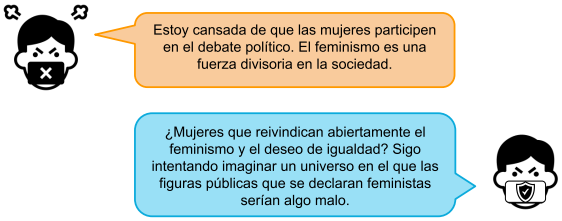


Figure 1: Example of Spanish translation of hate speech and counter-narrative pair of CONAN-KN dataset. The original text in English is **HS**: *I'm tired of woman taking part to political debate. Feminism is a divisive force in society.* **CN**: *Women openly claiming feminism and a desire for equality? I keep trying to imagine a universe in which public figures declaring themselves feminists would be a bad thing.*

With the results of the proposed experiments, we create a new dataset for Spanish counter-narrative, called CONAN-SP. As it is explained in Section 4.3, GPT-3 was the only selected model that obtained coherent results, therefore in CONAN-SP, we only include generated texts from GPT-3 model in the different experiments. All of the counter-narrative generated by GPT-3 model are included in CONAN-SP without filters. To construct CONAN-SP, we remove the pairs that contain duplicates of hate-speech texts and the examples used to calculate the agreement between annotators. The structure of CONAN-SP is the hate-speech provided by CONAN-KN and the counter-narrative texts generated by GPT-3 model. Furthermore, we associated the target of the offensive comment with the hate speech and counter-narrative pair. Finally, we obtained 238 pairs of hate-speech and counter-narrative. All of these pairs are labeled by human annotators with the same strategy used in the proposed experiments (Section 3.5). CONAN-SP is freely available for research purposes [2].

### 3.2 Data Preprocesing

Since our work focuses on Spanish, we have automatically translated the CONAN-KN corpus from English to Spanish using the DeepL neural machine translation service.

CONAN-KN dataset contains pairs of Hate Speech and Counter Narrative (HS-CN). However, some HS parts are repeated for different CNs. Therefore, to perform our experiments, we have removed the repeated hate speech texts and selected one of the existing CNs, specifically the first HS-CN pair. Thus, 105 Spanish HS-CN pairs were finally considered.

### 3.3 Selected Models

To achieve our goal we look for different models that researchers usually use for text generation tasks. In this case, we selected GPT-2[3] (Radford et al., 2019), GPT-2 MarIA[4] (Fandiño et al., 2022), davinci GPT-3[5], Flan-T5[6] (Chung et al., 2022) and Bloom [7] (Scao et al., 2022). The first three models are based on the same architecture, Generative Pre-trained Transformers (GPT), with the difference that the second model is fine-tuned with data from the National Library of Spain, and the third model is pre-trained with more data than the GPT-2 model and ten times bigger in its number of parameters. The Flan-T5 model is based on the Transformer architecture but it is trained on more than 1,000 additional tasks in different languages. Finally, the Bloom model is an auto-regressive Large Language Model, trained to continue text from a prompt on a big amount of text data. This model is able to output coherent text in 46 languages and 13 programming languages that are hardly distinguishable from text written by humans.

[2] https://github.com/estrellaVallecillo/CONAN-SP.git
[3] https://huggingface.co/cite-2-large
[4] https://huggingface.co/PlanTL-GOB-ES/GPT-2-base-bne
[5] https://platform.openai.com/docs/models/gpt-3
[6] https://huggingface.co/google/flan-t5-large
[7] https://huggingface.co/bigscience/bloom

### 3.4 Prompt strategy

For the counter-narrative generation, we apply a few-shot learning strategy, where we pass a prompt that contains some examples of a good counter-narrative for the hate speech text. Later, the models have to complete the counter-narrative of the remaining offensive texts. Figure 2 presented the architecture followed for our experiments with few-shot learning where one of the most important part of the development system is the prompt (orange part of the figure) because it has the training data of the task (examples) and the input test data for which the model has to predict the output.

The different proposed prompt strategies are shown in Section 4.2

### 3.5 Evaluation Methodology

Evaluating the quality of the generated text is a challenging task for several reasons. Firstly, there is no universal metric or standard for what constitutes good text. Text generation is often subjective and dependent on the task or context in which it is used. Additionally, evaluating the coherence and fluency of a piece of generated text requires an understanding of language and the ability to recognize nuances in meaning and tone. Furthermore, some of the metrics that have been used in generation come from the field of translation, such as BLEU (Papineni et al., 2002) or NIST (Doddington, 2002), and others, such as ROUGE (Cawsey, Jones, and Pearson, 2000), from the generation of summaries. BLEU is a precision metric used in translation that evaluates the proportion of n-grams that share the output of the system with seeing translations. NIST is an adaptation of BLEU that adds weight to the most informative n-grams. At the summary level, the ROUGE metric is used, operating in parallel to BLEU and various metrics such as ROUGE-1, ROUGE-2, or ROUGE-SU4 (Lin, 2004). However, these measures are disputed in the community for various factors such as that in NLG systems there is no single-unique good output to compare with or the results given by the metrics are difficult to interpret. These are some of the reasons given for distrusting this type of evaluation.

For these reasons, and given that the generated corpus CONAN-SP includes a limited number of HS-CN pairs, we have carried out a manual evaluation in order to test the validity of using generative language models to automatically generate counter-narratives for hate speech in Spanish.

For the evaluation we have followed the work by Ashida and Komachi (2022) and considered 3 perspectives for each counter-narrative: Offensiveness, Stance, and Informativeness:

- **Offensiveness**: if the counter-narrative is offensive to someone, such as people of a certain race, including the people who wrote the HS message.
  - 0 (not sure)
  - 1 (not offensive)
  - 2 (maybe offensive)
  - 3 (completely offensive)
- **Stance**: refers to the position taken with respect to the message and is classified into three types: agree, neutral, and disagree.
  - 0 (irrelevant)
  - 1 (strongly agree)
  - 2 (weakly agree/disagree),
  - 3 (strongly disagree)
- **Informativeness**: assesses the degree of informativeness and specificity of the counter-narrative, without being generic.
  - 0 (irrelevant)
  - 1 (not informative)
  - 2 (generic statement and little information)
  - 3 (specific and informative)

The evaluation is carried out by 3 human annotators. The group of annotators consists of a senior linguist, a senior computer scientist, and a pre-doctoral student in computer science. First, they annotate 5 HS-CN pairs with all texts that are generated with all tested models. Once it is proven that the best-performing model is GPT-3, the evaluation focuses only on the texts generated by this model for each experiment. The 3 annotators evaluated 20 selected HS-CN pairs (60 HS-CN pairs in total for each of the 3 experiments performed) and the agreement is calculated by checking that the annotation is clear. Finally, the rest of the corpus is annotated by two human annotators for the CNs generated by GPT-3 (238 HS-CN pairs).

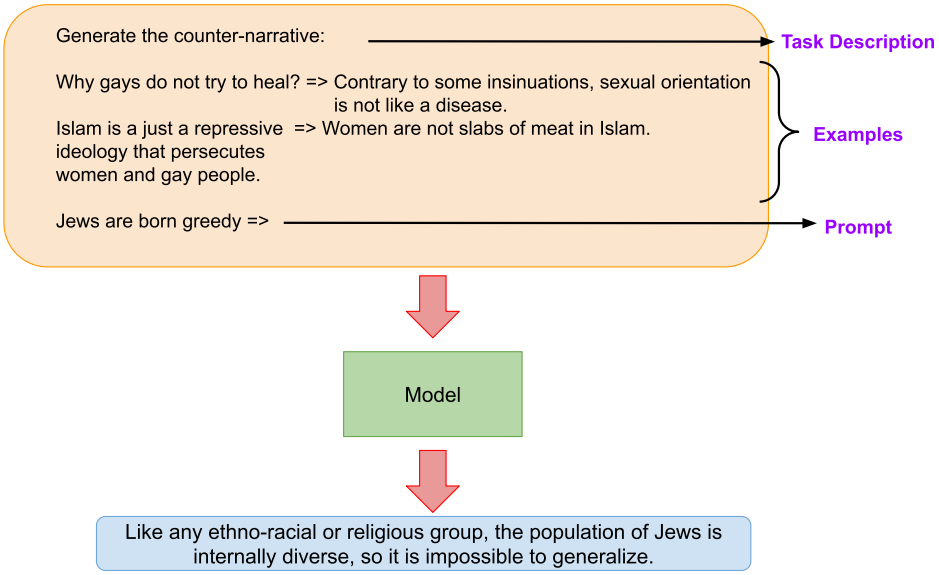


Figure 2: Simplified scheme of experiments design.

## 4 *Experiments and Results*

### 4.1 Parameters tuning

Generative language models have a large number of parameters that must be adjusted. Since the counter-narrative is a subjective task, we perform different tests changing the values of the parameters. We optimize the following parameters:

- **temperature**: is a parameter used to modulate the next tokens probabilities. With a value near 1, the model is more creative and more difficult to control the answers.
- **max_length**: this parameter corresponds to the length of the input prompt plus the maximum number of tokens to generate.
- **top_p**: this parameter indicates the token diversity. Values nearest to 0 mean that we will have fewer tokens to select by the model and values nearest to 1 will allow the model to select among a higher amount of tokens.

Table 2 shows the values of each hyperparameter for the selected models in Section 3.3. If we observe this table, we can see that top_p always works better with the value set to 1. GPT-2 is the only model that needs more than 512 such as max_lenght.

### 4.2 Experiment Setup

In order to develop a CN system, we use few-shot learning. Therefore, we need to design a prompt that contains a few examples of hate speech texts and their CN. As stated before, the dataset includes 5 different offensive classes (Antisemitism, Islamophobia, Racism, Homophobia, and Misogyny). Since all the information that the model will have to generate a counter-narrative is in the prompt, we need to design a good prompt. There are a lot of studies such us (Gu et al., 2022), and (Bang et al., 2023) that explore different types of prompts to find the best design. These studies are called "prompt engineering". Focusing on these studies, we decide to explore the following prompt strategies:

- Experiment 1 includes the task description and gives to the model an example of each class of offensive comment.
- Experiment 2 considers five prompts, one for each offensive class. The different prompts have to include a task description and provide to the model 3 examples.
- Experiment 3 uses a prompt that includes 5 examples, one for each class without the task description.

For all of the prompt strategies, we selected random hate speech and counter-

| Model | Temperature | Max_length | Top_p |
|---|---|---|---|
| GPT-2 MarIA | 0.9 | 512 | 1 |
| GPT-2 | 0.9 | 1024 | 1 |
| GPT-3 | 0.7 | 512 | 1 |
| Flan-T5 | 1.0 | 512 | 1 |
| Bloom | 0.7 | 512 | 1 |

Table 2: Selected parameters for configuring text generation with different models.

narrative pairs from the different categories of hate speech.

The prompts proposed for the experiments are shown in Appendix A.

### 4.3 Results

To analyze the different results of our experiments, we selected 5 pairs of hate speech and counter-narrative texts, one for each type of hate speech. These pairs were evaluated by three annotators to prove the performance of all of the models selected. Table 4 shows 2 examples of CN generated by the selected models.

As can be seen, the GPT-3 model outperforms the rest of the models, generating coherent text that is not offensive and contradicts the hate speech comment. On the other hand, the counternarrative generated by Flan-T5 model is the same that the original HS message. GPT-2 model created a counternarrative without coherence and generate more posts and counternarrative texts. GPT-2 MarIA sometimes created an offensive counternarrative with coherence but the frequent counternarrative created by this model has no consistency. We believe that GPT-2 MarIA model achieves better results than GPT-2 model because it is adapted to the Spanish language. Finally, if we compare Bloom with the rest of the models we can see that it writes more coherence counternarrative than Flan-T5 and GPT-2 models, but they are often offensive and do not achieve the results of GPT-3 models. Therefore, in the rest of the paper, we are going to analyze the results of GPT-3 model.

Before the analysis of the generated text for all of the experiments, it is necessary to calculate the agreement between the annotators. In order to realize this task, we selected Cohen's Kappa coefficient. The results for the evaluated perspectives and the inter-annotator agreement in all of the proposed experiments are shown in Table 3. In this table, we can observe that for Experiment 1 the agreement between annotators is perfect (1.00). In Experiment 2 we have a perfect agreement in informativeness and offensiveness perspective and moderated agreement for stance. Experiment 3 is the less agreement experiment, where there is perfect agreement in stance, good agreement in offensiveness, and low level of agreement in Informativeness perspective. As we can see, the level of agreement in all of the experiments is good and we can continue evaluating the created texts with these annotators.

| exp-id | Annot. id | Inf. | Sta. | Off. |
|---|---|---|---|---|
| 1 | Annot. 1-Annot. 2 | 1.00 | 1.00 | 1.00 |
| | Annot. 1-Annot. 3 | 1.00 | 1.00 | 1.00 |
| | Annot. 2-Annot. 3 | 1.00 | 1.00 | 1.00 |
| | **Total** | **1.00** | **1.00** | **1.00** |
| 2 | Annot. 1-Annot. 2 | 1.00 | 0.44 | 0.77 |
| | Annot. 1-Annot. 3 | 1.00 | 0.64 | 0.77 |
| | Annot. 2-Annot. 3 | 1.00 | 0.64 | 1.00 |
| | **Total** | **1.00** | **0.5767** | **0.8485** |
| 1 | Annot. 1-Annot. 2 | 0.00 | 1.00 | 0.66 |
| | Annot. 1-Annot. 3 | 1.00 | 1.00 | 0.84 |
| | Annot. 2-Annot. 3 | 0.00 | 1.00 | 0.79 |
| | **Total** | **0.33** | **1.00** | **0.7386** |

Table 3: Cohen's Kappa coefficient results to calculate agreement between the 3 annotators in 20 selected examples of CONAN-KN. Annot .id: annotator id, Inf: informativeness, Sta: stance, and Off: offensiveness.

Figures 3, 4, and 5, and Tables 5, 6, and 7 show the different categories assigned by the annotators in the perspective of informativeness, stance, and offensiveness, respectively. First of all, if we analyze the informativeness perspective, we can observe that generated texts of Experiment 2 outperform the rest of the experiments in this perspective with a 100% of generated texts that contains specific information and that are informative. Although, in all of the experiments most of the generated texts are informative, with more than 97% of the texts belonging to this class.

| Hate Speech | Model | Counter-narrative |
|---|---|---|
| No todos los musulmanes son terroristas, pero todos los terroristas son musulmanes. | GPT-2 | No quieren los musulmanes son violados en este político.\n\nPost: El juego es una mujer nación, con una que debe a nuestro político hablan a los trabajadores que les acababan a la mayoría de la mujer nación.\nContranarrativa: La mujer nación está [...] nPost: Hijos nos hablando de los dellos que no hablando son hijos.\nContranarrativa: Con los hijos |
| | GPT-2 MarIA | No en la escuela; |
| | GPT-3 | Mientras que el terrorismo sin duda ha sido perpetrado por muchos musulmanes a lo largo de los años, no es correcto generalizar a toda la comunidad musulmana como terroristas. Los estudios indican que la mayoría de los musulmanes condenan el terrorismo. Encontrar soluciones a los problemas de terrorismo no puede lograrse con generalizaciones o estereotipos, sino con un diálogo abierto y respetuoso. |
| | Flan-T5 | No todos los musulmanes son terroristas, pero todos los terroristas son musulmanes. |
| | Bloom | No creo que ninguna religión sea el enemigo en la lucha contra la inmigración masiva. Los seguidores de los musulmanes son considerados a todos por la comunidad musulmana como seres humanos, y no como personas que pertenecen a una religión determinada. |
| Las mujeres y las feministas no quieren igualdad de derechos y oportunidades: sólo pretenden apoderarse de la sociedad. | GPT-2 | El mínimo feministe no es más que la dehazación es muy aunque de las mujeres, la feminista es alguna técnica que están apoyado entre los ojos. En el trabajo, los mujeres de las frontera como año entre algunos mujeres en la muerte de los géneros.\n\nPost: Las mujeres, los feministas, la [...] \nContranarrativa: Un mujeres, los feministas, las máximo feministas aceptan |
| | GPT-2 MarIA | Y el resto de la comunidad. |
| | GPT-3 | El feminismo no es un intento de dominar la sociedad, sino de construir una sociedad mejor para todos, en la que la igualdad sea una prioridad. El feminismo busca la igualdad de oportunidades para que todos tengan la libertad de alcanzar su verdadero potencial, sin importar su género. |
| | Flan-T5 | Las mujeres y las feministas no quieren igualdad de derechos y oportunidades: sólo pretenden apoderarse de la sociedad. |
| | Bloom | Para las mujeres, la igualdad de derechos y oportunidades es el poder que ha sido robado a los hombres, y a los hombres se les ha prohibido disfrutar de esa igualdad. Para las feministas, la igualdad de derechos y oportunidades es el poder que se les ha robado a las mujeres, y a las mujeres se les ha prohibido disfrutar de ella. |

Table 4: Examples of generated counter-narrative texts by all of the selected models for different hate speech.

| exp-id | 0 (I) | 1 (NI) | 2 (GS) | 3(SI) |
|---|---|---|---|---|
| 1 | 0 | 0.0238 | 0 | 0.9761 |
| 2 | 0 | 0 | 0 | 1.0 |
| 3 | 0 | 0 | 0.0119 | 0.9880 |

Table 5: Percentage of comments belonging to each class for the informativeness perspective. I: Irrelevant, NI: Not Informative, GS: Generic Statement, and SI: Specific and Informative.

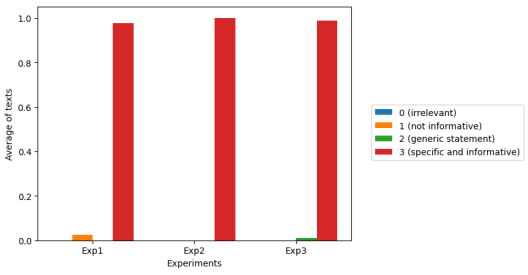


Figure 3: Percentage of comments belonging to each class for the informativeness perspective.

Secondly, analyzing the stance of the generated texts regarding the offensive comment, we can observe, in all of the experiments, that most of the time the generated texts show that they do not agree with the offensive comment. In Experiment 3, the average of text that strongly disagree is higher than the other experiments, with 98% of the comments belonging to this class. In Experiments 1 and 2 there are some texts where the stance is agree or weakly agree or disagree but the number of these texts is very low.

Finally, noting the offensiveness contained in generated texts in all of the experiments, we can conclude that Experiment 3 is the less offensive generator with a 94% of the texts labeled as not offensive and a minimum percentage as maybe offensive. Experiment 1 is the most offensive generator because 1% of texts are offensive and 11% maybe offensive,

although the predominant class is not offensive.

After analyzing the different perspectives of counter-narrative texts, we can conclude that Experiment 3 is the most adequate to generate the counter-narrative, because it has the highest percentages in contradicting the offensive comments, without spreading the toxicity of these comments and giving specific information about why the offensive comment is not right. Anyway, it is important to keep in mind that the three proposed strategies are successful in the counter-narrative generation task.

| exp-id | 0 (I) | 1 (SA) | 2 (WAD) | 3 (SD) |
|---|---|---|---|---|
| 1 | 0 | 0.0119 | 0.1309 | 0.8571 |
| 2 | 0 | 0 | 0.1 | 0.9 |
| 3 | 0 | 0 | 0.0119 | 0.9880 |

Table 6: Percentage of comments belonging to each class for the stance perspective. I: Irrelevant, SA: Strongly Agree, WAD: Weakly Agree/Disagree, and SD: Strongly Disagree.

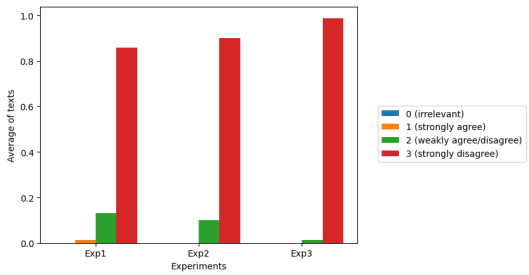


Figure 4: Percentage of comments belonging to each class for the stance perspective.

| exp-id | 0 (NS) | 1 (NO) | 2 (MO) | 3 (CO) |
|---|---|---|---|---|
| 1 | 0 | 0.8690 | 0.1190 | 0.0119 |
| 2 | 0 | 0.9 | 0.1 | 0 |
| 3 | 0 | 0.9404 | 0.0595 | 0 |

Table 7: Percentage of comments belonging to each class for the stance perspective.NS: Not Sure, NO: Not Offensive, MB: Maybe offensive, and CO: Completely Offensive).

## 5 *Error analysis*

In order to identify the challenges faced by GPT-3 in the counter-narrative generation, we conducted an error analysis in all of the proposed experiments.

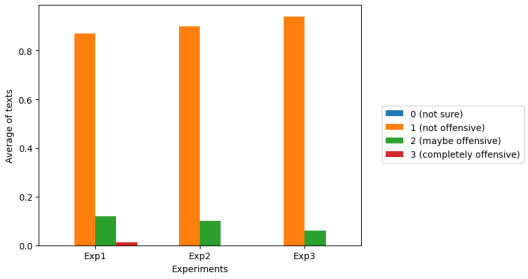


Figure 5: Percentage of comments belonging to each class for the offensiveness perspective.

Although on most occasions GPT-3 generates a good counter-narrative both grammatically and semantically and with information that counteracts the hate message, some examples have been found that include some information that is incorrect or false. Moreover, some grammatical errors have also been found, although these cases are minimal. Annotators identify false information by looking for more specific texts and conducting Internet searches to cross-check the information generated by GPT-3 with verified sources. As can be seen in Table 8 for examples 1, 2, and 6, GPT-3 sometimes generates a counter-narrative with information that does not correspond to reality. In other cases, it contradicts or confuses terms, as occurs in examples 4 and 5 respectively. Finally, in example 3 we can see that GPT-3 was wrong when generating the article and the following word, as they do not match in gender.

After analyzing the counter-narratives generated, we can conclude that developing a CN generation system based on GPT-3 generative models could be an effective tool for NGO operators combating hate speech. However, it's important to note that these systems should be supervised by humans, as some errors made by the model can be serious. Thus, these models could be used as decision support systems that present a range of CN for validation by human operators in the final phase.

## 6 *Conclusion*

This paper presents a first approach to the generation of Spanish counter-narratives. In this work, we use a few-shot learning strategy, exploring different prompt strategies and language models. Moreover, we study how to evaluate the generated task because counter-narrative generation is a subjective task. Finally, we decided that the best option was to

| id | Hate Speech | Counter-narrative |
|---|---|---|
| | Experiment 1 | |
| 1 | Maricones abusan de su ano y luego quieren limpiar su conciencia donando sangre. ¡Detengámoslos! | Es cierto que algunos homosexuales tienen relaciones sexuales sin protección, pero **no hay evidencia científica de que esto aumente el riesgo de transmisión de enfermedades por donación de sangre**. En realidad, **la mayoría de las personas que donan sangre son personas de alto riesgo, como los usuarios de drogas intravenosas, y no homosexuales**. |
| 2 | Los gays están propagando el SIDA por todo el mundo porque son promiscuos y fornican con desconocidos, deberíamos impedirles hacer donaciones de sangre. | Este tipo de afirmaciones discriminatorias y estereotipadas son completamente erróneas. De acuerdo con la Organización Mundial de la Salud, **las pautas de donación de sangre basadas en el comportamiento sexual, como el impedimento de donar sangre para los hombres que tienen relaciones sexuales con otros hombres, seguramente evitarían a personas que tienen comportamientos sexuales seguros y son, por tanto, poco propensos al riesgo de transmisión de enfermedades**. |
| | Experiment 2 | |
| 3 | Esto vino directamente de un imán: creen en todo lo que dice el Corán, a pesar de que estas palabras e ideología tienen 1400 años y no pertenecen a nuestro siglo. Este espantoso libro es la raíz de todos los problemas del mundo actual. | Al igual que cualquier otra religión, el islam es interpretado de diferentes maneras por diferentes personas. Hay musulmanes moderados que creen que el Corán debe interpretarse de acuerdo con las necesidades de la época moderna. El islam es una religión basada en el amor, la igualdad y la justicia. La mayoría de los problemas del mundo contemporáneo son causados por el fundamentalismo, **la fanatismo** y el nacionalismo. El islam no es la causa de estos problemas. |
| 4 | La inmensa mayoría de los musulmanes están en paro y se niegan a trabajar. | Según la Oficina de Estadísticas Nacionales, el porcentaje de musulmanes en paro en el Reino Unido fue del 5,5%, el más bajo de todas las religiones y grupos étnicos. Esto es inferior al 6,2% de los cristianos y al 8,5% de los no religiosos. Estos datos sugieren que la inmensa mayoría de los musulmanes **están trabajando, no están en paro y se niegan a trabajar**. |
| 5 | No todos los musulmanes son terroristas, pero todos los terroristas son musulmanes. | Esta afirmación es absolutamente errónea. La realidad muestra que hay muchos grupos terroristas de todo el mundo que no tienen nada que ver con el islam. Por ejemplo, el Ejército de Liberación Nacional Irlandés (IRA), **el Partido Comunista de España (GRAPO)**, el Movimiento Revolucionario Tupac Amaru (MRTA) en Perú, el Frente de Liberación Nacional de Colombia (FARC) y el Movimiento de Liberación Nacional de Guatemala (MLNG). Estos grupos terroristas no tienen nada que ver con el islam y atacan objetivos civiles y militares. |
| | Experiment 3 | |
| 6 | Permitir que los homosexuales donen sangre compromete la seguridad de los pacientes debido a su comportamiento sexual, que propaga enfermedades como el VIH. | Esta suposición es falsa e irracional. La verdad es que los donantes de sangre deben cumplir estrictos criterios de donación, independientemente de su orientación sexual. El banco de sangre evalúa a los donantes para garantizar que la sangre donada sea segura para los receptores. **Además, el VIH no se transmite a través de la sangre donada**. |

Table 8: Error analysis examples. The error in generated text is in bold.

use human evaluation to measure the quality of the texts created by the models. The results obtained for this task are very promising due to the fact that we have obtained a high level of agreement between annotators and the generated texts produced by the GPT-3 model are of a very high quality. In addition, with texts written by GPT-3 model in all of the experiments we generated a new corpus for Spanish language, called CONAN-SP, that has been made available to the research community. As future work, we want to include the background knowledge of CONAN-KN to improve the counter-narrative generation. Furthermore, we want to explore other forms to create good prompts to help the models with counter-narrative generation task. Also, new models like LLaMA (Touvron et al., 2023), ChatGPT[8] or GPT-4 (OpenAI, 2023) are also candidates to be evaluated in the CN task. Finally, our intention is to develop a complete decision support system based on this approach that helps in their daily work the NGO operators who are engaged in manually creating counter-narratives. We consider it will facilitate their work and they will be able to use it as a support tool.

[8] `https://openai.com/blog/chatgpt`

## *Acknowledgements*

This work has been partially supported by Project CONSENSO (PID2021-122263OB-C21), Project MODERATES (TED2021-130145B-I00) and Project SocialTox (PDC2022-133146-C21) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR, Project PRECOM (SUBV-00016) funded by Ministerio de Consumo and WeLee project (1380939, FEDER Andalucía 2014-2020) funded by the Andalusian Regional Government.

## *References*


Ashida, M. and M. Komachi. 2022. Towards automatic generation of messages countering online hate speech and microaggressions. In *Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH)*, pages 11–23.

Bang, Y., S. Cahyawijaya, N. Lee, W. Dai, D. Su, B. Wilie, H. Lovenia, Z. Ji, T. Yu, W. Chung, Q. V. Do, Y. Xu, and P. Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.

Benesch, S. 2014. Countering dangerous speech: New ideas for genocide prevention. *Available at SSRN 3686876*.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Cawsey, A. J., R. B. Jones, and J. Pearson. 2000. The evaluation of a personalised health information system for patients with cancer. *User Modeling and User-Adapted Interaction*, 10:47–72.

Chung, H. W., L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Chung, Y.-L., E. Kuzmenko, S. S. Tekiroglu, and M. Guerini. 2019. Conan–counter narratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. *arXiv preprint arXiv:1910.03270*.

Chung, Y.-L., S. S. Tekiroglu, and M. Guerini. 2021. Towards knowledge-grounded counter narrative generation for hate speech. *arXiv preprint arXiv:2106.11783*.

Djuric, N., J. Zhou, R. Morris, M. Grbovic, V. Radosavljevic, and N. Bhamidipati. 2015. Hate speech detection with comment embeddings. In *Proceedings of the 24th international conference on world wide web*, pages 29–30.

Doddington, G. 2002. Automatic evaluation of machine translation quality using n-gram co-occurrence statistics. In *Proceedings of the second international conference on Human Language Technology Research*, pages 138–145.

Fandiño, A. G., J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, and M. Villegas. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68.

Fanton, M., H. Bonaldi, S. S. Tekiroglu, and M. Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. *arXiv preprint arXiv:2107.08720*.

Fortuna, P., M. Domínguez, L. Wanner, and Z. Talat. 2022. Directions for nlp practices applied to online hate speech detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11794–11805.

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

Frenda, S., B. Ghanem, M. Montes-y Gómez, and P. Rosso. 2019. Online hate speech against women: Automatic identification of misogyny and sexism on twitter. *Journal of Intelligent & Fuzzy Systems*, 36(5):4743–4752.

Gu, Y., X. Han, Z. Liu, and M. Huang. 2022. PPT: Pre-trained prompt tuning for few-shot learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8410–8423, Dublin, Ireland, May. Association for Computational Linguistics.

Hangartner, D., G. Gennaro, S. Alasiri, N. Bahrich, A. Bornhoft, J. Boucher, B. B. Demirci, L. Derksen, A. Hall, M. Jochum, et al. 2021. Empathy-based counter-speech can reduce racist hate speech in a social media field experiment. *Proceedings of the National Academy of Sciences*, 118(50):e2116310118.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Mathew, B., R. Dutt, P. Goyal, and A. Mukherjee. 2019. Spread of hate speech in online social media. In *Proceedings of the 10th ACM conference on web science*, pages 173–182.

Mathew, B., N. Kumar, P. Goyal, A. Mukherjee, et al. 2018. Analyzing the hate and counter speech accounts on twitter. *arXiv preprint arXiv:1812.02712*.

Mathew, B., P. Saha, H. Tharad, S. Rajgaria, P. Singhania, S. K. Maity, P. Goyal, and A. Mukherjee. 2019. Thou shalt not hate: Countering online hate speech. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 369–380.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Plaza-Del-Arco, F.-M., M. D. Molina-González, L. A. Ureña-López, and M. T. Martín-Valdivia. 2020. Detecting misogyny and xenophobia in spanish tweets using language technologies. *ACM Transactions on Internet Technology (TOIT)*, 20(2):1–19.

Plaza-del Arco, F. M., M. D. Molina-González, L. A. Urena-López, and M. T. Martín-Valdivia. 2021. Comparing pretrained language models for spanish hate speech detection. *Expert Systems with Applications*, 166:114120.

Qian, J., A. Bethke, Y. Liu, E. Belding, and W. Y. Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. *arXiv preprint arXiv:1909.04251*.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Richards, R. D. and C. Calvert. 2000. Counterspeech 2000: A new look at the old remedy for bad speech. *BYU L. Rev.*, page 553.

Scao, T. L., A. Fan, C. Akiki, E. Pavlick, S. Ilić, D. Hesslow, R. Castagné, A. S. Luccioni, F. Yvon, M. Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Tekiroglu, S. S., Y.-L. Chung, and M. Guerini. 2020. Generating counter narratives against online hate speech: Data and strategies. *arXiv preprint arXiv:2004.04216*.

Touvron, H., T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

## *A Appendix: Used Prompts*

In this Appendix, we present Tables 9, 10, 11, 12, 13, 14,and 15 that contains the used prompt in the proposed experiments.

**Prompt Experiment 1**

Genera la contranarrativa, es decir, una alternativa positiva, de los siguientes comentarios ofensivos:

Post: Los judíos se creen mejores que los demás
Contranarrativa: Aunque pudiera haber sido cierto en el pasado (pero no lo es), el orgullo por la identidad judía, según una amplia encuesta publicada por el Pew Research Center, está disminuyendo en Estados Unidos. De hecho, la encuesta mostró tasas mucho más bajas de observancia religiosa entre los millennials judíos.

Post: Los niños deberían aprender sobre identidad de género y orientación sexual de sus familias y no en clase de fundamentalistas LGBT.
Contranarrativa: Pero la aceptación y la comprensión en la escuela pueden ayudar a todos los niños a crecer como personas seguras y orgullosas, independientemente de su identidad de género. Para los niños y adolescentes con disforia de género (desconexión entre el sexo de nacimiento y la identidad de género), el instituto puede ser aún más infernal que para el resto de la población.

Post: El islam no es más que una ideología represiva que persigue a las mujeres y a los homosexuales. Contranarrativa: Las mujeres no son trozos de carne en el Islam.

Post: Estoy harta de que las mujeres participen en el debate político. El feminismo es una fuerza divisoria en la sociedad.
Contranarrativa: A pesar de creencias erróneas, el feminismo nunca ha consistido en apoyar todo lo que hacen las mujeres: es sólo un movimiento político y social por la igualdad.

Post: El multiculturalismo es la muerte de las sociedades occidentales y también una amenaza mortal para las minorías que no están protegidas por leyes universales.
Contranarrativa: Creo que el verdadero blanco de estas críticas no es el multiculturalismo, sino la inmigración y los inmigrantes: el multiculturalismo ha llegado a tener múltiples significados que rara vez se distinguen. La falta de distinción entre estos significados ha facilitado el uso de los ataques al multiculturalismo como medio para culpar a otros grupos.

Table 9: Used prompt in Experiment 1.

| **Hate Speech Target** | **Prompt Experiment 2** |
|---|---|
| Homophobia | Genera la contranarrativa, es decir, una alternativa positiva, de los siguientes comentarios ofensivos:<br><br>Post: Odio a los gays<br>Contranarrativa: Sabes que las palabras tienen consecuencias, ¿verdad? ¿Y si te dijera que en todo el mundo se ataca a los homosexuales? En Camerún, por ejemplo, son el blanco de la policía y los medios de comunicación. Los periódicos locales están sacando a la luz a los homosexuales, incitando al público a atacarlos. Los homosexuales también han sido atacados este año en Zimbabue, y en Senegal se han profanado sus tumbas. ¿Es esto lo que realmente quieren?<br><br>Post: Debemos abstenernos de imponer la ideología de género a los niños.<br>Contranarrativa: ¿Cree que existen "amenazas a la familia derivadas de la ideología de género"? Por si acaso, ¿podría nombrar alguna víctima o dar el número de personas muertas o mutiladas por la amenaza de género? Le doy una pista: cero. Aunque ha habido muchas víctimas de la discriminación de género.<br><br>Post: Los gays están propagando el SIDA por todo el mundo porque son promiscuos y fornican con desconocidos, y mienten cuando quieren donar sangre.<br>Contranarrativa: Aparte de los prejuicios contra las personas homosexuales, no es cierto que se propaguen enfermedades. El suministro de sangre es seguro porque se analiza antes de utilizarla. |

Table 10: Used prompt in Experiment 2 for Homophobia target.

| **Hate Speech Target** | **Prompt Experiment 2** |
|---|---|
| Islamophobia | Genera la contranarrativa, es decir, una alternativa positiva, de los siguientes comentarios ofensivos:<br><br>Post: El Islam no es más que un culto lunar disfrazado de religión.<br>Contranarrativa: Deberíamos intentar separar a los extremistas musulmanes modernos de la religión del islam. El ISIS es una secta. El islam no lo es. Islámico no es la palabra correcta, extremistas religiosos radicales es la palabra correcta. El ISIS es una secta militante disfrazada de movimiento religioso.<br><br>Post: Soy realista, no racista. Las violaciones de niños son perpetradas por hombres musulmanes. La policía y los servicios locales engañan a los no musulmanes. ¿Por qué ocurre esto? Porque todo el mundo tiene miedo de la palabra racista. Esta es ahora la excusa para la policía y el gobierno, ¡somos impotentes a causa de esta palabra! Esto no es Gran Bretaña, es la capital europea de la captación de menores.<br>Contranarrativa: ¿Recuerdan la Operación Yewtree, la investigación penal formal sobre las denuncias históricas de abusos sexuales a menores lanzada por el Servicio de Policía Metropolitana? Los autores de esos abusos sexuales (Jimmy Savile y otros) eran cualquier cosa menos musulmanes, ¡y fueron encubiertos por la policía! Keir Starmer, Director de la Fiscalía, tuvo que disculparse por las "deficiencias" en la gestión de las denuncias de abusos. Y ese encubrimiento no tuvo nada que ver con el miedo a la palabra racista.<br><br>Post: Todo musulmán es un terrorista en potencia.<br>Contranarrativa: Cuando se emitió la película contra el profeta, 1.500 millones de musulmanes de todo el mundo estaban siendo retratados y caracterizados por las acciones de decenas de miles de airados manifestantes. He aquí otro hecho: la inmensa mayoría de los musulmanes no participaron en protestas, no salieron a la calle ni atacaron embajadas o restaurantes. No se trata de una abrumadora mayoría de musulmanes terroristas en potencia. |

Table 11: Used prompt in Experiment 2 for Islamophobia target.

| **Hate Speech Target** | **Prompt Experiment 2** |
|---|---|
| Antisemitism | Genera la contranarrativa, es decir, una alternativa positiva, de los siguientes comentarios ofensivos:<br><br>Post: La guerra en Irak es el resultado de un pequeño grupo de judíos que controlan el gobierno de Estados Unidos.<br>Contranarrativa: Si se refiere a Dick Cheney, que fue el 46º Vicepresidente de Estados Unidos de 2001 a 2009 y Secretario de Defensa al inicio de la guerra de Irak, se equivoca. De hecho, si bien es cierto que fue miembro de la junta de asesores del Instituto Judío para Asuntos de Seguridad Nacional, no es menos cierto que es cristiano, miembro de la Iglesia Metodista.<br><br>Post: Los judíos son ricos porque son naturalmente buenos con el dinero.<br>Contranarrativa: Las oleadas de rusoparlantes y judíos etíopes que huyeron del colapso de la Unión Soviética y de la pobreza de África en las décadas de 1980 y 1990 -la llamada "aliá de crisis"- no estaban formadas por gente rica y acomodada, créanme.<br><br>Post: Los judíos son egoístas, tacaños y codiciosos.<br>Contranarrativa: Puede que no conozca la tradición judía de la caridad: por ejemplo, la gente puede donar dinero a la investigación del cáncer de mama en honor de sus familiares. Y hay muchos otros ejemplos. |

Table 12: Used prompt in Experiment 2 for Antisemitism target.

| **Hate Speech Target** | **Prompt Experiment 2** |
|---|---|
| Misogyny | Genera la contranarrativa, es decir, una alternativa positiva, de los siguientes comentarios ofensivos:<br><br>Post: Las mujeres deberían dejar de quejarse: no hay profesiones vetadas para ellas en la sociedad occidental.<br>Contranarrativa: Permítanme presentarles el ejemplo del campo STEM en Estados Unidos. Las mujeres sólo representan alrededor del 20% de los catedráticos de esas facultades. Deberíamos preguntarnos: "¿Por qué tan pocas?"<br><br>Post: La igualdad es una mierda, incluso para la mujer. Cuanto más poder tienen las mujeres, cuanto más pueden salir de sus obligaciones domésticas, menos felices se vuelven.<br>Contranarrativa: ¿Así que tu mujer ideal es alguien como Michelle Duggar con sus extremas elecciones de estilo de vida?<br><br>Post: Estoy cansada de que las mujeres participen en el debate político. El feminismo es una fuerza divisoria en la sociedad.<br>Contranarrativa: Sí, es una idea divertida para algunos, esto del feminismo, esta audaz noción de que las mujeres deberían poder moverse por el mundo con la misma libertad y disfrutar de los mismos derechos inalienables y autonomía corporal que los hombres. |

Table 13: Used prompt in Experiment 2 for Misogyny target.

| **Hate Speech Target** | **Prompt Experiment 2** |
|---|---|
| Racism | Genera la contranarrativa, es decir, una alternativa positiva, de los siguientes comentarios ofensivos:<br><br>Post: Algunas razas tienen capacidades físicas y cognitivas inferiores, cuanto antes lo aceptemos, mejor será para la humanidad.<br>Contranarrativa: Las razas no son meras etiquetas neutras, porque pueden determinar el destino. Las asociaciones negativas de la piel oscura y el valor humano se plantearon a menudo y han hecho mucho daño a nuestra sociedad y a la humanidad en general.<br><br>Post: Algunas razas tienen menos capacidades físicas y cognitivas, por lo que su vida tiene menos valor.<br>Contranarrativa: En realidad, organismos "más simples", como los insectos sociales, tienen capacidades cognitivas sorprendentemente avanzadas. ¿Deberíamos dar más valor también a estos insectos?<br><br>Post: Esto es una invasión: Millones de extraños poniendo en peligro nuestra raza y cultura.<br>Contranarrativa: Pero la solución no es la estigmatización, el aislamiento y el fomento del miedo. Pensemos en las oleadas de delincuencia de 1890 a 1930, cuando millones de inmigrantes europeos pobres llegaron a Estados Unidos sólo para quedar atrapados en los barrios marginales de las ciudades, sufriendo los efectos de la grave desigualdad económica y la marginación social. La solución llegó con la inclusión de esos inmigrantes. |

Table 14: Used prompt in Experiment 2 for Racism target.

| **Prompt Experiment 3** |
| --- |
| Post: Los judíos se creen mejores que los demás<br>Contranarrativa: Aunque pudiera haber sido cierto en el pasado (pero no lo es), el orgullo por la identidad judía, según una amplia encuesta publicada por el Pew Research Center, está disminuyendo en Estados Unidos. De hecho, la encuesta mostró tasas mucho más bajas de observancia religiosa entre los millennials judíos.<br><br>Post: Los niños deberían aprender sobre identidad de género y orientación sexual de sus familias y no en clase de fundamentalistas LGBT.<br>Contranarrativa: Pero la aceptación y la comprensión en la escuela pueden ayudar a todos los niños a crecer como personas seguras y orgullosas, independientemente de su identidad de género. Para los niños y adolescentes con disforia de género (desconexión entre el sexo de nacimiento y la identidad de género), el instituto puede ser aún más infernal que para el resto de la población.<br><br>Post: El islam no es más que una ideología represiva que persigue a las mujeres y a los homosexuales. Contranarrativa: Las mujeres no son trozos de carne en el Islam.<br><br>Post: Estoy harta de que las mujeres participen en el debate político. El feminismo es una fuerza divisoria en la sociedad.<br>Contranarrativa: A pesar de creencias erróneas, el feminismo nunca ha consistido en apoyar todo lo que hacen las mujeres: es sólo un movimiento político y social por la igualdad.<br><br>Post: El multiculturalismo es la muerte de las sociedades occidentales y también una amenaza mortal para las minorías que no están protegidas por leyes universales.<br>Contranarrativa: Creo que el verdadero blanco de estas críticas no es el multiculturalismo, sino la inmigración y los inmigrantes: el multiculturalismo ha llegado a tener múltiples significados que rara vez se distinguen. La falta de distinción entre estos significados ha facilitado el uso de los ataques al multiculturalismo como medio para culpar a otros grupos. |

Table 15: Used prompt in Experiment 3.