CONAN-MT-SP: A Spanish Corpus for Counternarrative using GPT Models

M. Estrella Vallecillo-Rodríguez, M. Victoria Cantero-Romero, Isabel Cabrera-de-Castro, 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, vcantero, iccastro, amontejo, maite}@ujaen.es

Abstract

This paper describes the automated generation of CounterNarratives (CNs) for Hate Speech (HS) in Spanish using GPT-based models. Our primary objective is to evaluate the performance of these models in comparison to human capabilities. For this purpose, the English CONAN Multitarget corpus is taken as a starting point and we use the DeepL API to automatically translate into Spanish. Two GPT-based models, GPT-3 and GPT-4, are applied to the HS segment through a few-shot prompting strategy to generate a new CN. As a consequence of our research, we have created a high quality corpus in Spanish that includes the original HS-CN pairs translated into Spanish, in addition to the CNs generated automatically with the GPT models and that have been evaluated manually. The resulting CONAN-MT-SP corpus and its evaluation will be made available to the research community, representing the most extensive linguistic resource of CNs in Spanish to date. The results demonstrate that, although the effectiveness of GPT-4 outperforms GPT-3, both models can be used as systems to automatically generate CNs to combat the HS. Moreover, these models consistently outperform human performance in most instances.

Keywords: Counternarrative Generation, Hate-Speech in Spanish, Large Language Model, GPT Models, Linguistic analysis, Corpus generation

1. Introduction

The growing increase in social interactions through digital platforms has led to the emergence of inappropriate behaviors in the virtual environment. Among these behaviors is the dissemination of hate messages between users of these platforms known as Hate Speech (HS). Freedom of expression in these media has exposed users to publications that are sometimes used to belittle, insult, or harm using both subtle and offensive language based on personal characteristics such as gender, race, religion, and ideology, among others. Unfortunately, this form of communication can have harmful consequences and generate negative psychological effects on users, especially in young people, such as anxiety, cyberbullying, and even extreme cases of suicide (Hinduja and Patchin, 2018).

According to the UN Strategy and Plan of Action on Hate Speech[1], hate speech is defined as "any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor".

The most common strategy to combat HS on digital platforms is the removal of hate messages. However, this strategy can sometimes be counterproductive due to the fact that can generate a perception of censorship and limitation of freedom of expression, which can increase resistance and backlash from users. In addition, message removal does not address the underlying causes of HS and does not provide an opportunity for education and attitude change (Miller et al., 2020). For these reasons, a new strategy involving the generation of counternarratives (CNs) is explored. CN generation aims to challenge and dismantle the negative arguments and stereotypes promoted by HS by providing accurate and empathetic information. By offering an alternative and constructive perspective, CNs aim to encourage empathy, understanding, and tolerance among users, thus promoting a more inclusive and respectful online environment.

The current approach to countering HS involves manually generating CNs, often carried out by individuals or NGOs. However, this method has limitations in terms of effectiveness and practicality, as it is difficult to monitor all digital platforms manually. To address this problem, researchers are exploring a new approach that uses Natural Language Processing (NLP) and Machine Learning (ML) techniques to automatically generate CNs (Chung et al., 2023). This strategy could be

[1] https://www.un.org/en/hate-speech/understanding-hate-speech/what-is-hate-speech

LREC-COLING 2024, pages 3677–3688
20-25 May, 2024.

used by NGOs as a decision support system that can help human operators in their daily routine. This approach has advantages such as speed in generating CNs, scalability to accommodate massive volumes of content, and the ability to monitor online interactions in real-time. However, this method has some challenges and problems, such as the risk that these automatically generated CNs may be uninformative or lack precision in their content. Moreover, these systems can sometimes produce more offensive or problematic CNs than the HS they are intended to mitigate. Thus, it is necessary, firstly, to study the possibility of automatically generating these CNs, and secondly, to check if it is a good CN comparable to the one that could be created by a human. This article presents the work we are conducting in this area to combat HS through the automatic generation of CN using GPT-based language models. The main objective is focused on the development of a decision support system so that in a real environment, this computational system can help human operators in the generation of counter narratives.

The rest of the paper is structured as follows: In Section 2 we present some previous studies related to generating CN texts. The proposed experiments and the methodology of these experiments are explained in Section 3. In Section 3.4, we can see the evaluation methodology for the proposed experiments. A general analysis of the annotated corpus is provided in section 4. Section 5 offers a linguistic analysis of the generated CN. Finally, the conclusions are depicted in Section 6.

2. Related Work

In recent years, researchers have made efforts to generate corpus and datasets to elaborate automatic systems that are capable of detecting HS (Zampieri et al., 2019; Caselli et al., 2021; Plaza-del Arco et al., 2021).

While one of the main strategies to combat HS is to block HS, CN generation is seen as a more appropriate solution to combat this problem. Currently, there are some studies that are exploring the benefit of using it in user behaviour. Schieb and Preuss (2016) create a computational model that answers general questions about the effects that obstruct or favor the impact of counter-speech. Their results show that the factors that define the success of counter-speech are the proportion of the HS faction and the influence that counter-speakers can exert on the undecided.

The emergence of this task has led to the creation of new datasets aimed at combating HS through CNs. For example, the CONAN datasets, as described in previous research papers (Bonaldi et al., 2022; Fanton et al., 2021; Chung et al., 2019, 2021b), have been developed focused on the counternarrative task. These datasets contain HS messages directed to different hate targets. Additionally, there are other datasets that include comments from various social media platforms. For instance, the study by Garland et al. (2020) presents a corpus designed for classifying both HS and counter-speech, while another research (He et al., 2022) provides a dataset related to HS targeted at Asians and their corresponding CNs during the COVID-19 crisis. Furthermore, Mathew et al. (2018) have introduced a novel dataset that takes an interesting approach by analyzing messages from different user accounts, both hateful and counternarrative, in order to study both account types and their messages.

From a technical aspect, there are several works that explore different methodologies to generate CNs. Zhu and Bhat (2021) propose a system that generates different counterspeech candidates by a generative model, filters the ungrammatical CNs, and finally selects the most relevant with a retrieval method. Studies like Chung et al. (2021c) provided an online platform to monitor and counternarrative the HS against Islamophobia. Qian et al. (2019) achieve this task with three different methods (Seq2Seq, Variational Autoencoders, and Reinforcement Learning). Other studies use different language models, including pre-trained models and Large Language Models (LLMs) (Chung et al., 2021b; Tekiroglu et al., 2022; Lee et al., 2022). Furthermore, other works explore different prompt strategies in LLMs (Ashida and Komachi, 2022; Chung et al., 2021a). The first study includes external knowledge to avoid the hallucinations and the repetitive responses of the models. The second shows the good results of these methods in languages with fewer resources.

In (Chung et al., 2023), a comprehensive in-depth study of the current state of CNs is presented, covering everything from the systems that generate them and their evaluation to the datasets created to develop these systems (analyzing the languages of the resources and the sources from which the data is extracted). Although this review does not include any work in Spanish, a recent study makes a preliminary approach with a small corpus in this language (Vallecillo-Rodríguez et al., 2023). In that paper, experimentation on different LLMs and prompting strategies using a Spanish corpus is presented including GPT-3.5. The results are very promising although they are very limited. Our proposal takes this work as a basis but uses a much larger corpus on eight different targets in order to obtain more varied and robust results. Furthermore, our main objective is to test whether GPT-like LLMs could generate CNs similar to those generated by humans as well as

to compare the performance among existing GPT models. We employ and evaluate two cutting-edge GPT models, namely GPT-3 and GPT-4, for the generation of counter-narratives in Spanish. Our analysis highlights that these models excel not only in generating effective CNs but, on most occasions, outperform human operators in terms of performance.

3. Methodology

This section describes the corpus used as a basis for our experiments, the data processing as well as the models selected in our paper. Finally, we present the evaluation methodology including the different tested perspectives.

3.1. Dataset

To run our experiments, we use the MultiTarget CONAN dataset (CONAN-MT) (Fanton et al., 2021). CONAN-MT is an expert-based "Hate Speech - Counternarrative" dataset constructed through a semi-automatic mechanism. It contains 5,003 HS-CN pairs in English, covering multiple hate targets such as race, religion, country of origin, sexual orientation, disability, and gender. These targets represent various aspects of identity that are often subjected to online HS[2]. Figure 1 shows an example of a HS-CN pair of CONAN-MT dataset. Moreover, Table 1 shows the numbers of pairs for the different targets in CONAN-MT.

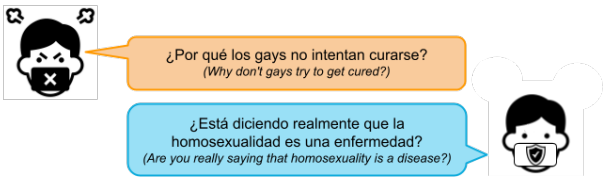

Figure 1: An example of Hate Speech - Counternarrative pair of CONAN-MT dataset.

The reason for using the CONAN-MT corpus is that it is based on the CONAN corpus, one of the reference corpora in this field of research. Moreover, one of the main advantages of CONAN-MT is the diversity and representativeness of the targets present in the corpus. By covering a wide range of targets, it has been possible to create a dataset that more accurately and comprehensively reflects the complexity of online HS. This means that the problem to be addressed becomes more robust and general, making it possible to develop models and algorithms that are able to deal with a variety of scenarios and contexts in which HS manifests itself while allowing for a target-based analysis.

[2]The dataset is publicly available and can be downloaded from the following link https://github.com/marcoguerini/CONAN

Target	#HS-CN Pairs
People with disabilities	220
Jews	594
LGBT+	617
Immigrants	957
Muslims	1,335
People of colour	352
Women	662
Others (Overweight, gypsies...)	266
Total	**5,003**

Table 1: Different targets present in CONAN-MT

3.2. Data Preprocesing

As our work focuses on Spanish, we have used the DeepL API, a Neural Machine Translation service with high accuracy (Hidalgo-Ternero, 2021), in order to automatically translate CONAN-MT from English into Spanish. To guarantee the validity and quality of the translations, a manual review is carried out to verify that the translation obtained is accurate and coherent, and maintains the original meaning of the CNs in the source language.

Because CONAN-MT contains HS-CN pairs in which the HS part is the same for different CNs, we select one of these pairs and eliminate the rest of the pairs as our aim is to generate new CNs.

3.3. Selected Models and Prompt Strategy

In this study, two different language models based on GPT are used to generate CNs. One of the models used is text-davinci-003 of GPT-3 model (Brown et al., 2020), which has proven to be highly effective in previous research, obtaining outstanding results in terms of accuracy for the CN generation task (Vallecillo-Rodríguez et al., 2023). The second model used is GPT-4 (OpenAI, 2023), which is introduced for the purpose of direct comparison with the recognized accuracy of GPT-3. The inclusion of GPT-4 in this study aims to assess whether this new model presents significant improvements in terms of quality and consistency in the generation of CNs, compared to its predecessor, GPT-3. This comparison will allow us to determine whether GPT-4 can overcome the already-known good accuracy obtained by GPT-3 and provide new insights into the generation of CNs.

The hyperparameter configuration for these two models is 0.7 for *temperature*, 1 for *top_p*, and 512 for *max_tokens*.

Since our main objective here is to compare different models and verify that they are valid for us-

ing as decision support systems in a real environment, we have not dedicated effort to designing a good promting strategy. Instead, we have decided to rely on the work of Vallecillo-Rodríguez et al. (2023), in which it was found that the different prompting strategies did not show significant differences in the results, and we have considered it more appropriate to use the most consistent of the strategies used in that work. Specifically, we have used as prompting the task definition together with 8 different examples, one for each of the targets that appear in the corpus that were randomly selected, and finally, the HS for which the model must generate a counternarrative. The prompt used is included in Appendix 8.

Due to the fact that we use GPT models, it is important to note that these experiments were conducted between April and May 2023, as we do not have control over the versions of those models and that could affect reproducibility.

3.4. Evaluation

Evaluating text generation tasks is challenging and lacks a simple solution. First, there is no standard metric for determining what constitutes quality text. In addition, text generation is often subjective and depends on the task or context in which it is used. Furthermore, in most cases, the metrics that have been used in the generation have come from the field of translation, such as BLEU (Papineni et al., 2002) or NIST (Doddington, 2002), or from summaries generation, such as ROUGE (Cawsey et al., 2000). However, these measures are the subject of controversy in the community due to several factors, such as the fact that in Natural Language Generation (NLG) systems, there is no single good result to compare with or that the results yielded by the metrics are difficult to interpret. There are some new metrics based on measuring the distance to the gold standard by embeddings distances, like BERTScore (Zhang et al., 2020) or MoverScore (Zhao et al., 2019). These two metrics overcome the problem of needing exact matching of terms, so they account for synonyms and paraphrasing. Anyhow, when dealing with CNs, the validity is difficult to measure according to a reference or gold standard. These are some of the reasons given for distrusting this type of evaluation. Consequently, a manual assessment of the dataset is carried out. For the evaluation, we followed the work of (Ashida and Komachi, 2022) and considered three perspectives for each CN: Offensiveness, Stance, and Informativeness within four-level scales.

Offensiveness: determines whether the CN is offensive to anyone (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 on the message:

- 0 (irrelevant)
- 1 (strongly agree)
- 2 (slightly agree/disagree)
- 3 (strongly disagree)

Informativeness: assesses how informative and specific the CN is, without being generic:

- 0 (irrelevant)
- 1 (not informative)
- 2 (generic and uninformative)
- 3 (specific and informative)

However, after an initial evaluation, we consider that it would be appropriate to incorporate additional measures to assess the veracity (Truthfulness), the need for possible edits (Editing required), and finally, the comparison between the quality of the CNs generated automatically by the GPT model and the CNs generated by humans (Comparison between H-M). These complementary measures will provide a more complete and accurate vision of the effectiveness and reliability of the generated CNs, as well as the ability of the GPT model to match or exceed the quality of human CNs in terms of coherence, contextual understanding, and relevant content.

Truthfulness: assesses whether what is said in the comment is truthful:

- 0 (not sure)
- 1 (not true)
- 2 (partially true)
- 3 (completely true)

Editing required: assesses whether human editing would be necessary to show CN:

- 0 (no editing)
- 1 (yes editing)

Comparison between H-M: assesses which CN, human or machine generated, is preferred:

- 0 (both CN are equally valid)
- 1 (human generates a better CN)
- 2 (machine generates a better CN)
- 3 (neither CN is good)

We have carried out a preliminary evaluation of 80 HS-CN pairs for the 2 generated datasets (GPT-3 and GPT-4) testing the different proposed perspectives. These 80 pairs are composed of 10 pairs randomly selected from each target (Table 4).

This evaluation is carried out by 3 human annotators (a senior linguist, a junior linguist, and a senior computer scientist). In Table 2, the percentage of agreement between the annotators in the 80 pairs of both models is shown. Analyzing these results we can see a high level of agreement among the annotators. Moreover, we can observe that metrics related to subjectivity such as "Offensiveness" and "Comparison between H-M" have a lower percentage of agreement than metrics that are not so subjective such as "Stance" or "Truthfulness".

Annot. id	Off.	Sta.	Inf.	Tru.	Edi.	H-M.
Annot. 1-Annot. 2	0.969	0,994	0,725	0,994	0,913	0,794
Annot. 1-Annot. 3	0.963	0,975	0,806	0,975	0,931	0,800
Annot. 2-Annot. 3	0.969	0,981	0,775	0,969	0,894	0,725
Annot. 1- Annot. 2-Annot. 3	0.950	0.975	0.869	0.969	0.869	0.869
Total (AVG)	**0.967**	**0.983**	**0.769**	**0.979**	**0.913**	**0.773**

Table 2: Percentage of coincidences between the 3 annotators for the 160 HS-CN selected pairs (80 generated by GPT-3 and 80 generated by GPT-4 model). Inf: informativeness, Sta: stance, Off: offensiveness, Tru: Truthfulness, Edi: Editing required, and H-M: Comparison between H-M.

Table 3 shows the agreement calculated using metrics of weighted agreement such us *Weighted Kappa, Multi Kappa, Krippendorff's alpha, Observed average agreement, and Kendall's W*. Since disagreement between the pair of annotations (0, 3) is not the same as disagreement between (0, 1). The *"observed average agreement"* is, in general, very high. However, the agreement values, except for multi-Kappa and the aspect "Stance" are generally low. This is mainly due to the low variability in the values used in the annotation, i.e., the data have been annotated with very similar values between them, so the possibility of getting it right by chance is high, as some values tend to predominate. Anyhow, considering the very high ratio of coincidences(Table 2), we can discard Kappa results and accept the agreement.

Metric	Off.	Sta.	Inf.	Tru.	Edi.	H-M.
Weighted Kappa	0.091	0.603	0.210	-0.006	0.130	0.228
Multi Kappa	0.707	0.864	0.729	0.664	0.674	0.558
Krippendorff's alpha	0.121	0.593	0.204	-0.008	0.147	0.222
Observed avg. agreement	0.983	0.994	0.919	0.992	0.964	0.847
Kendall's W	3.65e-04	5.21e-05	0.030	1.56e-04	6.77e-04	0.019

Table 3: Results of the weighted agreement metrics applied to our annotations. Inf: informativeness, Sta: stance, Off: offensiveness, Tru: Truthfulness, Edi: Editing required, and H-M: Comparison between H-M.

As seen in Table 4 in Section 4, the best results are obtained with the GPT-4 model, although the performance of GPT-3 is also very good, outperforming human-generated CNs. For this reason, and given that the cost of manual evaluation is very high, it has been decided to carry out only the evaluation of the corpus generated with the GPT-4 model. Thus, only the two annotators with a linguistic profile have performed the rest of the annotation of the corpus generated with the GPT-4 model, which we have named CONAN-MT-SP.

4. Annotation Analysis

This section includes an analysis of the annotation process.

4.1. GPT-3 and GPT-4 comparison

Table 4 shows the results of the 80 HS-CN pairs for the two GPT models. In particular, the GPT-4 model surpasses GPT-3 in all the evaluated parameters and can be regarded as an outstanding system with remarkable results. Nevertheless, GPT-3 also demonstrates a highly positive performance, with only a few minor cases that might be considered as errors.

In general, the CNs generated in GPT-3 exhibit a very good performance. They are generally non-offensive (with only one CN considered potentially offensive), tend to disagree with abusive comments (also only one CN is annotated as "Slightly agree"), and all the CNs are considered "Completely True". Furthermore, they require minimal editing (only 3 out of 80 CNs generated). Perhaps the worst measure is in informativeness, with 1 case being not informative and 10 cases being considered generic and uninformative compared to 69 cases being specific and informative.

A separate case can be considered the automatically generated or human-generated CN comparison measure. In the case of GPT-4 in 100% of the cases the generated CN is considered better than the one created by a human (the one found in the original CONAN-MT corpus and which has been automatically translated into Spanish). This is an exceptional result but it should be noted that the GPT-3 model also performs very well. Specifically, in only 2 out of 80 instances, the human is considered to have generated a better CN than the computational model.

Thus, although GPT-4 works best, it is important to note that either model can serve as an excellent decision support system, assisting human operators in CN generation.

4.2. CONAN-MT-SP corpus

Since the results with GPT-4 are outstanding, we have decided to create a high quality Spanish counter-narrative corpus of HS-CN pairs. For this, we have used the HS translated directly from the original CONAN-MT corpus and the CN generated by the GPT-4 model. In addition, we have included in the corpus the manual evaluation carried out by

Perspective	Label	GPT-3		GPT-4	
		#Instance	Percentage	#Instance	Percentage
Offensiveness	Not sure	0	0.00%	0	0.00%
	Not Offensive	79	98.75%	80	100.00%
	Maybe Offensive	1	1.25%	0	0.00%
	Completely offensive	0	0.00%	0	0.00%
Stance	Irrelevant	0	0.00%	0	0.00%
	Strongly agree	0	0.00%	0	0.00%
	Slightly agree/disagree	1	1.25%	0	0.00%
	Strongly disagree	79	98.75%	80	100.00%
Informativeness	Irrelevant	0	0.00%	0	0.00%
	Not Informative	1	1.25%	0	0.00%
	Generic and Uninformative	10	12.5%	2	2.50%
	Specific and Informative	69	86.25%	78	97.50%
Truthfulness	Not Sure	0	0.00%	0	0.00%
	Not True	0	0.00%	0	0.00%
	Partially True	0	0.00%	0	0.00%
	Completely True	80	100.00%	80	100.00%
Editing required	No Editing	77	96.25%	80	100.00%
	Yes Editing	3	3.75%	0	0.00%
Comparison between H-M	Both CN equally valid	10	12.25%	0	0.00%
	Human better than Computer	2	2.50%	0	0.00%
	Computer better than Human	67	83.75%	80	100.00%
	Neither CN good	1	1.25%	0	0.00%

Table 4: Manual evaluation of 80 pairs selected CN-HS

the annotators. The corpus generated has been named CONAN-MT-SP and it is composed of a total of 3,636 instances annotated. CONAN-MT-SP is publicly available and can be downloaded from the following link (will be included if the paper is accepted). Figure 2 shows the corpus generation process.

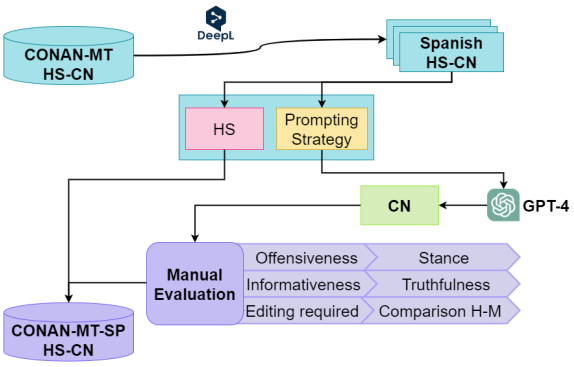

Figure 2: CONAN-MT-SP generation process.

The results of the manual evaluation of CONAN-MT-SP are presented in Table 5. As we can see in the percentage of the total corpus, the CNs generated in CONAN-MT-SP show very good results, where more than 98% are “Not Offensive”, ”Strongly disagree” with what the HS message says, provide “Completely True” information, and do not need editing (“No Editing”), being in these cases a minimum of counter-narratives (less than 1.8%) that are considered as “Maybe Offensive”, “Slightly agree”, include partially true information or need editing. Looking at the informativeness of the texts, we see that 79.51% of them are “Specific and Informative”, and only 0.06% of CNs are “Not informative”. Finally, in the comparison with the human, we see that in only 0.94% the human generates better CNs than those generated by GPT-4, which is less than 1%. Based on these values, we can conclude that CONAN-MT-SP is a high quality corpus for CN generation.

Examining the results based on the target percentages, we see that over 99% of CNs are ”Not Offensive”, ”Strongly Disagree”, and contain ”Completely True” information, for all classes except for Others where some CNs are considered as ”Maybe offensive” (2.29%), ”Slightly agree” (2.29%), and contain ”Partially True” information (1.71%). In terms of informativeness, Jews, LGBT+, POC, and Others have less than 85% of CNs being ”Specific Information”, with Jews containing the most generic information CNs (42.75%). Regarding editing, more than 93% of the generated texts do not need editing for all of the classes. POC target is the ones that need the most editing (6.19%) followed by Others (4.57%). In the comparison between H-M, over 94% favor the machine or are as good as humans, except for Muslims, POC, and Others, which have instances where neither is effective. In summary, the most effective CNs are generated for ”Migrants” and ”Muslims,” while ”Others” and ”POC” appear to yield less satisfactory results.

Perspective	Label	Percentage of each target (%)								% Total
		Disabled	Jews	LGBT+	Migrants	Women	Muslims	POC	Others	
Offensiveness	Not sure	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Not Offensive	99.40	100.00	99.78	99.84	99.82	99.69	99.31	97.71	99.64
	Maybe Offensive	0.60	0.00	0.22	0.16	0.18	0.31	0.69	2.29	0.36
	Completely offensive	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Stance	Irrelevant	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Strongly agree	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Slightly agree/disagree	0.60	0.25	0.89	0.00	0.18	0.31	0.69	2.29	0.44
	Strongly disagree	99.40	99.75	99.11	100.00	99.82	99.69	99.31	97.71	99.56
Informativeness	Irrelevant	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Not Informative	0.00	0.00	0.00	0.00	0.00	0.00	0.34	0.57	0.06
	Generic and Uninformative	14.88	42.75	36.00	9.95	10.65	14.77	21.99	30.86	20.43
	Specific and Informative	85.12	57.25	64.00	90.05	89.35	85.23	77.66	68.57	79.51
Truthfulness	Not Sure	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	Not True	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.57	0.03
	Partially True	0.00	0.25	0.67	0.32	0.36	0.93	0.69	1.71	0.61
	Completely True	100.00	99.75	99.33	99.68	99.64	99.07	99.31	97.71	99.37
Editing required	No Editing	99.40	100.00	99.78	99.52	99.10	97.00	93.81	95.43	98.21
	Yes Editing	0.60	0.00	0.22	0.48	0.90	3.00	6.19	4.57	1.79
Comparison between H-M	Both CN equally valid	5.95	1.72	0.00	1.12	14.62	6.10	4.12	1.71	4.92
	Human better than Computer	0.00	0.00	0.00	0.32	0.72	1.96	1.03	3.43	0.94
	Computer better than Human	94.05	98.28	100.00	98.56	84.66	91.21	94.16	93.71	93.84
	Neither CN good	0.00	0.00	0.00	0.00	0.00	0.72	0.69	1.14	0.30
# Instances		168	407	450	623	554	968	291	175	3636

Table 5: Manual evaluation results of the CONAN-MT-SP. These results show the percentage of counter-narratives per target that belong to each metric evaluated. Additionally, the total percentage of CONAN-MT-SP texts that have been labeled in each metric is included.

5. Linguistic Analysis

Additionally, we have conducted a linguistic and statistical analysis for the CN generated according to the model used and the hate targets. We have calculated the average number of words, characters, and the 6 main types of Part-Of-Speech (Nouns, Verbs, Adverbs, Adjectives, Subordinate Conjunctions, and Coordinate Conjunctions).

Tables 6 and 7 show the results of the evaluation of linguistic features for the GPT models and the hate target, respectively. As we can see in Table 6, GPT-4 produced more words than GPT-3, and this difference between the number of words affects the number of chars and POS tags. On average, GPT-3 has the middle of sentences than GPT-4. In addition, we can observe that both models generate more coordinated sentences than subordinates. This may indicate that the generated sentences are less complex. Moreover, the coordinate generation sentences are more frequent in GPT-4.

Analyzing the results between targets of Table 7, we can conclude the CN generated to answer posts directed to migrants and Muslims has more words than the rest and LGBT+ is the target with fewer words on average. If we observe the difference between the two types of conjunctions, the number of coordinates is higher than the subordinate. This could mean that in the sentences generated, coordinated sentences predominate over

	AVG Words	AVG Chars	AVG POS					
			VERB	NOUN	ADV	ADJ	CCONJ	SCONJ
GPT-3	29.91 ± 7.62	181.10 ± 44.51	2.57 ± 1.56	7.14 ± 2.43	1.21 ± 0.99	2.36 ± 1.45	1.65 ± 0.94	0.75 ± 0.88
GPT-4	47.94 ± 9.40	301.72 ± 58.26	4.84 ± 1.96	11.54 ± 2.87	1.84 ± 1.14	4.59 ± 2.08	3.90 ± 1.29	0.99 ± 0.94

Table 6: The average of the number of Words, Chars and POS Tags of the CNs generated by each GPT model. Next to the average is the standard deviation of these metrics.

	AVG words	AVG chars	AVG POS					
			VERB	NOUN	ADV	ADJ	CCONJ	SCONJ
Disabled	38.80 ± 11.73	248.85 ± 73.41	4.25 ± 2.00	9.74 ± 3.52	1.28 ± 1.05	3.38 ± 1.89	2.62 ± 1.52	0.93 ± 0.90
Jews	38.41 ± 13.11	239.02 ± 84.20	3.36 ± 1.97	9.23 ± 3.70	1.61 ± 1.02	3.86 ± 2.39	2.89 ± 1.72	0.80 ± 0.85
LGBT+	35.88 ± 11.35	223.74 ± 73.40	3.32 ± 2.00	8.42 ± 3.26	1.74 ± 1.13	3.36 ± 2.02	2.53 ± 1.51	0.81 ± 0.92
Migrants	40.24 ± 12.31	254.03 ± 80.30	4.32 ± 2.21	9.85 ± 3.47	1.40 ± 1.13	3.39 ± 2.00	2.72 ± 1.49	0.88 ± 0.95
Women	38.26 ± 11.72	231.56 ± 73.17	3.74 ± 2.02	9.82 ± 3.48	1.43 ± 1.14	2.97 ± 2.01	2.70 ± 1.54	0.88 ± 0.93
Muslims	40.74 ± 12.99	250.33 ± 83.03	3.59 ± 2.09	9.37 ± 3.36	1.51 ± 1.12	3.65 ± 2.13	2.88 ± 1.63	0.97 ± 0.95
POC	36.90 ± 11.72	229.49 ± 73.20	3.42 ± 2.06	8.75 ± 3.07	1.79 ± 1.10	3.61 ± 2.13	2.84 ± 1.64	0.63 ± 0.82
Others	38.77 ± 12.38	241.96 ± 80.79	3.77 ± 2.16	9.21 ± 3.52	1.44 ± 1.01	3.67 ± 2.12	2.93 ± 1.69	0.88 ± 0.90

Table 7: The average of the number of Words, Chars and POS Tags of the CNs generated for each hate target. Next to the average is the standard deviation of these metrics.

subordinate sentences and, consequently, their complexity decreases. In addition, to show the significance of the information reported by the calculated metrics (number of words, chars, and POS tags) in Appendix 8, we present the results of applying significance tests to these metrics.

With regard to linguistic errors in the CNs generated, it should be noted that minimal errors have been found including some lexical, orthographic, and grammar mistakes that affect the coherence and comprehension of CNs texts.

Finally, and after a general analysis of the corpus, we consider that the GPT models generate CNs that we can classify into 3 types: Not Valid (NV), Default (D), and High-Quality (HQ) CN. Table 8 shows examples for each type of CN.

First, we find some types of CN that we call "Not Valid CNs". These would be CNs that do not understand the HS and therefore do not generate a correct CN (see example NV1 in Table 8).

Secondly, we have "Default CN", in which the structure of the CN is similar regardless of the topic of the HS. Therefore, they would not be specific and would contain empty information, since they do not add new information. In this part, we can find two types of default CN. On the one hand, CNs in which a number of elements are substituted but the structure is the same in different topics. D1 example (Table 8) shows how the structure is repeated in the CNs and the same idea is also repeated. On the other hand, we can find CNs without a specific subject; therefore, they can serve several subjects (D2).

Finally, we consider "High-Quality CNs", which are those meeting specific criteria, including informativeness consistent with the HS they combat and adherence to the principle of linguistic economy(HQ1).

	id	Hate-Speech	GPT-4 Counternarrative
Not Valid (NV)	NV1	Los lameculos no son capaces de alimentarse correctamente, por lo que someten al servicio sanitario a un estrés innecesario (*The lard-asses are not able to eat properly, this way they put the healthcare service unnecessarily under stress.*)	Esta afirmación es injusta y carece de fundamento. Las personas tienen diferentes habilidades y estilos de vida, y no se puede generalizar sobre su capacidad para cuidar de sí mismas o el impacto que tienen en el sistema de salud. En lugar de juzgar y menospreciar a los demás, deberíamos fomentar la empatía y el apoyo mutuo. (*This statement is unfair and unfounded. People have different abilities and lifestyles, and one cannot generalize about their ability to care for themselves or the impact they have on the healthcare system. Instead of judging and belittling others, we should encourage empathy and mutual support.*)
Default (D)	D1	Los africanos no tienen derecho a hablar su lengua y son incapaces de establecer una civilización compleja como la nuestra (*Africans have no right to speak their language and are incapable of establishing a complex civilisation such as ours.*)	Este tipo de generalizaciones son incorrectas y perjudiciales. Las personas de cualquier raza o etnia pueden tener comportamientos buenos o malos, y no se debe juzgar a un grupo entero basándose en estereotipos negativos. En lugar de perpetuar estos mitos, debemos centrarnos en promover la comprensión y la aceptación entre las diferentes comunidades (*Such generalizations are incorrect and harmful. People of any race or ethnicity can have good or bad behaviors, and an entire group should not be judged based on negative stereotypes. Instead of perpetuating these myths, we should focus on promoting understanding and acceptance among different communities*)
	D2	Los negros son genéticamente diferentes de nosotros. Pero hay un nuevo racismo que empieza a arraigar en nuestro país. Se trata de seguir molestando a los blancos con cosas de discriminación de los negros (*Black people are genetically different from us. But there is a new racism starting to take root in our country. It is about keeping bothering whites with black discrimination stuff*)	Todos merecen derechos y oportunidades, independientemente de su género, orientación sexual, raza o religión. La diversidad enriquece nuestra sociedad y nos permite aprender unos de otros. Juntos, podemos trabajar en un mundo más inclusivo y equitativo. (*Everyone deserves rights and opportunities, regardless of gender, sexual orientation, race or religion. Diversity enriches our society and allows us to learn from each other. Together, we can work toward a more inclusive and equitable world.*)
High Quality (HQ)	HQ1	Los judios controlan los bancos, los medios de comunicacion, hollywood, incluso el gobierno de los estados unidos. (*Jews control the banks, media, hollywood, even the u.s. government.*)	El Holocausto fue una tragedia histórica en la que millones de personas, incluyendo judíos, fueron asesinados de manera sistemática e inhumana. Utilizar este término para difamar a una comunidad entera no solo es falso, sino también irrespetuoso y dañino. Debemos aprender de la historia y promover la tolerancia y la comprensión en lugar de perpetuar el odio y el prejuicio. (*The Holocaust was a historic tragedy in which millions of people, including Jews, were systematically and inhumanely murdered. Using this term to defame an entire community is not only untrue, but also disrespectful and harmful. We must learn from history and promote tolerance and understanding instead of perpetuating hate and prejudice.*)

Table 8: Examples of the three types of CN generated (Not Valid, Default, and High Quality).

6. Conclusions

We have conducted several experiments to compare two GPT-based LLMs for generating CNs, and both have demonstrated outstanding performance. While we have identified minor grammatical errors that can be easily rectified, these findings support the feasibility of using such systems as support tools for non-governmental organizations (NGOs) aiming to counter HS on social media. As a result, a new corpus for researching in counternarrative in Spanish has been released.

However, we must take into account that these GPT models present hard problems to tackle. Firstly, their non-open source nature limits accessibility, as not everyone can afford the usage fees associated with these models. Furthermore, their black-box operation prevents the ability to fine-tune or predict their outputs, significantly reducing their reliability and use cases.

As a next step, we plan to study other open-source systems such as LLaMA-2 (Touvron et al., 2023) or Vicuna (Zheng et al., 2023) that can be trained and adjusted using our own resources. Additionally, we will integrate the manually evaluated resource CONAN-MT-SP, which we believe to be of high quality. Finally, we will explore various prompting strategies in future experiments to determine if we can further improve our results.

These challenges and upcoming steps will allow us to move towards more accessible, adaptable, and reliable solutions in the fight against HS on social media.

7. 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. Thanks to the REAL-UP project funded by the European Commission (CERV-2021-EQUAL, ID: 101049673) and to the National Office for the Fight against Hate Crimes (ONDOD) for supporting this research.

8. Bibliographical References

Mana Ashida and Mamoru 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, Seattle, Washington (Hybrid). Association for Computational Linguistics.

Helena Bonaldi, Sara Dellantonio, Serra Sinem Tekiroğlu, and Marco Guerini. 2022. Human-machine collaboration approaches to build a dialogue dataset for hate speech countering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8031–8049, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH 2021)*, pages 17–25, Online. Association for Computational Linguistics.

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

Y. Chung, S. S. Tekiroğlu, and M. Guerini. 2021a. Italian counter narrative generation to fight online hate speech. In *Proceedings of the Seventh Italian Conference on Computational Linguistics*, Bologna, Italy. Torino: Accademia University Press. Dell'Orletta, F., Monti, J., & Tamburini, F. (Eds.).

Yi-Ling Chung, Gavin Abercrombie, Florence Enock, Jonathan Bright, and Verena Rieser. 2023. Understanding counterspeech for online harm mitigation.

Yi-Ling Chung, Elizaveta Kuzmenko, Serra Sinem Tekiroglu, and Marco Guerini. 2019. CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2819–2829, Florence, Italy. Association for Computational Linguistics.

Yi-Ling Chung, Serra Sinem Tekiroğlu, and Marco Guerini. 2021b. Towards knowledge-grounded counter narrative generation for hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 899–914, Online. Association for Computational Linguistics.

Yi-Ling Chung, Serra Sinem Tekiroğlu, Sara Tonelli, and Marco Guerini. 2021c. Empowering NGOs in countering online hate messages. *Online Social Networks and Media*, 24:100150.

George Doddington. 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.

Margherita Fanton, Helena Bonaldi, Serra Sinem Tekiroğ lu, and Marco Guerini. 2021. Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Association for Computational Linguistics.

Joshua Garland, Keyan Ghazi-Zahedi, Jean-Gabriel Young, Laurent Hébert-Dufresne, and Mirta Galesic. 2020. Countering hate on social media: Large scale classification of hate and counter speech. In *Proceedings of the Fourth Workshop on Online Abuse and Harms*, pages 102–112, Online. Association for Computational Linguistics.

Bing He, Caleb Ziems, Sandeep Soni, Naren Ramakrishnan, Diyi Yang, and Srijan Kumar. 2022. Racism is a virus: Anti-asian hate and counter-speech in social media during the covid-19 crisis. In *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '21, page 90–94, New York, NY, USA. Association for Computing Machinery.

Carlos Manuel Hidalgo-Ternero. 2021. Google translate vs. deepl: analysing neural machine translation performance under the challenge of phraseological variation. *MonTI. Monografías de Traducción e Interpretación*, pages 154–177.

Sameer Hinduja and Justin Patchin. 2018. Connecting adolescent suicide to the severity of bullying and cyberbullying. *Journal of School Violence*, 18:1–14.

Huije Lee, Young Ju Na, Hoyun Song, Jisu Shin, and Jong Park. 2022. ELF22: A context-based counter trolling dataset to combat Internet trolls. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3530–3541, Marseille, France. European Language Resources Association.

Binny Mathew, Navish Kumar, Ravina, Pawan Goyal, and Animesh Mukherjee. 2018. Analyzing the hate and counter speech accounts on twitter.

Richard Miller, Katrina Liu, and Arnetha F. Ball. 2020. Critical counter-narrative as transformative methodology for educational equity. *Review of Research in Education*, 44(1):269–300.

OpenAI. 2023. Gpt-4 technical report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing 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.

Flor Miriam Plaza-del Arco, M Dolores Molina-González, L Alfonso Ureña-López, and M Teresa Martín-Valdivia. 2021. Comparing pre-trained language models for spanish hate speech detection. *Expert Systems with Applications*, 166:114120.

Jing Qian, Anna Bethke, Yinyin Liu, Elizabeth Belding, and William Yang Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4755–4764, Hong Kong, China. Association for Computational Linguistics.

Carla Schieb and Mike Preuss. 2016. Governing hate speech by means of counterspeech on facebook.

Serra Sinem Tekiroglu, Helena Bonaldi, Margherita Fanton, and Marco Guerini. 2022. Using pre-trained language models for producing counter narratives against hate speech: a comparative study.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj

Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Maria Estrella Vallecillo-Rodríguez, Arturo Montejo-Raéz, and Maria Teresa Martín-Valdivia. 2023. Automatic counter-narrative generation for hate speech in spanish. *Procesamiento del Lenguaje Natural*, 71(0):227–245.

Zeerak Waseem and Dirk Hovy. 2016. Hateful symbols or hateful people? predictive features for hate speech detection on Twitter. In *Proceedings of the NAACL Student Research Workshop*, pages 88–93, San Diego, California. Association for Computational Linguistics.

Marcos Zampieri, Shervin Malmasi, Preslav Nakov, Sara Rosenthal, Noura Farra, and Ritesh Kumar. 2019. SemEval-2019 Task 6: Identifying and Categorizing Offensive Language in Social Media (OffensEval). In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 75–86, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. 2019. Moverscore: Text generation evaluating with contextualized embeddings and earth mover distance.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena.

Wanzheng Zhu and Suma Bhat. 2021. Generate, prune, select: A pipeline for counterspeech generation against online hate speech. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 134–149, Online. Association for Computational Linguistics.

Appendix A. Used Prompt

This section presents the prompt used to generate the counter-narratives with the GPT-3 and GPT-4 models. This prompt and its translation to English are shown in Table 9.

Appendix B. Significance Tests

In this section we present the results of the significance test of the calculated metrics of the counter narratives (number of Words, Characters and POS tags, such as Verbs, Nouns, Adjectives, Adverbs, Coordinating and Subordinating Conjunctions).

To calculate these tests, the first step is to see whether the distribution of the metrics follows a normal distribution. In affirmative case, we calculate three parametric significance tests, Student's t-test, paired Student's t-test, and ANOVA. Otherwise, we calculated three nonparametric tests, mannwhitneyu, wilcoxon and kruskal. We considered for each test a p-value of 0.05. The p-value of a significance test represents the significance level. If the p-value is below the significance level, the test indicates that there is sufficient evidence to reject the null hypothesis and that the samples are probably from populations with different distributions.

The results of the significance test calculation can be found at the following link: https://github.com/sinai-uja/CONAN-MT-SP/blob/main/significanceTest-experiments-LREC2024.xlsx. The results of these tests show that there are significant differences in the language used in almost all cases, with 95% confidence. Specifically, the counter-narratives associated with hate messages towards LGTB+, Muslims and Women, are the ones that show the most differences. Another aspect visualized by the significance tests is that the linguistic features that differ the most are in the use of verbs, adverbs and nouns (i.e., in the number of "content" words used).

Spanish prompt	English translation
Genera la contranarrativa, es decir, una alternativa positiva, de los siguientes comentarios ofensivos: Post: Las personas con síndrome de Down son retrasadas, por lo que son más propensas a abusar de las drogas y el alcohol. Contranarrativa: Nacer con síndrome de Down no está relacionado con el mal comportamiento. Exactamente igual que nacer en una silla de ruedas no está relacionado con un mal comportamiento. No hay correlación. Post: Los judíos no son una religión, nacen así y nunca se curarán de su enfermedad. Contranarrativa: No estoy de acuerdo, los judíos son personas religiosas que siempre han formado parte de la comunidad. Y han sido perseguidos por ello. Creo que esto es lo que querías decir con ”no se curarían”. Post: La homosexualidad y la bisexualidad son una enfermedad, la heterosexualidad es normal. Contranarrativa: Estas afirmaciones no son ciertas ni útiles y simplemente no reflejan la realidad tal y como yo la he vivido. Post: La inmigración es sólo una forma de crear más pobreza y desempleo en nuestro país. Contranarrativa: La pobreza y el desempleo son problemas reales en el Reino Unido y en todo el mundo. Dudo que esto tenga algo que ver con los inmigrantes. Post: La mayoría de las mujeres son infrahumanas, no se les puede confiar altos cargos. Contranarrativa: Me parece que es un comentario contra TODAS las mujeres, no sólo contra algunas. Post: Los musulmanes no tienen cabida en la Europa civilizada. Contranarrativa: En una sociedad democrática moderna como la europea, debe existir un equilibrio entre las distintas opiniones y el respeto a las diferentes creencias. Post: ¿Dónde está el amor en la Gran Bretaña blanca cuando lo necesitamos? Contranarrativa: Mi fe no dicta que no pueda amar y apoyar a mi pareja, pero no apoya que sea homófobo o racista. Post: El Gobierno del Reino Unido ha publicado recientemente un documento sobre ”Daños en línea”, en otras palabras, van a trabajar con organizaciones judías y musulmanas contra el odio para censurar Internet. (En referencia a Tell Mama, CST y Stop Hate UK) Contranarrativa: Internet se ha convertido en una plataforma de gritos para las opiniones odiosas de la gente. Por supuesto, hay que hacer algo, ya que la gente no parece capaz de moderar sus palabras por sí misma.	*Generate the counter-narrative, i.e., a positive alternative, to the following offensive comments:* *Post: People with Down syndrome are retarded, so they are more likely to abuse drugs and alcohol.* *Counter-narrative: Being born with Down syndrome is not related to bad behavior. Exactly like being born in a wheelchair is not related to bad behavior. There is no correlation.* *Post: Jews are not a religion, they are born that way and will never be cured of their disease.* *Counter-narrative: I disagree, Jews are religious people who have always been part of the community. And they have been persecuted for it. I think this is what you meant by ”they would not be cured”.* *Post: Homosexuality and bisexuality are a disease, heterosexuality is normal.* *Counter-narrative: These statements are not true or helpful and simply do not reflect reality as I have experienced it.* *Post: Immigration is just a way to create more poverty and unemployment in our country.* *Counter-narrative: Poverty and unemployment are real problems in the UK and around the world. I doubt this has anything to do with immigrants.* *Post: Most women are subhuman, they cannot be trusted with high positions.* *Counter-narrative: It seems to me that this is a comment against ALL women, not just some.* *Post: Muslims have no place in civilized Europe.* *Counter-narrative: In a modern democratic society like Europe, there must be a balance between different opinions and respect for different beliefs.* *Post: Where is the love in white Britain when we need it?* *Counter-narrative: My faith does not dictate that I cannot love and support my partner, but it does not support me being homophobic or racist.* *Post: The UK Government has recently published a document on ”Online Harm”, in other words, they are going to work with Jewish and Muslim anti-hate organizations to censor the Internet. (In reference to Tell Mama, CST and Stop Hate UK).* *Counter-narrative: the Internet has become a shouting platform for people’s hateful opinions. Of course, something has to be done, as people don’t seem to be able to moderate their words themselves.*

Table 9: Used prompt to generate the counternarratives with GPT-3.5 and GPT-4.