

# REPORT ON TOOLS AND STRATEGIES FOR GENERATING COUNTER NARRATIVES USING NLP



Catalogue of publications by the General State Administration  
<https://cpage.mpr.gob.es>



© Ministry of Inclusion, Social Security and Migration  
Madrid, 2024

Author: María Teresa Martín Valdivia

Published and distributed by: Spanish Observatory on Racism and Xenophobia  
María de Guzmán St. 52. 3rd floor. 28003 Madrid  
[oberaxe@inclusion.gob.es](mailto:oberaxe@inclusion.gob.es)  
<https://www.inclusion.gob.es/oberaxe/es/index.htm>

NIPO: 121-24-007-3

Design: Solana e Hijos, A. G., S. A. U

Layout: Diseño Gráfico Gallego y Asociados, S. L.

The information and opinions in this document are the responsibility of its author and do not necessarily reflect the official position of the Ministry of Inclusion, Social Security and Migration.

# TABLE OF CONTENTS

<b>1</b>	<b>Introduction</b>	<b>4</b>
<b>2</b>	<b>Previous work</b>	<b>7</b>
<b>3</b>	<b>Works in other languages</b>	<b>11</b>
<b>4</b>	<b>Works in Spanish</b>	<b>12</b>
<b>5</b>	<b>First article for generating counter narratives in Spanish</b>	<b>13</b>
5.1.	Language models	14
5.2.	Prompting strategies	14
5.3.	CONAN-SP	15
5.4.	Assessment	17
5.5.	Error analysis	19
<b>6</b>	<b>Preliminary work with other models</b>	<b>22</b>
6.1.	Experimenting with GPT-4	22
6.1.1.	Generating CONAN-MT-SP	22
6.1.2.	Assessment	24
6.1.3.	Results	26
6.1.4.	Conclusion and future work	31
6.2.	Experimentation with LLaMA (Large Language Model Meta AI)	32
<b>7</b>	<b>Related projects</b>	<b>33</b>
	<b>Bibliography</b>	<b>35</b>
	<b>Appendix 1</b>	<b>38</b>
	Prompt Experiment 1	38
	Prompt Experiment 2	39
	Prompt Experiment 3	41
	<b>Appendix 2</b>	<b>43</b>
	Prompt Experiment 1 for all the models	43

# 1 Introduction

This report seeks to summarise the strategies followed to generate counter narratives, paying special attention to automatic systems based on machine learning (ML) and Natural Language Processing (NLP).

The well-known expansion of social interactions through digital platforms has led to inappropriate behaviour on the Internet, such as the spread of hate speech among platform users. Freedom of expression on these media has exposed their users to publications that are sometimes used to denigrate, insult or hurt with language, whether mild or rude, on grounds of gender, race, religion, ideology or other personal characteristics. Unfortunately, this type of communication can be very harmful, causing negative psychological effects among users, especially among young people, in the form of anxiety, the feeling that they are being cyber-bullied, and even suicide in the most extreme cases.

This problem mainly involves governments and online platforms, which must adopt measures in the form of laws and policies to help promote healthy coexistence on these media. For example, since 2013, the European Council has promoted the “No Hate Speech”<sup>1</sup> movement to mobilise young people to combat hate speech and promote human rights on the Internet. In May 2016, the European Commission reached an agreement with Facebook, Microsoft, Twitter and YouTube, and signed a “Code of Conduct on Countering Illegal Hate Speech Online”.<sup>2</sup> Based on this Code, the “Protocol to Counter Illegal Hate Speech Online”<sup>3</sup> has been drafted in Spain. Between 2018 and 2020, other platforms such as Instagram, Snapchat, Dailymotion and TikTok signed up to the European Commission’s Code of Conduct.

According to the 2019 report of the Ministry of the Interior<sup>4</sup> on the trend in hate crimes in Spain, threats, insults and discrimination were the most frequent criminal acts, with the Internet (54.9%) and social media (17.2%) being the most used means to commit these actions.

It seems clear that the problem of detecting hate speech has worsened in recent years and that it is now necessary to study, analyse and implement solutions in all areas, including in the area of language technologies. Analysing this type of harmful content on the Internet requires automatic systems capable of processing and analysing human language. This is why detecting and analysing hate speech has become one of the main areas of research in Natural Language Processing (NLP). NLP is an important area of Artificial Intelligence that attempts to understand and generate language as humans do using computational methods. Moreover,

1 <https://www.coe.int/en/web/committee-on-combating-hate-speech/home>

2 [https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online\\_en](https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/eu-code-conduct-countering-illegal-hate-speech-online_en)

3 [https://www.inclusion.gob.es/oberaxe/ficheros/ejes/discursoodio/PROTOCOLO\\_DISCURSO\\_ODIO.pdf](https://www.inclusion.gob.es/oberaxe/ficheros/ejes/discursoodio/PROTOCOLO_DISCURSO_ODIO.pdf)

4 <http://www.interior.gob.es/documents/642012/3479677/informe+evolucion+2019/631ce020-f9d0-4feb-901c-c3ee0a777896>

the use of Machine Learning (ML) algorithms is making it possible to develop classification systems which, combined with advanced NLP techniques, help to tackle many of today's social problems, including the detection of offensive language and hate speech on social media.

In fact, the colossal quantity of social media data shared online each day requires hate mitigation to be addressed through reliable, efficient and scalable tools capable of generating discourse to combat this problem. Efforts have recently been made to collect anti-hate datasets and to automate the production of counter-discourses. However, this field of research has been little studied so far, and there remain many open questions about the most effective approaches and methods, and about how to assess them.

It is therefore important to highlight that so as to combat hate speech; attention must not only be paid to detecting it, but also to strategies to alleviate its consequences. One of the strategies traditionally used includes eliminating information deemed inappropriate or harmful (Roberts, 2016). Thus, for example, platforms such as Facebook remove millions of harmful publications each year, with the help of Artificial Intelligence (AI) tools. Although removing this content can immediately reduce the quantity of harmful messages, it can also lead to accusations of censorship and may not be effective at curbing hatred over the long term (Schieb and Preuss, 2016).

An alternative approach is to respond with an alternative discourse, that is, with responses that seek to refute hateful language using thoughtful and convincing reasons, and fact-based arguments. This has proven effective at influencing the behaviour of both aggressors and bystanders witnessing interactions, as well as in supporting victims (Benesh, 2014). This is what is known as **Counter Narrative**<sup>5</sup> and it is a polite and unaggressive response to hate speech that aims at countering extreme statements. A Counter Narrative (CN) is a direct response to hate speech and is considered an effective tool because:

- i. preserves the right to freedom of expression,
- ii. corrects stereotypes and misleading information with credible evidence,
- iii. can alter the viewpoints of perpetrators of hate and bystanders,
- iv. and it encourages the exchange of opinions among potential bystanders to change their perspectives.

In short, it is important to combat hate speech by not only detecting it but also through strategies to mitigate its consequences. Counter narratives are one of these strategies and consist of providing an alternative narrative to one that fosters hatred and violence. The counter narrative can include messages that foster tolerance, respect, and inclusion, and can be a powerful tool for challenging and dismantling hate speech (Mathew et al., 2019). This is why some Non-Governmental Organisations (NGOs) and public authorities make efforts to train operators to monitor social media platforms and to manually generate effective counter narratives where necessary. However, the fight against hate faces a series of challenges that are not easy to resolve. First, due to the enormous volume of hate speech that occurs daily on social media, manual intervention using counterarguments is not scalable and in most cases unfeasible. On the other hand, to

---

<sup>5</sup> Other terms used are Counter Stories, Counter Arguments and Counter Discourse.

compose a counter narrative, an expert operator typically reads hate messages, searches for and filters related information on the Internet, and then composes a counter narrative. This whole process can take several minutes, besides the psychological and mental harm of having people being exposed to this type of message for long periods of time.

This is a complex problem that has recently received the attention of researchers in the field of NLP as they have begun to study various computational techniques based on the latest developments in NLP and ML aimed at assisting these NGO operators to help them to generate counter narratives automatically, thus decreasing the time and effort in the fight against online hate. However, the problem entails many challenges and difficulties that we will discuss below.

This report examines the various methods that have been used to make the counternarrative and sets out the current situation both nationally and internationally. The use of large language models (LLM) for this task is also shown.



# 2 Previous work

Although there is a large bibliography concerned with detecting hate speech and offensive language, not only in English but also in other languages, including Spanish, in the case of generating counter narratives, there are few works, most of which are focused on the English language.

A summary of the state of play with the issue of detecting and tackling hate speech can be found in the papers of Poletto et al. (2020) and Jahan and Oussalah (2023). For languages other than English, the work of Plaza-del-Arco et al. (2021) reviews various pre-trained language models to detect hate speech in Spanish.

On the topic of generating counter narratives, it must be noted that the research is very recent and there are not yet many studies that address the problem, and those that have done so focus mainly on generating counter narratives in English; there are only a few on other languages.

To get an overview of the problem, mention might be made of two works that compile research on the subject. The first of these is the thesis supervised by Professor Marco Guerini of the Bruno Kessler Foundation (in Trento, Italy) and defended by Yi Ling Chung (Chung, 2022), which provides an in-depth study of the question and sets out many of the ideas and models that have been used to date. Secondly, the compilation of works presented in Alsagheer et al. (2022) includes more than 60 publications related to the counter narrative on social media.

One of the first works to posit the benefits of using counter narratives to mitigate the effects of hate speech can be found in Benesh (2014). In this paper, several strategies opposed to the removal and censorship of hate speech can be found, as well as clear definitions in the matter.

From the point of view of NLP, there are several studies that explore the possibility of automatically generating counter narratives or of using strategies that combine human intervention to counteract hate and harmful speech on the Internet.

Qian et al. (2019) were among the first to attempt automatic generation of counter narratives (CN: Counter Narratives). They created a resource of 10,243 counter narratives to respond to 5,257 instances of hate speech (HS: Hate Speech) taken from 5,020 conversations containing 22,324 Reddit comments and 31,487 counter narratives to respond to 14,614 examples of hate speech in 11,825 conversations that contained 33,776 Gab posts. They used crowdsourcing<sup>6</sup> to make counter narratives and used it to train neural models.

---

<sup>6</sup> The term crowdsourcing refers to the act of compiling services, ideas or content from the contributions of a large group of people.

Chung et al. (2019) describes how one of the first counter narrative corpus, called CONAN (COunter NARRatives through Nichesourcing), was generated. This corpus includes 6,645 hate speech-counter narrative (HS-CN) pairs in English, including 2,781 pairs translated from French and Italian. It is the first work in which languages other than English are tackled. In principle, CONAN focuses on 3 specific communities (targets) (Muslims, Jews, LGBTI), although this resource has been used as the basis for generating other resources with a greater number of targets (Multi-Target).

Mathew et al. (2019) analyse the results on various interesting points, such as counter narrative comments receiving twice as many likes as comments that are not counter narrative. For certain communities, most of the comments that are not counter narrative tend to be hate speech; the various types of counter narratives are not all equally effective, and language choices by users who publish counter narratives is very different from that of those posting comments that are not counter narrative, as revealed by a detailed psycholinguistic analysis. The interesting thing about their study is that they use machine learning models to detect counter narratives in YouTube videos, with an F1 score of 0.73. Moreover, they provide a dataset for detecting counter narratives using YouTube comments to conduct a measurement study that characterises the linguistic structure of the counter narrative. Overall, their study offers valuable information about the impact of the counter narrative on online interactions, the difference in language use between comments with and without a counter narrative, and it provides important resources for future research in this field.

Tekiroglu, et al. (2020), describes various strategies for generating counter narratives, and for the first time proposed using NLP-based tools; specifically, in this case, GPT-2 is used, and while the direct output of the computational system is not excellent, they do serve as the basis for review by various non-experts, who perform an initial filtering and, finally, NGO experts validate the final result.

After this work, in 2021, Chung et al. (2021a) proposed a more elaborate architecture based first on generating automatic queries and extracting phrases from a knowledge base. And, secondly, it generates counter narratives based on the phrases extracted. The CONAN corpus is used as the basis, but this resource is enriched by a series of phrases obtained by means of a module for extracting, generating and selecting knowledge, creating the CONAN-KN corpus. This paper explores several experimentation models, including GPT-2, an improved version of GPT, and XNLG.

The same methodology for collecting data about hate speech directed at other religions, races and gender is used to fine-tune GPT2 for automatic generation of counter narratives (Fanton et al., 2021). A corpus in English is generated from CONAN called CONAN Multi-Target in which expert NGO operators review the counter narratives generated by the computer system and post-edit them. This work is particularly valuable because it produces a high-quality resource generated by automatic models but manually reviewed by human operators, and because it includes eight different classes of hate speech targets (see table 1).

**Table 1. Distribution of “Hate Speech-Counter-Narrative” pairs by target in the CONAN-Multi-Target corpus**

Target of hate speech	Number of “Hate Speech - Counter Narrative” pairs
People with disabilities	220
Jews	594
LGBT+	617
Immigrants	957
Muslims	1,335
People of colour	352
Women	662
Other (Overweight people, Gypsies, etc.)	266
<b>Total</b>	<b>5,003</b>

All of this previous work has helped Chung et al. (2021b) to develop a tool to help NGOs propose counter narratives to combat hate speech using this automated decision-making support system.

On the other hand, Zhu and Bhat (2021) proposed a procedure to generate candidate counter narratives by using a generative model based on a recurrent neural network (RNN), trained with this dataset and which selected the most relevant candidate.

The paper of Bonaldi et al. (2022) covers a very interesting aspect of this, exploring as it does the generation of the counter narrative from a dialogue perspective as opposed to a simple hate-counter narrative discourse pair. A hybrid approach to data collection through dialogues is presented, which combines the intervention of expert human annotators with machine-generated dialogues obtained using different configurations. Furthermore, a freely available corpus called DIALOCONAN (DIALogicalCOunter-NArrativescollectionN) is generated which includes a dataset with more than 3,000 fictitious multi-turn dialogues between a hater and an NGO operator.<sup>7</sup> Moreover, six hate targets are covered, such that it is a new resource for combating hate speech towards different target groups (table 2).

<sup>7</sup> <https://github.com/marcoguerini/CONAN>

**Table 2. Distribution of “Hate Speech-Counter Narrative” pairs by target in the DIALOCONAN corpus**

<b>Target / Target Group</b>	<b>No. of pairs</b>
Jews	468
LGBT+	591
Immigrants	534
Muslims	505
African or African-descended people	493
Women	462
Other (Overweight people, Gypsies, etc.)	6
<b>Total</b>	<b>3,059</b>

Finally, Ashida and Komachi (2022) examine how pre-trained models can be used to automatically generate messages to counteract offensive texts on social media. They use the GPT2, GPT2-Neo and GPT3 models to generate the counter narratives, which are then manually assessed using Amazon Mechanical Turk; this shows that GPT-3 is the model that produces the highest quality messages. The generated corpus CHASM: Countering HAtE Speech and Microaggressions) is available on the website <https://github.com/tmu-nlp/CHASM>

# 3

## Works in other languages

As mentioned above, works on generating counter narratives are not very abundant, but as regards languages other than English, the bibliography is not much more than token. The first paper to look at other languages was published by Bartlett and Krasodomski-Jones (2015), which explores the effect of counter narratives on Facebook. The authors argue for the importance of maintaining the principle of freedom of expression on the Internet and that it be a place where people feel they can say what they think openly and freely even when it comes to extreme or radical content. They therefore focus on the effect that counternarratives have in comparison to removal of content. The study was conducted in various countries (France, United Kingdom, Morocco and Tunisia, Indonesia and India) and concludes that, depending on the region, the context in which the counter narrative is made and shared changes and the content of the counter narrative works differently.

Garland et al. (2020) also explore how people from different countries react to counter narrative content on Facebook and tries to identify what types of content are most likely to attract users. The study shows that users engage with the counter narrative according to their location in the world, which indicates that there is no broad approach covering the whole of Facebook. Specific approaches are required for different places and countries in which Facebook can offer an important platform for sharing messages that challenge hate narratives and incitement to violence.

Chung et al. (2020) focus on generating the first corpus of counter narratives in Italian by means of the automated translation of counter narratives in English based on the CONAN corpus.

Miškolci et al. (2020) explore hate speech against the Roma community in Slovakia on Facebook between April 2016 and January 2017. In total, they examine 60 debates on Facebook with more than 7,500 comments on matters related to the Roma community, which were published by the profiles of several members of the Slovak Parliament and the most popular online media.

Finally, Garland et al. (2022) assessed the effectiveness of the counter narrative using various measures at the macro- and micro-levels to analyse 180,000 political conversations in German held on Twitter over four years. The results suggest that organised hate speech is associated with changes in public discourse and that the counter narrative, especially when it is organised, can have an important impact on online hate rhetoric.

# 4 Works in Spanish

As regards works in Spanish, until 2023 we find only one attempt to tackle the problem, in the article by Furman et al. (2022). Although this research focuses more on mining argumentation, and the corpus we work on for Spanish, it is hardly representative. This work enriches the HatEval corpus by adding hate speech tweets (Basile et al., 2019) mainly in English, but some also in Spanish. The authors annotate tweets with information about argumentation to facilitate the automated generation of counter narratives as they posit that such argumentation could help build more persuasive and effective counter narratives against hate speech. It is a very interesting work with a very innovative goal, but it only manages to enrich a total of 970 tweets in English and only 296 in Spanish.

Based mainly on the work of Chung et al. (2021a), we have conducted a first investigation focusing on the Spanish language, which was published in the SEPLN journal (Vallecillo et al., 2023). It discusses the use of linguistic models to automatically generate counter narratives to hate speech in Spanish. The article shows that GPT-3 outperforms other models in generating inoffensive and informative counter-narratives sometimes including persuasive arguments. Various few-shot learning algorithms have been used, applying several prompting strategies and analysing the results for each. Moreover, a new corpus called CONAN-SP<sup>8</sup>, consisting of 238 hate speech and counter narrative pairs in Spanish, has been made available to the research community to assist with further research in this area. These results underscore the potential in language modelling to combat hate speech in Spanish by generating counter narratives. Given the relevance of this work, a summary of it is set out below.

---

<sup>8</sup> <https://github.com/estrellaVallecillo/CONAN-SP.git>

# 5 First article for generating counter narratives in Spanish

In this first article tackling the automatic generation of counter narratives in Spanish, the following main contributions were made:

1. Study automatic generation of counter narratives to hate speech in Spanish.
2. Compare various models for generating counter narratives in Spanish.
3. Generate a new corpus in Spanish with hate speech-counter narrative pairs using generative language models (CONAN-SP).
4. Assess different prompting strategies to automatically generate counter narratives.

In this paper, the CONAN-KN is used as the basis for experimentation because it represents a first subset of CONAN, with 195 hate speech - counter narrative pairs and also includes external knowledge that has supported manual generation of the counter narrative. First, an automatic translation of the corpus is made using the DeepL tool. Although the corpus has been translated in full, only the Hate Speech (HS) part will be used and the aim is to generate the counter narrative (CN) part. It should be noted that the original CONAN-KN corpus included in the HS-CN pairs some hate speech that was repeated and different counter narratives were obtained. In our experiments we have only selected one of those pairs in which the hate speech was repeated, specifically the first one to appear, such that in the end in the corpus in Spanish only 105 HS-CN pairs have been considered which have been automatically translated into Spanish. Since the goal is to generate the counter narrative automatically, only the hate speech part of the CONAN-KN corpus is in fact used to generate the counter narrative using different prompting strategies and different language models.

## 5.1. Language models

Specifically, the models used for automatic generation are as follows:

- GPT-2<sup>9</sup> (Radford et al., 2019).
- MarIA GPT-2 (Fandiño et al., 2022).
- Flan T5<sup>10</sup> (Chung et al., 2022).
- Bloom<sup>11</sup> (Scao et al., 2022).
- Davinci GPT-3<sup>12</sup>.

The article describes how the various parameters have been adjusted to train each model.

## 5.2. Prompting strategies

As for prompting strategies (Liu et al., 2023), we used a few-shot learning strategy that includes some examples of “Hate Speech- Counter Narrative” (HS-CN) along with the description of the task to be conducted (to generate a counter narrative). Figure 1 summarises the process followed.

As mentioned above, the dataset in the base corpus (CONAN-KN) includes five offensive classes (anti-Semitism, Islamophobia, racism, homophobia and misogyny). Since all the information the model will use to generate a counter narrative is in the prompt, we decided to explore different prompting strategies and to compare the results. Specifically, the following alternatives have been followed:

- **Experiment 1:** this includes a description of the task and provides the model with an example of each of the five classes of offensive comments.
- **Experiment 2:** five prompts are considered, one for each class of offensive comment. The various instructions must include a description of the task and provide the model with three examples of each class.
- **Experiment 3:** a prompt is used that includes five examples, one for each class, without the description of the task.

The prompts we used are included in Appendix 1.

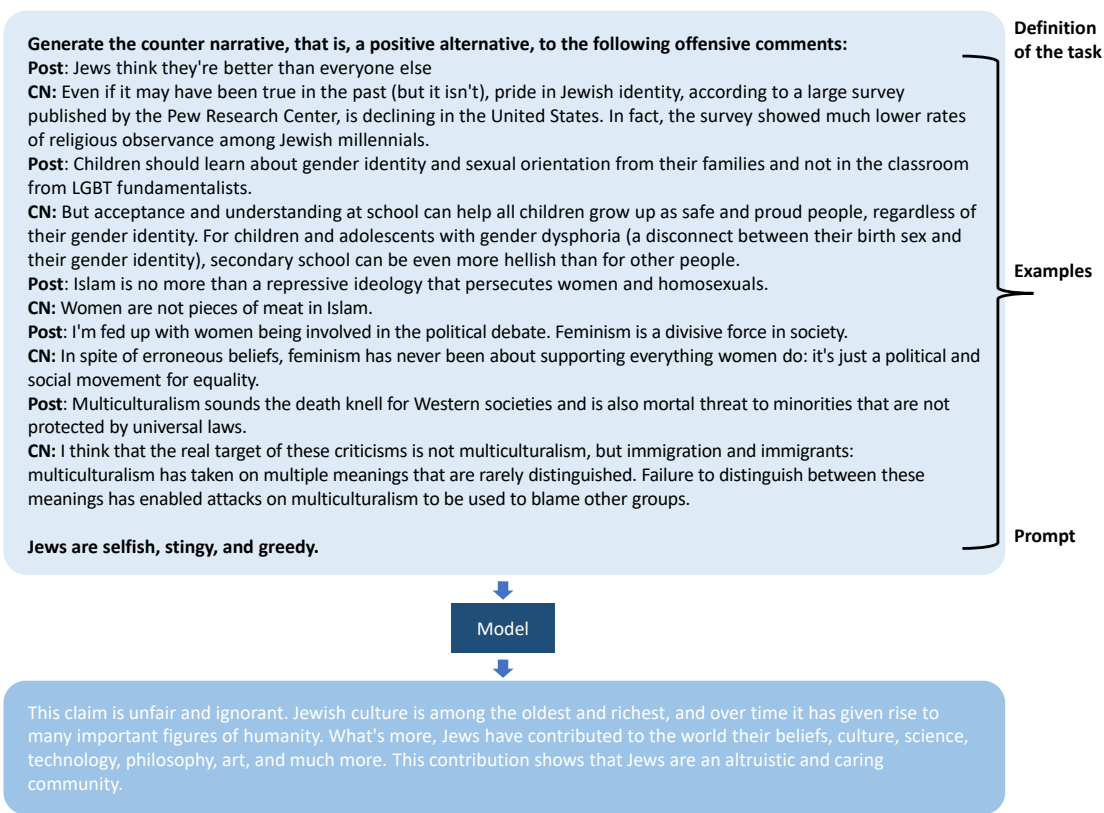
9 <https://huggingface.co/GPT-2-large>

10 <https://huggingface.co/google/flan-t5-large>

11 <https://huggingface.co/bigscience/bloom>

12 <https://platform.openai.com/docs/models/gpt-3>

**Figure 1: Prompting strategy followed for generating counter narratives**



### 5.3. CONAN-SP

With the results of the proposed experiments, we created a new dataset for counter narratives in Spanish, called CONAN-SP. Of all the models used, only the responses generated with GPT-3 are included in the CONAN-SP corpus, since it is the only one that generates counter narratives with a certain syntactic quality, albeit, as we will see later, it also makes some significant errors in terms of the veracity of the information. However, the CONAN-SP corpus includes the GPT-3 generations without conducting any filtering.

The following table shows that GPT2-based models do not generate usable counter narratives. On the other hand, in Flan T5, although it is syntactically usable, the messages generated are even more offensive than the hate speech. As regards Bloom, although it seems it can be used on some occasions, when it is compared with GPT-3 the result is that the latter is the model that works best. It is therefore the one we used to generate the CONAN-SP corpus.

**Table 3. Examples of counter narratives generated with the different models**

Hate Speech	Model	Counter Narrative
Not all Muslims are terrorists, but all terrorists are Muslims.	GPT-2	They don't want Muslims to be raped in this politician. Post: The game is a woman nation, with one that must our politician talk to the workers that ended up most of the woman nation. Counter narrative: The female nation is [...] n. Post: Children talking to us about those of them who not talking are children. Counter Narrative: With the children.
	GPT-2 MarIA	Not at school.
	GPT-3	While terrorism has undoubtedly been perpetrated by many Muslims over the years, it is wrong to generalise and say that the entire Muslim community are terrorists. Studies indicate that most Muslims condemn terrorism. Solutions cannot be found to the problems of terrorism by generalising or through stereotypes, but rather with open and respectful dialogue.
	FlanT5	Not all Muslims are terrorists, but all terrorists are Muslims.
	Bloom	I don't think any religion is the enemy in the fight against mass immigration. The followers of the Muslims are considered by the whole Muslim community to be human beings, and not people belonging to a certain religion.
Women and feminists don't want equal rights and opportunities: they just want to take over society.	GPT-2	The feminist minimum is just the de hazation is very although of women, the feminist is some technique that are supported between the eyes. At work, frontier women as a year among some women in the death of the genders. Post: Women, feminists, [...] \nCounter Narrative: A women, feminists, the top feminists accept.
	GPT-2 MarIA	And the rest of the community.
	GPT-3	Feminism is not an attempt to dominate society, but rather to build a better society for all, in which equality is a priority. Feminism seeks equal opportunities to ensure that everyone has the freedom to achieve their true potential, regardless of their sex.
	FlanT5	Women and feminists don't want equal rights and opportunities: they just want to take over society.
	Bloom	For women, equal rights and opportunities is the power that has been stolen from men, and men have been forbidden from enjoying that equality. For feminists, equal rights and opportunities is the power that has been stolen from women, and women have been forbidden from enjoying it.

In the article, we have made a manual assessment by three annotators of five Hate Speech - Counter Narrative pairs for each of the experiments we will analyse below.

To build CONAN-SP, we removed the pairs containing duplicate hate speech texts and the examples used by annotators to make a manual assessment, besides those used to generate the prompt strategy. So, finally, 238 hate speech- counter narrative pairs are obtained.

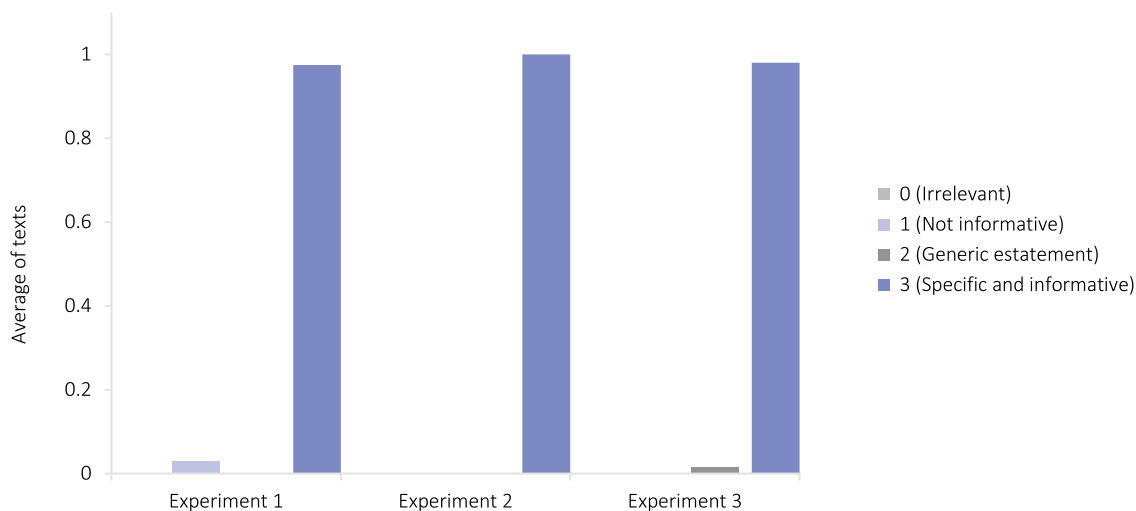
## 5.4. Assessment

The question of assessment in generative language models is a challenge with no easy solution. In fact, what we have done is to perform a manual assessment by three annotators with different profiles (linguist, junior computer technician, senior computer technician). For the assessment we have followed the work of Ashida and Komachi and we have considered three perspectives for each counter narrative: offensiveness, stance and informativeness:

- **Offensiveness:** this determines whether the counter narrative is offensive to anyone (for example, for people of a certain ethnic origin) including people who wrote the hate speech message.
  - 0 (not sure).
  - 1 (not offensive).
  - 2 (perhaps offensive).
  - 3 (completely offensive).
- **Stance:** this refers to the position adopted with respect to the message and is classified into three types: agrees, neutral, disagrees.
  - 0 (irrelevant).
  - 1 (totally agree).
  - 2 (slightly agree/disagree).
  - 3 (totally disagree).
- **Informativeness:** it 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).

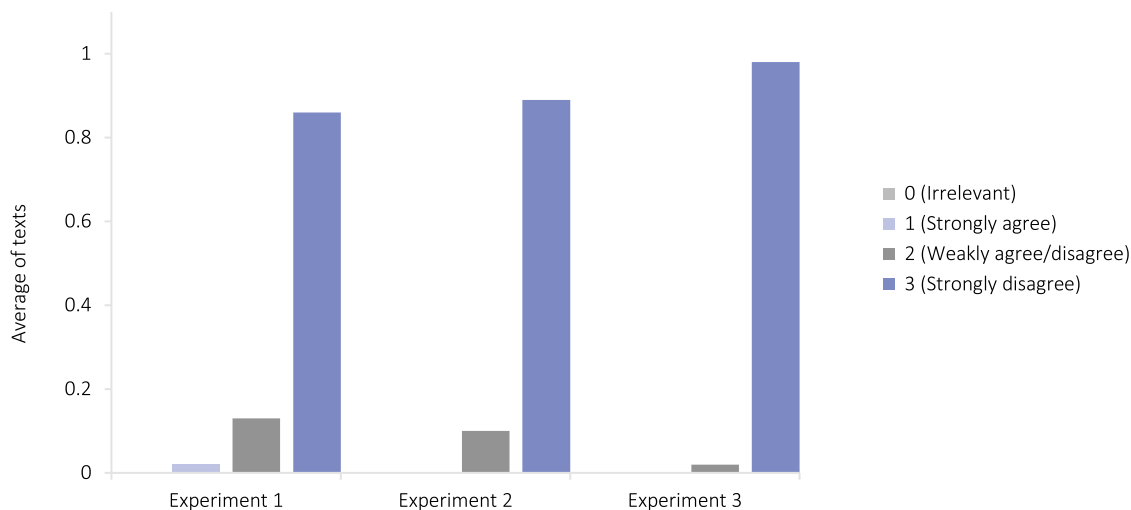
The group of assessing annotators first note five pairs of hate speech-counter narratives with all the texts generated with all the models tested. Having proven that the GPT-3 is the best system, the assessment focuses only on the texts generated by this model in each experiment. The three annotators re-assessed 20 selected hate speech- counter narrative pairs (60 pairs in total for each of the three experiments performed) and the agreement achieved clearly shows that the GPT3 model works very well, with the prompting strategy used in experiment 3 being the most suitable. Finally, the rest of the corpus is annotated by two human annotators for the counter narratives generated by GPT-3 (238 pairs).

**Figure 2: Informativeness results in the counter narratives generated**



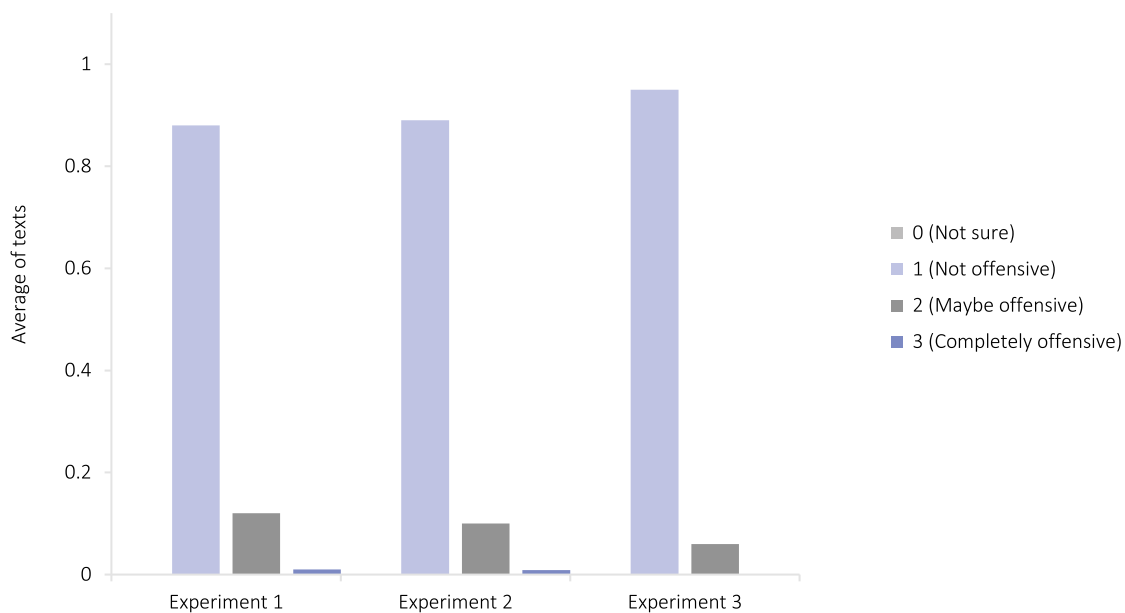
If we analyse the perspective of informativeness, we can see that the texts in Experiment 2 outperform the other experiments with 100% of the counter narratives generated being considered specific and informative. However, in all these experiments, the texts generated are informative: more than 97% (see figure 2).

**Figure 3. Results of the Stance metric in the counter narratives generated**



When analysing the stance of the texts generated in response to the offensive comment, we see in all the experiments that most of the time the generated texts disagree with offensive comment. In Experiment 3, the average number of texts that totally disagree is higher than in the other experiments, with 98% of the counter narratives generated. In Experiments 1 and 2 there are some texts where the stance is “slightly agree/disagree” or “completely agree”, but the number of these texts is very low (see figure 3).

**Figure 4: Offensiveness results for the counter narratives generated**



Finally, when analysing the offensive character of the texts generated in all the experiments, we may conclude that Experiment 3 is that which generates less offensive counter narratives, with 94% of the texts labelled as “inoffensive” and a minimal percentage as “possibly offensive.” On the contrary, Experiment 1 is that which generates the most offensive counter narratives because 1% of the texts are offensive and 11% “possibly offensive” although the predominant category is “inoffensive” (see figure 4).

After analysing the various perspectives to assess the counter narratives generated, we may conclude that Experiment 3 is the most suitable for creating the counter narrative, because it has the highest percentages to counteract the offensive comments, without disseminating the toxicity of these comments and giving specific information about why the offensive comment is wrong. However, it is important to keep in mind that the three proposed strategies are successful in the task of generating counter narratives.

## 5.5. Error analysis

To identify the challenges facing GPT-3 in generating counter narratives, we performed an error analysis in all the proposed experiments.

Although in most cases GPT-3 generates a good counter narrative in both grammatical and semantic terms and contains information that counteracts the hate message, some examples have been found that include some incorrect or false information. Moreover, some grammatical errors have also been found. The annotators identified the false information by searching for more specific texts and by making Internet searches to compare the information generated by GPT-3 with verified sources.

As may be seen in table 4 below for examples 1, 2 and 6, GPT-3 sometimes generates a counter narrative with information that does not match reality. In other cases, it contradicts or confuses terms, as in examples 4 and 5, respectively. Finally, in example 3 we can see that GPT-3 made an error in generating the article and the next word, since the genders do not match.

After analysing the counter narratives generated, we may conclude that a system based on GPT-3 could be an effective tool for NGO operators who combat hate speech.

However, it is important to bear in mind that these systems must be supervised by humans in the final phase, since some errors committed by the model may be serious.

**Table 4. Examples of delusions generated with the GPT3.5 model**

Hate Speech	Counter Narrative
<b>Experiment 1</b>	
Homos abuse their anus and then want to cleanse their conscience by donating blood. Let's stop them!	It's true that some homosexuals have unprotected sex, but <b>there's no scientific evidence that this increases the risk of transmission due to donating blood.</b> In reality, most people who donate blood are high-risk people, such as intravenous drug users, and not homosexuals.
Gays are spreading AIDS around the world because they're promiscuous and fornicate with strangers; we should prevent them from donating blood.	These types of discriminatory and stereotyped statements are entirely erroneous. According to the World Health Organization, <b>blood donation guidelines based on sexual behaviour, such as prohibiting men who have sex with men from donating blood, are likely to exclude people who engage in safe sexual behaviour and are therefore unlikely to be at risk of disease transmission.</b>
<b>Experiment 2</b>	
This came directly from an imam: they believe everything in the Quran, even though these words and ideology are 1400 years old and have no place in the twenty-first century. This appalling book is the root of all the problems in the world today.	Like any other religion, Islam is interpreted differently by different people. There are moderate Muslims who believe that the Quran should be interpreted according to the needs of the modern era. Islam is a religion based on love, equality and justice. Most of the problems in the contemporary world are caused by fundamentalism, fanaticism and nationalism. Islam is not the cause of these problems.
The vast majority of Muslims are unemployed and refuse to work.	According to the Office for National Statistics, 5.5% of Muslims were unemployed in the United Kingdom, the lowest of all religions and ethnic groups. This is lower than 6.2% of Christians and 8.5% of non-religious people. These data suggest that the vast majority of Muslims are working, are not unemployed and do not <b>refuse to</b> work.

**Table 4. Examples of delusions generated with the GPT3.5 model**

Hate Speech	Counter Narrative
<p>Not all Muslims are terrorists, but all terrorists are Muslims.</p>	<p>This statement is absolutely wrong. In fact there are many terrorist groups around the world that have nothing to do with Islam. For example, the Irish National Liberation Army (IRA), the <b>Communist Party of Spain (GRAPO)</b>, the Tupac Amaru Revolutionary Movement (MRTA) in Peru, the National Liberation Front of Colombia (FARC) and the National Liberation Movement of Guatemala (MLNG). These terrorist groups have nothing to do with Islam and attack civil and military targets.</p>
<p>Are you an English girl living in a Muslim area? Remember to stay away from Muslim men. Don't accept gifts such as sweets, cigarettes, drinks, or drugs. Don't accept. Accept a lift or invitations to parties. Don't trust the police or social workers.</p>	<p>If you're an English girl living in a Muslim area, remember that most Muslims are respectable people who want to live in peace and harmony with others. There is no need to be afraid of Muslim men. <b>If someone offers you something like sweets, cigarettes, drinks or drugs, don't accept. If someone invites you to a party or offers you a lift, don't accept.</b> Always trust the police and social workers for help and protection.</p>
<p><b>Experiment 3</b></p>	
<p>Allowing homosexuals to donate blood compromises patients' safety because of their sexual behaviour, which spreads diseases such as HIV.</p>	<p>This assumption is false and irrational. The truth is that blood donors must fulfil strict donation criteria, regardless of their sexual orientation. The blood bank assesses donors to ensure that donated blood is safe for recipients. <b>Moreover, HIV is not transmitted through donated blood.</b></p>

# 6 Preliminary work with other models

## 6.1. Experimenting with GPT-4

As mentioned, experimenting with GPT3 is very useful but, though they are minimal, it entails some errors. That's why we decided to perform a similar experiment with GPT4, mainly to compare both models. To do this, we took the CONAN Multi-Target corpus in English as a starting point and conducted an automatic translation using the API of DeepL to obtain the CONAN-MT-SP (CONAN Multi-Target in Spanish) corpus. Two models were applied based on GPT technologies, namely, GPT3 and GPT4, to the hate speech part of this corpus, which is provided as a guide together with eight counter narrative examples.

The main purpose of this study is to generate a high-quality corpus for the Spanish language and to test its validity by means of a manual assessment of the corpus. The CONAN-MT-SP corpus generated along with its assessment shall be made available to the scientific community for their use. Each instance in the corpus has the hate speech - counter narrative part translated directly to Spanish with DeepL from the CONAN Multi-Target corpus, plus the counter narrative generated by GPT4. Moreover, assessments made by human experts have also been included as a separate document in line with the CONAN-MT-SP corpus. The results show that, while the effectiveness of GPT4 is superior to that of GPT3, both models can be used to automatically generate counter narratives against hate speech.

### 6.1.1. Generating CONAN-MT-SP

The **CONAN Multi-Target (CONAN-MT)** is a corpus generated by human experts: a hate speech- counter narrative dataset built using a semiautomatic mechanism (Fantón et al., 2021). It contains 5,003 hate speech-counter narrative pairs in English covering multiple targets of hate speech such as racial origin, religion, country of origin, sexual orientation, disability and gender. These targets represent various aspects of identity that are often the target of online hate (see table 5).

**Table 5. Instance distribution in the CONAN Multi-Target corpus**

Target group of hate speech	#HS-CN Pairs
People with disabilities	220
Jewish people	594
LGBTI	617
Immigrants	957
Muslims	1,335
African or African-descended people	352
Women	662
Others (Overweight people, Gypsies, etc.)	266
<b>Total</b>	<b>5,003</b>

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

The reason the CONAN-MT corpus has been used is because it is based on the CONAN corpus, one of the reference corpora in this area of research. One of the main advantages of CONAN-MT lies in the diversity and representativeness of the targets in the corpus. By covering a wide range of targets, such as gender, race, religion, ideology and other personal characteristics, it was possible to create a dataset that more accurately and fully reflects the complexity of online hate speech. This means that the problem to be tackled has become more robust and general, enabling the development of models and algorithms capable of addressing a variety of scenarios and contexts in which hate speech appears.

Based on the CONAN-MT corpus, it is automatically translated into Spanish using the DeepL API. This automatic translation makes it possible to obtain a set of 5,003 pairs of hate speech phrases and their respective counter narratives in Spanish, thus forming the first part of the CONAN-MT-SP corpus. To guarantee the validity and quality of the translations, a manual review is performed to verify that the translation obtained is accurate, consistent and that it reproduces the original meaning of the counter narratives in the target language.

The starting point used are the two most successful prompting strategies based on the results of prior work with GPT3. In particular, the strategy that uses only the examples (“Experiment 1”) and that which also includes the definition of the task along with the examples of counter narrative (“Experiment 2”). In this case, we used eight examples as a prompt such that an example of each target is included. The prompt used for Experiment 1 are in Appendix 2. For Experiment 2, the same prompt is used but the definition of the task is added at the beginning. Specifically, the following text is included: *“It generates the counter narrative, that is, a positive alternative, to the following offensive comments”*.

In this study, two different language models are used to generate counter narratives based on the selected <sup>13</sup>prompts. One of the models used is GPT3, which has been shown to be highly effective in previous research, where outstanding results were obtained in terms of accuracy. The second model used is GPT4, which is presented to make a comparison with the recognised accuracy of GPT3. The purpose of including GPT4 in this study is to assess whether it offers significant improvements in terms of quality and consistency in generating counter narratives, in comparison with its predecessor, GPT3.

### 6.1.2. Assessment

To make the assessment, we followed the work of Ashida and Komachi (2022) which was used in the previous paper with GPT3 and we again considered three perspectives for each counter narrative: Offensiveness, Stance and Informativeness (see section 5.4):

However, after an initial assessment, we believe that it would be appropriate to incorporate further measures to assess the veracity, the quality of the text generated and the need for possible editing, and finally, the comparison between the quality of the counter narratives generated automatically by the GPT model and those generated by humans (Comparison between H-M). These complementary measures will yield a fuller and more accurate picture of the effectiveness and reliability of the counter narratives generated, as well as the ability of the GPT model to match or exceed the quality of human counter narratives with respect to coherence, contextual understanding and relevant content.

- **Veracity:** assess whether what is said in the comment is true.
  - 0 (not sure).
  - 1 (not true).
  - 2 (partially true).
  - 3 (totally true).
- **Editing required:** assess whether human editing is necessary to show a counter narrative.
  - 0 (unedited).
  - 1 (with edition).
- **Comparison between H-M:** it assesses which counter narrative would be chosen between the human one or that of the wizard.
  - 0 (both counter narratives are equally valid).
  - 1 (the human generates a better counter narrative).
  - 2 (the machine generates a better counter narrative).
  - 3 (no counter narrative is good).

---

<sup>13</sup> A prompt is an instruction or a request given to a language model in artificial intelligence, such as chatGPT or Google Bard, to generate a response or to complete a specific task.

The assessment is made in several steps. First, the first 50 examples of each of the four generated datasets (GPT3-exp1, GPT3-exp2, GPT4-exp1, GPT4-exp2) are assessed. This first assessment was conducted by three human annotators (one senior linguist, one junior linguist, one senior computer technician) using the indicators discussed above and it determined that GPT4 clearly performs much better than GPT3 with very high concordance. Moreover, it was found that there is scarcely any difference between Experiment 1 and Experiment 2. Thus, and given that the cost in effort of making the manual assessment is very great, it was decided to only make the assessment of the corpus generated with Experiment 1 using the GPT4 model.

This hate speech-counter narrative corpus, along with the assessment made, shall be made available to the scientific community.

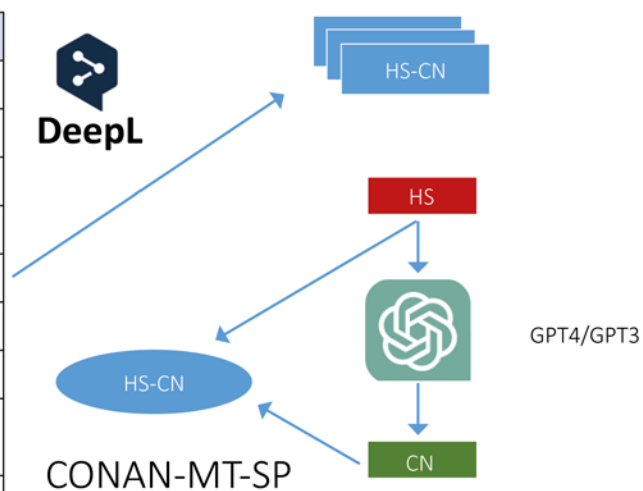
Fully to annotate the corpus, only the two annotators with a linguistic profile performed the rest of the labelling of the corpus generated with GPT4 using the prompt in Experiment 1. This corpus has been called CONAN-MT-SP and contains a total of 3,635 examples of “Hate Speech-Counter Narrative”. Figure 5 shows the process of generating the corpus.

**Figure 5. Methodology for generating the new CONAN-MT-SP corpus**

**PRELIMINARY WORK WITH GPT4/GPT3**

**CONAN-Multi-Target**

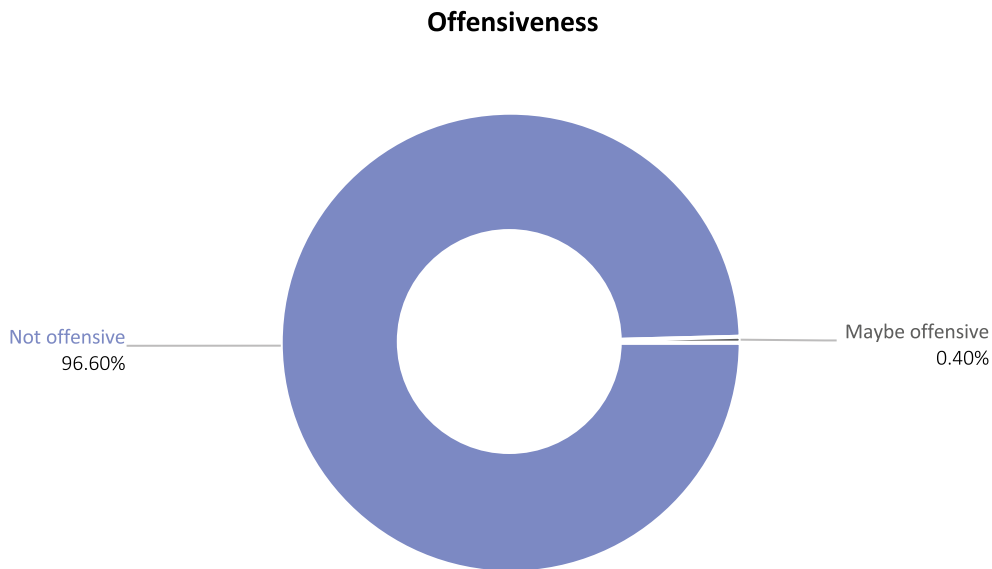
Target	No. of pairs
People with disabilities	220
Jews	594
LGBTB+	617
Immigrants	957
Muslims	1,335
People of colour	352
Women	662
Other (Overweight people, Gypsies, etc.)	266
<b>Total</b>	<b>5,003</b>



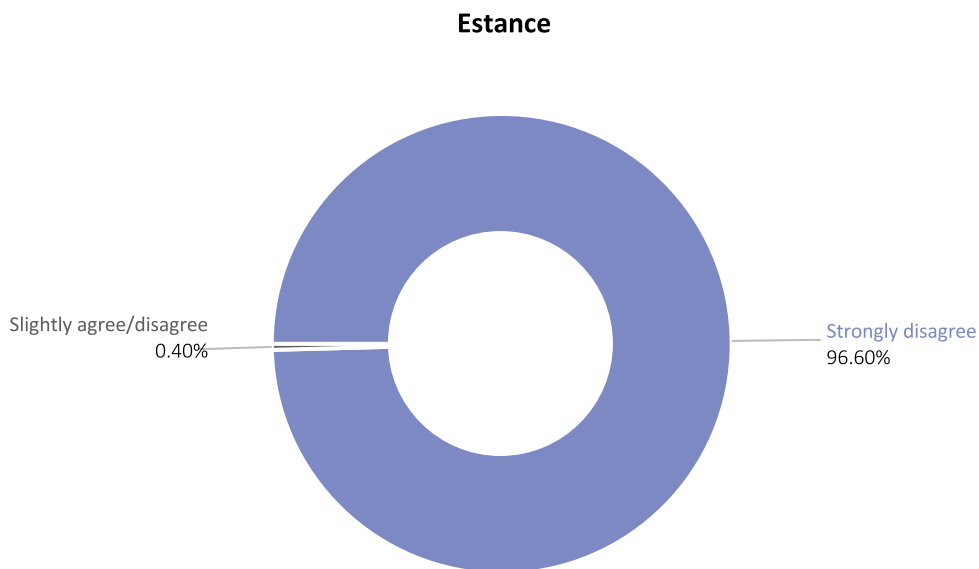
### 6.1.3. Results

This section includes the results obtained during the annotation. As can be seen in the results of the assessment set out in table 6 and figures 6, 7, 8, 9, 10 and 11, although there are some cases in which the text is imperfect, the quality of the counter narratives generated with GPT4 is very high.

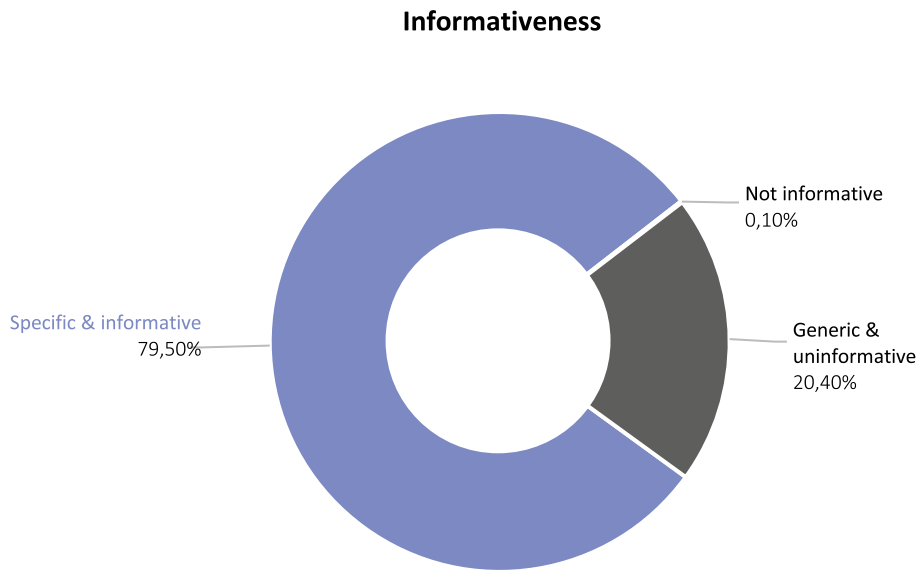
**Figure 6. Offensive results in the counter narratives generated**



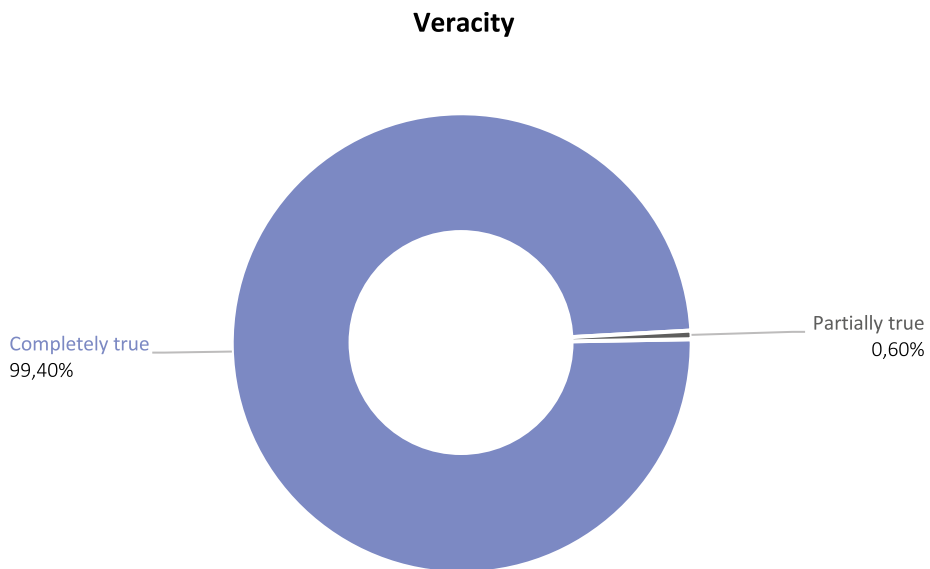
**Figure 7. Stance results in the counter narratives generated**



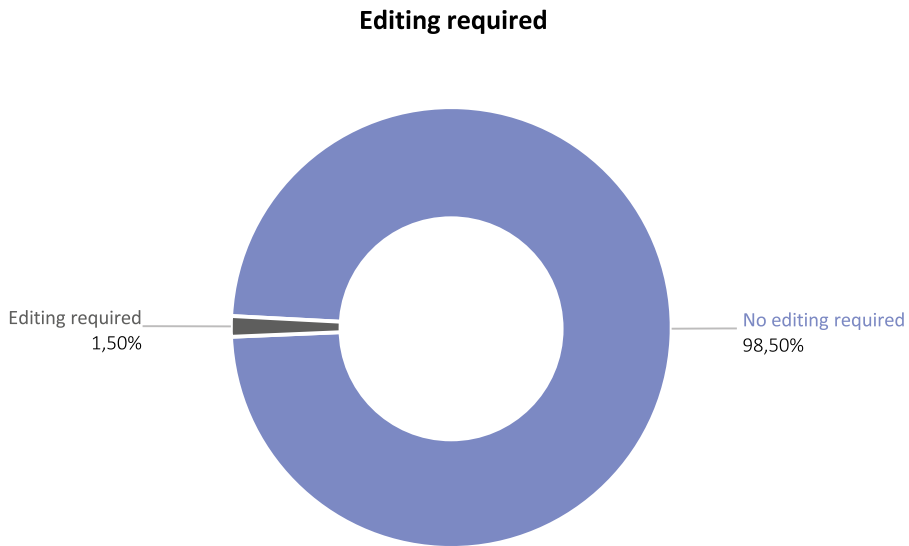
**Figure 8. Informativeness results in the counter narratives generated**



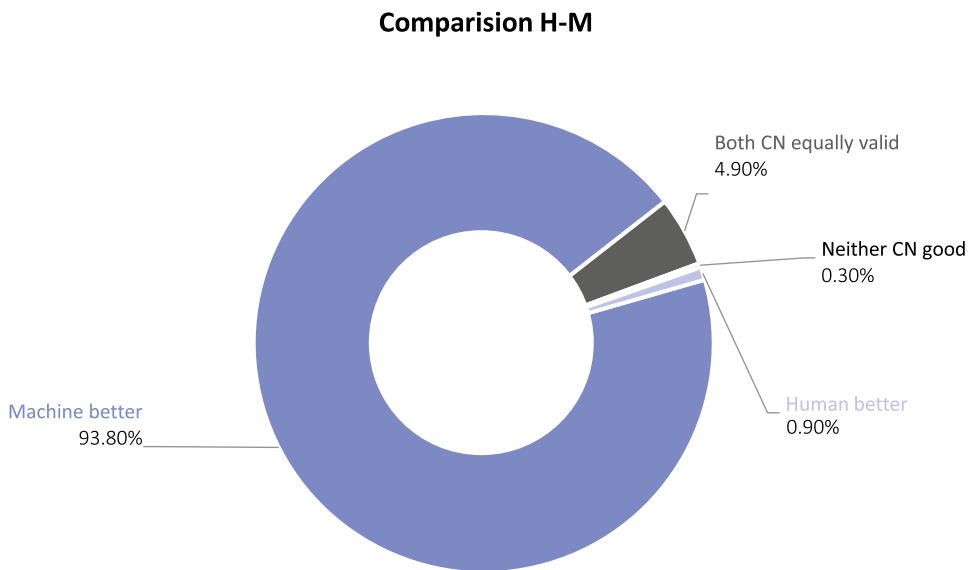
**Figure 9. Veracity results in the counter narratives generated**



**Figure 10. Result of the need for editing in the counter narratives generated**



**Figure 11. Comparative result between the counter narrative generated by humans and by automatic models**



**Table 6. Summary of results in the generation of the CONAN-MT-SP corpus**

<b>Offensiveness</b>		
I'm not sure	0	0.00%
It's not offensive	3623	99.64%
Perhaps offensive	13	0.36%
Completely offensive	0	0.00%
<b>Stance</b>		
Irrelevant	0	0.00%
Totally agree	0	0.00%
Slightly Agree/Disagree	16	0.44%
I totally disagree	3620	99.56%
<b>Informativeness</b>		
Irrelevant	0	0.00%
Not informative	2	0.06%
Generic and not informative	743	20.43%
Specific and informative	2891	79.51%
<b>Veracity</b>		
I'm not sure	0	0.00%
Not true	1	0.03%
Partly true	22	0.61%
Completely true	3613	99.37%
<b>Edition</b>		
No Edition	3581	98.49%
Yes Edition	55	1.51%
<b>Comparison between H-M generation</b>		
Both CNs equally valid	179	4.92%
The Human is better than the Machine	34	0.94%
Better the machine than the human	3412	93.84%
Neither of the two CNs is good	11	0.30%

Below (see table 7), there are some examples (one for each hate speech target population group considered in the corpus) with the hate speech message and the counter narrative that has been automatically translated with the DeepL API, along with the counter narrative generated by the GPT4 model. Moreover, the assessment was made for each of the metrics taken into account in the study.

**Table 7. Examples of the assessment made of each target in the corpus. OF: Offensiveness; ST: Stance; IN: Informativeness; VE: Veracity; ED: Required edition; HM: Human-Machine Comparison**

Hate Speech	Human counter narrative	Counter narrative GPT4	OF	ST	IN	VE	ED	HM
The disabled are a burden on society.	They aren't; they only need a place to live free of discrimination and a little support.	People with disabilities may face additional challenges in life, but that doesn't mean they're a burden for society. In fact, many people with disabilities contribute significantly to our communities and enrich our lives with their unique perspectives and skills. It's important to remember that everyone deserves respect and opportunities, regardless of their physical or mental abilities.	1	3	3	3	0	2
Jews are manipulating our world for their benefit.	I think this comment indicates ignorance of the subject in question.	I think that generalising and blaming an entire group of people for global problems is wrong and unfair. Instead of defaming Jews, we must work together of a society to tackle and solve the problems we are facing.	1	3	3	3	0	2
Homos are violent: we must get rid of them.	It has been shown time and again that homosexuals are very peaceful people.	Violence has nothing to do with a person's sexual orientation. Stigmatising a group of people based on their orientation is unfair and wrong. We must focus on addressing the causes of the violence rather than discriminating against people because of their identity.	1	3	3	3	0	2
We have been invaded by immigrants from the beginning. It's time for this to end.	Our country was founded on the principles of equality and freedom. We cannot allow racism to undermine it.	We have been culturally, economically and socially enriched by the diversity that the immigrants have contributed throughout history. It's important to recognise and celebrate the positive contributions they've made rather than focusing on negative aspects.	1	3	3	3	0	2

**Table 7. Examples of the assessment made of each target in the corpus. OF: Offensiveness; ST: Stance; IN: Informativeness; VE: Veracity; ED: Required edition; HM: Human-Machine Comparison**

Hate Speech	Human counter narrative	Counter narrative GPT4	OF	ST	IN	VE	ED	HM
Women must be at the service of men. They must serve to please men. This is their proper place in life.	Women are human beings with dignity. They must have the opportunity to live with this dignity.	Women, like men, are individuals with their own skills, desires and goals in life. It is neither fair nor appropriate to limit women to a role of servitude. In an egalitarian and just society, all people have the same opportunities and rights to achieve their goals and to contribute to the community according to their talents and abilities.	1	3	3	3	0	2
Every Muslim is a potential terrorist.	Any attempt to prohibit, isolate or criminalise a Muslim is an act of terror.	This statement is unfair and erroneous. Most Muslims are peaceful and respectful and condemn terrorism and violence. We must not judge an entire group of people on the basis of the actions of a few extremists.	1	3	3	3	0	2
No black person should have the right to vote.	If someone is not allowed to vote, he/she cannot influence the legislative or executive powers of a government and ends up revolting against injustice and oppression. Is this what they want: more ethnic clashes?	This statement is completely unacceptable and goes against the fundamental principles of equality and democracy. All people, whatever their race, deserve a say in society.	1	3	3	3	0	2
Shaming fat people is good because they will adopt healthier habits.	A shamed person doesn't care about the intention of the shameful action. They may internalise the shame and believe they are a bad person.	Shaming someone for their physical appearance not an effect or respectful way to motivate someone to change their habits. Instead, we should focus on supporting people to make healthy decisions and on creating a supporting environment for everyone.	1	3	2	3	0	2

#### 6.1.4. Conclusion and future work

By way of final conclusion, the use of GPT4 appears to be a very good solution to tackle the problem of automatically generating counter narratives. As may be seen, the quality of these counter narratives is exceptional and in practically 99.5% of cases it is equal to or even better than that which a human would

generate, although there are some cases in which attention ought to be paid. We think that the biggest drawback is using a closed and paid model such as GPT4. That's why our next task is to test systems that seem to be performing well, which are free and which we can adjust ourselves, such as LLaMA (Large Language Model Meta AI).

## 6.2. Experimentation with LLaMA (Large Language Model Meta AI)

We are currently working with the LLaMA (Large Language Model Meta AI) model to automatically generate counter narratives. LLaMA is an artificial intelligence language model that was created by Meta AI and is based on the transformer architecture. Although it is based on the same architecture as GPT4, this model has the advantage of having been released so it can be adjusted and fine-tuned to prevent biases and other problems by using customised data, and it could be a more appropriate solution for automatic generation of counter narratives. The quality of the generated text will not be as high as with GPT4 but because it is a released model we can adjust it using our own data besides trying to understand how the system works.

Besides the LLaMA model, we continue to focus on exploring and experimenting with other language models in order to improve still more our ability automatically to generate counter narratives. Moreover, we are assessing new data sources and pre-processing strategies to address biases and to improve the quality of the text generated. The results of these investigations and developments shall be set out in future reports, where we will share our advances and the impact these advances may have in combating disinformation and hate speech.

# 7 Related projects

Finally, it is worth considering some international projects related to the topic of this report. Since 2006, the organisation [Stop Hate UK](#) has combated hate and discrimination in the business, legal and community sectors and currently manages a free services for reporting hate crimes that operates 24 hours a day in the United Kingdom.

Another European Union project that has been pioneering and fully oriented to hate speech applying NLP techniques is the [Hatemeter](#) project, although it focuses on Islamophobia. The aim is to systematise, expand and share knowledge about online anti-Muslim hatred, and to improve the efficiency and effectiveness of NGOs in preventing and combating Islamophobia at EU level, by developing and testing a technology platform to automatically monitor and analyse internet and social media data on the phenomenon, and to produce computer-assisted responses and suggestions to support counter narratives and awareness-raising campaigns.

Finally, it should be stressed that between 2018 and 2021, the Ministry of Inclusion, Social Security and Migration, through OBERAXE (Spanish Observatory on Racism and Xenophobia) led the [ALRECO](#) project (Hate speech, racism and xenophobia: Alert and coordinated response mechanisms) which seeks to improve the capacities of State authorities to identify, analyse, monitor and assess hate speech on social media, in order to design common strategies against racist, xenophobic, Islamophobic, anti-Semitic and anti-Gypsy speech.

As a continuation of this project, and also led by OBERAXE, the [REAL-UP](#) project “Hate Discourse, Racism and Xenophobia: Alert and Response Mechanisms, Upstander Discourse Analysis” was launched, in order to improve the capacities of national authorities to identify, analyse, supervise and assess online hate speech and to develop and strengthen counter narrative strategies. Both in this project and in ALRECO, the ONDOD (National Office for Combating Hate Crimes. Ministry of Interior) participates actively by leading the WP3, which aims to automate monitoring of hate speech and generation of counter narrative using artificial intelligence tools.

As regards counter narratives, we may highlight the project run by the Alan Turing Institute in London (UK) [Counterspeech: a better way of tackling online hate?](#), which focuses specifically on examining and analysing counter narratives to combat hate speech.

As for research forums that worth considering, it is useful to review the articles that have been published in the seven editions of the [Workshop on Online Abuse and Harms \(WOAH\)](#), which contains some works related to counter narrative and hate speech.

Finally, a new forum focused specifically on counter narratives has been held last September 2023, [Workshop Counter Speech for Online Abuse \(CS4OA\)](#).



# Bibliography

- Alsagheer, D., Mansourifar, H., & Shi, W. (2022). Counter hate speech in social media: A survey. arXiv preprint arXiv:2203.03584.
- Ashida, M., & Komachi, M. (2022, July). Towards Automatic Generation of Messages Countering Online Hate Speech and Microaggressions. In Proceedings of the Sixth Workshop on Online Abuse and Harms (WOAH) (pp. 11-23).
- Bartlett, J., & Krasodonski-Jones, A. (2015). Counter-speech examining content that challenges extremism online. DEMOS, October.
- Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F. M. R., ... & Sanguinetti, M. (2019, June). Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In Proceedings of the 13th international workshop on semantic assessment (pp. 54-63).
- Benesch, S. (2014). Countering dangerous speech: new ideas for genocide prevention. Washington, DC: US Holocaust Memorial Museum.
- Bonaldi, H., Dellantonio, S., Tekiroglu, S. S., & Guerini, M. (2022). Human-Machine Collaboration Approaches to Build a Dialogue Dataset for Hate Speech Countering. arXiv preprint arXiv:2211.03433.
- Chung, Y. L. (2022). Counter Narrative Generation for Fighting Online Hate Speech. la tesis de esta última es muy completa e ilustrativa, recoge todo el estado del arte.
- Chung, Y.-L., Kuzmenko, E., Tekiroglu, S. S., and Guerini, M. (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.
- Chung, Y.-L., Tekiroglu, S. S., and Guerini, M. (2020). Italian counter narrative generation to fight online hate speech. In Proceedings of the Seventh Italian Conference on Computational Linguistics, Online.
- Chung, Y.-L., Tekiroglu, S. S., and Guerini, M. (2021a). 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.

- Chung, Y.-L., Tekiroglu, S. S., Tonelli, S., and Guerini, M. (2021b). Empowering ngos in countering online hate messages. *Online Social Networks and Media*, 24:100150.
- Fanton, M., Bonaldi, H., Tekiroglu, S. S., & Guerini, M. (2021). Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech. *arXiv preprint arXiv:2107.08720*.
- Furman, D. A., Torres, P., Rodriguez, J. A., Martinez, L., Alemany, L. A., Letzen, D., & Martinez, M. V. (2022). Parsimonious Argument Annotations for Hate Speech counter narratives. *arXiv preprint arXiv:2208.01099*.
- Garland, J., Ghazi-Zahedi, K., Young, J. G., Hébert-Dufresne, L., & Galesic, M. (2020). Countering hate on social media: Large scale classification of hate and counter speech. *arXiv preprint arXiv:2006.01974*.
- Garland, J., Ghazi-Zahedi, K., Young, J. G., Hébert-Dufresne, L., & Galesic, M. (2022). Impact and dynamics of hate and counter speech online. *EPJ data science*, 11(1), 3.
- Gutiérrez-Fandiño, A., Armengol-Estapé, J., Pàmies, M., Llop-Palao, J., Silveira-Ocampo, J., Carrino, C. P., ... & Villegas, M. (2021). Maria: Spanish language models. *arXiv preprint arXiv:2107.07253*.
- Jahan, M. S., & Oussalah, M. (2023). A systematic review of Hate Speech automatic detection using Natural Language Processing. *Neurocomputing*, 126232.
- Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9), 1-35.
- Mathew, B., Saha, P., Tharad, H., Rajgaria, S., Singhanian, P., Maity, S. K., Goyal, P., and Mukherjee, A. (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.
- Miškolci, J., Kováčová, L., & Rigová, E. (2020). Countering hate speech on Facebook: The case of the Roma minority in Slovakia. *Social Science Computer Review*, 38(2), 128-146.
- Plaza-del-Arco, F. M., Molina-González, M. D., Urena-López, L. A., & Martín-Valdivia, M. T. (2021). Comparing pre-trained language models for Spanish hate speech detection. *Expert Systems with Applications*, 166, 114120.
- Poletto, F., Basile, V., Sanguinetti, M., Bosco, C., & Patti, V. (2021). Resources and benchmark corpora for hate speech detection: a systematic review. *Language Resources and Evaluation*, 55, 477-523.
- Qian, J., Bethke, A., Liu, Y., Belding, E., and Wang, W. Y. (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.

- Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, 1(8), 9.
- Roberts, S. T. (2016). *Commercial Content Moderation: Digital Laborers' Dirty Work*, chapter The Intersectional Internet: Race, Sex, Class and Culture Online, pages 147–160. New York: Peter Lang.
- Scao, T. L., Fan, A., Akiki, C., Pavlick, E., Ilić, S., Hesslow, D., ... & Manica, M. (2022). Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.
- Schieb, C., & Preuss, M. (2016, June). Governing hate speech by means of counterspeech on Facebook. In 66th ica annual conference, at fukuoka, japan (pp. 1-23).
- Tekiroglu, S. S., Chung, Y.-L., and Guerini, M. (2020). Generating counter narratives against online hate speech: Data and strategies. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1177–1190, Online. Association for Computational Linguistics.
- Vallecillo, E., Montejo, A. & Martín, M.T. (2023) Automatic counter narrative generation for hate speech in Spanish. *Procesamiento del Lenguaje Natural*, 71.
- 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.
- Zhu, W. and Bhat, S. (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 1

## Prompt Experiment 1

**Definition of the task:** Generate the counter narrative, that is, a positive alternative, to the following offensive comments:

**Post:** Jews think they are better than others.

**Counter Narrative:** Although it may have been true in the past (but it isn't), pride in Jewish identity, according to a large survey published by the Pew Research Center, is declining in the United States. In fact, the survey showed much lower rates of religious observance among Jewish millennials.

**Post:** Children should learn about gender identity and sexual orientation from their families and not in the classroom from LGBT fundamentalists.

**Counter Narrative:** But acceptance and understanding at school can help all children grow up as safe and proud people, regardless of their gender identity. For children and adolescents with gender dysphoria (a disconnect between their birth sex and their gender identity), secondary school can be even more hellish than for other people.

**Post:** Islam is no more than a repressive ideology that persecutes women and homosexuals.

**Counter Narrative:** Women are not pieces of meat in Islam.

**Post:** I am tired of women participating in political debate. Feminism is a divisive force in society.

**Counter Narrative:** In spite of erroneous beliefs, feminism has never been about supporting everything women do: it's just a political and social movement for equality.

**Post:** Multiculturalism sounds the death knell for Western societies and is also a mortal threat to minorities that are not protected by universal laws.

**Counter Narrative:** I think that the real target of these criticisms is not multiculturalism, but immigration and immigrants: multiculturalism has taken on multiple meanings that are rarely distinguished. Failure to distinguish between these meanings has enabled attacks on multiculturalism to be used to blame other groups.

## Prompt Experiment 2

### TARGET: HOMOPHOBIA

Generate the counter narrative, that is, a positive alternative, to the following offensive comments:

**Post:** I hate gays.

**Counter Narrative:** You know words have consequences, don't you? What if I told you that homosexuals are being attacked all over the world? In Cameroon, for example, they are the target of the police and the media. Local newspapers are outing homosexuals, inciting the public to attack them. Homosexuals have also been attacked this year in Zimbabwe, and in Senegal their tombs have been desecrated. Is this what they really want?

**Post:** We must stop imposing gender ideology on children.

**Counter Narrative:** Do you think there are "threats to the family from gender ideology"? Just in case you do, could you name a victim or give the number of people killed or mutilated by the gender threat? I'll give you a clue: zero. Although there have been many victims of gender discrimination.

**Post:** Gays are spreading AIDS throughout the world because they're promiscuous and fornicate with strangers, and they lie when they go to donate blood.

**Counter Narrative:** Aside from prejudice against gay people, it is not true that diseases are spread. The blood supply is safe because it is tested before it is used.

### OBJECTIVE OF HATE SPEECH: ISLAMOPHOBIA

Generate the counter narrative, that is, a positive alternative, to the following offensive comments:

**Post:** Islam is nothing more than a lunar cult disguised as a religion.

**Counter Narrative:** We should try to separate modern Islamic extremists from the religion of Islam. ISIS is a sect. Islam is not. Islamic is not the right word; radical religious extremists is the right expression. ISIS is a militant cult disguised as a religious movement.

**Post:** I'm realist, not racist. The rapes of children are perpetrated by Muslim men. Police and local services deceive non-Muslims. Why is this happening? Because everyone is afraid of the word racist. This is now the excuse for the police and the government; we're impotent because of this word! This is not Great Britain, it's the grooming capital of Europe.

**Counter Narrative:** Remember Operation Yewtree, the criminal investigation into historic allegations of child sexual abuse launched by the Metropolitan Police Service? The perpetrators of these sexual abuses (Jimmy Savile and others) were anything but Muslim, and that was covered up by the police! Keir Starmer,

the Director of Public Prosecutions, had to apologise for “deficiencies” in managing reports of abuse. And that cover-up had nothing to do with fear of the word racist.

**Post:** Every Muslim is a potential terrorist.

**Counter Narrative:** When the film against the prophet was shown, 1.5 billion Muslims around the world portrayed and characterised by the actions of tens of thousands of angry protesters. Here is another fact: the overwhelming majority of Muslims did not participate in protests, take to the streets or target embassies or restaurants. This is not an overwhelming majority of potential Muslim terrorists.

## **OBJECTIVE OF HATE SPEECH: ANTISEMITISM**

Generate the counter narrative, that is, a positive alternative, to the following offensive comments:

**Post:** The war in Iraq is the result of a small group of Jews who control the government of the United States.

**Counter Narrative:** If you’re referring to Dick Cheney, who was the 46<sup>th</sup> Vice President of the United States from 2001 to 2009 and Secretary of Defense at the start of the Iraq war, you’re wrong. In fact, while he was a member of the advisory board of the Jewish Institute for National Security Affairs, it is equally true that he is a Christian, a member of the Methodist Church.

**Post:** Jews are rich because they are naturally good with money.

**Counter Narrative:** The waves of Russian-speakers and Ethiopian Jews who fled from the collapse of the Soviet Union and poverty in Africa in the 1980s and 1990s- the so-called “aliyah crisis”- were not wealthy and comfortable people, believe me.

**Post:** Jews are selfish, stingy, and greedy.

**Counter Narrative:** You may not be aware of the Jewish tradition of charity: for example, people can donate money to breast cancer research in honour of their relatives. And there are many other examples.

## **OBJECTIVE OF HATE SPEECH: MISOGYNY**

Generate the counter narrative, that is, a positive alternative, to the following offensive comments:

**Post:** Women should stop complaining: no profession is closed to them in Western society.

**Counter Narrative:** Allow me to introduce you to the example of the STEM field in the United States. Women are only about 20% of the professors in these faculties. We should ask ourselves: “Why so few?”

**Post:** Equality is shit, even for women. The more power women have, the more they can avoid their domestic duties, the less happy they become.

**Counter Narrative:** So your ideal woman is someone like Michelle Duggar with her extreme lifestyle choices?

**Post:** I am fed up with women participating in political debate. Feminism is a divisive force in society.

**Counter Narrative:** Yes, it's a funny notion some notion, this feminism business, this bold notion that women should be able to live with the same freedom and enjoy the same inalienable rights and bodily autonomy as men.

### **OBJECTIVE OF HATE SPEECH: RACISM**

Generate the counter narrative, that is, a positive alternative, to the following offensive comments:

**Post:** Some races have lesser physical and cognitive abilities; the sooner we accept it, the better it will be for humanity.

**Counter Narrative:** Races are not mere neutral labels, because they can determine destiny. The negative associations between dark skin and human value were raised often and have done a great deal of damage to our society and humanity in general.

**Post:** Some races have lesser physical and cognitive abilities, so their lives are less valuable.

**Counter Narrative:** In reality, "simpler" organisms, such as social insects, have surprisingly advanced cognitive abilities. Should we place more value on these insects as well?

**Post:** This is an invasion: Millions of foreigners endangering our race and culture.

**Counter Narrative:** But the solution is not stigmatisation, isolation and creating fear. Think of the crime waves of the 1890s to 1930s, when millions of poor European immigrants went to the United States only to become mired in urban slums, suffering the effects of severe economic inequality and social marginalisation. The solution arrived with the inclusion of these immigrants.

### **Prompt Experiment 3**

**Post:** Jews think they are better than others.

**Counter Narrative:** Although it may have been true in the past (but it isn't), pride in Jewish identity, according to a large survey published by the Pew Research Center, is declining in the United States. In fact, the survey showed much lower rates of religious observance among Jewish millennials.

**Post:** Children should learn about gender identity and sexual orientation from their families and not in the classroom from LGBT fundamentalists.

**Counter Narrative:** But acceptance and understanding at school can help all children grow up as safe and proud people, regardless of their gender identity. For children and adolescents with gender dysphoria (a disconnect between their birth sex and their gender identity), secondary school can be even more hellish than for other people.

**Post:** Islam is no more than a repressive ideology that persecutes women and homosexuals.

**Counter Narrative:** Women are not pieces of meat in Islam.

**Post:** I am tired of women participating in political debate. Feminism is a divisive force in society.

**Counter Narrative:** In spite of erroneous beliefs, feminism has never been about supporting everything women do: it's just a political and social movement for equality.

**Post:** Multiculturalism sounds the death knell for Western societies and is also a mortal threat to minorities that are not protected by universal laws.

**Counter Narrative:** I think that the real target of these criticisms is not multiculturalism, but immigration and immigrants: multiculturalism has taken on multiple meanings that are rarely distinguished. Failure to distinguish between these meanings has enabled attacks on multiculturalism to be used to blame other groups.

# Appendix 2

## Prompt Experiment 1 for all the models

**Post:** People with Down's syndrome are retarded, so they are more likely to abuse drugs and alcohol.

**Counter Narrative:** Being born with Down's syndrome is not related to misbehaviour. Just like being born in a wheelchair, it's not related to misbehaviour. There is no correlation.

**Post:** Jews are not a religion; they are born that way and they will never be cured of their illness.

**Counter Narrative:** I don't agree, Jews are religious people who've always been part of the community. And they've been persecuted for it. I think this is what you meant by "they wouldn't be cured."

**Post:** Homosexuality and bisexuality are a disease, heterosexuality is normal.

**Counter Narrative:** These statements are neither true nor useful and simple fail to reflect reality as I've 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 United Kingdom and around the world. I doubt this has anything to do with immigrants.

**Post:** Most women are subhuman; they cannot be trusted with senior positions.

**Counter Narrative:** I think it's a comment against ALL women, not just some.

**Post:** Muslims have no place in civilised Europe.

**Counter Narrative:** In a modern democratic society such as 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 doesn't dictate that I can't love and support my partner, but it doesn't make me homophobic or racist.

**Post:** The United Kingdom Government recently published a document on “Online Harm;” in other words, they are going to work with Jewish and Muslim anti-hate organisations to censor the Internet. (Referring to Tell Mama, CST and Stop Hate UK).

**Counter Narrative:** The Internet has become a platform in which people shout their hateful opinions. Of course, something needs to be done, since people don’t seem to be capable of moderating their opinions for themselves.



