An Initial Exploration of How Argumentative Information Impacts Automatic Generation of Counter-Narratives Against Hate Speech

Damián Ariel Furman[1,2], Pablo Torres[3], José A. Rodríguez[3], Diego Letzen[3], Vanina Martínez[4] and Laura Alonso Alemany[3]

[1]University of Buenos Aires (UBA)), Intendente Güiraldes 2160 - Ciudad Universitaria, Buenos Aires, Argentina
[2]CONICET, Godoy Cruz 2290, Buenos Aires, Argentina
[3]Universidad Nacional de Córdoba, Argentina
[4]Artificial Intelligence Research Institute (IIIA-CSIC), Barcelona, Spain

Abstract

Fighting hate speech through automatic counter-narrative generation is gaining interest because of the increasing capabilities of Large Language Models. However, counter-narrative generation is a challenging task that can benefit from insightful analyses of text. In this work, we present an approach to improve the generation of counter-narratives by providing Large Language Models with high-quality examples. In addition, we show that enhancing the original hate speech with an argumentative analysis, identifying justifications and conclusions, together with collectives and the properties associated to them, seems to produce some improvements, specially with with smaller training datasets, helping to orient the generation towards a particular response strategy. The dataset of counter-narratives with argumentative information is made publicly available.

Warning: This work contains offensive and hateful text that may be distressing. It does not represent the views of the authors.

Keywords

Counter-narrative generation, Hate speech, Argument mining, Large Language Models

1. Introduction

In social media platforms, hate speech is amplified beyond human scale, spreading faster and increasing their reach, with negative impacts in societies, like polarization or an increase in violent episodes against targeted communities or individuals. It is because of these known consequences that many legal systems typify it as a crime, at least in some of its forms.

The predominant strategy adopted so far to counter hate speech in social media is to recognize, block and delete these messages and/or the users that generated it. This strategy has two main disadvantages. The first one is that blocking and deleting may prevent a hate message from spreading, but does not counter its consequences on those who were already reached by it.

Arg&App 2023: International Workshop on Argumentation and Applications, September 2023, Rhodes, Greece
damian.a.furman@gmail.com (D. A. Furman)
https://damifur.github.io/ (D. A. Furman)
0000-0002-0877-7063 (D. A. Furman)

The second one is that there is no place for subtleties or shades while defining hate speech: it must be done as a binary classification because the consequence of that classification is binary. This can generate accusations of overblocking or censorship, and not just because of errors in automated systems, which have been shown to be highly biased [1], but because blocking seems to be an overly simplistic approach to deal with the inherent complexity of hate speech.

An alternative to blocking that has been gaining attention in the last years, is to "*oppose hate content with counter-narratives (i.e. informed textual responses)*" [2, 3][1]. This way, the consequences of errors in the hate classification are minimized, overblocking is avoided, and it helps to spread a message against hate that can reach people that are not necessarily convinced, or even not involved in the conversation.

However, the huge volume of online hate messages makes the manual generation of counter-narratives an impossible task. In this scenario, automating the generation of counter-narratives is an appealing avenue, but the task poses a great challenge due to the complex linguistic and communicative patterns involved in argumentation.

Traditional machine learning approaches have typically produced less than satisfactory results for argumentation mining and generation. However, the recent availability of Large Language Models (LLMs) provides a promising approach to address the task of counter-narrative generation. Indeed, LLMs seem capable of generating satisfactory text for many tasks. Thorburn and Kruger [4] showed that a version of ChatGPT can tackle 6 argumentative reasoning tasks with some degree of success. They also find that finetuning the LLM parameters outperforms prompt-only based approaches.

However, as Hinton and Wagemans [5] show in their in-depth analysis of the argumentative capabilities of GPT-3, the argumentative text generated by LLMs tends to show some weaknesses. Although the language they use is clearly argumentative, as is the structure of arguments they create, most of them are not considered acceptable by humans, falling in fallacies like 'begging the question' and providing mostly irrelevant information.

In this paper we present an initial exploration of the impact of argumentative information in improving the quality of arguments generated by LLMs, more concretely, in improving the quality of automatically generated counter-narratives against hate speech. We compare different scenarios: LLMs without any specific adaptation to the task or domain, with fine-tuning using a dataset of counter-narratives, in a few-shot approach, and providing additional information about some of the argumentative aspects of the hate speech.

To assess the quality of the counter-narratives generated in the different scenarios, we carry out a preliminary evaluation with human judges, who achieved moderate agreement between each other. Based on those judgements, we can say that argumentative information by itself does not produce an improvement in the counter-narratives, but high-quality, specifically targeted fine-tuning seems to have a positive impact. Argumentative information does produce improvements in scenarios with very small training data and very specific fine-tuning, which seems promising to produce highly tailored counter-narratives, as in Gupta et al. [6].

The rest of the paper is organized as follows. In the next section, we review relevant work related to automated counter-narrative generation and argumentative analysis of hate speech. Then in Section 3 we describe our dataset of counter-narratives, with which we carry out the

[1]No Hate Speech Movement Campaign: http://www. nohatespeechmovement.org/

comparison of scenarios described in Section 4, where we also describe extensively our approach to the evaluation of generated counter-narratives, based on human judgements, and the prompts used to obtain the counter-narratives. Results analyzed in Section 5 show how fine-tuned LLMs and argumentative information provide better results, which we illustrate with some examples.

2. Related work

Automated counter-narrative generation has been recently tackled by leveraging the rapid advances in neural natural language generation. As with most natural language generation tasks in recent years, the basic machine learning approach has been to train or fine-tune a generative neural network with examples specific to the target task.

The CONAN dataset [3] is, to our knowledge, the first dataset with counter-narratives. It has 4078 Hate Speech – Counter Narrative original pairs manually written by NGO operators, translated to three languages: English, French and Italian. Data was augmented using automatic paraphrasing and translations between languages to obtain 15024 final pairs of hate speech – counter-narrative. Unfortunately, this dataset is not representative of the language in social media.

Similar approaches were carried out by Qian et al. [7] and Ziems et al. [8]. Qian et al. [7]'s dataset consists of reddit and Gab conversations where Mechanical Turkers identified hate speech and wrote responses.Ziems et al. [8] did not produce new text, but labeled COVID-19 related tweets as hate, counter-speech or neutral based on their hatefulness towards Asians.

In follow-up work to the seminal CONAN work, Tekiroğlu et al. [9] applied LLMs to assist experts in creating the corpus, with GPT-2 generating a set of counter-narratives for a given hate speech and experts editing and filtering them. Fanton et al. [10] iteratively refined a LLM where the automatically generated counter-narratives were filtered and post-edited by experts and then fed them to the LLM as further training examples to fine-tune it, in a number of iterations. Bonaldi et al. [11] apply this same approach to obtain a machine-generated dataset of dialogues between people producing hate speech and experts in hate countering. As a further enhancement in the LLM-based methodology, Chung et al. [12] enhanced the LLM assistance with a knowledge-based retrieval architecture to enrich counter-narrative generation.

Ashida and Komachi [13] use LLMs for generation with a *prompting* approach, instead of fine-tuning them with manually created or curated examples. They also propose a methodology to evaluate the generated output, based on human evaluation of some samples. This same approach is applied by Vallecillo-Rodríguez et al. [14] to create a dataset of counter-narratives for Spanish. Both these approaches are targeted to user-generated text, closely related to social media.

However, none of the aforementioned datasets or approaches to counter-narrative generation includes or integrates any additional annotated information apart from the hate message, possibly its context, and its response. That is why we consider an alternative approach that aims to reach generalization not by the sheer number of examples, but by providing a richer analysis of such examples that guides the model in finding adequate generalizations. We believe that information about the argumentative structure of hate speech, may be used as constraints for automatic counter-narrative generation.

Chung et al. [15] address an argumentative aspect of hate speech countering. They classify counter-narratives by type, using a LLM, and showing that knowledge about the type of counter-narratives can be successfully transferred across languages, but they do not use this information to generate counter-narratives.

To our knowledge, ours is the only corpus where tweets of hate speech have been annotated with argumentative information: ASOHMO [16], based on the Hateval corpus [17]. This dataset enriches the argumentative tweets of Hateval [17] with a manual analysis of their argumentative aspects, following an adaptation of the proposal of Wagemans [18], an analytic approach to represent the semantics of the core schemes proposed by Walton et al. [19], with fewer categories based on a limited set of general argument features. The following argumentative aspects are manually identified in tweets:

- **Justifications** and **Conclusions**.
- **Type** of Justification and Conclusion: Fact, Policy or Value.
- A **Pivot** signalling the argumentative relation between Justification and Premise.
- Two domain-specific components: the **Collective** which is the target of hate, and the **Property** that is assigned to such Collective.

In this work, we present counter-narratives manually associated to the hate tweets in ASOHMO and present an initial exploration of the impact of different kinds of information (counter-narratives, counter-narratives by subtype and information about argumentative components) in improving automatic generation of counter-narratives.

3. Creating counter-narratives associated to argumentative aspects of hate speech

Here we present CONEAS (Counter-Narratives Exploiting Argumentative Structure), a dataset of counter-narratives defined according to the argumentative information labeled on tweets from ASOHMO [16]. Each argumentative tweet is paired with counter-narratives of three different types defined by applying systematic transformations over argumentative components of the tweet, and a fourth type consisting of any counter-narrative that does not fall under any of the other three.

All counter-narratives, regardless of their type, also follow the guidelines of the Get The Trolls Out project[2]: *don't be aggressive or abusive, don't spread hate yourself, try to de-escalate the conversation, respond thinking on a wider audience than the person posting the original tweet and try to build a narrative.* Annotators were suggested to try to write at least one counter-narrative of each type but only if they came naturally, otherwise they could leave it blank.

The instructions to generate each type of counter-narrative are as follows:

Negate Relation Between Justification And Conclusion (Type A) Negate the implied relation between the justification and the conclusion.

[2] https://getthetrollsout.org/stoppinghate

HATE TWEET:
`user must deport all illegal migrants india already reeling under constant threat of muslim radicals curb population`

Justification: `india already reeling under constant threat of muslim radicals curb population` **(fact)**
Conclusion: `must deport all illegal migrants` **(policy)**
Collective: `illegal migrants`
Property: `muslim radicals`

COUNTER NARRATIVE A (Negate relation between justification and conclusion)
Deporting illegal migrants will not mitigate the problems with muslim radicals.
COUNTER NARRATIVE B (Negate relation between collective and property)
Illegal migrants are not necessarily muslim radicals.
COUNTER NARRATIVE C (Negate justification based on type)
It is not true that India is reeling under threat of muslim radicals.
FREE COUNTER NARRATIVE (Free)
Deporting illegal migrants without consideration to their circumstances is an inhumane move.

Figure 1: Examples of each type of counter narratives.

Negate association between Collective and Property (type B) Attack the relation between the property, action or consequence that is being assigned to the targeted group and the targeted group itself.

Attack Justification based on it is type (Type C) If the justification is a fact, then the fact must be put into question or sources must be asked to prove that fact. If it is of type "value", it must be highlighted that the premise is actually an opinion, possibly relativizing it as a xenophobous opinion. If it is a "policy", a counter policy must be provided.

Free Counter-Narrative (type D) All counter-narratives that the annotator comes up with and do not fall within any of the other three types.

An example of each type of counter-narrative can be seen in Figure 1. Our dataset[3] consists of a total of 1722 counter-narratives for 725 argumentative tweets in English and 355 counter-narratives for 144 tweets in Spanish (an average of 2.38 and 2.47 per tweet respectively). Table 1 shows the percentage of tweets that has a counter-narrative of each type.

4. Experiments

We designed a series of experiments to assess the impact of high-quality examples and argumentative information in the automatic generation of counter-narratives via prompting LLMs. We want to explore the following approaches:

Fine-tuned vs Few-shot Use a LLM that has been trained for general purposes to generate counter-narratives by prompting the LLM with some examples of the desired input-output,

[3] https://github.com/ConeasDataset/CONEAS/

as shown in the left column of Figure 3, or take a general LLM and fine-tune it with the examples of hate tweets associated to manually generated counter-narratives.

With or without argumentative information We want to assess the impact of different combinations of argumentative information provided within the input of the model: Collective and Property; Justification, Conclusion and Pivot; and all types.

With specific kinds of counter-narratives We pretrained two models for each type of counter-narrative using only that type: one without extra information and another adding argumentative information relevant for the correspondent type (Justification and Conclusion for type A, Collective and Property for type B and Justification for type C).

Small or Big size of the same kind of LLM We want to compare performance of a larger model with higher hardware requirements against a smaller one, fine-tuned, cheaper to run but requiring a specific annotated dataset. After testing behavior of similar alternatives (Bloom, GPT-J and GPT2), we chose Flan-T5 [12], an open model with base (250M parameters) and XL (3B parameters) versions that is instrution-fine-tuned.

Few-shot experiments were conducted for Flan-T5 Base (small) and XL (larger) models. fine-tuning was only conducted on Flan-T5 Base due to computational resource constraints.

We conducted some manual evaluation of prospective to find optimal parameters for generation, and we found that using Beam Search with 5 beams yielded the best results, so this is the configuration we used throughout the paper.

4.1. Fine-tuning of the LLM with counter-narratives

To fine-tune FLAN-T5 with our dataset of counter-narratives, we randomly split our dataset in training, development and test partitions, assuring that all counter-narratives for the same hate tweet are contained into the same partition. Details can be seen on Table 1.

	English						Spanish					
	#Tweets	#CNs	% corpus	A	B	C	#Tweets	#CNs	% corpus	A	B	C
Train	509	1201	69.8%	496	238	467	105	257	72.4%	101	59	97
Dev	71	173	10.0%	67	38	68	12	27	7.6%	12	8	7
Test	145	348	20.2%	138	74	136	27	71	20%	27	21	23
Proportion of tweets with counter-narrative												
				96%	47%	90%				97%	61%	89%

Table 1
Size of dataset partitions of English and Spanish datasets. Columns A, B and C show the amount of counter-narratives used for each partition when training only with counter-narratives of a given type.

All models were trained starting from Flan-T5-Base, in a multilingual setting using mixed English and Spanish examples, with a learning rate of 2e-05 for 8 epochs.

4.2. Experiments based on few-shot

For the few-shot experiments, the prompt has an instruction followed by two random examples taken from the test partition of the dataset. For each example, the hate tweet and its

corresponding counter-narrative are enclosed in special tokens defining the start and end.

4.3. Evaluation method for generated counter-narratives

Evaluation of counter-narratives is not straightforward. So far, no automatic technique has been found satisfactory for this specific purpose. Automatic metrics proposed for other NLP tasks, like BLEU [20] for automatic translation or ROUGE [21] for summarization, are not adequate for this task because they rely strongly on word or n-gram overlap with manually generated examples. These measures are disputed in the NLP community because, among other factors, they can't be adapted to cases where there can be many possible good outputs of the model, with significant differences between themselves, such as our case. We discarded these measures after comparing different counter-narratives of a same tweet from our dataset and noting that many of them scored 0 on both.

Faced with the lack of appropriate automatic metrics adequate for the task, many authors have conducted manual evaluations for automatically generated counter-narratives. Manual evaluations typically distinguish different aspects of the adequacy of a given text as a counter-narrative for another. Chung et al. [12] evaluate three aspect of the adequacy of counter-narratives: *Suitableness* (if the counter-narrative was suited as a response to the original hate message), *Informativeness* (how specific or generic the response is) and *Intra-coherence* (internal coherence of the counter-narrative regardless of the message it is responding to). Ashida and Komachi [13], on the other hand, assess these three other aspects: *Offensiveness*, *Stance* (towards the original tweet) and *Informativeness* (same as Chung et al. [12]).

Based on these previous works, we have put together a first version of criteria to manually evaluate[4] the adequacy of counter-narratives, considering four different aspects:

- **Offensiveness**: if the tweet is offensive to either the target group, the author of the tweet or any other group or person. Possible values are: Offensive; Possibly Offensive/Not clear; Not offensive.
- **Stance**: if the tweet supports or counters the specific message of the hate tweet. Possible values are: Supports the original message; Not clear/Changes subject wrt original tweet; Counters the original message. Stance incorporates a certain notion of suitableness, since it assigns value "Changes the subject" if the counter-narrative is not responding specifically to the standpoint of the original tweet.
- **Informativeness**: Evaluates the complexity and specificity of the generated text. Only counter-narratives with a "Counters" Stance are evaluated. Possible values are:
    1. **Generic statement**: replies that don't incorporate any information mentioned on the tweet and could counter many different hate messages (e.g "I don't think so" or "That is not true").
    2. **Specific but not argumentative**: the reply is a simple statement, possibly composed of a single sentence without providing justification for the stance but referring to some specific aspect of the original tweet. Usually they comply with a formula composed of a prefix (like "I don't think that" or "Do you have proof that") and a verbatim copy of some part of the hate tweet.

[4]Results of the evaluation can be found on https://shorturl.at/aetFZ

3. **Specific and Argumentative**: counter-narratives with some degree of elaboration of the information contained on the hate message. We identified three common patterns that we associate with this value:
   A - replies that take more than one element from the original message and stablish some relation between them (e.g. "I don't see the relation between {*element from the original message*} and {*other element from the original message*}").
   B - A simple statement declaring stance over a single element from the original tweet but adding a second coordinated statement with personal appreciations about it (e.g. "I don't think we should {*some policy mentioned on the tweet*}. It is a bad idea").
   C - An argumentative reply based on information not mentioned explicitly on the original tweet, but necessarily inferred, showing a comprehensive understanding of the meaning of the hate message (e.g. a reply to a tweet concluding with #BuildTheWall saying "*Building a wall would cost the taxpayers more*" or "*Building a wall won't give you more control over illegal trafficking*").

- **Felicity**: This category is related to Chung et al. [12]'s Intra-Coherence, but also considering additional dimensions like syntactical and semantic correctness. It evaluates independently of the original tweet, if the generated text sounds, by itself, fluent and correct. There are three possible values: The text is incoherent or semantic or syntactically incorrect; The text is coherent with small errors like incoordination of genre/tense/etc. or repeating parts of the original text without adapting them to the text being generated; The text is fluent and sounds correct.

Aggregating the results for these four categories, we define two extra concepts: Good and Excellent counter-narratives. Good counter-narratives will be those with optimal values on Offensiveness, Stance and Felicity. Excellent counter-narratives will be those that also have the optimal value for Informativeness. We believe Informativeness is the most valuable of the four categories, that is why it is determinant in characterizing Excellent counter-narratives. The Good indicator shows that productions are not harmful or totally random.

We are planning to improve the kind of information that is currently captured in the Informativeness category in a second version of the evaluation criteria.

4.4. Annotation environment and agreement

To properly evaluate the quality of the generated counter-narratives with the presented method, we conducted a preliminary manual evaluation. We evaluated three random subsets of 20 hate tweets in English and 10 in Spanish. One contains only tweets associated with counter-narratives of both types A and C on our dataset, and was used to evaluate models fine-tuned only with these kinds of counter-narratives. Another contains only tweets associated with counter-narratives of type B and was also used to evaluate models fine-tuned only with this type of counter-narratives. The last subset contains tweets with counter-narrative pairs of all types, and was used for all the rest of the experiments.

We generated one counter-narrative for each tweet in the corresponding evaluation subset for each combination of features to be assessed: few-shot, fine-tuned, with different kinds of

	1 vs 2	2 vs 3	1 vs 3
Offensiveness	0.47	0.40	0.41
Stance	0.63	0.58	0.63
Informativeness	0.49	0.42	0.54
Felicity	0.67	0.37	0.36

Table 2
Agreement scores between annotators 1, 2 and 3 using Cohen's Kappa.

argumentative information, with different sizes of LLM. For the larger version of FLAN-T5 we only applied the few-shot approach, and, after assessing no improvement on the smaller version, we aborted the rest of experiments with this version of the LLM to reduce the carbon footprint of our experiments. The results for the 18 experiments can be seen in Table 3.

Then, three annotators labeled each tweet according to the four categories described above. The final value for each category was obtained by calculating the value with more votes (at least two annotators agreed on the value). In total, each annotator labeled 540 hate tweet/counter-narrative pairs. Of all these, there were 10 cases where each of the three annotators labeled a different value. In these cases, we adopted a conservative criterion and assigned the worst of the three possible values.

Table 2 shows the agreement scores between the three annotators, calculated using Cohen's Kappa [22]. In most cases, agreement ranges from Moderate ($0.41 < \kappa < 0.60$) to Substantial ($0.61 < \kappa < 0.80$), except for the agreement achieved by annotator 3 against the other two on the category of Felicity which is just Fair ($0.21 < \kappa < 0.40$)[5]. As can be expected for such an interpretative task, agreement between annotators can be improved. However, this initial assessment served as a starting approach to assess the impact of different factors in the quality of generated counter-arguments.

We are currently working on a second version of the evaluation criteria, with more insightful categories, expanding on Informativeness and trying to capture argument acceptability, relevance and persuasiveness. We will check whether this improved criteria improve inter-annotator agreement. If so, we will engage a higher number of judges and aim to obtain a more reliable assessment of the quality of automatically generated counter-narratives.

5. Analysis of results

Results of the manual evaluation of different strategies for counter-narrative generation for English can be seen in Table 3. A summary of this table can be seen in Figure 2, which displays the aggregated proportion of Good and Excellent counter-narratives for each strategy.

We can clearly see that the larger versions of the model (XL) produce counter-narratives that are less satisfactory in general, and that argumentative information only decreases the quality of the generated text. Fine-tuned models produce better counter-narratives in general, even if smaller. A very valuable conclusion that can be obtained from these results is that a small number of high quality examples produce a much bigger improvement in performance than

[5]The interpretation of the ranges of values of the kappa coefficient is according to Landis and Koch [23].

Approaches	Offensiveness		Stance		Informative		Felicity	
	Off	NotOff	Supp	Count	Gen	Arg	Infel	Felic
Few-shot Approaches								
Base	10%	**60%**	15%	40%	40%	0%	15%	70%
Base All	40%	35%	45%	25%	10%	5%	10%	45%
Base Collective	5%	50%	15%	20%	15%	5%	5%	**90%**
Base Premises	30%	35%	30%	35%	20%	5%	5%	70%
XL	60%	25%	60%	25%	10%	0%	10%	45%
XL All	80%	10%	80%	10%	0%	10%	5%	15%
XL Collective	55%	25%	55%	15%	10%	5%	20%	0%
XL Premises	55%	10%	60%	0%	0%	0%	30%	25%
Fine-tuned Approaches								
Base	10%	**65%**	10%	**65%**	25%	35%	15%	**80%**
Base All	15%	45%	15%	30%	0%	10%	15%	**80%**
Base Collective	0%	55%	0%	**60%**	0%	35%	5%	**85%**
Base Premises	10%	40%	10%	45%	0%	30%	10%	**80%**
Base CNs A	10%	45%	10%	35%	5%	25%	35%	60%
Base CNs A Premises	10%	**60%**	10%	45%	0%	**40%**	25%	65%
Base CNs B	30%	20%	25%	5%	5%	0%	80%	10%
Base CNs B Collective	0%	15%	0%	15%	5%	10%	85%	15%
Base CNs C	0%	50%	5%	25%	20%	5%	65%	25%
Base CNs C Justification	10%	30%	10%	35%	5%	20%	25%	55%

Table 3
Manual evaluation of automatically generated counter-narratives for English hate tweets, using different sizes of the model (Base and XL), two learning techniques (few-shot and fine-tuning), two different training settings (all counter-narratives or only one kind: A, B or C) and different combinations of argumentative information (no information, Collective and Property, Premises and pivot and all the information available). We report the percentage of counter-narratives for the two extreme values of our four analysis categories: Offensiveness, Stance, Informativeness and Felicity.

using larger models, which are also more taxing.

If we focus on Informativeness (third dimension of evaluation in Table 3, we can see that the approaches that produce most informative counter-narratives are fine-tuned (lower half of the Table), without a detriment in any of the other dimensions of evaluation. Interestingly, when fine-tuned only with counter-narratives of a single type, providing argumentative information consistently improves the informativeness of the counter-narratives, even if only slightly. We have to take into account that such approaches use a much smaller number of counter-narratives, as can be seen in Table 1. Even in the case of type B counter-narratives, with extremely few examples to fine-tune, argumentative information produces an improvement in informativeness.

When we make a qualitative analysis of the generated counter-narratives, we can see that providing argumentative information about the hate tweet does yield counter-narratives that are more specific and informative, as can be seen in Figure 3. Models counting with this information frequently use it by negating the relation between Collective and Property or between Justification and Conclusion.

Results obtained for counter-narratives for Spanish hate tweets were much worse, as could be expected given the much smaller number of examples for fine-tuning and that base LLMs

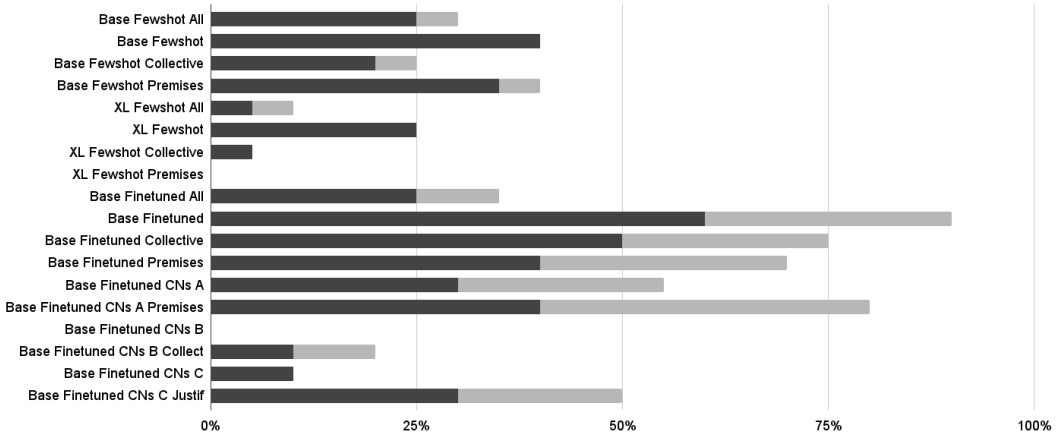

Figure 2: Proportion of Excellent (light gray) and Good (light + dark gray) counter-narratives produced by each approach.

Tweet with argumentative information:	Tweet without argumentative information:
street interview whit italians "send all migrants back to where they came from they block streets to pray " - free speech time - https://t co/d5dqr8pg3r @user | Justification: street interview whit italians "send all migrants back to where they came from they block streets to pray " (fact) | Conclusion: "send all migrants back to where they came from they block streets to pray " (policy) | Pivot: migrants - they - they	street interview whit italians "send all migrants back to where they came from they block streets to pray " - free speech time - https://t co/d5dqr8pg3r @user
Counter-narrative:	Counter-narrative:
I don't think it's a good idea to send all migrants back to where they came from.	I don't think it's the right thing to do.

Figure 3: Counter-narratives obtained for the same tweet with different strategies: including argumentative information (above) and without argumentative information (below).

perform worse for tasks in Spanish in general. Indeed, values for Informativeness and Felicity almost never reach more than 10% positive, and Stance and Offensiveness are almost never beyond 30% positiveness. However, the same tendency as for English could be observed: fine-tuned models perform better than non-fine-tuned models, even if the latter are bigger. Moreover, argumentative information seems to make a bigger impact in improving the generated counter-narratives than in the case of English, with increases in the range of 30%-50% in the reduction of negative scores for Offensiveness and Stance, although a decrease in Felicity. Given these

encouraging results with such few examples, we will be increasing the number of examples with argumentative information in future work.

6. Conclusions and future work

We have presented an approach to generate counter-narratives against hate speech in social media by prompting large language models with information about some argumentative aspects of the original hate speech. We have carried out a small manual evaluation of the quality of generated counter-narratives. This evaluation is preliminary, with a small number of judgements and moderate to substantial inter-annotator agreement, but we have found promising tendencies.

We have shown that argumentative information by itself does not improve the quality of counter-narratives generated by LLMs, on the contrary, it may even be detrimental, specially in the case of bigger models. However, fine-tuning a smaller model with a small corpus of high-quality examples of pairs hate speech – counter-narrative yields some improvement in performance. This finding has a significant impact both because smaller language models are more accessible to low-budget scenarios, and because of their smaller carbon footprint.

We have also shown that some kinds of argumentative information do have some positive impact in generating more specific, more informative counter-narratives. In particular, we have found that the types of counter-narrative that negate the relation between the Justification and the Conclusion and that negate the Justification have an improvement in performance if argumentative information about the Justification and the Conclusion is provided.

Moreover, we have also found that argumentative information makes a positive impact in scenarios with very few tweets, as shown by our experiments for Spanish. Although the quality of the counter-narratives generated for Spanish is much lower than for English, the fact that argumentative information has a positive impact is encouraging, and we will continue to annotate examples for Spanish to improve the generation of counter-narratives.

We will also explore other aspects of the quality of counter-narratives, with a more insightful, more extensive human evaluation. We will also explore the interaction between argumentative information and other aspects, like vocabulary, level of formality, and culture.

Finally, the evaluation of counter-narratives is still far from being solved. We are currently considering different avenues to improve it, as it is a crucial step to advance the field. We are working on obtaining a higher number of judgements, but also on more insightful guidelines that reflect more valuable aspects of counter-narratives, more related to argument acceptability.

7. Acknowledgments

This work was funded in part by Secretaría de Investigación Científica y Tecnológica FCEN–UBA (RESCS-2020-345-E-UBA-REC), CONICET under the PIP (grant 11220200101408CO), Agencia Nacional de Promoción Científica y Tecnológica, Argentina under grants PICT-2018-0475 (PRH-2014-0007), PICT-2020- SERIEA-01481, and the NAACL Regional Americas Fund (2022). This work used computational resources from CCAD – UNC (https://ccad.unc.edu.ar/), which are part of SNCAD – MinCyT, Argentina. We specially want to thank two anonymous reviewers that contributed to improve this work with their thoughtful and constructive comments.

References

[1] T. Davidson, D. Bhattacharya, I. Weber, Racial bias in hate speech and abusive language detection datasets, in: Proceedings of Third Workshop on Abusive Language Online, 2019.

[2] S. Benesch, Countering dangerous speech: New ideas for genocide prevention, United States Holocaust Memorial Museum, 2014.

[3] Y.-L. Chung, E. Kuzmenko, S. S. Tekiroglu, M. Guerini, CONAN - COunter NArratives through nichesourcing: a multilingual dataset of responses to fight online hate speech, in: ACL, 2019.

[4] L. Thorburn, A. Kruger, Optimizing language models for argumentative reasoning, in: Proceedings of the 1st Workshop on Argumentation & Machine Learning co-located with 9th International Conference on Computational Models of Argument (COMMA 2022), 2022.

[5] M. Hinton, J. H. M. Wagemans, How persuasive is ai-generated argumentation? an analysis of the quality of an argumentative text produced by the GPT-3 AI text generator, Argument Comput. 14 (2023) 59–74. URL: https://doi.org/10.3233/AAC-210026. doi:10.3233/AAC-210026.

[6] R. Gupta, S. Desai, M. Goel, A. Bandhakavi, T. Chakraborty, M. S. Akhtar, Counterspeeches up my sleeve! intent distribution learning and persistent fusion for intent-conditioned counterspeech generation, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 5792–5809. URL: https://aclanthology.org/2023.acl-long.318.

[7] J. Qian, A. Bethke, Y. Liu, E. M. Belding, W. Y. Wang, A benchmark dataset for learning to intervene in online hate speech, CoRR abs/1909.04251 (2019).

[8] C. Ziems, B. He, S. Soni, S. Kumar, Racism is a virus: Anti-asian hate and counterhate in social media during the covid-19 crisis, 2020.

[9] S. S. Tekiroğlu, Y.-L. Chung, M. Guerini, Generating counter narratives against online hate speech: Data and strategies, in: ACL, 2020.

[10] M. Fanton, H. Bonaldi, S. S. Tekiroğlu, M. Guerini, Human-in-the-loop for data collection: a multi-target counter narrative dataset to fight online hate speech, in: ACK, 2021.

[11] H. Bonaldi, S. Dellantonio, S. S. Tekiroğlu, M. Guerini, Human-machine collaboration approaches to build a dialogue dataset for hate speech countering, in: EMNLP, 2022.

[12] Y.-L. Chung, S. S. Tekiroğlu, M. Guerini, Towards knowledge-grounded counter narrative generation for hate speech, in: Findings of the ACL-IJCNLP 2021, 2021.

[13] M. Ashida, M. Komachi, Towards automatic generation of messages countering online hate speech and microaggressions, in: Proceedings of Sixth Workshop on Online Abuse and Harms (WOAH), 2022.

[14] M. Vallecillo-Rodríguez, A. Montejo Ráez, M. Martín-Valdivia, Automatic counter-narrative generation for hate speech in spanish, Procesamiento del Lenguaje Natural 71 (2023).

[15] Y.-L. Chung, M. Guerini, R. Agerri, Multilingual counter narrative type classification, in: Proceedings of the 8th Workshop on Argument Mining, Association for Computational Linguistics, Punta Cana, Dominican Republic, 2021, pp. 125–132. URL: https://aclanthology.org/2021.argmining-1.12. doi:10.18653/v1/2021.argmining-1.12.

[16] D. Furman, P. Torres, J. Rodríguez, D. Letzen, V. Martínez, L. Alonso Alemany, Which argumentative aspects of hate speech in social media can be reliably identified?, in: Proceedings of Fourth International Workshop on Designing Meaning Representations, co-located with IWCS 2023, 2023.
[17] V. Basile, C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, M. Sanguinetti, SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter, in: Proceedings of 13th International Workshop on Semantic Evaluation, 2019.
[18] J. H. M. Wagemans, Constructing a periodic table of arguments, in: Proceedings of 11th International Conference of the Ontario Society for the Study of Argumentation, 2016.
[19] D. Walton, C. Reed, F. Macagno, Argumentation Schemes, CUP, 2008.
[20] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: ACL, 2002.
[21] C.-Y. Lin, ROUGE: A package for automatic evaluation of summaries, in: Text Summarization Branches Out, 2004.
[22] J. Cohen, A Coefficient of Agreement for Nominal Scales, Educational and Psychological Measurement 20 (1960) 37.
[23] J. R. Landis, G. G. Koch, The measurement of observer agreement for categorical data, Biometrics 33 (1977) 159–174.