Procesamiento del Lenguaje Natural, Revista nº 73, septiembre de 2024, pp. 449-459 recibido 01-07-2024 revisado 22-07-2024 aceptado 25-07-2024

Overview of RefutES at IberLEF 2024: Automatic Generation of Counter Speech in Spanish

Resumen de la tarea RefutES en IberLEF 2024: Generación automática de refutaciones en español

María Estrella Vallecillo-Rodríguez[1], **María Victoria Cantero-Romero**[1],
Isabel Cabrera-de-Castro[1], **Luis Alfonso Ureña-López**[1],
Arturo Montejo-Ráez[1], **María Teresa Martín-Valdivia**[1]
[1]Department of Computer Science, Advanced Studies Center in ICT (CEATIC)
Universidad de Jaén, Campus Las Lagunillas, 23071, Jaén, Spain
{mevallec, vcantero, iccastro, laurena, amontejo, maite}@ujaen.es

Abstract: This paper presents an overview of RefutES 2024, organized at IberLEF 2024 and co-located with the 40th International Conference of the Spanish Society for Natural Language Processing (SEPLN 2024). The main purpose of RefutES is to promote research on the automatic generation of counter speech in Spanish. Counter speech generation is a new strategy developed to combat hate speech on social media that involves generating a response that negates the offensive message. In this shared task, participants must be able to generate a response to hate speech messages directed at various targets of offense in Spanish. The response should be reasoned, respectful, non-offensive, and contain specific and truthful information. Moreover, we asked participants to submit measurements of carbon emissions for their systems, emphasizing the need for sustainable NLP practices. In this first edition, a total of 6 teams signed up to participate in the task, 1 submitted official runs on the test data, and 1 submitted system description papers.

Keywords: Counter speech Generation, Hate-Speech in Spanish, Large Language Models, Text Generation.

Resumen: Este artículo presenta la tarea RefutES 2024, organizada en IberLEF 2024 junto a la 40ª Conferencia Internacional de la Sociedad Española para el Procesamiento del Lenguaje Natural (SEPLN 2024). El objetivo principal de RefutES es promover la investigación sobre la generación automática de contranarrativas en español. La generación de contranarrativas es una nueva estrategia desarrollada para combatir los mensajes de odio en redes sociales que consiste en la generación de una respuesta que niega el mensaje offensivo. En esta tarea compartida, los participantes deben generar una respuesta a mensajes de odio que están dirigidos a diferentes colectivos en español. Esta respuesta debe de ser argumentada, respetuosa, no ofensiva y contender información específica y veraz. Además, los participantes tienen que presentar mediciones de las emisiones de carbono de sus sistemas, haciendo hincapié en la necesidad de prácticas de PNL sostenibles. En esta primera edición, un total de 6 equipos se registration en la tarea, 1 subió los resultados de las ejecuciones realizadas sobre los datos de test y 1 escribió el artículo con la descripción de su sistema.

Palabras clave: Generación de contranarrativas, Discurso de odio en español, Modelos grandes del lenguaje, Generación de textos.

1 *Introduction*

The generation of counter-narratives, or counterspeech, to combat hate messages involves creating responses to offensive messages that reject and deconstruct the narratives behind them. This strategy aims to avoid the censorship that occurs when users who write offensive messages are blocked or removed from social networks. The challenge lies in denying the offensive message and providing arguments explaining why the comment is offensive and inappropriate. This is essential to promote a safe, inclusive, and respectful online environment (Benesch, 2014;

ISSN 1135-5948 DOI 10.26342/2024-73-34

Schieb and Preuss, 2016).

This strategy to address offensive messages has traditionally been carried out manually by individuals or NGOs. However, this method is not effective because it is difficult to monitor all digital platforms manually, and it can negatively impact the mental health of those moderating the content. Therefore, researchers are exploring a new approach that uses Natural Language Processing (NLP) and Machine Learning (ML) techniques to automatically generate counter-narratives (Bonaldi et al., 2024). Most existing work focuses on collecting and generating counter-narratives for English (Qian et al., 2019; Tekiroglu et al., 2022; Halim et al., 2023; Mathew et al., 2018), highlighting the need for efforts to develop counter-narrative datasets and systems for Spanish, where only a few studies have been conducted (Furman et al., 2023; Vallecillo-Rodríguez, Montejo-Raéz, and Martín-Valdivia, 2023; Vallecillo-Rodríguez et al., 2024; Bengoetxea et al., 2024).

RefutES is a novel task on the refutation of hate-speech messages directed to different targets of offensive organized within the Iberian Languages Evaluation Forum (IberLEF 2024) (Chiruzzo, Jiménez-Zafra, and Rangel, 2024). In this task, participants must be able to generate a response to the offensive message in Spanish. The response should be reasoned, respectful, non-offensive, and contain information that is specific and truthful.

2 Task Description

The aim of RefutES 2024 shared task is to promote the automatic generation of counter-narratives in Spanish. It consists of the generation of a response to hate speech messages directed to different offended collectives. The response should be reasoned, respectful, non-offensive, and contain specific and truthful information. The offended targets are disabled, jews, LGBT+, migrants, muslims, people of colour, women, and other groups.

The challenges faced in this task are:

1 Identify what is being offensive in the message received.

2 Generate neutral and respectful responses, avoiding falling into the same offensive tone of the original message.

3 Informative and well-argued answers, presenting reliable information with coherent arguments and avoiding hallucinations or confabulations.

4 Adaptation to the large number of topics that contain offensive messages. Offenses vary for each offended group and each type may require a different approach. It is also essential to avoid bias.

5 Apply generative language techniques to Spanish and prevent answers from containing grammatical errors.

To develop their systems, participants will receive dataset partitions for training and development of their systems. In the final phase, the participants will receive the test partition with which we will evaluate their systems. To evaluate the systems we will take reference counter-narratives that will be taken as gold standard for the automatic evaluation and in addition 5 people will manually evaluate a subset of the test data. With these metrics, the winner will be determined.

2.1 Evaluation Metrics

We have multiple metrics to measure systems's behavior arranged in three main categories: performance, efficiency, and manual annotations.

2.1.1 Performance metrics

Performance metrics are intended to measure how well systems achieve the proposed task regarding prediction quality. Each submission must be evaluated with the following metrics:

- **Sentence-MoverScore (MS) (Zhao et al., 2019):** Calculates the semantic similarity between two sentences using contextualized embeddings, allowing many-to-one pairings of embeddings. This improves the evaluation of the similarity between generated texts and human references.
- **BERTScore (BS)(Zhang et al., 2019):** Measures the token-level semantic similarity between a generated sentence and a reference sentence using contextual embeddings, allowing one-to-one matchings of semantically similar tokens. BERTScore includes BERT-precision, BERT-recall, and BERT-F1 score.

These metrics are designed to evaluate the semantic similarity between two sentences (a reference and a candidate). To calculate this, we use the XLMRoBERTa-large (Conneau et al., 2019) model.

2.1.2 Efficiency metrics

Efficiency metrics are intended to measure the impact of the system in terms of resources needed and environmental issues. We want to recognize those systems that can perform the task with minimal demand for resources. This will allow us to, for instance, identify those technologies that could run on a mobile device or a personal computer, along with those with the lowest carbon footprint. Each submission must contain the following information:

- The time in seconds to make a prediction.
- The kilograms of CO2 emissions generated when making a prediction.
- The energy per CPU or/and GPU (kW) used when making a prediction.
- The energy used per RAM (kW) when making a prediction
- A sum of CPU energy, GPU energy, and RAM energy (kW) consumed
- The number of CPU or/and GPU used, their models, and the total RAM size needed.

For this, the Code Carbon tool (Courty et al., 2024) will be used.

2.1.3 Manual annotations metrics

At the end of the evaluation campaign, we will make a random sampling. This sampling consists of a random selection of a few Hate-Speech and Counter-narrative (HS-CN) pairs from the submission of the participants (the sampling includes pairs with higher and low results in our performance metrics). Specifically, this sample will contain 10 HS-CN pairs for each hate target in the dataset, 5 in the CONAN-MT-DP dataset, and 5 obtained from X (Twitter) and human-generated counter-narratives. To this sampling, we will apply manual annotation metrics to measure how well proposed automatic metrics align with manual annotation metrics. Human annotators must evaluate each pair of random sampling with the same metrics as the CONAN-MT-SP Corpus:

- **Offensiveness:** 0 (not sure), 1 (not offensive), 2 (maybe offensive), 3 (completely offensive).
- **Stance:** 0 (irrelevant), 1 (strongly agree), 2 (slightly agree/disagree), 3 (strongly disagree).
- **Informativeness:** 0 (irrelevant), 1 (not informative), 2 (generic and uninformative statement), 3 (specific and informative).
- **Truthfulness:** 0 (not sure), 1 (not true), 2 (partially true), 3 (completely true).
- **Editing required:** 0 (no editing), 1 (yes editing).

3 Dataset and Baselines

3.1 CONAN-MT-SP

A new dataset has been created for this task. We are going to release the corpus CONAN-MT-SP, which consists of 3635 HS-CN pairs covering 8 different hate targets (disabled, jews, LGBT+, migrants, muslims, people of colour (POC), women, and other groups).

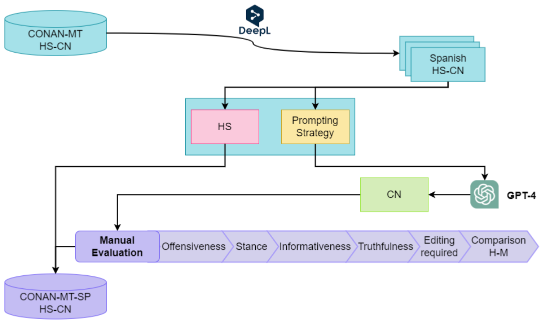

Figure 1: Schema of the creation of CONAN-MT-SP Dataset.

As we can see in Figure 1, to build CONAN-MT-SP, we use the hate speech of the English MultiTarget CONAN (CONAN-MT) corpus (Fanton et al., 2021) that collected its HS-CN pairs by niche sourcing from two different NGOs and subsequently used these pairs to generate more HS-CN with GPT-4 with human review integrated into the process. Due to the fact that the hate speech message is in English in CONAN-MT, we translate it into Spanish using the DeepL API. All translations were reviewed by our annotators, and in those pairs where the translations were erroneous, they were edited. The associated counter-narrative to each hate-speech message is generated by the GPT-4 model using a prompt strategy. The strategy used consisted in a Few Shot Learning Strategy, where the model was prompted with a task description and 8 examples of HS-CN pairs (one for each target). In addition, the counter-narrative generated by GPT-4 has been evaluated by human experts in terms of Offensiveness, Stance, Informativeness, Truthfulness, Editing required, and Comparison between Human Model.

In RefutES, we selected from this corpus the "perfect" counter-narratives, in other words, those that are non-offensive, in complete disagreement, specific and informative, completely truthful, do not need editing, and are equal or better than the original CONAN-MT counter-narrative. Furthermore, in order not to have only data coming from an English-to-Spanish translation, we extracted 78 Spanish comments from *X* and with human annotators generated the counter-narrative associated with these messages. This subset of data together with a partition of CONAN-MT-SP will be used to evaluate the systems. The corpus is divided into three subsets, each related to a different part of the competition. In Table 1 we present the distribution of the dataset, indicating the number of instances of the dataset for each split and label.

Label	**Train**	**Test**		**Development**	**Total**
		GPT-4	**Human**		
Disabled	120	9	10	13	152
Jews	201	9	10	23	243
LGBT+	250	10	10	28	298
Migrants	493	10	10	55	568
Muslims	711	10	10	79	810
POC	187	10	11	21	229
Women	432	10	10	48	500
Other groups	101	10	7	11	129
Total	2495	78	78	278	2929

Table 1: Distribution of instances of the RefutES dataset for each label and split.

3.2 Baselines

To establish a benchmark for the RefutES corpus, we conducted two experiments using the LLaMA2-Chat-13B model (Touvron et al., 2023). As a generative model, we explored two strategies: Zero-Shot Learning (ZSL) and fine-tuning using QLoRA (Dettmers et al., 2023).

For the ZSL strategy, we provided the model with a task description followed by the offensive message, prompting it to generate a counter-narrative. For the second strategy, we fine-tuned the LLaMA2-Chat-13B model using QLoRA. QLoRA combines Low-Rank Adaptation (LoRA) fine-tuning, which introduces trainable rank decomposition matrices in each transformer layer, with quantization to reduce memory usage by quantizing weight parameters while keeping pre-trained weights frozen.

> Eres un experto en contranarrativa, es decir, en elaborar respuestas informativas en español que contesten a mensajes ofensivos. Dichas respuestas deben de ser respetuosas y breves.
> ###Post: {texto mensaje ofensivo}
> ###Contranarrativa: {texto contranarrativa}
>
> (You are an expert in counter-narrative, that is, in elaborating informative responses in Spanish that answer offensive messages. Such responses should be respectful and brief.
> ###Post: {offensive text message}
> ###Counter-narrative {counter-narrative text})

Figure 2: Structure of the prompt used for the implementation of the baselines.

The LLaMA2-Chat-13B model was trained on an RTX 6000 GPU for 10 epochs with an early stopping set to 3 epochs. The learning rate was 2e-4. Quantization was applied using 4-bit NormalFloat for weight storage and 16-bit bfloat16 for computations. The rank (r) was set to 16, with a scaling factor $(alpha)$ of 64, and a dropout rate of 0.05. The maximum number of tokens to be generated was 150.

In both experiments, we used the prompt structure shown in the Figure 2. The text following "###Contranarrative: " ("*###Contranarrative:* ") will only be provided to the models during training, as it is the one they must generate during the evaluation.

4 Participant approaches

A total of 6 teams from 4 countries (Spain, Mexico, Colombia, and Vietnam) signed up for RefutES 2024. Among them only 1 submitted runs for the shared task. Each team had the opportunity to present a maximum of 3 runs, demonstrating their experience and

Rank	**Team run**	**BERTScore**			**MoverScore**	**(F1$_{\mathbf{BS}}$ + MS)/2**
		f1	**precision**	**recall**		
1	Ixa-run3	0.8923	0.8974	0.8948	0.6325	0.8923
2	Llama RefutES	0.89203	0.90456	0.89819	0.62645	0.8920
3	LLaMA ZSL	0.8723	0.8920	0.8820	0.6060	0.8723
4	Ixa-run2	0.8606	0.8859	0.8729	0.5956	0.8606
5	Ixa-run1	0.8569	0.8897	0.8729	0.5885	0.8569

Table 2: Global results in Performance metrics. Ranking metric: $(F1_{BS} + MS)/2$.

Rank	**Team run**	**Label**	**Percentage**				
			Offensiveness	**Stance**	**Informativeness**	**Truthfulness**	**Editing required**
1	Ixa-run3	0	0,00	0,00	0,00	1,25	**58,75**
		1	**98,75**	0,00	2,50	1,25	41,25
		2	1,25	1,25	68,75	6,25	-
		3	0,00	**98,75**	**28,75**	**91,25**	-
2	Ixa-run1	0	0,00	1,25	2,50	2,50	**22,50**
		1	**80,00**	1,25	6,25	6,25	77,50
		2	13,75	1,25	51,25	28,75	-
		3	6,25	**96,25**	**40,00**	**62,50**	-
3	Ixa-run2	0	6,25	11,25	8,75	8,75	**18,75**
		1	**75,00**	0,00	10,00	5,00	81,25
		2	15,00	3,75	41,25	22,5	-
		3	3,75	**85,00**	**40,00**	**63,75**	-

Table 3: Global results in manual annotation metrics. The position of the label that in a perfect system should have the highest score is shown in bold type.

strategies in the challenge. Below we describe the approaches of the team that participated in the competition.

Ixa (Zubiaga, Soroa, and Agerri, 2024) explored open LLMs such as Mistral-Instruct-7B, Zephyr, and a quantified version of Command-R model of 35B parameters to generate counter-narratives. Moreover, to evaluate their systems, they studied the use of JudgeLM (a vicuna-based scalable language model judge designed to evaluate LLMs in open-ended scenarios). Concretely, for the first run, they generate counter-narratives using ZSL with the 3 selected models and use the JudgeLM model to determine the best counter-narrative among the three candidates. In the second run use the results of the counter-narratives used by the 3 models and each tournament of JudgeLM. Moreover, they used the Maximum Likelihood Estimate ELO to establish a hierarchy of models based on their performance as determined by JudgeLM. Subsequently, they selected the top-performing model and leveraged it to generate counter-narratives in a ZSL. In run 3 they use the model that achieved the most effective counter-narrative in the last experiments thinking that this model obtained better results if it is finetuned. So they train a MistralInstruct 7B model using QLoRA, an efficiency technique to train LLMs.

5 *Results*

The results of the automatic and manual evaluation of the systems presented in our task are shown in this section. We have grouped the results according to the different aspects we wanted to evaluate, the results in general, grouped according to the hate target of the offensive messages to be counter-narratives and according to the origin of the reference counter-narrative with which the generated counter-narratives are compared or, in other words, who generates these counter-narratives taken as a reference.

5.1 Global results

The results obtained by the Ixa Team in each type of Evaluation are presented in Tables 2 and 3. As we can see in Table 2 which refers to the automatic results, the fine-tuning model (Ixa-run3) is the best-performing system followed closely by the LLaMA baseline trained with the RefutES data. Both models differ from the rest by more than 0.02 in the average of F1-BERTScore and MoverScore. The results of the baseline applying ZSL and the other two Ixa team runs are worse, as the models that have been trained with the task data are closer to the task data than the models that have not obtained that information.

In the manual annotation results (Table 3), Ixa-run 3, which corresponds to a model trained with the task data, again achieves the best results across all evaluation metrics

Team run	**Label**	**BERTScore**			**MoverScore**	**(F1$_{BS}$+ MS)/2**
		f1	**precision**	**recall**		
Ixa-run3	Disabled	0.8888	0.8965	0.8924	0.6346	0.8888
	Jews	0.8903	0.8938	0.8920	0.6300	0.8903
	LGBT+	0.8938	0.8959	0.8948	0.6328	0.8938
	Migrants	0.8906	0.8961	0.8933	0.6325	0.8906
	Women	0.8943	0.8975	0.8959	0.6395	0.8943
	Muslims	0.8944	0.9022	0.8982	0.6339	**0.8944**
	POC	0.8923	0.8988	0.8955	0.6266	0.8923
	Others	0.8942	0.8987	0.8963	0.6303	0.8942
Llama RefutES	Disabled	0.89333	0.90635	0.89976	0.62239	0.8933
	Jews	0.88969	0.90029	0.89489	0.62573	0.8897
	LGBT+	0.89126	0.90284	0.89692	0.61944	0.8913
	Migrants	0.89366	0.9084	0.90092	0.63233	0.8937
	Women	0.89015	0.90092	0.89549	0.63325	0.8902
	Muslims	0.89308	0.90492	0.89889	0.62645	0.8931
	POC	0.89023	0.90786	0.8989	0.62104	0.8902
	Others	0.89538	0.9046	0.89991	0.63184	**0.8954**
LLaMA ZSL	Disabled	0.8721	0.8924	0.8821	0.6093	0.8721
	Jews	0.8686	0.8912	0.8797	0.6046	0.8686
	LGBT+	0.8714	0.8900	0.8805	0.6036	0.8714
	Migrants	0.8753	0.8927	0.8838	0.6103	**0.8753**
	Women	0.8742	0.8914	0.8827	0.6084	0.8742
	Muslims	0.8704	0.8896	0.8798	0.6021	0.8704
	POC	0.8717	0.8935	0.8824	0.6019	0.8717
	Others	0.8748	0.8956	0.8850	0.6088	0.8748
Ixa-run2	Disabled	0.8607	0.8871	0.8736	0.5993	0.8607
	Jews	0.8598	0.8828	0.8711	0.5928	0.8598
	LGBT+	0.8578	0.8819	0.8695	0.5902	0.8578
	Migrants	0.8678	0.8929	0.8801	0.6066	**0.8678**
	Women	0.8612	0.8893	0.8750	0.5993	0.8612
	Muslims	0.8594	0.8849	0.8718	0.5882	0.8594
	POC	0.8568	0.8828	0.8695	0.5894	0.8568
	Others	0.8613	0.8849	0.8729	0.5995	0.8613
Ixa-run1	Disabled	0.8596	0.8904	0.8747	0.5937	0.8596
	Jews	0.8554	0.8893	0.8720	0.5822	0.8554
	LGBT+	0.8564	0.8910	0.8733	0.5877	0.8564
	Migrants	0.8575	0.8912	0.8740	0.5894	0.8575
	Women	0.8506	0.8866	0.8682	0.5880	0.8506
	Muslims	0.8603	0.8893	0.8745	0.5902	0.8603
	POC	0.8551	0.8896	0.8720	0.5854	0.8551
	Others	0.8614	0.8900	0.8754	0.5922	**0.8614**

Table 4: Results of performance metrics grouped by each offended target in hate speech messages within the test split of the RefutES dataset.

except for informativeness, where it scores lower than the other runs. Additionally, Ixa team's run 1 outperforms run 2, highlighting the challenges of evaluating such systems using automatic metrics. It is also notable that some LLMs, such as JudgeLM, demonstrate the ability to evaluate human-preferred counter-narratives.

5.2 Grouping by hate target

This section shows the analysis of the results obtained according to the hate label of the offensive messages for which the counter-narratives are elaborated. Looking at the results of the automatic evaluation (Table 6), no significant differences between the counter-narratives generated for each hate target within the different systems developed are apparent. However, in Table 7, which shows the results of the manual evaluation, the situation is different.

The system developed for Ixa-run1 shows that the counter-narratives generated for POC and migrants are more offensive, while the informativeness and veracity of the counter-narratives directed at Jews are higher than in the rest of the labels. In addition, it is

Origin CN	**Label**	**Ixa-run1 (%)**					**Ixa-run2 (%)**					**Ixa-run3 (%)**				
		Off	**Sta**	**Inf**	**Tru**	**Edi**	**Off**	**Sta**	**Inf**	**Tru**	**Edi**	**Off**	**Sta**	**Inf**	**Tru**	**Edi**
Disabled	0	0	0	0	0	**10**	10	10	10	10	**10**	0	0	0	0	**60**
	1	**90**	0	0	10	90	**80**	0	0	10	90	**100**	0	0	0	40
	2	0	0	70	30	-	0	0	60	30	-	0	0	70	0	-
	3	10	**100**	**30**	**60**	-	10	**90**	**30**	**50**	-	0	**100**	**30**	**100**	-
Jews	0	0	0	0	0	**30**	10	30	10	10	**40**	0	0	0	0	**40**
	1	**90**	0	0	0	70	**80**	0	20	0	60	**100**	0	10	0	60
	2	10	0	40	10	-	10	0	30	30	-	0	0	50	20	-
	3	0	**100**	**60**	**90**	-	0	**70**	**40**	**60**	-	0	**100**	**40**	**80**	-
LGBT+	0	0	0	0	0	**20**	20	20	30	20	**20**	0	0	0	0	**70**
	1	**90**	0	10	0	80	**60**	0	0	0	80	**100**	0	0	0	30
	2	10	0	70	30	-	20	0	40	10	-	0	0	50	0	-
	3	0	**100**	**20**	**70**	-	0	**80**	**30**	**70**	-	0	**100**	**50**	**100**	-
Migrants	0	0	0	0	0	**10**	0	0	0	10	**10**	0	0	0	10	**80**
	1	**70**	0	0	0	90	**60**	0	10	0	90	**100**	0	10	0	20
	2	20	10	60	40	-	40	10	40	20	-	0	0	50	0	-
	3	10	**90**	**40**	**60**	-	0	**90**	**50**	**70**	-	0	**100**	**40**	**90**	-
Women	0	0	10	10	0	**20**	0	10	10	0	**20**	0	0	0	0	**60**
	1	**80**	0	0	20	80	**70**	0	10	20	80	**100**	0	0	0	40
	2	10	0	50	0	-	20	0	40	0	-	0	0	70	0	-
	3	10	**90**	**40**	**80**	-	10	**90**	**40**	**80**	-	0	**100**	**30**	**100**	-
Muslims	0	0	0	10	10	**0**	0	0	0	0	**10**	0	0	0	0	**50**
	1	**80**	0	0	0	100	**100**	0	0	0	90	**100**	0	0	0	50
	2	20	0	50	60	-	0	0	60	30	-	0	0	80	10	-
	3	0	**100**	**40**	**30**	-	0	**100**	**40**	**70**	-	0	**100**	**20**	**90**	-
POC	0	0	0	0	10	**50**	0	0	0	10	**40**	0	0	0	0	**60**
	1	**50**	0	20	10	50	**70**	0	30	0	60	**90**	0	0	0	40
	2	30	0	50	30	-	20	10	40	40	-	10	0	80	10	-
	3	20	**100**	**30**	**50**	-	10	**90**	**30**	**50**	-	0	**100**	**20**	**90**	-
Others	0	0	0	0	0	**40**	10	20	10	10	**0**	0	0	0	0	**50**
	1	**90**	10	20	10	60	**80**	0	10	10	100	**100**	0	0	10	50
	2	10	0	20	30	-	10	10	20	20	-	0	10	100	10	-
	3	0	**90**	**60**	**60**	-	0	**70**	**60**	**60**	-	0	**90**	**0**	**80**	-

Table 5: Results of manual annotation metrics grouped by each offended target in hate speech messages within the test split of the RefutES dataset. The position of the label that in a perfect system should have the highest score is shown in bold type. Off: Offensiveness, Sta: Stance, Inf: Informativeness, Tru: Truthfulness, Edi: Edition required.

observed that counter-narratives about hate messages towards POC and other groups require more editing.

In this team's run 2, counter-narratives targeting LGBT+ and migrants are more offensive than those targeting other hate groups. On the other hand, counter-narratives targeting other groups and women have very high values in terms of informativeness and truthfulness, respectively. As for the need for editing, all counter-narratives require quite a lot of editing, except those against POC and jews.

Finally, when analyzing Ixa-run3, significant differences in terms of informativeness are appreciated, with very generic values for all counter-narratives except for LGBT+. In addition, counter-narratives targeting jews require a lot of editing, while those targeting migrants and LGBT+ do not require as much editing.

With all this, we can conclude that the manual evaluation has been fundamental to detecting the different biases in the language models, appreciating differences according to the hate target of the offensive messages. All these biases found may be due to prior knowledge of the models used, to biases contained in the task dataset itself, or to the unbalancing of the task dataset.

5.3 Grouping by the origin of reference counter-narrative

This section presents the results of the evaluation performed automatically and manually, grouped according to the origin of the reference counter-narratives. Looking at the results of the automatic evaluation (Table 8), it can be seen that the counter-narratives generated by the presented systems are more similar to those generated by GPT-4, belonging to the CONAN-MT-SP dataset, than to those introduced to evaluate the task and generated by humans. Table 9 shows the results according to the manual evaluation. When analyzing this table, we observe the same trend as in the previous one: systems generate better counter-narratives in terms of of-

Team run	**Label**	**BERTScore**			**MoverScore**	**(F1$_{\text{BS}}$+ MS)/2**
		f1	**precision**	**recall**		
Ixa-run3	Disabled	0.8888	0.8965	0.8924	0.6346	0.8888
	Jews	0.8903	0.8938	0.8920	0.6300	0.8903
	LGBT+	0.8938	0.8959	0.8948	0.6328	0.8938
	Migrants	0.8906	0.8961	0.8933	0.6325	0.8906
	Women	0.8943	0.8975	0.8959	0.6395	0.8943
	Muslims	0.8944	0.9022	0.8982	0.6339	**0.8944**
	POC	0.8923	0.8988	0.8955	0.6266	0.8923
	Others	0.8942	0.8987	0.8963	0.6303	0.8942
Llama RefutES	Disabled	0.89333	0.90635	0.89976	0.62239	0.8933
	Jews	0.88969	0.90029	0.89489	0.62573	0.8897
	LGBT+	0.89126	0.90284	0.89692	0.61944	0.8913
	Migrants	0.89366	0.9084	0.90092	0.63233	0.8937
	Women	0.89015	0.90092	0.89549	0.63325	0.8902
	Muslims	0.89308	0.90492	0.89889	0.62645	0.8931
	POC	0.89023	0.90786	0.8989	0.62104	0.8902
	Others	0.89538	0.9046	0.89991	0.63184	**0.8954**
LLaMA ZSL	Disabled	0.8721	0.8924	0.8821	0.6093	0.8721
	Jews	0.8686	0.8912	0.8797	0.6046	0.8686
	LGBT+	0.8714	0.8900	0.8805	0.6036	0.8714
	Migrants	0.8753	0.8927	0.8838	0.6103	**0.8753**
	Women	0.8742	0.8914	0.8827	0.6084	0.8742
	Muslims	0.8704	0.8896	0.8798	0.6021	0.8704
	POC	0.8717	0.8935	0.8824	0.6019	0.8717
	Others	0.8748	0.8956	0.8850	0.6088	0.8748
Ixa-run2	Disabled	0.8607	0.8871	0.8736	0.5993	0.8607
	Jews	0.8598	0.8828	0.8711	0.5928	0.8598
	LGBT+	0.8578	0.8819	0.8695	0.5902	0.8578
	Migrants	0.8678	0.8929	0.8801	0.6066	**0.8678**
	Women	0.8612	0.8893	0.8750	0.5993	0.8612
	Muslims	0.8594	0.8849	0.8718	0.5882	0.8594
	POC	0.8568	0.8828	0.8695	0.5894	0.8568
	Others	0.8613	0.8849	0.8729	0.5995	0.8613
Ixa-run1	Disabled	0.8596	0.8904	0.8747	0.5937	0.8596
	Jews	0.8554	0.8893	0.8720	0.5822	0.8554
	LGBT+	0.8564	0.8910	0.8733	0.5877	0.8564
	Migrants	0.8575	0.8912	0.8740	0.5894	0.8575
	Women	0.8506	0.8866	0.8682	0.5880	0.8506
	Muslims	0.8603	0.8893	0.8745	0.5902	0.8603
	POC	0.8551	0.8896	0.8720	0.5854	0.8551
	Others	0.8614	0.8900	0.8754	0.5922	**0.8614**

Table 6: Results of performance metrics grouped by each offended target in hate speech messages within the test split of the RefutES dataset.

fensiveness, stance, informativeness, truthfulness, and editing when the offending message and its counter-narrative are more similar to those generated by a model such as GPT-4 than to those generated by humans. In conclusion, as expected, the systems generate content more similar to that of other systems than to human-generated content.

6 Conclusions

This paper presents the first shared task on the automatic generation of counter-speech in Spanish, organized within the IberLEF workshop as part of the SEPLN 2024 conference. Participants were tasked with developing respectful and reasoned responses to offensive messages. Although six teams registered for the task, only one team ultimately participated due to the complexity of the task. The participating team used large language models (LLMs) and applied Zero-Shot Learning techniques and finetuning of these models. They also explored using another LLM to evaluate the generated counter-narratives, but the best results were obtained with a model specifically adapted to the task.

Origin CN	**Label**	**Ixa-run1 (%)**					**Ixa-run2 (%)**					**Ixa-run3 (%)**				
		Off	**Sta**	**Inf**	**Tru**	**Edi**	**Off**	**Sta**	**Inf**	**Tru**	**Edi**	**Off**	**Sta**	**Inf**	**Tru**	**Edi**
Disabled	0	0	0	0	0	**10**	10	10	10	10	**10**	0	0	0	0	**60**
	1	**90**	0	0	10	90	**80**	0	0	10	90	**100**	0	0	0	40
	2	0	0	70	30	-	0	0	60	30	-	0	0	70	0	-
	3	10	**100**	**30**	**60**	-	10	**90**	**30**	**50**	-	0	**100**	**30**	**100**	-
Jews	0	0	0	0	0	**30**	10	30	10	10	**40**	0	0	0	0	**40**
	1	**90**	0	0	0	70	**80**	0	20	0	60	**100**	0	10	0	60
	2	10	0	40	10	-	10	0	30	30	-	0	0	50	20	-
	3	0	**100**	**60**	**90**	-	0	**70**	**40**	**60**	-	0	**100**	**40**	**80**	-
LGBT+	0	0	0	0	0	**20**	20	20	30	20	**20**	0	0	0	0	**70**
	1	**90**	0	10	0	80	**60**	0	0	0	80	**100**	0	0	0	30
	2	10	0	70	30	-	20	0	40	10	-	0	0	50	0	-
	3	0	**100**	**20**	**70**	-	0	**80**	**30**	**70**	-	0	**100**	**50**	**100**	-
Migrants	0	0	0	0	0	**10**	0	0	0	10	**10**	0	0	0	10	**80**
	1	**70**	0	0	0	90	**60**	0	10	0	90	**100**	0	10	0	20
	2	20	10	60	40	-	40	10	40	20	-	0	0	50	0	-
	3	10	**90**	**40**	**60**	-	0	**90**	**50**	**70**	-	0	**100**	**40**	**90**	-
Women	0	0	10	10	0	**20**	0	10	10	0	**20**	0	0	0	0	**60**
	1	**80**	0	0	20	80	**70**	0	10	20	80	**100**	0	0	0	40
	2	10	0	50	0	-	20	0	40	0	-	0	0	70	0	-
	3	10	**90**	**40**	**80**	-	10	**90**	**40**	**80**	-	0	**100**	**30**	**100**	-
Muslims	0	0	0	10	10	**0**	0	0	0	0	**10**	0	0	0	0	**50**
	1	**80**	0	0	0	100	**100**	0	0	0	90	**100**	0	0	0	50
	2	20	0	50	60	-	0	0	60	30	-	0	0	80	10	-
	3	0	**100**	**40**	**30**	-	0	**100**	**40**	**70**	-	0	**100**	**20**	**90**	-
POC	0	0	0	0	10	**50**	0	0	0	10	**40**	0	0	0	0	**60**
	1	**50**	0	20	10	50	**70**	0	30	0	60	**90**	0	0	0	40
	2	30	0	50	30	-	20	10	40	40	-	10	0	80	10	-
	3	20	**100**	**30**	**50**	-	10	**90**	**30**	**50**	-	0	**100**	**20**	**90**	-
Others	0	0	0	0	0	**40**	10	20	10	10	**0**	0	0	0	0	**50**
	1	**90**	10	20	10	60	**80**	0	10	10	100	**100**	0	0	10	50
	2	10	0	20	30	-	10	10	20	20	-	0	10	100	10	-
	3	0	**90**	**60**	**60**	-	0	**70**	**60**	**60**	-	0	**90**	**0**	**80**	-

Table 7: Results of manual annotation metrics grouped by each offended target in hate speech messages within the test split of the RefutES dataset. The position of the label that in a perfect system should have the highest score is shown in bold type.

Team run	**Origin of CN References**	**BERTScore**			**MoverScore**	**(F1$_{\text{BS}}$+ MS)/2**
		f1	**precision**	**recall**		
Ixa-run3	GPT-4	0.9084	0.9165	0.9124	0.6521	0.9084
	Human	0.8763	0.8784	0.8772	0.6129	0.8763
Llama RefutES	GPT-4	0.90848	0.92566	0.91696	0.64547	0.9085
	Human	0.87558	0.88346	0.87943	0.60744	0.8756
LLaMA ZSL	GPT-4	0.8793	0.9044	0.8916	0.6128	0.8793
	Human	0.8652	0.8796	0.8723	0.5992	0.8652
Ixa-run2	GPT-4	0.8668	0.8987	0.8824	0.6003	0.8668
	Human	0.8543	0.8730	0.8635	0.5908	0.8543
Ixa-run1	GPT-4	0.8635	0.9023	0.8825	0.5927	0.8635
	Human	0.8503	0.8770	0.8634	0.5843	0.8503

Table 8: Results of performance metrics grouped by the origin of reference counter-narratives within the test split of the RefutES dataset.

Origin CN	**Label**	**Ixa-run1 (%)**					**Ixa-run2 (%)**					**Ixa-run3 (%)**				
		Off	**Sta**	**Inf**	**Tru**	**Edi**	**Off**	**Sta**	**Inf**	**Tru**	**Edi**	**Off**	**Sta**	**Inf**	**Tru**	**Edi**
GPT-4	0	0.0	0.0	0.0	0.0	**20.0**	7.5	7.5	7.5	7.5	**20.0**	0.0	0.0	0.0	0.0	**70.0**
	1	**95.0**	0.0	2.5	2.5	80.0	**87.5**	0.0	5.0	0.0	80.0	**100.0**	0.0	0.0	2.5	30.0
	2	2.5	0.0	45.0	27.5	-	5.0	2.5	32.5	12.5	-	0.0	0.0	65.0	5.0	-
	3	2.5	**100.0**	**52.5**	**70.0**	-	0.0	**90.0**	**55.0**	**80.0**	-	0.0	**100.0**	**35.0**	**92.5**	-
Human	0	0.0	2.5	5.0	5.0	**25.0**	5.0	15.0	10.0	10.0	**17.5**	0.0	0.0	0.0	2.5	**47.5**
	1	**65**	2.5	10.0	10.0	75.0	**62.5**	0.0	15.0	10.0	82.5	**97.5**	0.0	5.0	0.0	52.5
	2	25.0	2.5	57.5	30.0	-	25.0	5.0	50.0	32.5	-	2.5	2.5	72.5	7.5	-
	3	10.0	**92.5**	**27.5**	**55.0**	-	7.5	**80.0**	**25.0**	**47.5**	-	0.0	**97.5**	**22.5**	**90.0**	-

Table 9: Results of manual annotation metrics grouped by the origin of reference counter-narratives within the test split of the RefutES dataset. The position of the label that in a perfect system should have the highest score is shown in bold type. Off: Offensiveness, Sta: Stance, Inf: Informativeness, Tru: Truthfulness, Edi: Edition required.

As future work, we plan to expand our dataset with more counter-narratives to provide a broader set of valid responses for the same offensive comment. This will allow us to further explore effective methods for refuting offensive messages.

Acknowledgments

This work has been partially supported by Project CONSENSO (PID2021-122263OB-C21), Project MODERATES (TED2021-130145B-I00), and Project SocialTox (PDC2022-133146-C21) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR.

References

Benesch, S. 2014. Countering dangerous speech: New ideas for genocide prevention. URL: https://ssrn.com/abstract=3686876.

Bengoetxea, J., Y.-L. Chung, M. Guerini, and R. Agerri. 2024. Basque and Spanish counter narrative generation: Data creation and evaluation. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2132–2141, Torino, Italia, May. ELRA and ICCL. URL: https://aclanthology.org/2024.lrec-main.192.

Bonaldi, H., Y.-L. Chung, G. Abercrombie, and M. Guerini. 2024. Nlp for counterspeech against hate: A survey and how-to guide.

Chiruzzo, L., S. M. Jiménez-Zafra, and F. Rangel. 2024. Overview of IberLEF 2024: Natural Language Processing Challenges for Spanish and other Iberian Languages. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS.org.*

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Courty, B., V. Schmidt, S. Luccioni, Goyal-Kamal, MarionCoutarel, B. Feld, J. Lecourt, LiamConnell, A. Saboni, Inimaz, supatomic, M. Léval, L. Blanche, A. Cruveiller, ouminasara, F. Zhao, A. Joshi, A. Bogroff, H. de Lavoreille, N. Laskaris, E. Abati, D. Blank, Z. Wang, A. Catovic, M. Alencon, M. Stęchły, C. Bauer, Lucas-Otavio, JPW, and MinervaBooks. 2024. mlco2/codecarbon: v2.4.1, May.

Dettmers, T., A. Pagnoni, A. Holtzman, and L. Zettlemoyer. 2023. QLoRA: Efficient Finetuning of Quantized LLMs, May. arXiv:2305.14314 [cs].

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

Furman, D., P. Torres, J. Rodríguez, D. Letzen, M. Martinez, and L. Alemany. 2023. High-quality argumentative information in low resources approaches improve counter-narrative generation. In H. Bouamor, J. Pino, and K. Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2942–2956, Singapore, December. Association for Computational Linguistics.

Halim, S. M., S. Irtiza, Y. Hu, L. Khan, and B. Thuraisingham. 2023. Wokegpt: Improving counterspeech generation against online hate speech by intelligently augmenting datasets using a novel metric. In *2023 International Joint Conference on Neural Networks (IJCNN)*, pages 1–10. IEEE.

Mathew, B., H. Tharad, S. Rajgaria, P. Singhania, S. K. Maity, P. Goyal, and A. Mukherjee. 2018. Thou shalt not hate: Countering online hate speech. In *International Conference on Web and Social Media.*

Qian, J., A. Bethke, Y. Liu, E. Belding, and W. Y. Wang. 2019. A benchmark dataset for learning to intervene in online hate speech. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on*

Natural Language Processing (EMNLP-IJCNLP), pages 4755–4764, Hong Kong, China, November. Association for Computational Linguistics.

Schieb, C. and M. Preuss. 2016. Governing hate speech by means of counterspeech on facebook. In *66th ica annual conference, at fukuoka, japan*, pages 1–23.

Tekiroglu, S., H. Bonaldi, M. Fanton, and M. Guerini. 2022. Using pre-trained language models for producing counter narratives against hate speech: a comparative study. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 3099–3114.

Touvron, H., L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models, July. arXiv:2307.09288 [cs].

Vallecillo-Rodríguez, M.-E., M.-V. Cantero-Romero, I. Cabrera-De-Castro, A. Montejo-Ráez, and M.-T. Martín-Valdivia. 2024. CONAN-MT-SP: A Spanish corpus for counternarrative using GPT models. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, and N. Xue, editors, *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3677–3688, Torino, Italy, May. ELRA and ICCL.

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

Zhang, T., V. Kishore, F. Wu, K. Q. Weinberger, and Y. Artzi. 2019. BERTScore: Evaluating Text Generation with BERT. September.

Zhao, W., M. Peyrard, F. Liu, Y. Gao, C. M. Meyer, and S. Eger. 2019. MoverScore: Text Generation Evaluating with Contextualized Embeddings and Earth Mover Distance. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *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 563–578, Hong Kong, China, November. Association for Computational Linguistics.

Zubiaga, I., A. Soroa, and R. Agerri. 2024. Ixa at refutes 2024: Leveraging language models for counter narrative generation. In *IberLEF (Working Notes)*. CEUR Workshop Proceedings.