Procesamiento del Lenguaje Natural, Revista nº 74, marzo de 2025, pp. 241-252 recibido 05-12-2024 revisado 19-01-2025 aceptado 05-02-2025

Optimizing Few-Shot Learning through a Consistent Retrieval Extraction System for Hate Speech Detection

Optimización de Few-Shot Learning mediante un Sistema de Extracción Coherente para la Detección Del Discurso de Odio

Ronghao Pan, José Antonio García-Díaz, Rafael Valencia-García
Facultad de Informática, Universidad de Murcia, Murcia, España
{ronghao.pan, joseantonio.garcia8, valencia}@um.es

Abstract: Hate speech is a growing phenomenon on social media, posing significant risks to social cohesion and online safety. Its detection is crucial to mitigate these effects, but fine-tuning-based approaches are costly and prone to overfitting due to biases in the training data. In-context learning, which uses pre-trained models with instructions and examples during inference, is emerging as a promising alternative, although it lacks clear strategies for selecting relevant examples. This work proposes an intelligent example selection system for Few-Shot Learning (FSL) based on diversity and uncertainty metrics, which optimizes recognition compared to Zero-Shot Learning (ZSL) and Random FSL methods. Our approach was evaluated on four Spanish hate speech datasets. This strategy consistently improves the results, with the Gemma-2-2b and Gemma-2-9b models excelling across different datasets. In specific cases, the pre-trained knowledge of certain models benefits ZSL, but overall our proposal proves to be an effective and adaptable solution.
Keywords: Hate-speech detection, Zero-shot learning, Few-shot learning, Document Classification.

Resumen: El discurso de odio es un fenómeno presente en redes sociales que supone un grave riesgo para la cohesión social y la seguridad en Internet. Su detección es fundamental para mitigar estos efectos, pero los enfoques basados en ajustar grandes modelos del lenguaje son costosos y propensos al sobreajuste debido a los sesgos de los datos de entrenamiento. El in-context learning, que utiliza modelos preentrenados con instrucciones y ejemplos durante la inferencia, es una alternativa prometedora. Sin embargo, el in-context learning carece de estrategias claras para seleccionar qué ejemplos son relevantes. En este trabajo se propone un sistema de selección inteligente para seleccionar ejemplos basado en diversidad e incertidumbre, mejorando los resultados de elegir estos ejemplos al azar o un baseline de evaluar el modelo sin ejemplos. Nuestra propuesta se ha evaluado en cuatro corpus de discurso de odio en español y los resultados mejoran consistentemente, destacando los modelos Gemma-2-2b y Gemma-2-9b. En casos específicos, el conocimiento preentrenado de ciertos modelos beneficia al aprendizaje sin ejemplos, pero, en general, nuestra propuesta demuestra ser una solución eficaz y adaptable.
Palabras clave: Detección de Discurso de Odio, Zero-shot learning, Few-shot learning, Clasificación Automática de Odio.

1 Introduction

While social networks offer unprecedented freedom of communication and expression, they have also become hotspots for harmful content, including hate speech. Hate speech includes any communication that disparages individuals or groups based on characteristics such as race, ethnicity, gender, sexual orientation, or political affiliation (Zhang and Luo, 2019). Its prevalence on social media poses significant risks to social cohesion and online safety (Castaño-Pulgarín et al., 2021). As a result, identifying and mitigating toxic discourse has become a pressing challenge, prompting extensive research into automated detection systems.

ISSN 1135-5948 DOI 10.26342/2025-74-16

However, hate speech detection remains inherently complex due to its context-dependent nature and linguistic variability, ranging from explicit hostility to subtle stereotyping (Jahan and Oussalah, 2023). Traditional supervised learning methods often require extensive annotated datasets and fine-tuned models, which are costly to develop and susceptible to domain biases. Furthermore, the need for separate models and datasets for each type of hate speech further complicates the development of adaptive and efficient detection systems.

To address these challenges, Large Language Models (LLMs) have emerged as a promising alternative, offering greater adaptability in NLP tasks through their ability to generalize across domains. A key capability of LLMs is in-context learning (ICL), which allows models to perform new tasks by processing task demonstrations at inference time, without requiring additional fine-tuning or parameter updates (Brown, 2020). This flexibility has made approaches such as Zero-Shot Learning (ZSL) and Few-Shot Learning (FSL) viable alternatives to traditional supervised methods, particularly in low-resource settings such as hate speech detection in underrepresented languages.

Among these approaches, FSL presents an additional challenge: selecting the most representative examples to effectively train the model. While randomly selected examples are often used, they may not always be informative, leading to suboptimal performance. A key open problem in FSL is how to systematically identify the most useful examples that improve generalization across different datasets.

To address this issue, we propose a novel retrieval system that optimizes FSL by selecting examples based on diversity and uncertainty criteria. This system ensures that the selected examples maximize linguistic coverage while reducing model uncertainty, leading to improved hate speech detection performance across multiple Spanish language datasets. This retrieval-based approach to optimizing FSL in hate speech detection is evaluated on four Spanish hate speech detection corpora and compared to two baseline approaches: ZSL and random FSL, both of which have demonstrated effectiveness in previous studies (Pan, Antonio García-Díaz, and Valencia-García, 2024; Mozafari, Farahbakhsh, and Crespi, 2022; Alkhamissi et al., 2022). In addition, we test five state-of-the-art LLMs, including models from the LLaMa (Touvron et al., 2023), Gemma (Team et al., 2024a), and Mistral (Jiang et al., 2023) families, ranging from 3 billion to 9 billion parameters.

The rest of the paper is organized as follows. First, section 2 defines the problem of hate speech detection, gives a summary of techniques, and explores in-context strategies for hate speech detection. Second, section 3 describes our proposal and summarizes the evaluation pipeline and the evaluated corpora. Third, section 4 presents and discusses the results obtained by our method compared to the baselines and explores the relevance of varying the number of examples. Finally, section 5 outlines the conclusions of the paper and presents further work.

2 State-of-the-art

Hate speech detection has advanced significantly in recent years but remains challenging due to its context-dependent nature and linguistic variability (Jahan and Oussalah, 2023). Traditional deep learning-based approaches often require fine-tuning on domain-specific annotated datasets, which can be costly and computationally intensive (García-Díaz et al., 2023). These methods also suffer from biases and poor adaptability to new contexts, limiting their effectiveness.

LLMs have transformed NLP with their ability to generalize across tasks through in-context learning (ICL) (Brown, 2020). Prominent models such as GPT, LLaMA (Touvron et al., 2023), and Gemma (Team et al., 2024a; Team et al., 2024b) achieve strong performance without task-specific training. This flexibility makes them suitable for hate speech detection, especially in low-resource settings where annotated data is scarce.

The use of ICL for hate speech detection is still underexplored. Studies such as (García-Díaz, Pan, and Valencia-García, 2023; Plaza del Arco, Nozza, and Hovy, 2023) have demonstrated the potential of LLMs in ZSL scenarios, showing their ability to classify hate speech without fine-tuning. Similarly, FSL has been successfully applied in other domains, such as medical text classification (Ge et al., 2023), and has been extended to several low-resource languages (Cahyawijaya, Lovenia, and Fung, 2024).

However, an open challenge in FSL is the selection of the most informative examples to maximize performance, which is the focus of our work.

Recent research has explored strategies for improving FSL by selecting more informative examples rather than relying on random sampling. Some work has investigated how selecting relevant feature representations from multiple domains improves generalization in FSL tasks (Dvornik, Schmid, and Mairal, 2020). Other studies have proposed active learning-based techniques, such as entropy and confidence margin, to prioritize instances that maximize model performance (Wang et al., 2020). In addition, recent approaches propose to dynamically combine multiple selection strategies to adapt to different datasets and task requirements, achieving superior results over static selection methods (Lu et al., 2023). These results are consistent with our work, where we focus on optimizing example selection using uncertainty and diversity criteria in the context of hate speech detection.

3 Methods and experiments

Figure 1 shows the general architecture of our hate speech classification system, evaluated with corpora from different international conferences or shared tasks as validation set. The presented approach explores three main classification strategies using LLMs with different configurations, with the aim to analyzing the impact of ZSL and FSL in-context learning techniques of LLMs on hate speech detection.

For the ZSL and FSL approaches, a base template was used to generate prompts adapted to different corpora from different domains. Figure 1 illustrates the structure of the prompts used for classification in ZSL and FSL. The template contains the following fields: (1) **Instructions**: Where the task instructions for the LLMs are defined; (2) **Labels description**: Where the possible labels are defined together with their description; (3) **Text for analysis**: Where the text to be analyzed is entered; and (4) **Output format**: Where the expected output format is defined. In the case of FSL, an additional field called ***Representative Examples*** is added, which contains a set of examples along with their corresponding labels. These examples allow the model to implicitly learn the classification pattern from them, providing a more task-oriented approach with concrete examples. The post processing method used in this work involves a series of steps to ensure the accuracy and consistency of the predictions. Once the models generate predictions, a standardization process is applied to align the predicted labels with a predefined vocabulary. If a prediction does not match any of the expected labels, it is adjusted to the closest match within the predefined set, such as "hatespeech", "non_hatespeech", "sexist", "non-sexist", and others.

The standard FSL approach is based on selecting n random examples from each label in the training set for LLMs to implicitly learn classification patterns. However, this strategy does not always outperform the ZSL approach. This is because sometimes the set of selected examples is not sufficiently representative or relevant to the task. Instead of helping the LLM identify classification patterns, these examples can confuse the model and reduce performance. To address this issue, a new retrieval system was implemented to extract the most significant and relevant examples based on diversity and uncertainty values for each label.

Figure 2 shows the architecture of the retrieval system implemented for the optimized FSL approach. The first step was to fine-tune a pre-trained model for Spanish, called MarIA-large (Gutiérrez-Fandiño et al., 2022). This model is based on the RoBERTa-large architecture, specifically adapted to the Spanish language. It is a masked language model trained on the largest available Spanish text corpus, with a total of 570 GB of data. These texts have been cleaned, deduplicated and processed from web crawls carried out by the National Library of Spain between 2009 and 2019. The fine-tuning process is crucial to adapt a pre-trained model to specific tasks, improve its ability to respond to new data, and reduce the uncertainty in its predictions. By training the model in this way, we aim to minimize error margins and increase prediction confidence.

After training the model, the next step is to create clusters for each label to enhance diversity while ensuring representativeness through centroids. To achieve this, a single example per cluster is selected—the text whose embedding is closest to the centroid. This guarantees that chosen examples cap-

ture essential characteristics while minimizing redundancy. Extracting one representative per cluster broadens the range of semantic features, improving the model's generalization and adaptability in few-shot learning scenarios. To prevent biased representations due to repetitive selections, the experiments employed a fine-tuned model with the K-means algorithm (Ikotun et al., 2023), setting **k**=15 to generate 15 clusters. This unsupervised approach groups texts based on similarity, with each cluster's centroid representing the midpoint of its embeddings. The selection of the nearest embedding ensures that each chosen example best represents its cluster while avoiding redundancy.

The next step is to calculate the uncertainty for each example. This value indicates how confident the model is in classifying a text. In this context, a lower uncertainty value implies greater confidence in the model's prediction. Uncertainty is measured by the entropy of the probability distributions predicted by the model, as obtained by the softmax function. Entropy measures the amount of "uncertainty" in the model's predictions. A low value of entropy indicates that the model has a high degree of certainty in its prediction, i.e. it assigns a high probability to a particular class. Conversely, a high entropy value means that the model is less certain about its classification, since it distributes the probabilities more evenly among the different classes.

Finally, a hybrid approach was implemented that takes into account both diversity and uncertainty of the selected examples. Diversity ensures that the selected examples come from different clusters in the vector space, covering a wide range of semantic features and reducing redundancy. Uncertainty is assessed using the entropy of the model predictions: lower uncertainty indicates higher confidence in the classification, while higher uncertainty indicates more ambiguous predictions.

Diversity is measured by calculating how far each selected example is from the average representation of all examples. These distances are then normalized to ensure comparability. The final selection score is a weighted combination of diversity and uncertainty, where a predefined factor determines the relative importance of each. This approach balances the selection of representative and challenging examples, improving data efficiency in the learning process.

Therefore, the combined score has the following formula, where diversity and uncertainty are the values obtained for diversity and uncertainty for each text:

$$\text{Combined\_score} = w_d \cdot \text{diversity} + w_u \cdot \text{uncertainty}$$

Here, w_d and w_u are the diversity and uncertainty weights, respectively. For the experiments, *diversity_weight* and *uncertainty_weight* were configured to 0.7 and 0.3, respectively, which gives more weight to diversity. In this case, when using the fine-tuned MarIA model to obtain the uncertainty value, there is a possibility that some instances may be misclassified. Therefore, more weight is given to diversity to mitigate some of this problem and avoid over-reliance on uncertainty, which could lead to biased or incorrect predictions. Once the combined score is computed for all selected examples, the n examples with the highest scores are selected. This ensures that the final selection includes both diverse and uncertain examples, optimizing the learning process by covering a wide semantic space while focusing on the most uncertain predictions. For the experiments, diversity was given more weight than uncertainty, prioritizing a wider range of representations while still including uncertain cases.

Note that due to their probabilistic nature, LLMs will generate a new random response each time they receive the same input. This happens because by default they use stochastic sampling to select words based on their probabilities. This behavior is controlled by the `do_sample` parameter which, when set to `True`, enables this random sampling. To ensure reproducibility and mitigate this problem, `do_sample=False` has been set to disable random sampling. This ensures that the model always selects the word with the highest probability at each step, so that consistent results are obtained in each run. Additionally, when `do_sample=False` is set, other parameters that influence randomness, such as `temperature` and `top_p`, are no longer applicable.

In order to evaluate the developed method, we selected several corpora related to the detection of hate speech in the Spanish language. We chose (1) **HateEval** (Basile et al., 2019) (SemEval 2019), focused on the de-

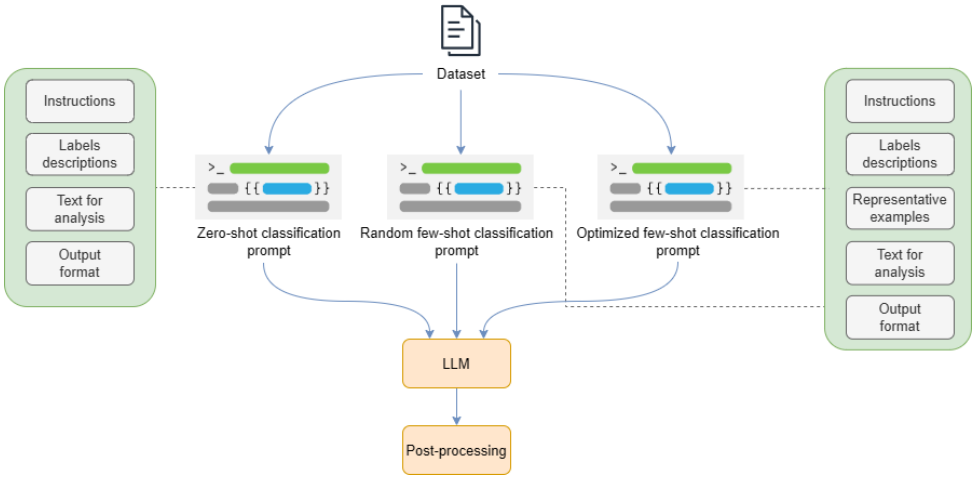

Figure 1: General system pipeline.

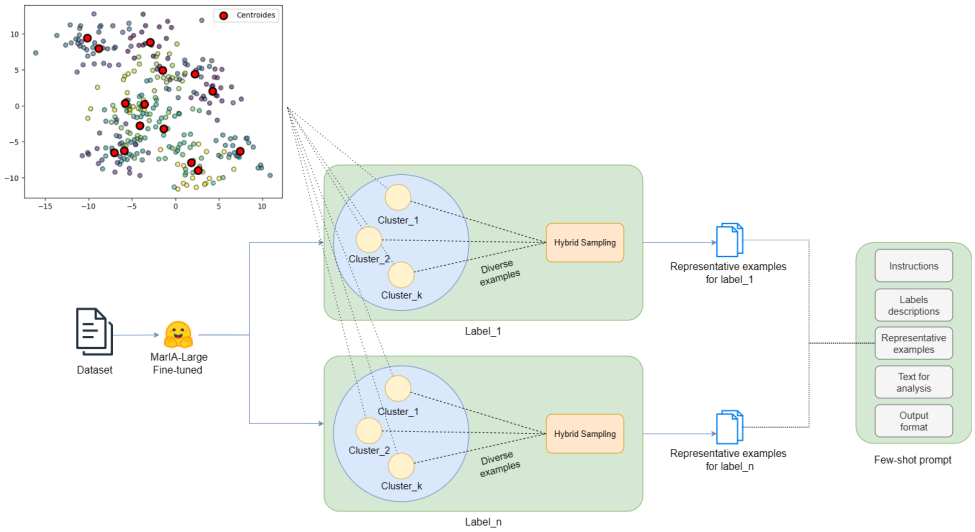

Figure 2: Retrieval system pipeline.

tection of hate content targeting immigrants and women with texts in English and Spanish; (2) **EXIST** (Rodríguez-Sanchez et al., 2021; Rodríguez-Sánchez et al., 2022; Plaza et al., 2023; Plaza et al., 2024) (IberLEF, CLEF), focused on the identification of sexism, covering the concept of sexism from explicit expressions of misogyny to more subtle manifestations of implicit sexist behavior; (3) **DESTEST-DIS** (Ariza-Casabona et al., 2022) (IberLEF), whose main objective is to identify and classify explicit and implicit stereotypes in social media and news comments. Stereotypes reinforce toxic and hateful discourses, often in subtle or implicit ways; and (4) **HOMO-MEX-2024** (Beltagy, Peters, and Cohan, 2020; Gómez-Adorno et al., 2024) (IberLEF), which focused on detecting of LGBTQ+phobic messages in different types of social content.

Table 1 shows the distribution of the training dataset and Table 2 shows the distribution of the corpora used for validation. We mainly used four corpora designed to detect inappropriate language in social networks, covering different domains such as sexism, hate speech and stereotypes. It should be noted that the EXIST corpus does not have a test set with definitive labels (golden labels). For this reason, a validation set was created by extracting 20% of the training set.

For the experiment, several medium-sized models were evaluated, ranging from 2 billion to 9 billion parameters. Models from different families were included, such as Llama, Gemma and Mistral. The evaluated models are detailed in the table 3, where the model size in terms of parameters, the main lan-

Dataset	Label_0	Label_1	Label_2	Total
HateEval-spanish	Hatespeech (1,857)	Non-hatespeech (2,643)	-	4,500
EXIST-2022	Sexist (2,257)	Non-sexist (2,303)	-	4,560
Detests-DIS-2022	Stereotype (2,605)	Non-stereotype (7,301)	-	9,906
HOMO-MEX-2023	P (689)	NONE (1,422)	NP (3,488)	5,599

Table 1: Corpora used to train the proposed approach, where P denotes LGBT+phobic, and NP denotes Not-LGBT+phobic.

Dataset	Label_0	Label_1	Label_2	Total
HateEval-spanish	Hatespeech (660)	Non-hatespeech (939)	-	1,599
EXIST-2022	Sexist (607)	Non-sexist (534)	-	1.141
Detests-DIS-2022	Stereotype (921)	Non-stereotype (1,274)	-	2,205
HOMO-MEX-2023	P (173)	NONE (356)	NP (872)	1,041

Table 2: Corpora used to validate the proposed approach, where P denotes LGBT+phobic, and NP denotes Not-LGBT+phobic.

guage of the model and the token limit in the context are specified. Regarding the limit of allowed context tokens, it can be observed that the Gemma and Llama families have the same limit, which is 8,192 tokens. On the other hand, the Mistral model has a much higher capacity, allowing up to 32,768 tokens.

4 Results

In this section, we present and compare Optimized FSL with the baselines based on ZSL and FSL for each of the corpora evaluated. For ZSL, the model makes predictions without labeled examples, using only the instructions and label descriptions in the prompt. For Random FSL, five examples per label are randomly selected from the training set to provide context to the model and facilitate implicit pattern learning. Optimized FSL, on the other hand, uses a retrieval system that selects the five most meaningful and representative examples (n=5) based on diversity and uncertainty criteria. We also evaluate different values of n (n=2 and n=10), where n is the number of most relevant examples extracted for each label. The results are evaluated using three main metrics: macro precision, macro recall and macro F1 score.

4.1 HateEval

In the analysis of the results for the HateEval corpus, the approaches ZSL, Random FSL and Optimized FSL were evaluated. Table 4 shows the results obtained. In the case of ZSL, the models do not use labeled examples and generate predictions based only on prompts with label descriptions. Among the models evaluated, Llama-3.1-8b achieved the best performance in terms of F1 score (62.116) and the highest recall, highlighting its ability to correctly identify a greater number of positive cases.

On the other hand, in Random FSL, the random selection of 5 examples per label negatively affected the stability of the results. The variability in the quality and representativeness of the selected examples led to inconsistent performance between the models. Although Gemma-2-2b achieved the best F1 score within this approach, its accuracy was comparable to ZSL, reflecting the challenges of a random selection strategy. The Optimized FSL approach, which selects examples based on diversity and uncertainty, showed more consistent results.

Furthermore, as n increased (see Table 8), with n=10, Optimized FSL consistently outperformed the other methods across all evaluated models, standing out for its ability to select representative and diverse examples that maximize semantic coverage while reducing prediction uncertainty. Overall, the results show that Optimized FSL is a more robust and effective approach, especially for complex tasks such as hate speech classification, because it balances diversity and uncertainty criteria to optimize model performance.

4.2 EXIST

Table 5 shows the results for ZSL, Random FSL and Optimized ZSL for the EXIST corpus. In ZSL, Gemma-2-9b achieved the high-

Models	Parameters	Language	Limit of tokens
Llama-3.2-3b	3B	Multilingual	8,192
Llama-3.1-8b	8B	Multilingual	8,192
Gemma-2-2b	2B	Multilingual	8,192
Gemma-2-9b	9B	Multilingual	8,192
Mistral-7b	7B	Multilingual	32,768

Table 3: LLMs used for the experiments.

LLM	P	R	F1	P	R	F1	P	R	F1
	Zero-shot			Random Few-shot			Optimized FSL (n=5)		
Llama-3.2-3b	69.001	53.637	37.354	66.949	56.951	44.991	69.033	57.344	44.990
Llama-3.1-8b	70.165	**66.981**	**62.116**	70.542	64.128	56.914	**70.929**	67.100	61.846
Gemma-2-2b	**71.348**	58.194	45.865	**70.983**	**67.290**	**62.142**	70.104	**68.165**	**64.258**
Gemma-2-9b	71.182	61.262	51.552	69.937	62.407	54.160	70.659	63.960	56.561
Mistral-7b	68.215	60.232	50.965	68.722	60.353	50.950	69.170	59.880	49.849

Table 4: Macro results of the ZSL, Random FSL and Optimized FSL for HateEval corpus.

LLM	P	R	F1	P	R	F1	P	R	F1
	Zero-shot			Random Few-shot			Optimized FSL (n=5)		
Llama-3.2-3b	61.201	60.277	59.892	64.655	63.307	62.951	65.095	64.352	64.256
Llama-3.1-8b	66.915	66.468	66.483	64.655	63.307	62.951	69.441	66.684	66.158
Gemma-2-2b	69.447	60.237	56.177	69.691	65.650	64.651	67.335	67.398	67.337
Gemma-2-9b	**76.203**	**70.439**	**69.687**	**72.057**	**70.216**	**70.113**	**72.825**	**71.036**	**70.975**
Mistral-7b	66.776	59.855	56.411	69.576	60.728	56.983	70.680	63.754	61.558

Table 5: Macro results of the ZSL, random FSL and Optimized FSL for EXIST corpus.

LLM	P	R	F1	P	R	F1	P	R	F1
	Zero-shot			Random Few-shot			Optimized FSL (n=5)		
Llama-3.2-3b	57.226	57.425	57.012	39.018	38.880	38.948	58.899	58.508	58.546
Llama-3.1-8b	59.912	59.250	59.266	62.750	62.919	61.568	62.481	62.753	62.493
Gemma-2-2b	62.152	58.021	50.556	39.476	39.345	37.995	59.455	59.669	58.731
Gemma-2-9b	**64.160**	**64.102**	**62.266**	**65.156**	**65.407**	**64.095**	**66.343**	**66.667**	**65.437**
Mistral-7b	60.767	60.049	57.072	63.182	58.126	49.958	62.926	59.864	54.070

Table 6: Macro results of the ZSL, random FSL and Optimized FSL for Detests-DIS corpus.

LLM	P	R	F1	P	R	F1	P	R	F1
	Zero-shot			Random Few-shot			Optimized FSL (n=5)		
Llama-3.2-3b	37.133	32.453	22.669	59.034	47.912	33.099	58.278	46.823	30.859
Llama-3.1-8b	**62.218**	**68.168**	**58.992**	61.415	57.030	47.166	64.902	56.054	46.409
Gemma-2-2b	52.913	53.895	39.217	50.925	56.894	44.121	49.166	54.882	42.161
Gemma-2-9b	60.551	64.541	50.289	**65.535**	**69.968**	**59.314**	**65.710**	**68.374**	**57.572**
Mistral-7b	54.120	53.864	36.022	53.070	54.281	41.508	52.364	54.479	41.274

Table 7: Macro results of the ZSL and random FSL for HOME-MEX corpus.

LLM	P	R	F1
n=2			
Llama-3.2-3b	**70.703**	50.266	29.811
Llama-3.1-8b	70.527	66.819	61.605
Gemma-2-2b	69.272	**67.360**	**63.427**
Gemma-2-9b	70.657	63.769	56.229
Mistral-7b	69.187	62.188	54.137
n=10			
Llama-3.2-3b	62.399	59.498	53.497
Llama-3.1-8b	70.954	**67.458**	**62.448**
Gemma-2-2b	69.458	67.061	62.743
Gemma-2-9b	**71.309**	63.638	55.689
Mistral-7b	68.849	61.138	52.364

Table 8: Macro results of the Optimized FSL for HateEval corpus.

LLM	P	R	F1
n=2			
Llama-3.2-3b	66.370	66.214	66.246
Llama-3.1-8b	71.816	68.923	68.553
Gemma-2-2b	68.309	67.172	67.076
Gemma-2-9b	**73.557**	**70.961**	**70.777**
Mistral-7b	70.086	62.293	59.429
n=10			
Llama-3.2-3b	64.329	64.162	64.184
Llama-3.1-8b	66.202	64.981	64.769
Gemma-2-2b	66.753	66.780	66.763
Gemma-2-9b	**69.840**	**68.441**	**68.341**
Mistral-7b	68.151	63.661	62.195

Table 9: Macro results of the Optimized FSL for EXIST corpus.

est performance, obtaining the best F1 score (69.686) and demonstrating high precision (76.202) and recall (70.439). This indicates its superior ability to correctly identify and label cases compared to other models in this setup.

In contrast, Random FSL introduced randomness in the selection of 5 examples per label, which led to some improvements, but at the cost of inconsistency. Again, Gemma-2-9b outperformed the other models, achieving an F1 score of 70.113. This result reflects the advantage of having even a few labeled examples, although the randomness in the selection limited the overall effectiveness of the method. For the Optimized FSL approach, the results improved as n increased (see Table 9). In particular, Gemma-2-9b excelled in this setup, achieving the highest scores across all metrics and n values. With n=2 it achieved an F1 score of 70.777; with n=5 it improved to 70.975; and with n=10 it maintained a robust performance of 68.341.

Overall, the Optimized FSL approach consistently outperformed both ZSL and Random FSL across all configurations and models, confirming its effectiveness in exploiting diversity and uncertainty to maximize model performance.

4.3 Detests-DIS

Table 6 shows the results for ZSL, Random FSL and Optimized FSL for the Detests DIS corpus. For ZSL, the models rely only on prompts with label descriptions, without using labeled examples. Among the models tested, we observed that Gemma-2-9b achieved the best performance with an F1 score of 62.266, along with high precision (64.160) and recall (64.102), while other models, such as Mistral-7b and Llama-3.1-8b, showed weaker F1 scores (57.072 and 59.266, respectively), indicating a comparative disadvantage in using only label descriptions.

In the random FSL, Gemma-2-9b continued to perform best, achieving an F1 score of 64.095, reflecting its robustness even under suboptimal selection methods. Llama-3.1-8b followed closely with an F1 score of 61.568; however, the lower scores for Llama-3.2-3b (38.948) and Gemma-2-2b (37.995) illustrate the importance of informed selection strategies, as random sampling often fails to provide representative examples. In terms of the performance of the Optimized FSL approach, Llama-3.1-8b gave a better result, outperforming ZSL and Random FSL. Increasing n to 5 and 10 further improved the results (see Table 10). For n=10, Gemma-2-9b achieved an F1 score of 66.243, with precision and recall consistently leading all models. With n=2, Gemma-2-9b achieved an F1 score of 64.070, significantly improving recall and maintaining high precision. It draws attention to Mistral-7b, whose results for ZSL outperform those for both Optimized FSL and Random FSL across all configurations of n. The likely reason for Mistral-7b's superior performance in ZSL compared to FSL approaches lies in the specific nature of the model's pre-trained knowledge, which is better able to detect stereotypes without

the need for examples at the prompt.

LLM	P	R	F1
	n=2		
Llama-3.2-3b	58.491	57.967	57.959
Llama-3.1-8b	63.081	63.081	63.081
Gemma-2-2b	40.782	40.934	40.285
Gemma-2-9b	**66.383**	**66.166**	**64.070**
Mistral-7b	62.802	59.685	53.784
	n=10		
Llama-3.2-3b	41.215	41.230	41.220
Llama-3.1-8b	65.613	65.446	65.515
Gemma-2-2b	61.651	61.936	61.612
Gemma-2-9b	**67.523**	**67.744**	**66.243**
Mistral-7b	64.335	60.334	53.874

Table 10: Macro results of the Optimized FSL for Detests-DIS corpus.

4.4 HOMO-MEX

The results for the HOMO-MEX corpus using ZSL, random FSL and Optimized FSL are shown in Table 7. The analysis of different values of n for the Optimized FSL is shown in Table 11. In the case of ZSL, Llama-3.1-8b obtained the best performance in terms of F1 score (58.992) and the highest recall, highlighting its ability to correctly identify a larger number of relevant instances. However, Gemma-2-9b showed a very competitive performance, with an F1 score of 50.289 and the second highest recall (64.54), making it another strong contender in ZSL. On the other hand, the random FSL had a negative impact on the stability of the results. The variability in the quality and representativeness of the selected examples led to uneven performance among the models. Although Gemma-2-9b obtained the best F1 score (59.314) with this approach, its accuracy (P = 65.54) was comparable to that of the other models, reflecting the difficulties of a random selection strategy, especially in terms of balancing accuracy and recall.

The Optimized FSL consistently outperformed the other methods in most of the models evaluated, except in the case of Llama-3.1-8b in F1-score. Although Llama-3.1-8b had the best results in ZSL for this dataset, the Optimized selection strategy did not always manage to improve its performance compared to previous configurations. This could be due to the nature of the model in relation to the HOMO-MEX corpus, which may contain specific biases or features that influence the effectiveness of example selection. If the dataset has high variability, noise in the data, or unevenly distributed categories, even an optimized strategy may not achieve significant improvement. In particular, Llama-3.1-8b already performs well in ZSL, so adding additional examples selected in an optimized manner may not provide enough new information to have a positive impact on the F1 score. However, Llama-3.1-3b and Mistral-7b improved significantly with higher n values. In addition, Gemma-2-9b and Gemma-2-2-2b increased their accuracy with respect to the random FSL from 65.535 to 65.710 (n=5 in Optimized FSL) for Gemma-2-9b and from 50.925 to 52.414 for Gemma-2-2-2b with n=2.

Overall, the results show that Optimized FSL is a more effective approach than Random FSL and ZSL because it balances the diversity and uncertainty criteria to optimize model performance. Gemma-2-9b proved to be the most reliable model in different scenarios, while Llama-3.1-8b showed great potential, especially in zero-shot tasks.

LLM	P	R	F1
	n=2		
Llama-3.2-3b	56.815	41.711	23.353
Llama-3.1-8b	63.975	59.007	52.532
Gemma-2-2b	52.414	55.306	41.424
Gemma-2-9b	**65.068**	**69.487**	**57.992**
Mistral-7b	52.895	50.444	36.660
	n=10		
Llama-3.2-3b	56.445	52.470	39.048
Llama-3.1-8b	66.081	60.065	52.318
Gemma-2-2b	47.477	56.197	37.212
Gemma-2-9b	**63.518**	**67.570**	**55.617**
Mistral-7b	50.728	54.120	41.610

Table 11: Macro results of the Optimized FSL for HOMO-MEX corpus.

5 Conclusions and future lines

In this paper, the effectiveness of a new retrieval system that extracts meaningful examples from a corpus to perform FSL has been implemented. This strategy has been compared with ZSL and Random FSL for the detection of different types of hate speech

in social networks in Spanish has been investigated, using public corpora from different conferences or international shared tasks such as HateEval, EXIST, Detests-DIS and HOMO-MEX as a validation set.

In the presented experiments, approaches such as ZSL, Random FSL and Optimized FSL were evaluated with different values of n (n=2, n=5 and n=10), where n represents the number of examples selected per label using the proposed retrieval system. In addition, a systematic exploration and comparison of several recent LLMs of different sizes, ranging from 3 billion to 9 billion parameters, was performed. The results showed that LLMs have the ability to identify hate speech in different domains, with the Gemma family models being the most suitable for this task. Gemma-2-2b performed best on HateEval, while Gemma-2-9b excelled on the other corpora. Regarding the methods used for classification, it was observed that our Optimized FSL approach, which uses uncertainty values obtained by the fine-tuned MarIA model and diversity values obtained by the K-means clustering technique, showed good results. By combining these values to find meaningful examples, diversity ensures that the data covers a wide range of semantic features and reduces redundancy, while uncertainty is assessed by the entropy of the model predictions, which indicates the likelihood that the text is relevant for classification. The results show that this method improves almost all the proposed approaches (ZSL and Random FSL) for all models and corpora, except for Mistral-7b in the Detests-DIS corpus and Llama-3.1-8b in HOMO-MEX. This is because these models show superior performance in ZSL compared to the FSL approaches, due to the specific nature of the pre-trained knowledge of the model, which is more effective in detecting hate speech from Detests-DIS and HOME-MEX without the need for additional examples at the prompt.

As future work, we plan to explore the applicability of the proposed Optimized FSL approach to other languages and cultural contexts, with a particular focus on low-resource languages. In addition, we aim to extend the system to multimodal datasets, incorporating textual, visual and auditory information to address the increasingly diverse nature of online hate speech. Finally, we intend to investigate interpretability techniques to better understand how the selected examples influence model decisions, thereby increasing transparency and trust in sensitive applications.

Acknowledgements

This work is part of the research project LT-SWM (TED2021-131167B-I00) funded by MCIN/AEI/10.13039/501100011033 and by the European Union NextGenerationEU/PRTR. This work is also part of the research project LaTe4PoliticES (PID2022-138099OB-I00) funded by MCIN/AEI/10.13039/
501100011033 and the European Fund for Regional Development (ERDF)-a way to make Europe, and the research project "Services based on language technologies for political microtargeting" (22252/PDC/23) funded by the Autonomous Community of the Region of Murcia through the Regional Support Program for the Transfer and Valorization of Knowledge and Scientific Entrepreneurship of the Seneca Foundation, Science and Technology Agency of the Region of Murcia.

References

Alkhamissi, B., F. Ladhak, S. Iyer, V. Stoyanov, Z. Kozareva, X. Li, P. Fung, L. Mathias, A. Celikyilmaz, and M. Diab. 2022. Token: Task decomposition and knowledge infusion for few-shot hate speech detection. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2109–2120.

Ariza-Casabona, A., W. Schmeisser-Nieto, M. Nofre, M. Taulé, E. Amigó, B. Chulvi, and P. Rosso. 2022. Overview of detests at iberlef 2022: Detection and classification of racial stereotypes in spanish. *Proces. del Leng. Natural*, 69:217–228.

Basile, V., C. Bosco, E. Fersini, D. Nozza, V. Patti, F. M. Rangel Pardo, P. Rosso, and M. Sanguinetti. 2019. SemEval-2019 task 5: Multilingual detection of hate speech against immigrants and women in Twitter. In J. May, E. Shutova, A. Herbelot, X. Zhu, M. Apidianaki, and S. M. Mohammad, editors, *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 54–63, Minneapo-

lis, Minnesota, USA, June. Association for Computational Linguistics.

Beltagy, I., M. E. Peters, and A. Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Brown, T. B. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Cahyawijaya, S., H. Lovenia, and P. Fung. 2024. Llms are few-shot in-context low-resource language learners. *arXiv preprint arXiv:2403.16512*.

Castaño-Pulgarín, S. A., N. Suárez-Betancur, L. M. T. Vega, and H. M. H. López. 2021. Internet, social media and online hate speech. systematic review. *Aggression and violent behavior*, 58:101608.

Dvornik, N., C. Schmid, and J. Mairal. 2020. Selecting relevant features from a multi-domain representation for few-shot classification. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part X 16*, pages 769–786. Springer.

García-Díaz, J. A., S. M. Jiménez-Zafra, M. A. García-Cumbreras, and R. Valencia-García. 2023. Evaluating feature combination strategies for hate-speech detection in spanish using linguistic features and transformers. *Complex & Intelligent Systems*, 9(3):2893–2914.

García-Díaz, J. A., R. Pan, and R. Valencia-García. 2023. Leveraging zero and few-shot learning for enhanced model generality in hate speech detection in spanish and english. *Mathematics*, 11(24).

Ge, Y., Y. Guo, S. Das, M. A. Al-Garadi, and A. Sarker. 2023. Few-shot learning for medical text: A review of advances, trends, and opportunities. *Journal of Biomedical Informatics*, 144:104458.

Gómez-Adorno, H., G. Bel-Enguix, H. Calvo, S.-L. Ojeda-Trueba, S. T. Andersen, J. Vásquez, T. Alcántara, M. Soto, and C. Macias. 2024. Overview of homomex at iberlef 2024: Hate speech detection towards the mexican spanish speaking lgbt+ population. *Proces. del Leng. Natural*, 73:393–405.

Gutiérrez-Fandiño, A., J. Armengol-Estapé, M. Pàmies, J. Llop-Palao, J. Silveira-Ocampo, C. P. Carrino, C. Armentano-Oller, C. Rodriguez-Penagos, A. Gonzalez-Agirre, and M. Villegas. 2022. Maria: Spanish language models. *Procesamiento del Lenguaje Natural*, 68:39–60.

Ikotun, A. M., A. E. Ezugwu, L. Abualigah, B. Abuhaija, and J. Heming. 2023. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, 622:178–210.

Jahan, M. S. and M. Oussalah. 2023. A systematic review of hate speech automatic detection using natural language processing. *Neurocomputing*, 546:126232.

Jiang, A. Q., A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. d. l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Lu, J., S. Wang, X. Zhang, Y. Hao, and X. He. 2023. Semantic-based selection, synthesis, and supervision for few-shot learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 3569–3578.

Mozafari, M., R. Farahbakhsh, and N. Crespi. 2022. Cross-lingual few-shot hate speech and offensive language detection using meta learning. *IEEE Access*, 10:14880–14896.

Pan, R., J. Antonio García-Díaz, and R. Valencia-García. 2024. Comparing fine-tuning, zero and few-shot strategies with large language models in hate speech detection in english. *CMES - Computer Modeling in Engineering and Sciences*, 140(3):2849–2868.

Plaza, L., J. Carrillo-de Albornoz, R. Morante, J. Gonzalo, E. Amigó, D. Spina, and P. Rosso. 2023. Overview of exist 2023: sexism identification in social networks. In *Proceedings of ECIR'23*, pages 593–599.

Plaza, L., J. Carrillo-de Albornoz, V. Ruiz, A. Maeso, B. Chulvi, P. Rosso, E. Amigó, J. Gonzalo, R. Morante, and D. Spina. 2024. Overview of exist 2024 – learning

with disagreement for sexism identification and characterization in tweets and memes. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9–12, 2024, Proceedings, Part II*, page 93–117, Berlin, Heidelberg. Springer-Verlag.

Plaza del Arco, F. M., D. Nozza, and D. Hovy. 2023. Respectful or toxic? using zero-shot learning with language models to detect hate speech. In *The 7th Workshop on Online Abuse and Harms (WOAH)*, pages 60–68.

Rodríguez-Sánchez, F., J. C. de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, and T. Donoso. 2022. Overview of exist 2022: sexism identification in social networks. *Proces. del Leng. Natural*, 69:229–240.

Rodríguez-Sanchez, F. J., J. C. de Albornoz, L. Plaza, J. Gonzalo, P. Rosso, M. Comet, and T. Donoso. 2021. Overview of exist 2021: sexism identification in social networks. *Proces. del Leng. Natural*, 67:195–207.

Team, G., T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, et al. 2024a. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

Team, G., M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, A. Ramé, et al. 2024b. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

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

Wang, Y., Q. Yao, J. T. Kwok, and L. M. Ni. 2020. Generalizing from a few examples: A survey on few-shot learning. *ACM computing surveys (csur)*, 53(3):1–34.

Zhang, Z. and L. Luo. 2019. Hate speech detection: A solved problem? the challenging case of long tail on twitter. *Semantic Web*, 10(5):925–945.