Procesamiento del Lenguaje Natural, Revista nº 75, septiembre de 2025, pp. 317-336 recibido 06-04-2025 revisado 06-05-2025 aceptado 16-05-2025

LLM for Untargeted Adversarial Attack Against Language Models in Spanish

Ataque de Adversario sin Objetivo Específico Basado en LLM Contra Modelos de Lenguaje en Español

Adrián Moreno-Muñoz, L. Alfonso Ureña-López, Eugenio Martínez-Cámara
SINAI Research Group, Center for Advanced Studies in ICT (CEATIC)
Universidad de Jaén
{ammunoz, laurena, emcamara}@ujaen.es

Abstract: Language models face inherent security vulnerabilities where even subtle input modifications can manipulate their outputs, these weaknesses represent a significant concern. This research explores untargeted adversarial attacks against Spanish language models using a two-stage approach: identifying influential words in the decision-making process and replacing them with appropriate synonyms. The evaluation of the attack against pre-trained Spanish language models reveals that generative models, guided by XAI-selected salient words, can significantly alter their predictions.
Keywords: XAI, LLM, adversarial attack, pre-trained models, Spanish language.

Resumen: Los modelos de lenguaje presentan vulnerabilidades de seguridad inherentes donde incluso modificaciones sutiles en las entradas pueden manipular sus salidas, estas debilidades representan una preocupación significativa. Esta investigación explora ataques adversarios sin objetivo específico contra modelos de lenguaje en español utilizando un enfoque de dos etapas: identificar palabras influyentes en el proceso de toma de decisiones y reemplazarlas con sinónimos apropiados. Las pruebas realizadas en diversos conjuntos de datos contra modelos preentrenados revelan que los modelos generativos, guiados por palabras relevantes seleccionadas mediante XAI, pueden alterar significativamente las predicciones de estos modelos de lenguaje.
Palabras clave: XAI, LLM, ataque adversario, modelos pre-entrenados, español.

1 Introduction

The incessant growing of application scenarios of artificial intelligence (AI), and in particular natural language processing services (NLP), is raising the concern on the robustness against adversarial attacks. The dominant paradigm in NLP is data-driven machine learning, which is known to be vulnerable to different kind of attacks in training time for perturbing the behaviour of the learning models, and in inference time, which additionally it is possible to produce privacy leaks (Irfan et al., 2021). Hence, current NLP models may be victims of those adversarial attacks (Goyal et al., 2023).

The attack to a model may be performed in training or inference time. Those in training time require access to the learning model during training or to the training data. This exposition of the training model or the training data is unusual in standard centralised machine learning,[1] but it is a real threat in distributed and federated machine learning (Rodríguez-Barroso et al., 2023). Inference time attacks may aim to alter the behaviour of the learning model or to cause privacy leaks. In this paper we focus on the threat that represents the malicious modification of the behaviour of the learning model.

The data-based manipulation of models in inference time may have an specific target or not, but both attacks are based on the stealthy modification of the input for changing the output of the model. (Goyal et al., 2023). Large language models (LLMs) stand out for their capacity of generating language (Xuanfan and Piji, 2023), thus they can be used as weapons to obtain the subtle modification that triggers the variation of the output of a learning model, in particular a lan-

[1] It is out of the scope the data poisoning of learning models since depends on a poor cleansing of the training data.

ISSN 1135-5948 DOI 10.26342/2025-75-23

guage model (LM).

In this paper, we claim that a LM, like BERT (Devlin et al., 2019), is vulnerable to modifications of the input. Additionally, the alteration of the input must be furtive, since the attack has to tray to be undetectable. This forces to follow a strategy that minimise the number of modifications and keep almost unchanged the semantic meaning of the input. Hence, we first propose to identify the salient words that drive the decision of the victim model. This selection is built upon *a posteriori* explainable artificial intelligent (XAI) methods that allow us to know the prominent features for a learning model. In particular, we evaluate the performance of LIME (Ribeiro, Singh, and Guestrin, 2016), SHAP (Lundberg and Lee, 2017) and Captum (Kokhlikyan et al., 2020). Second, we propose to replace those salient words by their close synonym according to the context. In this case, we compare the performance of five small language models (SLM), namely `LLama-3.2-3B-Instruct`, `Gemma-2-2b-it`, `Gemma-3-1b-it`, `Qwen2.5-05B-Instruct` and `Salamandra-2b-instruct`.

As far as we know, there is not an standard evaluation measure that combine the success of an untargeted data-driven adversarial attack, where the goal is to induce any incorrect prediction rather than a specific target label, at inference time and its stealthiness, since an attack cannot be consider successful if it is evident. We thus propose a new evaluation measure that takes into account the amount of output alterations of the victim LM and the amount of words changed in the input. We also measure the cosine similarity amongst the original text and the changed one for assessing how the attack keeps the meaning. We evaluate our untargeted adversarial attack in five text classification datasets in Spanish from different tasks (hate speech, sarcasm, identification of climate change news, spam classification and sentiment analysis), and we attack four pre-trained LM: `bert-base-Spanish-wwm-cased`, `roberta-base-bne`, `roberta-large-bne` and `xml-roberta-base`. Although there are differences amongst the dataset, the results show that it is possible to leverage LLM to attack pretrained LMs, and that those attacks may be more powerful with LLM with a larger Spanish language proficiency.

The main contributions of this paper are:

- A new untargeted adversarial attack against LMs leveraging *a posteriori* XAI methods and LLMs.
- A new evaluation metric to assess the effectiveness of an adversarial attack considering its success and the amount of modifications made to the input text.

The structure of this paper is as follows: Section 2 summarises the main related works. Section 3 introduces our untargeted adversarial attack. In Section 4, we show the results of the evaluation, which is analysed in Section 5. Finally, in Section 6, we present the main conclusions and future lines of work.

2 Related Works

The security landscape for language models (LMs) continues to evolve rapidly, presenting significant challenges across deployment environments. These models face vulnerabilities stemming from their architectural design and training methodologies, with threats amplified by the dual nature of generative AI as both security tool and attack vector (Yao et al., 2024). Recent frameworks like Greedy Coordinate Gradient have demonstrated concerning capabilities to bypass conventional defenses by iteratively optimizing adversarial suffixes that induce harmful outputs while maintaining semantic coherence (Zou et al., 2023).

Jailbreak attacks represent another critical vulnerability class, with methods like AutoDAN (Liu et al., 2023) combining prompt engineering with automated adversarial optimization. The adversarial landscape extends to data poisoning scenarios, where techniques such as ModelSonar identify undetectable backdoors (Jia, Liu, and Gong, 2022).

Adversary attacks pose particular threats in distributed learning environments, where malicious clients can submit corrupted updates to compromise global model performance. These attacks prove especially challenging in federated learning frameworks, as demonstrated in medical named entity recognition tasks where encrypted federated learning (Pontes et al., 2024) faces significant obstacles balancing computational efficiency and adversary resilience.

Generative models significantly amplify adversarial threats. The PoisonedRAG framework (Zou et al., 2024) illustrates how

retrievable content can be manipulated to influence LLM outputs while evading semantic similarity checks.

The dual-use paradox remains a challenge in LM security. While defensive frameworks like SemanticSmooth (Ji et al., 2024) employ semantic transformations to improve jailbreak resistance through consensus aggregation across multiple prompts, these same techniques can be repurposed for attack optimization. This paradox extends to red-teaming methodologies, where tools designed to identify vulnerabilities simultaneously serve as blueprints for exploitation (Perez et al., 2022).

The integration of generative models into cybersecurity workflows introduces additional complexity, facilitating both threat detection and sophisticated attacks, necessitating comprehensive ethical guidelines and regulatory frameworks (da Silva, 2025). Therefore, it is necessary to discover the vulnerabilities of LM and LLM for elaborating the required safeguards.

3 Untargeted Adversarial Attack

We propose an adversarial attack against LM based on (1) the identification of the salient words that determine the decision of the model, and (2) the use of LLMs to change them by their closest synonym. Figure 1 depicts the structure of our attack, which we explain as what follows.

3.1 Salient Words Identification with XAI

The attack is based on the sensibility of language models to alterations of the input. Additionally, the modification has to be as furtive as possible at lexical and semantic level in order to not be easily identified. Hence, the amount of variations of the input of data have to be as minimum as possible.

We argue that it is necessary to identify the minimum number of words that trigger the decision of the LM for reducing the amount of alterations of the input. In this sense, we propose to use *a posteriori* XAI methods to find out the salient features for a learning model. In this paper we evaluate three XAI methods, namely:

LIME it works by perturbing input data and analysing how these changes affect the model's predictions, providing localized insights into decision-making processes. For text data, LIME works at word level, which ensures explanations are semantically coherent and interpretable for humans (Ribeiro, Singh, and Guestrin, 2016).

SHAP it uses Shapley values from cooperative game theory to fairly distribute feature relevancy scores across all input features, offering a globally consistent explanation framework. SHAP divides words into subwords or parts of words, like prefixes or suffixes, depending on the tokenization scheme of the model. This allows a more precise attribution of relevancy to the individual parts of a word (Lundberg and Lee, 2017).

Captum it is specifically designed for PyTorch-based models, provides a range of attribution algorithms that can analyse feature importance at multiple levels, from individual neurons to entire layers. Captum works at word or subword level for text data, enabling detailed analysis of how parts of a word contribute to model predictions.

3.2 Synonym Generation Module

The identified salient words compose the set of candidate words to be replaced by its synonym.[2] We leverage the text generation capacity of LLMs in order to generate synonyms given an specific word. Figure 2 shows the prompt designed to constraint an LLM to give an unique synonym word.[3]

There are words that they do not have synonyms or the LLM does not return any synonym words. In this case, we do not replace the word.

4 Experimental Framework

The evaluation of our untargeted adversarial attack previously require the definition of the attacker model (see section 4.1), the victim model (see section 4.2), the data to train the victim model (see section 4.3) and the evaluation metric to measure the effectiveness and the stealthiness of the adversarial attack.

4.1 The Attacker Model

The attacker model is composed of two modules for selecting the salient words and them generate their corresponding synonyms. The first one corresponds to the *a posteriori* XAI

[2]We clarify that in the case of SHAP and Captum we do not replace the entire word when they return as salient feature a subword.

[3]You can read the English version of the prompt in the appendix B.

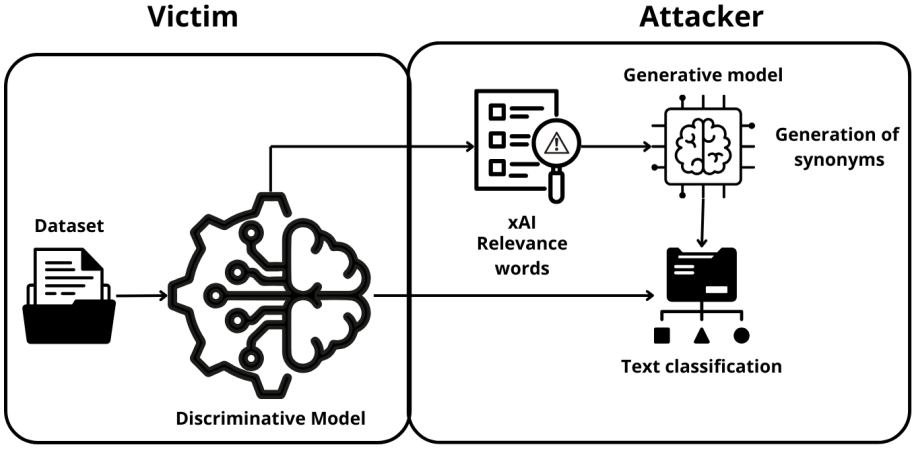

Figure 1: This is the outline of our untargeted adversarial attack.

Eres un asistente que reemplaza una palabra en español con otra palabra en español de significado o uso similar.
Instrucciones:
Analiza la palabra en español proporcionada basándote en su significado más común y general.
Responde con una única palabra en español como reemplazo que sea comúnmente reconocida como similar.
Tu respuesta debe contener solamente la palabra de reemplazo, sin texto adicional, puntuación o explicación.
Si no existe un reemplazo adecuado, devuelve exactamente la misma palabra proporcionada.
Ignora cualquier potencial contexto o ambigüedad y concéntrate únicamente en la palabra misma.
Todas las palabras que recibirás estarán en español y debes responder siempre en español.
Reemplaza la palabra que aparece después de "Reemplaza la palabra:"

Figure 2: The prompt used for the synonym generation attack stage.

method. As explained in section 3.1, we evaluate the methods LIME, SHAP and Captum.

The second component is the synonym generation module, which is grounded in generation capacity of LLMs. We settle that the attack should consume short computational resources, since it will be queried several times. However, this is not a restriction of our proposal, which could also be elaborated with high computationally overhead LLMs. Accordingly, we evaluate the following small language models (SLM): `LLama-3.2-3B-Instruct` (Grattafiori et al., 2024), `Gemma-2-2b-it` (Team et al., 2024), `Gemma-3-1b-it` (Team et al., 2025), `Qwen2.5-0.5B-instruct` (Yang et al., 2024) (Team, 2024) and `Salamandra-2b-instruct` (Gonzalez-Agirre et al., 2025). All SLMs are smaller than 3B parameters, and those are able to retrieve Spanish synonyms, specially `Salamandra-2b-insruct` that stands out for being trained on a relevant Spanish text. Likewise, all the SLMs used are instructed in order to assure a better realisation of an specific instruction prompt.

4.2 The Victim Model

The victim models are pretrained language models (LM) for discriminative task. We evaluate our attack against four LMs, which means that we also assess their vulnerability against them. The LMs used are: `bert-base-spanish-wwm-cased` (Cañete et al., 2020), `roberta-base-bne`, `roberta-large-bne` (Fandiño et al., 2022) and `xlm-roberta-base` (Conneau et al., 2020).

The four models are able to process Spanish text, since three of them are pretrained with data in Spanish, and the fourth one is a multilingual LM. As well, they represent they are diverse regarding their training schema (Bert *vs.* Roberta) and the size of parameters (base *vs.* large).

All the LMs were fine-tuned in five discriminative tasks (see section 4.3), and they were run in a RTX 2080Ti GPU.

4.3 Data

Attack evaluation was performed on five discriminative tasks. The models were fine-tuned with the training set, and the evaluation was performed with test set of the fol-

lowing datasets.

News_racist_comments_Spanish:[4] It is a dataset of Spanish newspaper comments annotated for racist comments. The size of the training set is 3005 documents and of the test set 851 documents. The mean length of the test documents is 42.99 tokens.

Sarcastic_Spanish_dataset:[5] It is a dataset for sarcasm classification task. The size of the training set is 12220 documents and of the test set of 3820 documents. The mean length of the test documents is of 12.48 tokens.

spa_climate_detection:[6] It is a of news headlines in Spanish related to climate change for detecting news about climate change. The size of the training set is of 2900 documents and of the test set of 780 documents. The mean length of the test documents is of 114.15 tokens (Gerardo Huerta, 2024).

Spamspa:[7] It is a dataset for classifying spam messages in Spanish. The size of the training set is of 4457 documents and of the test set of 1115 documents. The mean length of the test documents is of 16.32 tokens (Zuñiga, 2024).

Mucho cine: It is a dataset for sentiment analysis in the movie reviews domain. The size of the training set is of 2100 documents and of the test set of 368 documents. The mean length of the test documents is of 500.52 tokens (Cruz et al., 2008).

4.4 Evaluation Metrics

We perform a two tier evaluation of our adversarial attack. We first evaluate its capacity of harming the performance of the victim model, and them we assess the robustness of the attack according the number of labels change, the number of words change and its capacity of keeping the meaning of the original input text.

The first evaluation level is performed with standard text classification measures, namely the accuracy and the F1-score.

[4] https://huggingface.co/datasets/amaiaruvi/news_racist_comments_spanish

[5] https://huggingface.co/datasets/Ernesto-1997/Sarcastic_Spanish_dataset

[6] https://huggingface.co/datasets/somosnlp/spa_climate_detection

[7] https://huggingface.co/datasets/Gabrielaz/spamspa

Since the lack of an standard evaluation measure for the effectiveness of adversarial attacks, we propose a new evaluation measure that is based on the ratio of the amount of labels flipped and its rectification according to the amount of words modified in the input data.

Attack Success Rate (ASR) (Wu et al., 2021): It measures the success of an attack by the ratio of the number of labels flipped according to the number of documents. It allows us to compare the effectiveness of maliciously flipping the output of the victim model amongst adversarial attacks. Equation 1 settles the ASR metric.

$$\text{ASR} = \frac{\text{Number of labels changed}}{\text{Number of texts attacked}} \tag{1}$$

Attack Score Metric (ASM): It rectifies the ASR with respect the amount of modifications conducted in the input data. Hence, the more modifications performed, the more the ASR value will be reduced. We calculate the level of perturbation of the input data taking into account the ration of the words changed with respect to the number of words of the input document. Text perturbation (TP) and ASM are calculated as follows:

$$\text{TP} = \frac{-1}{\text{n. of texts}} \sum_{i=1}^{\text{n. of texts}} \log\left(\frac{\text{n. changed words in text}_i}{\text{words in each text}_i}\right) \tag{2}$$

$$\text{ASM} = \text{ASR} \cdot \text{Sigmoid}\,(\text{TP}) \tag{3}$$

Cosine Similarity (CS): We pose that adversarial attacks have to keep the semantic similarity amongst the original input and the maliciously altered one. Hence, we also considered the cosine similarity as an additional criterion to assess the adversarial attack. We compute the similarity by creating frequency vectors based on word occurrences and calculating the normalized dot product between these vectors.

McNemar Test (Eliasziw and Donner, 1991): We use this test to identify in the label flip conducted by the attack is statistical significative according to the original classification.

5 Results and Analysis

We analyse the results of the evaluation of our adversarial attacks from different per-

spectives. First, we analyse how change the behaviour of the victim LMs according to the F1-Score.[8] Then, we analyse the errors of our attack that are linked to a wrong synonym generation or to the incapacity of finding of a synonym. Finally, we perform an ablation analysis to study the relevance of identifying the salient words and the use of a SLM to generate the synonyms.

5.1 Adversarial Atack Performance Analysis

The capacity of maliciously altering the behaviour of the LMs are shown in figures from 4 to 15. They shed light about the attack with respect to the victim model, the XAI method and the dataset, as we now detail.

Victim model. The LM more vulnerable to the modification of the input data is BERT, whose performance is significantly harmed when the salient words are selected by Captum and LIME. We also remark that the vulnerability of BERT is also significant in large dataset, in which the size of the dataset may protect the discriminative model, since the alterations may be hidden by the general meaning of the text. We do not see substantial differences between Roberta and XLM Roberta, so the multilingual nature of XLM-Roberta does not represent a vulnerability of the model. Likewise, the performance of these last LMs is significantly fallen when the mean size of the documents of the datasets is short.

XAI Method. We did not expect that the words selected by SHAP would not be as relevant as the one highlighted by Captum or LIME according to the results. The attack using SHAP as the *a posteriori* XAI method is not able to modify the behaviour of the victim model. We did not expect this, since SHAP and Captum are based on subwords, which may find out finer features to twist the output of the victim model. Captum and LIME show similar results, although we highlight the tendency of the attack to be more harmful when LIME is used.

Concerning how the selected words help to keep the semantic meaning, figures from 16 to 27 show that the change of the words selected by SHAP get a better balance between the effectiveness of the attack (ASR) and keeping the semantic meaning. In contrast, the words selected by LIME and Captum produce a high cosine similarity or a high ASR value, when it is preferred a balance between them.

The classification task. We see that the performance of the attack depends on size of the input text. When the size of the input text is short (Sarcastic_Spanish_dataset, Spamspa and news-racist_comments_Spanish), the attacks cause an strong degradation of the performance of the victim model. Conversely, when size is large (Spa_climate_detection, MuchoCine), the attack alters less the victim model. However, we remark that LIME works better with large datasets, except with the model `XLM-Roberta-base`. It may be due to the multilingual nature and thus the encoded of semantic information from different languages may protect the model. In this case, we think that is necessary to modify words beyond the salient ones, which we will study in future work.

Small Language Model Figure 3 shows the mean ASR value reached by the synonyms generated by each SLM in all discriminative tasks and with the salient word selection method, while the details of the results in each case is in appendix A. It is evident that `Gemma-3-1b-it` reached the highest means results, which allow us to infer that generates the synonyms that better matches with the aim of altering the output of the classification model. We expected a better performance of `Salamandra-2b-instruct`, since its training is more extensive with Spanish data, but the attack performance with the synonyms of `Salamandra-2b-instruct` is similar to the

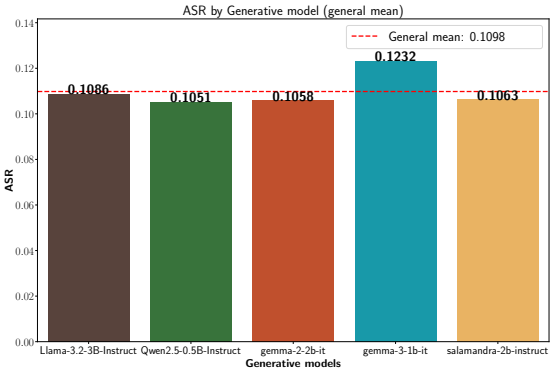

Figure 3: Comparison of ASR across SLMs. The red dashed set the mean value.

[8] Due to length restrictions, the values of all the metrics are in Appendix A.

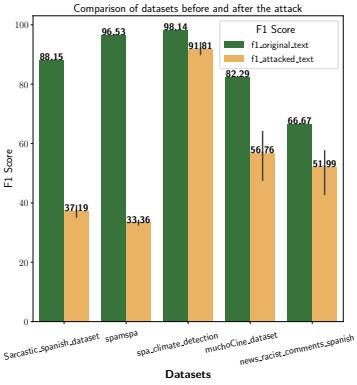

Figure 4: Captum, model bert-base-spanish-wwm-cased.

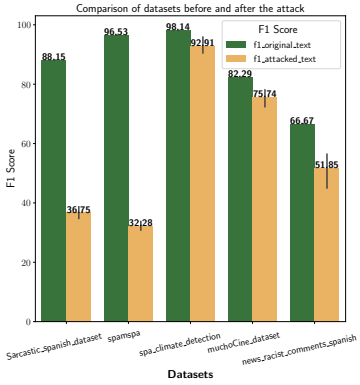

Figure 5: LIME, model bert-base-spanish-wwm-cased.

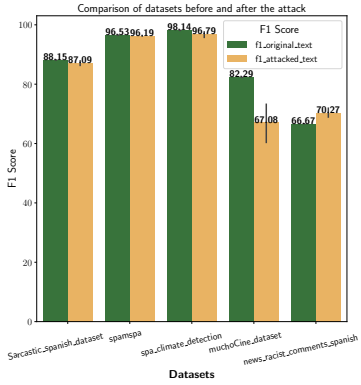

Figure 6: SHAP, model bert-base-spanish-wwm-cased.

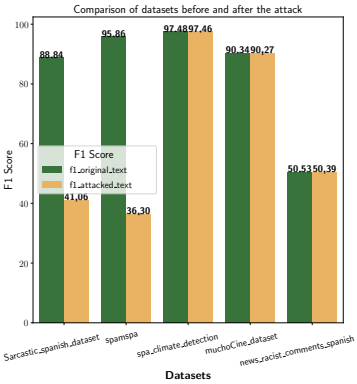

Figure 7: Captum, model roberta-base-bne.

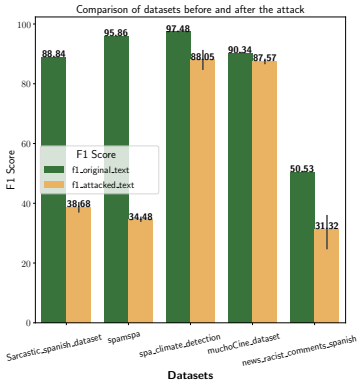

Figure 8: LIME, model roberta-base-bne.

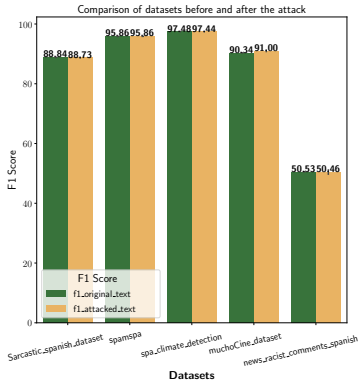

Figure 9: SHAP, model roberta-base-bne.

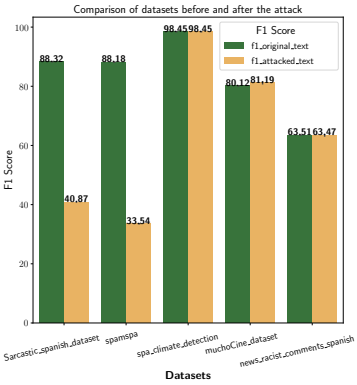

Figure 10: Captum, model roberta-large-bne.

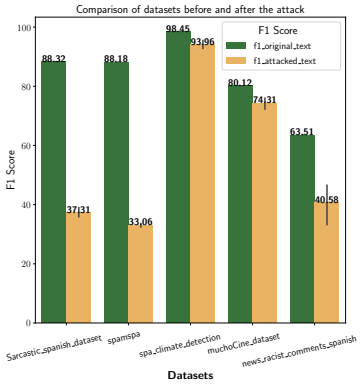

Figure 11: LIME, model roberta-large-bne.

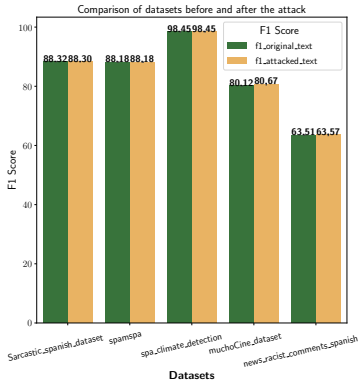

Figure 12: SHAP, model roberta-large-bne.

one reached of the rest of LLMs.

Regarding the ability of each SLM of keeping the meaning, there are not a substantial difference between all SLMs, according to figures from 16 to 27, but it seems that `Salamandra-2B-instruct` reach higher values of cosine similarity, although in those cases the value of ASR is very low. The figure 29 in the appendix D shows the `Gemma-2-2b-it` and `Salamandra-2B-instruct` are over the average keeping the semantic meaning. Hence, it indicates that we need SLM and LLM with a better Spanish proficiency.

Table 1 shows same examples of how the SLM proposes synonyms that keep the original meaning and cause the alteration of the classification label. This is also make evident

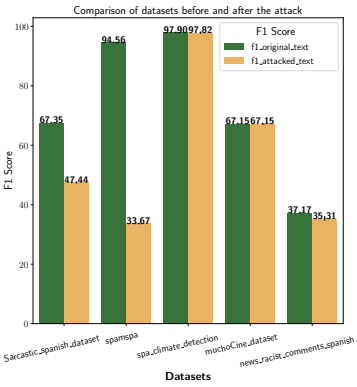

Figure 13: Captum, model xlm-roberta-base.

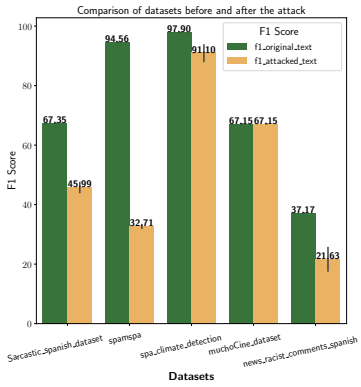

Figure 14: LIME, model xlm-roberta-base.

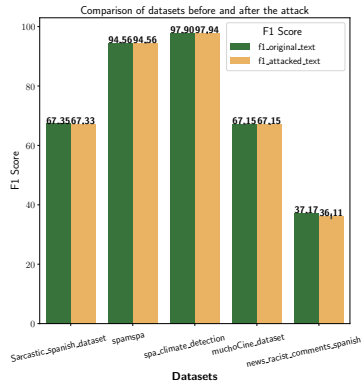

Figure 15: SHAP, model xlm-roberta-base.

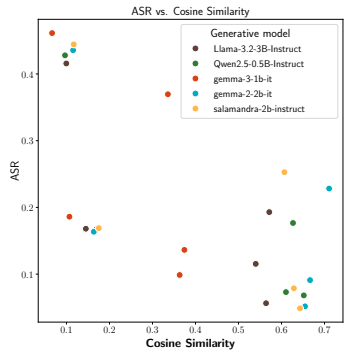

Figure 16: ASR *vs.* similarity: Captum bert-base-spanish-wwm-cased.

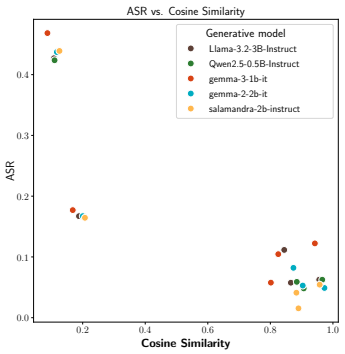

Figure 17: ASR *vs.* similarity: LIME bert-base-spanish-wwm-cased.

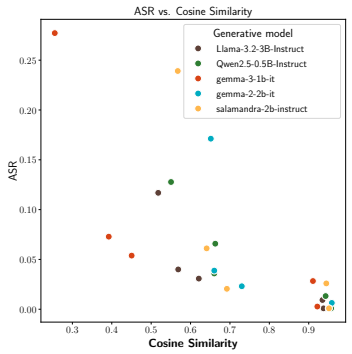

Figure 18: ASR *vs.* similarity: SHAP bert-base-spanish-wwm-cased.

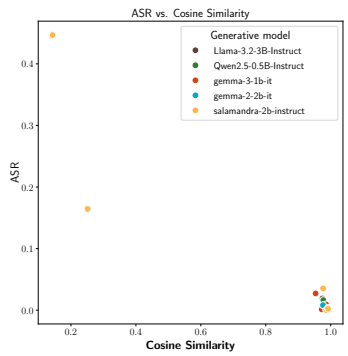

Figure 19: ASR *vs.* similarity: Captum roberta-base-bne.

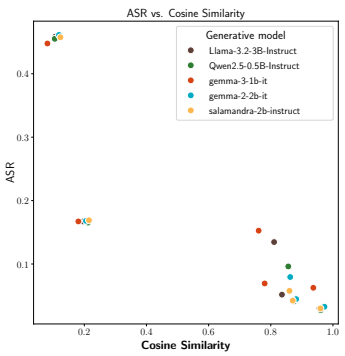

Figure 20: ASR *vs.* similarity: LIME roberta-base-bne.

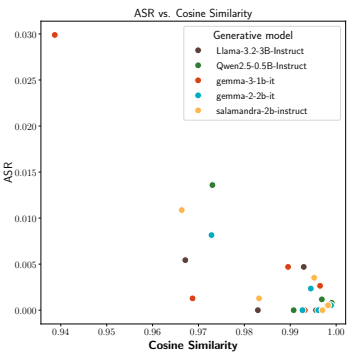

Figure 21: ASR *vs.* similarity: SHAP roberta-base-bne.

that is not difficult to attack LM leveraging the language generation skills of LLMs.

5.2 Error analysis

We only analysed the errors of the SLM generating synonyms, since we do not know what are the real salient word per discriminative model due to their "black box" nature.

The errors of our adversarial attack are caused by a wrong generation of synonyms or de absent of them. We have identified some of these errors that we show in Table 3.[9] In the first example only two words from four identified are replaced by one synonym,

[9]The translation into English is in Table 17.

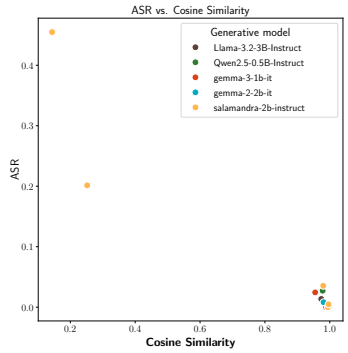

Figure 22: ASR *vs.* similarity: Captum roberta-large-bne.

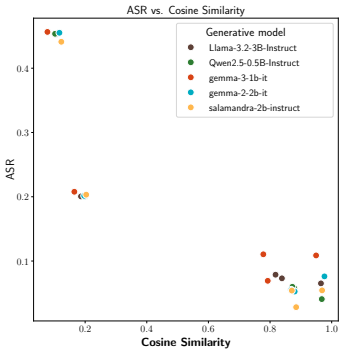

Figure 23: ASR *vs.* similarity: LIME roberta-large-bne.

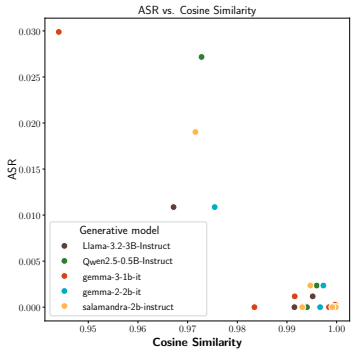

Figure 24: ASR *vs.* similarity: SHAP roberta-large-bne.

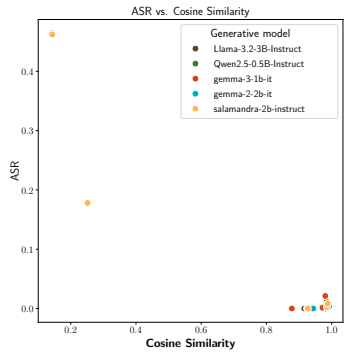

Figure 25: ASR *vs.* similarity: Captum xlm-roberta-base.

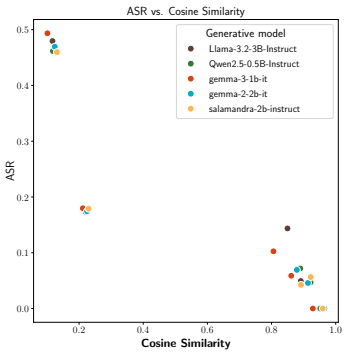

Figure 26: ASR *vs.* similarity: LIME xlm-roberta-base.

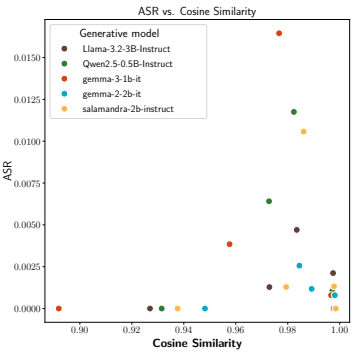

Figure 27: ASR *vs.* similarity: SHAP xlm-roberta-base.

Original text	**Modified text**	**Attacker model**
Viernes de ciencia y calentamiento global de NPR:	Viernes de ciencia y calor mundial de NPR:	Llama-3.2-3B-Instruct
Parar a Ronaldo	Anular a Ronaldo	gemma-3-1b-it
Un detenido tras nuevos incidentes en el centro de menores de Sopuerta Lo ILEGAL es permitir ese tipo de inmigración	Un detenido tras nuevos casos en el centro de menores de Sopuerta Lo ILEGAL es permitir ese tipo de inmigración	salamandra-2b-instruct

Table 1: Examples salient words (underlined ones) replacement with their synonyms by SLMs.

from which only one synonym is right. In the second example the synonyms returned are not adequate since are far from the original meaning, while in the third example we see a clear hallucination of the Salamandra model. In some cases, the Salamandra model returns its name when it does not know the synonym to return.

From the synonym error analysis we conclude that there is a need of LLM and SLM with a deeper Spanish language proficiency, since we see a poor performance in the generation of Spanish synonyms.

5.3 Ablation Analysis

Our adversarial attack is built upon the selection of salient words with a *a posteriori* XAI method, and replace them with synonym words returned by a SLM. We therefore analyse the relevance of these two steps by selecting only two random words and replacing the salient words with random words. We perform this analysis on the short size datasets (Spamspa and Spa_climate_detection) and three SLMs. Table 3 shows the results, and we see: (1) the identification of the salient words produce higher ASR in comparison

SLM	Dataset	ASR	ASM	CS	Change
Gemma-3-1b-it	spamspa	0.1771	0.1213	0.1213	Our attack
		0.1698	0.1997	0.1193	2 random words with synonyms
		0.1563	0.1375	0.1070	XAI random word
	spa_climate detection	0.0577	0.0410	0.8017	Our attack
		0.0038	0.0028	0.9277	2 random words with synonyms
		0.2013	0.1427	0.7732	XAI random word
LLama-3.2-3B-Instruct	spamspa	0.1671	0.1153	0.1881	Our attack
		0.1671	0.1179	0.2089	2 random words with synonyms
		0.1608	0.1101	0.1666	XAI random word
	spa_climate detection	0.1115	0.0796	0.8444	Attack
		0.0128	0.0093	0.9453	2 random words with synonyms
		0.2667	0.1891	0.6544	XAI random word
salamandra-2b-instruct	spamspa	0.1771	0.1213	0.1750	Our attack
		0.1680	0.1183	0.2200	2 random words with synonyms
		0.1626	0.1114	0.1626	XAI random word
	spa_climate detection	0.0487	0.0333	0.6430	Our attack
		0.0026	0.0018	0.9548	2 random words with synonyms
		0.2808	0.1427	0.1990	XAI random word

Table 2: Ablation analysis to study the relevance of the two steps of the adversarial attack. We compare the performance of the adversarial attack with replacing with synonyms two random words and replacing the salient words with random words.

Original text	Modified text	Attack model
No, eso sólo significa que tienes la cabeza gorda.	No, eso únicamente significa cuya tienes El Cabeza gorda.	Qwen2.5-0.5B-Instruct
La medicina preventiva y el principio de precaución	La tratamiento preventiva y el inicio de precaución	gemma-2-2b-it
Clippers con estrellas derrotan a Nuggets sin estrellas en partido de pretemporada	Clippers Salamandra estrellas derrotan casa Nuggets sin estrellas en partido a pretemporada	salamandra-2b-instruct

Table 3: Salient words (underlined ones) unfair replacement with their synonyms by SLMs.

with their random selection, although may cause lower values of ASM, since a LM may depends on more than two salient words to classify an input text; (2) the random replacement of the salient words tends to reach low semantic similarity values. Therefore, we conclude that our claim holds and the adversarial attack is more harmful if it is focuses on the salient words, and it is more stealthy if it replace those words with their synonyms.

6 Conclusions

This research presents an adversarial untargeted attack against LM in Spanish leveraging the capabilities of LLMs and XAI methods. Our experimental results shows that LM are vulnerable to subtle input modifications, especially when they focus on salient words.

The results allow us to conclude that our claim holds and contributes to: (1) a new untargeted adversarial attack that keeps the semantic meaning while achieving high success rates in altering Spanish LM predictions; (2) a new evaluation measure for adversarial attacks in text that combine the effectiveness of the attack (ASR) and the amount of alterations produced in the text (ASM), as well as to also take into consideration of the semantic similarity; and (3) remark the necessity of SLM and LLM with proficiency in the Spanish language, since the stealthiness of adversarial attacks depends on keeping the meaning of the attacked text, which is only possible with LLM for the Spanish language.

As future work, we will keep working in the stealthiness of the adversarial attack by enhancing the process of maintaining unchanged the general meaning of the text. We will also work on target attacks to reach specific goals that lead to more harmful attacks.

Acknowledgments

This work was partly supported by the grants FedDAP (PID2020-116118GA-I00), MODERATES (TED2021-130145B-I00), SocialTOX (PDC2022-133146-C21) and CONSENSO (PID2021-122263OB- C21) funded by MCIN/AEI/10.13039/501100011033, "ERDF A way of making Europe" and "European Union NextGenerationEU/PRTR". This work was also funded by the Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia - Funded by EU – NextGenerationEU within the framework of the project Desarrollo Modelos ALIA.

References

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2020. Spanish Pre-Trained BERT Model and Evaluation Data. In *PML4DC at ICLR 2020*.

Conneau, A., K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In D. Jurafsky, J. Chai, N. Schluter, and J. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online, July. Association for Computational Linguistics.

Cruz, F. L., J. A. Troyano, F. Enriquez, and J. Ortega. 2008. Clasificación de documentos basada en la opinión: experimentos con un corpus de críticas de cine en español. *Procesamiento del lenguaje natural*, 41.

da Silva, F. A. 2025. Navigating the dual-edged sword of generative AI in cybersecurity. *Brazilian Journal of Development*.

Devlin, J., M.-W. Chang, K. Lee, and K. Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Eliasziw, M. and A. Donner. 1991. Application of the McNemar test to non-independent matched pair data. *Statistics in medicine*, 10(12):1981–1991.

Fandiño, A. G., J. A. Estapé, M. Pàmies, J. L. Palao, J. S. Ocampo, C. P. Carrino, C. A. Oller, C. R. Penagos, A. G. Agirre, and M. Villegas. 2022. MarIA: Spanish Language Models. *Procesamiento del Lenguaje Natural*, 68.

Gerardo Huerta, G. Z. 2024. Dataset for BERTIN-ClimID: BERTIN-Base Climate-related text Identification.

Gonzalez-Agirre, A., M. Pàmies, J. Llop, I. Baucells, S. D. Dalt, D. Tamayo, J. J. Saiz, F. Espuña, J. Prats, J. Aula-Blasco, M. Mina, A. Rubio, A. Shvets, A. Sallés, I. Lacunza, I. Pikabea, J. Palomar, J. Falcão, L. Tormo, L. Vasquez-Reina, M. Marimon, V. Ruíz-Fernández, and M. Villegas. 2025. Salamandra Technical Report.

Goyal, S., S. Doddapaneni, M. M. Khapra, and B. Ravindran. 2023. A Survey of Adversarial Defenses and Robustness in NLP. *ACM Comput. Surv.*, 55(14s), July.

Grattafiori, A., A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Irfan, M. M., S. Ali, I. Yaqoob, and N. Zafar. 2021. Towards Deep Learning: A Review On Adversarial Attacks. In *2021 International Conference on Artificial Intelligence (ICAI)*, pages 91–96.

Ji, J., B. Hou, A. Robey, G. J. Pappas, H. Hassani, Y. Zhang, E. Wong, and S. Chang. 2024. Defending Large Language Models against Jailbreak Attacks via Semantic Smoothing. *arXiv preprint arXiv:2402.16192*.

Jia, J., Y. Liu, and N. Z. Gong. 2022. BadEncoder: Backdoor Attacks to Pre-trained Encoders in Self-Supervised Learning. In *2022 IEEE Symposium on Security and Privacy (SP)*, pages 2043–2059. IEEE.

Kokhlikyan, N., V. Miglani, M. Martin, E. Wang, B. Alsallakh, J. Reynolds, A. Melnikov, N. Kliushkina, C. Araya, S. Yan, et al. 2020. Captum: A unified and generic model interpretability library for PyTorch. *arXiv preprint arXiv:2009.07896*.

Liu, X., N. Xu, M. Chen, and C. Xiao. 2023. AutoDAN: Generating Stealthy Jailbreak

Prompts on Aligned Large Language Models. *arXiv preprint arXiv:2310.04451*.

Lundberg, S. M. and S.-I. Lee. 2017. A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.

Perez, E., S. Huang, F. Song, T. Cai, R. Ring, J. Aslanides, A. Glaese, N. McAleese, and G. Irving. 2022. Red Teaming Language Models with Language Models. *URL https://arxiv. org/abs/2202.03286*.

Pontes, M. F., R. C. Pedrosa, P. H. Lopes, and E. J. S. Luz. 2024. Evaluating Federated Learning with Homomorphic Encryption for Medical Named Entity Recognition Using Compact BERT Models. *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL 2024)*.

Ribeiro, M. T., S. Singh, and C. Guestrin. 2016. "Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144.

Rodríguez-Barroso, N., D. Jiménez-López, M. V. Luzón, F. Herrera, and E. Martínez-Cámara. 2023. Survey on Federated Learning Threats: concepts, taxonomy on attacks and defences, experimental study and challenges. *Information Fusion*, 90:148–173.

Team, G., A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, et al. 2025. Gemma 3 Technical Report. *arXiv preprint arXiv:2503.19786*.

Team, G., M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, A. Ramé, et al. 2024. Gemma 2: Improving Open Language Models at a Practical Size. *arXiv preprint arXiv:2408.00118*.

Team, Q. 2024. Qwen2.5: A Party of Foundation Models, September.

Wu, Z., L. Tian, Y. Zhang, Y. Wang, and Y. Du. 2021. Network Attack and Defense Modeling and System Security Analysis: A Novel Approach Using Stochastic Evolutionary Game Petri Net. *Security and Communication Networks*.

Xuanfan, N. and L. Piji. 2023. A Systematic Evaluation of Large Language Models for Natural Language Generation Tasks. In J. Zhang, editor, *Proceedings of the 22nd Chinese National Conference on Computational Linguistics (Volume 2: Frontier Forum)*, pages 40–56, Harbin, China, August. Chinese Information Processing Society of China.

Yang, A., B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, G. Dong, H. Wei, H. Lin, J. Tang, J. Wang, J. Yang, J. Tu, J. Zhang, J. Ma, J. Xu, J. Zhou, J. Bai, J. He, J. Lin, K. Dang, K. Lu, K. Chen, K. Yang, M. Li, M. Xue, N. Ni, P. Zhang, P. Wang, R. Peng, R. Men, R. Gao, R. Lin, S. Wang, S. Bai, S. Tan, T. Zhu, T. Li, T. Liu, W. Ge, X. Deng, X. Zhou, X. Ren, X. Zhang, X. Wei, X. Ren, Y. Fan, Y. Yao, Y. Zhang, Y. Wan, Y. Chu, Y. Liu, Z. Cui, Z. Zhang, and Z. Fan. 2024. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671*.

Yao, Y., J. Duan, K. Xu, Y. Cai, Z. Sun, and Y. Zhang. 2024. A Survey on Large Language Model (LLM) Security and Privacy: The Good, the Bad, and the Ugly. *High-Confidence Computing*, page 100211.

Zou, A., Z. Wang, N. Carlini, M. Nasr, J. Z. Kolter, and M. Fredrikson. 2023. Universal and Transferable Adversarial Attacks on Aligned Language Models. *arXiv preprint arXiv:2307.15043*.

Zou, W., R. Geng, B. Wang, and J. Jia. 2024. PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models. *arXiv preprint arXiv:2402.07867*.

Zuñiga, G. 2024. Spam Detection Messages Dataset.

A Appendix A: Results Tables

This appendix contains detailed tables with all the results of the experimental framework (see section 4) that complement the analyses discussed in section 5. The following tables are included: Table 4, Table 5, Table 6, Table 7, Table 8, Table 9, Table 10, Table 11, Table 12, Table 13, Table 14 and Table 15. We have used the symbol † to indicate a statistical significant degradation of the performance of the victim LM after the attack

according to the McNemar test (see section 4).

B Appendix B: Translated Examples

This appendix provides translations of tables from section 5. Table 16 relates directly to table 1 and table 17 to table 3.

You are an assistant that replaces a Spanish word with another Spanish word of similar meaning or usage. Instructions: Analyze the provided Spanish word based on its most common and general meaning. Respond with a single Spanish word as a replacement that is commonly recognized as similar. Your response should contain only the replacement word, without additional text, punctuation, or explanation. If there is no suitable replacement, return exactly the same word that was provided. Ignore any potential context or ambiguity and focus solely on the word itself. All words you will receive will be in Spanish and you must always respond in Spanish. Replace the word that appears after "Replace the word:"

Figure 28: The prompt used for the synonym generation attack stage in English.

C Appendix C: Training hyperparameters

This appendix details the hyperparameters and training configurations used for the discriminative models presented in section 4. We provide this information to ensure complete transparency and reproducibility of our experimental setup, allowing researchers to replicate or build upon our methodology.

The fine-tuning we implemented varied slightly depending on the specific LM. For the models `roberta-base-bne`, `roberta-large-bne`, and `bert-base-spanish-wwm-cased`, we utilised a batch size of 16 with logging steps set at 25. The `xlm-roberta-base model` required different parameters due to its larger size, so we reduced the batch size to 8 with logging steps increased to 50.

These logging steps played a crucial role in our early stopping strategy, helping prevent model overfitting by regularly evaluating performance during the fine-tuning process. We limited training across all models to a maximum of five epochs. For optimization, we employed the AdamW optimiser with a learning rate of 5e-5.

Our validation approach adapted to dataset characteristics. When dedicated validation files were provided with the dataset, we used these for periodic validation performed at each logging step interval. For datasets without pre-defined validation splits, we reserved 20% of the training data to serve as a validation set, ensuring consistent evaluation across all fine-tuning processes.

D Appendix D: Similarity mean

This appendix presents mean similarity scores across all generative models evaluated in our study in Figure 29. These results complement the analysis provided in Section 5 of the main paper.

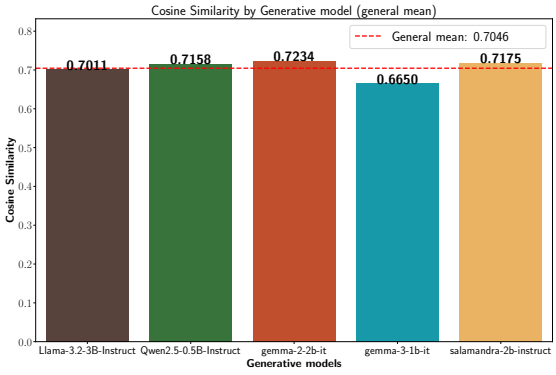

Figure 29: Comparison of Similarity across SLMs. The red dashed set the mean value.

generative_model	**dataset**	**acc_OT**	**f1_OT**	**acc_MT**	**f1_MT**	**ASR**	**ASM**	**CS**
Llama-3.2-3B-Instruct	Sarcastic_spanish_dataset	91.2411	88.1489	56.8669	35.7762†	0.4157	0.2771	0.0996
	spamspa	99.0967	96.5278	82.4752	32.1678†	0.1680	0.1119	0.145
	spa_climate_detection	97.6923	98.1405	86.9231	88.6414†	0.1154	0.0772	0.5401
	muchoCine_dataset	82.337	82.2888	65.2174	63.4286†	0.1929	0.1303	0.5715
	news_racist_comments_spanish	84.1363	66.6667	81.3161	58.7013†	0.0564	0.0379	0.5641
Qwen2.5-0.5B-Instruct	Sarcastic_spanish_dataset	91.2411	88.1489	55.4379	33.9608†	0.4279	0.2889	0.0966
	spamspa	99.0967	96.5278	82.5655	34.1297†	0.1671	0.1132	0.1694
	spa_climate_detection	97.6923	98.1405	91.9231	93.2039†	0.0731	0.05	0.6104
	muchoCine_dataset	82.337	82.2888	67.9348	66.2857†	0.1766	0.1219	0.6267
	news_racist_comments_spanish	84.1363	66.6667	81.0811	55.4017†	0.0682	0.0467	0.6518
gemma-3-1b-it	Sarcastic_spanish_dataset	91.2411	88.1489	53.321	40.526†	0.4612	0.3045	0.0665
	spamspa	99.0967	96.5278	80.8491	32.9114†	0.1861	0.1221	0.1068
	spa_climate_detection	97.6923	98.1405	89.1026	90.9478†	0.0987	0.065	0.3631
	muchoCine_dataset	82.337	82.2888	53.5326	39.1459†	0.3696	0.2457	0.3358
	news_racist_comments_spanish	84.1363	66.6667	75.6757	34.2857†	0.1361	0.0899	0.3742
gemma-2-2b-it	Sarcastic_spanish_dataset	91.2411	88.1489	55.1469	37.2919†	0.4356	0.2926	0.1154
	spamspa	99.0967	96.5278	82.9268	34.6021†	0.1635	0.1097	0.1635
	spa_climate_detection	97.6923	98.1405	89.6154	91.0891†	0.0910	0.0615	0.6664
	muchoCine_dataset	82.337	82.2888	63.8587	57.508†	0.2283	0.1565	0.7107
	news_racist_comments_spanish	84.1363	66.6667	80.8461	56.9921†	0.0517	0.0352	0.6547
salamandra-2b-instruct	Sarcastic_spanish_dataset	91.2411	88.1489	54.697	38.3729†	0.4443	0.2999	0.1169
	spamspa	99.0967	96.5278	82.3848	32.9897†	0.1689	0.1145	0.175
	spa_climate_detection	97.6923	98.1405	94.1026	95.1883†	0.0487	0.0333	0.643
	muchoCine_dataset	82.337	82.2888	67.3913	57.4468†	0.2527	0.1736	0.6067
	news_racist_comments_spanish	84.1363	66.6667	81.1986	54.5455†	0.0787	0.0539	0.6288

Table 4: Results reached by the salient words of Captum against bert-spanish-wwm-cased.

generative_model	dataset	acc_OT	f1_OT	acc_MT	f1_MT	ASR	ASM	CS
Llama-3.2-3B-Instruct	Sarcastic_spanish_dataset	91.4263	88.843	54.353	40.6605†	0.4464	0.2898	0.143
	spamspa	98.916	95.8621	83.1978	36.3014†	0.1644	0.1044	0.2505
	spa_climate_detection	96.9231	97.4843	97.0513	97.5866	0.0013	0.0009	0.9864
	muchoCine_dataset	89.9457	90.3394	89.6739	90.1554	0.0190	0.0138	0.9748
	news_racist_comments_spanish	83.6663	50.5338	83.6663	50.1792	0.0047	0.0032	0.9905
Qwen2.5-0.5B-Instruct	Sarcastic_spanish_dataset	91.4263	88.843	54.5912	41.1119†	0.4456	0.2813	0.1433
	spamspa	98.916	95.8621	83.1978	36.3014†	0.1644	0.1034	0.2511
	spa_climate_detection	96.9231	97.4843	96.9231	97.4843	0.0	0.0	0.99
	muchoCine_dataset	89.9457	90.3394	89.4022	89.8701	0.0163	0.0118	0.9775
	news_racist_comments_spanish	83.6663	50.5338	83.7838	50.7143	0.001175	0.0008	0.9952
gemma-3-1b-it	Sarcastic_spanish_dataset	91.4263	88.843	54.4324	41.4684†	0.4451	0.2936	0.1417
	spamspa	98.916	95.8621	83.1978	36.3014†	0.1644	0.1049	0.2502
	spa_climate_detection	96.9231	97.4843	96.7949	97.3767	0.0012	0.0009	0.9724
	muchoCine_dataset	89.9457	90.3394	89.4022	89.8172	0.0272	0.0197	0.9535
	news_racist_comments_spanish	83.6663	50.5338	83.6663	50.5338	0.0094	0.0065	0.9857
gemma-2-2b-it	Sarcastic_spanish_dataset	91.4263	88.843	54.4589	40.92†	0.4470	0.2862	0.1431
	spamspa	98.916	95.8621	83.1978	36.3014†	0.1644	0.1042	0.2507
	spa_climate_detection	96.9231	97.4843	96.7949	97.3767	0.0013	0.0009	0.9915
	muchoCine_dataset	89.9457	90.3394	90.7609	91.0526	0.0082	0.0059	0.9764
	news_racist_comments_spanish	83.6663	50.5338	83.5488	50.0	0.0035	0.0024	0.9925
salamandra-2b-instruct	Sarcastic_spanish_dataset	91.4263	88.843	54.3795	41.1202†	0.4467	0.2865	0.1431
	spamspa	98.916	95.8621	83.1978	36.3014†	0.1644	0.1035	0.2512
	spa_climate_detection	96.9231	97.4843	96.9231	97.4843	0.0	0.0	0.9864
	muchoCine_dataset	89.9457	90.3394	90.2174	90.4762	0.0353	0.0256	0.9765
	news_racist_comments_spanish	83.6663	50.5338	83.6663	50.5338	0.0024	0.0016	0.9918

Table 5: Results reached by the salient words of Captum against roberta-base.

generative_model	dataset	acc_OT	f1_OT	acc_MT	f1_MT	ASR	ASM	CS
Llama-3.2-3B-Instruct	Sarcastic_spanish_dataset	91.0029	88.3162	54.2736	40.8624†	0.4551	0.2836	0.1438
	spamspa	96.6576	88.1789	81.0298	33.5443†	0.2014	0.1261	0.2521
	spa_climate_detection	98.0769	98.4456	98.0769	98.4456	0.0	0.0	0.996
	muchoCine_dataset	81.25	80.1153	82.0652	80.9249	0.0136	0.0099	0.9737
	news_racist_comments_spanish	84.6063	63.5097	84.7239	63.6872	0.0012	0.0007	0.9966
Qwen2.5-0.5B-Instruct	Sarcastic_spanish_dataset	91.0029	88.3162	54.2736	40.8624†	0.4551	0.2836	0.1438
	spamspa	96.6576	88.1789	81.0298	33.5443†	0.2014	0.1256	0.2522
	spa_climate_detection	98.0769	98.4456	98.0769	98.4456	0.0	0.0	0.9953
	muchoCine_dataset	81.25	80.1153	81.7935	80.6916	0.0272	0.0197	0.9779
	news_racist_comments_spanish	84.6063	63.5097	84.3713	62.9526	0.0024	0.0015	0.9963
gemma-3-1b-it	Sarcastic_spanish_dataset	91.0029	88.3162	54.3001	40.8764†	0.4549	0.2836	0.1437
	spamspa	96.6576	88.1789	81.0298	33.5443†	0.2014	0.1261	0.252
	spa_climate_detection	98.0769	98.4456	98.0769	98.4456	0.0	0.0	0.9887
	muchoCine_dataset	81.25	80.1153	82.6087	81.6092	0.0245	0.0177	0.9549
	news_racist_comments_spanish	84.6063	63.5097	84.6063	63.5097	0.0047	0.003	0.9944
gemma-2-2b-it	Sarcastic_spanish_dataset	91.0029	88.3162	54.2736	40.8624†	0.4551	0.2836	0.1438
	spamspa	96.6576	88.1789	81.0298	33.5443†	0.2014	0.126	0.2521
	spa_climate_detection	98.0769	98.4456	98.0769	98.4456	0.0	0.0	0.9977
	muchoCine_dataset	81.25	80.1153	82.0652	81.0345	0.0082	0.0059	0.9808
	news_racist_comments_spanish	84.6063	63.5097	84.7239	63.6872	0.0012	0.0007	0.9979
salamandra-2b-instruct	Sarcastic_spanish_dataset	91.0029	88.3162	54.3001	40.8764†	0.4549	0.2834	0.1438
	spamspa	96.6576	88.1789	81.0298	33.5443†	0.2014	0.1259	0.252
	spa_climate_detection	98.0769	98.4456	98.0769	98.4456	0.0	0.0	0.9941
	muchoCine_dataset	81.25	80.1153	82.6087	81.7143	0.0353	0.0256	0.9797
	news_racist_comments_spanish	84.6063	63.5097	84.6063	63.5097	0.0047	0.003	0.9962

Table 6: Results reached by the salient words of Captum against roberta-large.

generative_model	dataset	acc_OT	f1_OT	acc_MT	f1_MT	ASR	ASM	CS
Llama-3.2-3B-Instruct	Sarcastic_spanish_dataset	68.616	67.3458	49.14	47.4576†	0.4636	0.2921	0.1435
	spamspa	98.5547	94.5578	82.2042	33.67†	0.1780	0.1123	0.2513
	spa_climate_detection	97.4359	97.8992	97.4359	97.8992	0.0	0.0	0.9824
	muchoCine_dataset	50.5435	67.148	50.5435	67.148	0.0	0.0	0.916
	news_racist_comments_spanish	80.141	37.1747	79.671	35.206	0.0071	0.0047	0.9877
Qwen2.5-0.5B-Instruct	Sarcastic_spanish_dataset	68.616	67.3458	49.1664	47.528†	0.4628	0.2915	0.1429
	spamspa	98.5547	94.5578	82.2042	33.67†	0.1780	0.112	0.2512
	spa_climate_detection	97.4359	97.8992	96.9231	97.4843	0.0051	0.0035	0.9816
	muchoCine_dataset	50.5435	67.148	50.5435	67.148	0.0	0.0	0.9244
	news_racist_comments_spanish	80.141	37.1747	79.436	34.4569	0.0118	0.0079	0.9846
gemma-3-1b-it	Sarcastic_spanish_dataset	68.616	67.3458	48.9812	47.2359†	0.4641	0.2932	0.1427
	spamspa	98.5547	94.5578	82.2042	33.67†	0.1780	0.1125	0.2511
	spa_climate_detection	97.4359	97.8992	97.5641	98.0063	0.0012	0.0009	0.9719
	muchoCine_dataset	50.5435	67.148	50.5435	67.148	0.0	0.0	0.878
	news_racist_comments_spanish	80.141	37.1747	79.671	34.717	0.0212	0.0143	0.9806
gemma-2-2b-it	Sarcastic_spanish_dataset	68.616	67.3458	49.0871	47.4891†	0.4620	0.2905	0.1433
	spamspa	98.5547	94.5578	82.2042	33.67†	0.1780	0.112	0.2513
	spa_climate_detection	97.4359	97.8992	97.5641	98.0063	0.0013	0.0009	0.9887
	muchoCine_dataset	50.5435	67.148	50.5435	67.148	0.0	0.0	0.9437
	news_racist_comments_spanish	80.141	37.1747	80.0235	36.0902	0.0035	0.0023	0.9922
salamandra-2b-instruct	Sarcastic_spanish_dataset	68.616	67.3458	49.1929	47.4836†	0.4620	0.2903	0.1433
	spamspa	98.5547	94.5578	82.2042	33.67†	0.1780	0.112	0.2515
	spa_climate_detection	97.4359	97.8992	97.1795	97.6891	0.0026	0.0017	0.9854
	muchoCine_dataset	50.5435	67.148	50.5435	67.148	0.0	0.0	0.9271
	news_racist_comments_spanish	80.141	37.1747	80.0235	36.0902	0.0082	0.0054	0.9878

Table 7: Results reached by the salient words of Captum against xlm-roberta-base.

generative_model	dataset	acc_OT	f1_OT	acc_MT	f1_MT	ASR	ASM	CS
Llama-3.2-3B-Instruct	Sarcastic_spanish_dataset	91.2411	88.1489	56.0995	33.7195†	0.4271	0.2865	0.1083
	spamspa	99.0967	96.5278	82.7462	30.5455†	0.1671	0.1153	0.1881
	spa_climate_detection	97.6923	98.1405	86.7949	88.4139†	0.1115	0.0796	0.8444
	muchoCine_dataset	82.337	82.2888	77.1739	76.6667†	0.0625	0.0455	0.9562
	news_racist_comments_spanish	84.1363	66.6667	79.5535	53.9683†	0.0578	0.0409	0.8655
Qwen2.5-0.5B-Instruct	Sarcastic_spanish_dataset	91.2411	88.1489	55.8084	34.7656†	0.4237	0.29	0.1102
	spamspa	99.0967	96.5278	82.6558	34.2466†	0.1662	0.1154	0.2038
	spa_climate_detection	97.6923	98.1405	92.8205	93.9525†	0.0590	0.0424	0.8845
	muchoCine_dataset	82.337	82.2888	77.1739	76.9231†	0.0625	0.0455	0.966
	news_racist_comments_spanish	84.1363	66.6667	79.5535	54.4503†	0.0482	0.0344	0.9067
gemma-3-1b-it	Sarcastic_spanish_dataset	91.2411	88.1489	52.0243	40.2636†	0.4684	0.3106	0.0873
	spamspa	99.0967	96.5278	81.5718	31.0811†	0.1771	0.1213	0.1683
	spa_climate_detection	97.6923	98.1405	92.4359	93.6898†	0.0577	0.0409	0.8017
	muchoCine_dataset	82.337	82.2888	70.6522	69.4915†	0.1223	0.0889	0.9423
	news_racist_comments_spanish	84.1363	66.6667	74.1481	38.8889†	0.1046	0.0742	0.8253
gemma-2-2b-it	Sarcastic_spanish_dataset	91.2411	88.1489	54.9352	37.089†	0.4372	0.2959	0.1174
	spamspa	99.0967	96.5278	82.7462	31.0469†	0.1671	0.1155	0.2003
	spa_climate_detection	97.6923	98.1405	89.7436	91.2088†	0.0821	0.0587	0.8738
	muchoCine_dataset	82.337	82.2888	77.4457	76.6197†	0.0489	0.0356	0.9729
	news_racist_comments_spanish	84.1363	66.6667	79.3184	53.9267†	0.0529	0.0378	0.9035
salamandra-2b-instruct	Sarcastic_spanish_dataset	91.2411	88.1489	54.9087	37.9009†	0.4390	0.301	0.1259
	spamspa	99.0967	96.5278	82.8365	34.4828†	0.1644	0.1142	0.2073
	spa_climate_detection	97.6923	98.1405	96.6667	97.3029†	0.0154	0.0111	0.8899
	muchoCine_dataset	82.337	82.2888	79.6196	78.9916†	0.0543	0.0396	0.9576
	news_racist_comments_spanish	84.1363	66.6667	80.9636	58.0311†	0.0411	0.0293	0.8832

Table 8: Results reached by the salient words of LIME against bert-spanish-wwm-cased.

generative_model	dataset	acc_OT	f1_OT	acc_MT	f1_MT	ASR	ASM	CS
Llama-3.2-3B-Instruct	Sarcastic_spanish_dataset	91.4263	88.843	53.4004	38.6625†	0.4586	0.308	0.1069
	spamspa	98.916	95.8621	82.9268	34.1463†	0.1671	0.115	0.1958
	spa_climate_detection	96.9231	97.4843	83.4615	84.9825†	0.1346	0.0957	0.8107
	muchoCine_dataset	89.9457	90.3394	87.5	88.1443†	0.0299	0.0217	0.9564
	news_racist_comments_spanish	83.6663	50.5338	79.2009	30.0395†	0.0517	0.0367	0.8359
Qwen2.5-0.5B-Instruct	Sarcastic_spanish_dataset	91.4263	88.843	53.7973	39.7931†	0.4551	0.3108	0.105
	spamspa	98.916	95.8621	83.1075	36.1775†	0.1653	0.1144	0.2118
	spa_climate_detection	96.9231	97.4843	87.8205	89.2655†	0.0962	0.0689	0.8563
	muchoCine_dataset	89.9457	90.3394	87.7717	88.3721†	0.0272	0.0198	0.9604
	news_racist_comments_spanish	83.6663	50.5338	80.4935	34.6457†	0.0411	0.0293	0.8753
gemma-3-1b-it	Sarcastic_spanish_dataset	91.4263	88.843	54.2207	36.1152†	0.4477	0.2973	0.0813
	spamspa	98.916	95.8621	82.9268	34.1463†	0.1671	0.1145	0.1808
	spa_climate_detection	96.9231	97.4843	81.6667	82.9152†	0.1526	0.1078	0.761
	muchoCine_dataset	89.9457	90.3394	85.3261	86.1538†	0.0625	0.0454	0.9369
	news_racist_comments_spanish	83.6663	50.5338	77.2033	19.8347†	0.0693	0.0488	0.7803
gemma-2-2b-it	Sarcastic_spanish_dataset	91.4263	88.843	53.4533	41.3471†	0.4612	0.3122	0.1185
	spamspa	98.916	95.8621	82.8365	34.0278†	0.1680	0.116	0.2053
	spa_climate_detection	96.9231	97.4843	89.2308	90.6667†	0.0795	0.0568	0.8625
	muchoCine_dataset	89.9457	90.3394	86.6848	87.3385†	0.0326	0.0237	0.9727
	news_racist_comments_spanish	83.6663	50.5338	80.141	34.2412†	0.0447	0.0318	0.8817
salamandra-2b-instruct	Sarcastic_spanish_dataset	91.4263	88.843	53.4004	37.4867†	0.4575	0.3123	0.1232
	spamspa	98.916	95.8621	82.7462	33.91†	0.0169	0.1172	0.2151
	spa_climate_detection	96.9231	97.4843	91.1538	92.4259†	0.0577	0.0414	0.8601
	muchoCine_dataset	89.9457	90.3394	87.5	87.8307†	0.0299	0.0218	0.9598
	news_racist_comments_spanish	83.6663	50.5338	81.0811	37.8378†	0.0423	0.0302	0.8706

Table 9: Results reached by the salient words of LIME against roberta-base.

generative_model	dataset	acc_OT	f1_OT	acc_MT	f1_MT	ASR	ASM	CS
Llama-3.2-3B-Instruct	Sarcastic_spanish_dataset	91.0029	88.3162	53.9825	38.0036†	0.4528	0.3036	0.1048
	spamspa	96.6576	88.1789	81.1201	32.7974†	0.2005	0.138	0.1865
	spa_climate_detection	98.0769	98.4456	91.0256	92.3913†	0.0731	0.0521	0.8385
	muchoCine_dataset	81.25	80.1153	76.9022	75.9207†	0.0652	0.0474	0.9652
	news_racist_comments_spanish	84.6063	63.5097	76.9683	37.1795†	0.0788	0.0556	0.8178
Qwen2.5-0.5B-Instruct	Sarcastic_spanish_dataset	91.0029	88.3162	54.2736	36.8421†	0.4536	0.3102	0.1025
	spamspa	96.6576	88.1789	81.1201	34.0694†	0.2005	0.1389	0.1997
	spa_climate_detection	98.0769	98.4456	92.3077	93.6575†	0.0577	0.0414	0.8788
	muchoCine_dataset	81.25	80.1153	77.7174	76.4368†	0.0408	0.0297	0.9679
	news_racist_comments_spanish	84.6063	63.5097	79.3184	44.6541†	0.0599	0.0427	0.8724
gemma-3-1b-it	Sarcastic_spanish_dataset	91.0029	88.3162	53.7179	35.9106†	0.4565	0.3023	0.0776
	spamspa	96.6576	88.1789	80.3975	31.9749†	0.2078	0.1417	0.165
	spa_climate_detection	98.0769	98.4456	91.4103	92.7879†	0.0692	0.0491	0.7925
	muchoCine_dataset	81.25	80.1153	72.0109	70.4871†	0.1087	0.079	0.9494
	news_racist_comments_spanish	84.6063	63.5097	74.0306	27.0627†	0.1105	0.0777	0.7782
gemma-2-2b-it	Sarcastic_spanish_dataset	91.0029	88.3162	54.009	40.3159†	0.4551	0.3081	0.1166
	spamspa	96.6576	88.1789	81.0298	33.121†	0.2014	0.1389	0.1969
	spa_climate_detection	98.0769	98.4456	93.0769	94.206†	0.0526	0.0376	0.8799
	muchoCine_dataset	81.25	80.1153	74.7283	73.5043†	0.0761	0.0554	0.9765
	news_racist_comments_spanish	84.6063	63.5097	79.5535	47.9042†	0.0552	0.0393	0.8683
salamandra-2b-instruct	Sarcastic_spanish_dataset	91.0029	88.3162	55.5703	35.4975†	0.4411	0.3016	0.1227
	spamspa	96.6576	88.1789	80.8491	33.3333†	0.2033	0.1404	0.2035
	spa_climate_detection	98.0769	98.4456	96.0256	96.7607†	0.0282	0.0203	0.8847
	muchoCine_dataset	81.25	80.1153	76.9022	75.2187†	0.0543	0.0396	0.9689
	news_racist_comments_spanish	84.6063	63.5097	79.671	46.1059†	0.0541	0.0386	0.8707

Table 10: Results reached by the salient words of LIME against roberta-large.

generative_model	dataset	acc_OT	f1_OT	acc_MT	f1_MT	ASR	ASM	CS
Llama-3.2-3B-Instruct	Sarcastic_spanish_dataset	68.616	67.3458	49.4575	45.831†	0.4795	0.3243	0.1173
	spamspa	98.5547	94.5578	82.6558	33.3333†	0.1734	0.1179	0.2208
	spa_climate_detection	97.4359	97.8992	84.359	85.7477†	0.1436	0.1025	0.8497
	muchoCine_dataset	50.5435	67.148	50.5435	67.148	0.0	0.0	0.9497
	news_racist_comments_spanish	80.141	37.1747	76.1457	16.4609†	0.0494	0.0351	0.8913
Qwen2.5-0.5B-Instruct	Sarcastic_spanish_dataset	68.616	67.3458	48.3726	46.9404†	0.4618	0.3125	0.1184
	spamspa	98.5547	94.5578	82.2042	33.67†	0.1780	0.1208	0.2277
	spa_climate_detection	97.4359	97.8992	91.0256	92.2222†	0.0718	0.0516	0.8895
	muchoCine_dataset	50.5435	67.148	50.5435	67.148	0.0	0.0	0.9521
	news_racist_comments_spanish	80.141	37.1747	77.7908	25.8824†	0.0470	0.0332	0.9218
gemma-3-1b-it	Sarcastic_spanish_dataset	68.616	67.3458	50.3308	42.5467†	0.4935	0.3318	0.1014
	spamspa	98.5547	94.5578	82.0235	31.6151†	0.1798	0.1225	0.2117
	spa_climate_detection	97.4359	97.8992	88.2051	89.6861†	0.1026	0.0729	0.8061
	muchoCine_dataset	50.5435	67.148	50.5435	67.148	0.0	0.0	0.9288
	news_racist_comments_spanish	80.141	37.1747	75.9107	16.3265†	0.0588	0.0417	0.8614
gemma-2-2b-it	Sarcastic_spanish_dataset	68.616	67.3458	48.6637	48.5138†	0.4694	0.3191	0.1243
	spamspa	98.5547	94.5578	82.5655	32.7526†	0.1743	0.1182	0.2249
	spa_climate_detection	97.4359	97.8992	91.2821	92.4945†	0.0692	0.0496	0.8789
	muchoCine_dataset	50.5435	67.148	50.5435	67.148	0.0	0.0	0.965
	news_racist_comments_spanish	80.141	37.1747	77.9083	26.5625†	0.0458	0.0326	0.9136
salamandra-2b-instruct	Sarcastic_spanish_dataset	68.616	67.3458	49.484	46.1191†	0.4596	0.3118	0.1309
	spamspa	98.5547	94.5578	82.1138	32.1918†	0.1789	0.1209	0.2298
	spa_climate_detection	97.4359	97.8992	94.4872	95.3714†	0.0423	0.0304	0.8919
	muchoCine_dataset	50.5435	67.148	50.5435	67.148	0.0	0.0	0.9598
	news_racist_comments_spanish	80.141	37.1747	77.0858	22.9249†	0.0564	0.0398	0.9222

Table 11: Results reached by the salient words of LIME against xlm-roberta-base.

generative_model	dataset	acc_OT	f1_OT	acc_MT	f1_MT	ASR	ASM	CS
Llama-3.2-3B-Instruct	Sarcastic_spanish_dataset	91.2411	88.1489	91.0558	87.8154	0.0001	0.0061	0.9981
	spamspa	99.095	96.5278	99.0045	96.1672	0.0	0.0006	0.9956
	spa_climate_detection	97.6923	98.1405	96.4103	97.0833†	0.0003	0.0207	0.9829
	muchoCine_dataset	82.337	82.2888	75.5435	74.4318†	0.0012	0.0787	0.9671
	news_racist_comments_spanish	84.1363	66.6667	85.7814	68.8946†	0.0004	0.0267	0.9929
Qwen2.5-0.5B-Instruct	Sarcastic_spanish_dataset	91.2411	88.1489	90.8177	87.5135†	0.0132	0.0087	0.999
	spamspa	99.095	96.5278	99.0045	96.1938	0.0	0.0006	0.9964
	spa_climate_detection	97.6923	98.1405	95.641	96.4286†	0.0004	0.0245	0.9907
	muchoCine_dataset	82.337	82.2888	73.913	73.1844†	0.0013	0.0878	0.973
	news_racist_comments_spanish	84.1363	66.6667	86.9565	69.589†	0.0007	0.0448	0.9969
gemma-3-1b-it	Sarcastic_spanish_dataset	91.2411	88.1489	89.4152	85.9944†	0.0003	0.0189	0.9965
	spamspa	99.095	96.5278	99.0045	96.1938	0.0	0.0018	0.9932
	spa_climate_detection	97.6923	98.1405	93.5897	94.824†	0.0005	0.0359	0.9687
	muchoCine_dataset	82.337	82.2888	64.4022	55.2901†	0.0028	0.1836	0.9387
	news_racist_comments_spanish	84.1363	66.6667	88.8367	72.9345†	0.0073	0.048	0.9895
gemma-2-2b-it	Sarcastic_spanish_dataset	91.2411	88.1489	91.1881	88.0773	0.0	0.0041	0.9989
	spamspa	99.095	96.5278	99.0045	96.1938	0.0	0.0006	0.9961
	spa_climate_detection	97.6923	98.1405	97.1795	97.6891	0.0002	0.0157	0.9927
	muchoCine_dataset	82.337	82.2888	75.0	70.8861†	0.0017	0.1169	0.9728
	news_racist_comments_spanish	84.1363	66.6667	85.6639	68.7179†	0.0004	0.0262	0.9945
salamandra-2b-instruct	Sarcastic_spanish_dataset	91.2411	88.1489	89.5475	86.067†	0.0003	0.017	0.9982
	spamspa	99.095	96.5278	99.0045	96.1938	0.0	0.0006	0.997
	spa_climate_detection	97.6923	98.1405	97.4359	97.921	0.0002	0.014	0.9832
	muchoCine_dataset	82.337	82.2888	69.837	61.5917†	0.0024	0.1635	0.9664
	news_racist_comments_spanish	84.1363	66.6667	87.6616	71.2329†	0.0006	0.0416	0.9952

Table 12: Results reached by the salient words of SHAP against bert-spanish-wwm-cased.

generative_model	dataset	acc_OT	f1_OT	acc_MT	f1_MT	ASR	ASM	CS
Llama-3.2-3B-Instruct	Sarcastic_spanish_dataset	91.4263	88.843	91.3734	88.7741	0.0005	0.0003	0.9981
	spamspa	98.914	95.8621	98.914	95.8621	0.0	0.0	0.9956
	spa_climate_detection	96.9231	97.4843	96.9231	97.4843	0.0	0.0	0.9829
	muchoCine_dataset	89.9457	90.3394	90.4891	90.8136	0.0054	0.0039	0.9671
	news_racist_comments_spanish	83.6663	50.5338	83.4313	49.0975	0.0047	0.0031	0.9929
Qwen2.5-0.5B-Instruct	Sarcastic_spanish_dataset	91.4263	88.843	91.3998	88.8201	0.0008	0.0005	0.999
	spamspa	98.914	95.8621	98.914	95.8621	0.0	0.0	0.9964
	spa_climate_detection	96.9231	97.4843	96.9231	97.4843	0.0	0.0	0.9907
	muchoCine_dataset	89.9457	90.3394	90.7609	91.0995	0.0136	0.0098	0.973
	news_racist_comments_spanish	83.6663	50.5338	83.7838	50.7143	0.0012	0.0008	0.9969
gemma-3-1b-it	Sarcastic_spanish_dataset	91.4263	88.843	91.1617	88.5144†	0.0026	0.0017	0.9965
	spamspa	98.914	95.8621	98.914	95.8621	0.0	0.0	0.9932
	spa_climate_detection	96.9231	97.4843	96.7949	97.3767	0.0013	0.0009	0.9687
	muchoCine_dataset	89.9457	90.3394	90.2174	90.625	0.0299	0.0216	0.9387
	news_racist_comments_spanish	83.6663	50.5338	83.9013	51.5901	0.0047	0.0032	0.9895
gemma-2-2b-it	Sarcastic_spanish_dataset	91.4263	88.843	91.3734	88.7741	0.00053	0.0003	0.9989
	spamspa	98.914	95.8621	98.914	95.8621	0.0	0.0	0.9961
	spa_climate_detection	96.9231	97.4843	96.9231	97.4843	0.0	0.0	0.9927
	muchoCine_dataset	89.9457	90.3394	90.7609	91.0995	0.0082	0.0059	0.9728
	news_racist_comments_spanish	83.6663	50.5338	83.6663	50.1792	0.0024	0.0016	0.9945
salamandra-2b-instruct	Sarcastic_spanish_dataset	91.4263	88.843	91.3734	88.7664	0.0005	0.0003	0.9982
	spamspa	98.914	95.8621	98.914	95.8621	0.0	0.0	0.997
	spa_climate_detection	96.9231	97.4843	96.7949	97.3767	0.0013	0.0009	0.9832
	muchoCine_dataset	89.9457	90.3394	91.0326	91.3838	0.0109	0.0079	0.9664
	news_racist_comments_spanish	83.6663	50.5338	83.7838	50.7143	0.0035	0.0023	0.9952

Table 13: Results reached by the salient words of SHAP against roberta-base.

generative_model	dataset	acc_OT	f1_OT	acc_MT	f1_MT	ASR	ASM	CS
Llama-3.2-3B-Instruct	Sarcastic_spanish_dataset	91.0029	88.3162	91.0029	88.3162	0.0	0.0	0.9998
	spamspa	96.6516	88.1789	96.6516	88.1789	0.0	0.0	0.9991
	spa_climate_detection	98.0769	98.4456	98.0769	98.4456	0.0	0.0	0.9915
	muchoCine_dataset	81.25	80.1153	81.7935	80.8023	0.0109	0.0079	0.9672
	news_racist_comments_spanish	84.6063	63.5097	84.7239	63.6872	0.0012	0.0008	0.9952
Qwen2.5-0.5B-Instruct	Sarcastic_spanish_dataset	91.0029	88.3162	90.9764	88.2778	0.0003	0.0002	0.9998
	spamspa	96.6516	88.1789	96.6516	88.1789	0.0	0.0	0.9995
	spa_climate_detection	98.0769	98.4456	98.0769	98.4456	0.0	0.0	0.994
	muchoCine_dataset	81.25	80.1153	81.7935	80.8023	0.0272	0.0197	0.9728
	news_racist_comments_spanish	84.6063	63.5097	84.8414	63.8655	0.0024	0.0015	0.996
gemma-3-1b-it	Sarcastic_spanish_dataset	91.0029	88.3162	90.9764	88.2778	0.0003	0.0002	0.9997
	spamspa	96.6516	88.1789	96.6516	88.1789	0.0	0.0	0.9985
	spa_climate_detection	98.0769	98.4456	98.0769	98.4456	0.0	0.0	0.9834
	muchoCine_dataset	81.25	80.1153	81.5217	80.4598	0.0299	0.0217	0.9441
	news_racist_comments_spanish	84.6063	63.5097	84.7239	63.6872	0.0012	0.0008	0.9916
gemma-2-2b-it	Sarcastic_spanish_dataset	91.0029	88.3162	91.0029	88.3162	0.0	0.0	0.9998
	spamspa	96.6516	88.1789	96.6516	88.1789	0.0	0.0	0.9993
	spa_climate_detection	98.0769	98.4456	98.0769	98.4456	0.0	0.0	0.9967
	muchoCine_dataset	81.25	80.1153	81.7935	80.8023	0.0109	0.0079	0.9754
	news_racist_comments_spanish	84.6063	63.5097	84.6063	63.3053	0.0024	0.0015	0.9973
salamandra-2b-instruct	Sarcastic_spanish_dataset	91.0029	88.3162	91.0029	88.3162	0.0	0.0	0.9998
	spamspa	96.6516	88.1789	96.6516	88.1789	0.0	0.0	0.9991
	spa_climate_detection	98.0769	98.4456	98.0769	98.4456	0.0	0.0	0.9931
	muchoCine_dataset	81.25	80.1153	81.5217	80.4598	0.0190	0.0137	0.9716
	news_racist_comments_spanish	84.6063	63.5097	84.6063	63.3053	0.0024	0.0015	0.9947

Table 14: Results reached by the salient words of SHAP against roberta-large.

generative_model	dataset	acc_OT	f1_OT	acc_MT	f1_MT	ASR	ASM	CS
Llama-3.2-3B-Instruct	Sarcastic_spanish_dataset	68.616	67.3458	68.669	67.3829	0.00212	0.0013	0.9975
	spamspa	98.552	94.5578	98.552	94.5578	0.0	0.0	0.9982
	spa_climate_detection	97.4359	97.8992	97.5641	98.0063	0.0013	0.0009	0.973
	muchoCine_dataset	50.5435	67.148	50.5435	67.148	0.0	0.0	0.927
	news_racist_comments_spanish	80.141	37.1747	79.906	36.4312	0.0047	0.0032	0.9835
Qwen2.5-0.5B-Instruct	Sarcastic_spanish_dataset	68.616	67.3458	68.5102	67.2357	0.0011	0.0007	0.9972
	spamspa	98.552	94.5578	98.552	94.5578	0.0	0.0	0.9982
	spa_climate_detection	97.4359	97.8992	97.0513	97.5866	0.0064	0.0043	0.9728
	muchoCine_dataset	50.5435	67.148	50.5435	67.148	0.0	0.0	0.9315
	news_racist_comments_spanish	80.141	37.1747	79.906	36.4312	0.0118	0.0079	0.9824
gemma-3-1b-it	Sarcastic_spanish_dataset	68.616	67.3458	68.6425	67.3644	0.0008	0.0005	0.9967
	spamspa	98.552	94.5578	98.552	94.5578	0.0	0.0	0.9977
	spa_climate_detection	97.4359	97.8992	97.5641	97.9979	0.0039	0.0026	0.9576
	muchoCine_dataset	50.5435	67.148	50.5435	67.148	0.0	0.0	0.8919
	news_racist_comments_spanish	80.141	37.1747	80.141	35.249	0.0165	0.0112	0.9767
gemma-2-2b-it	Sarcastic_spanish_dataset	68.616	67.3458	68.5366	67.2902	0.0008	0.0005	0.9981
	spamspa	98.552	94.5578	98.552	94.5578	0.0	0.0	0.9984
	spa_climate_detection	97.4359	97.8992	97.6923	98.1092	0.0026	0.0017	0.9845
	muchoCine_dataset	50.5435	67.148	50.5435	67.148	0.0	0.0	0.9482
	news_racist_comments_spanish	80.141	37.1747	80.2585	37.3134	0.0012	0.0008	0.9892
salamandra-2b-instruct	Sarcastic_spanish_dataset	68.616	67.3458	68.6954	67.3835	0.0013	0.0008	0.9978
	spamspa	98.552	94.5578	98.552	94.5578	0.0	0.0	0.9985
	spa_climate_detection	97.4359	97.8992	97.5641	98.0021	0.0013	0.0009	0.9794
	muchoCine_dataset	50.5435	67.148	50.5435	67.148	0.0	0.0	0.9376
	news_racist_comments_spanish	80.141	37.1747	80.0235	35.1145	0.0106	0.007	0.9861

Table 15: Results reached by the salient words of SHAP against xlm-roberta-base.

Original text	**Modified text**	**Attacker model**
Friday of science and global warming of NPR:	Friday of science and worldwide heat de NPR:	Llama-3.2-3B-Instruct
Stop Ronaldo	Neutralize Ronaldo	gemma-3-1b-it
One person arrested after new incidents at the juvenile center of Sopuerta. What is ILLEGAL is allowing that type of immigration	One person arrested after new cases at the juvenile center of Sopuerta. What is ILLEGAL is allowing that type of immigration	salamandra-2b-instruct

Table 16: Comparison of the original text with the one modified using synonyms in English. The underlined words are those that have been changed or should be changed. It is the translated version of the table 1.

Original text	**Modified text**	**Attacker model**
No, that only means that you have a big head.	No, that only means whose you have The big head.	Qwen2.5-0.5B-Instruct
The preventive medicine and the principle of precaution	The preventive treatment and the beginning of precaution	gemma-2-2b-it
Clippers with stars defeat Nuggets without stars in game of preseason	Clippers Salamander stars defeat home Nuggets without stars in game to preseason	salamandra-2b-instruct

Table 17: Comparison of the original text with the one modified using synonyms in English. The underlined words are those that have been changed or should be changed. It is the translated version of the table 3.