Subtle Shifts, Significant Threats: Leveraging XAI methods and LLMs to undermine Language Models Robustness

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 exhibit inherent security vulnerabilities, which may be related to several factors, among them the malicious alteration of the input data. Such weaknesses compromise the robustness of language models, which is more critical when adversarial attacks are stealthy and do not require high computational resources. In this work, we study how vulnerable English language models are to adversarial attacks based on subtle modifications of the input of pretrained English language models. We claim that the attack may be more effective if it is targeted to the most salient words for the discriminative task of the language models. Accordingly, we propose a new attack built upon a two-step approach: first, we use *a posteriori* explainability methods to identify the most influential words for the classification task, and second, we replace them with contextual synonyms retrieved by a small language model. Since the attack has to be as stealthy as possible, we also propose a new evaluation measure that combines the effectiveness of the attack with the number of modifications performed. The results show that pretrained English language models are vulnerable to minimal semantic changes, which makes the design of countermeasure methods imperative.

1 Introduction

The incessant growing of application scenarios of artificial intelligence (AI), and in particular natural language processing (NLP) services, 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 types of attacks in training that perturb the behaviour of the learning models, and even in inference time, in which in addition it is possible to raise privacy leaks (Irfan et al., 2021). Hence, current NLP models may be victims of these 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 exposing 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 learning model behaviour or to cause privacy leaks. In this work, we focus on the threat posed by the harmful modification of the learning model behaviour.

The data-based manipulation of learning models in inference time may have a specific target or not, but both attacks are based on the stealthy modification of the input to manipulate 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 pretrained language model (LM).

In this work, we claim that an adversarial attack against LM that only seeks to alter the outcome of LM in inference time may be more effective and sneaky if it is targeted to the most salient words for the discriminative task. Accordingly, we propose a new adversarial attack grounded in a strategy that minimise the number of modifications and keep almost unchanged the semantic meaning of the input. Hence, we first identify the salient words that drive the decision of the victim model. This selection is built upon *a posteriori* explainable artificial intelligence (XAI) methods that allow us to know the prominent features for the victim LM in this case. In particular, we evaluate the performance of LIME (Ribeiro et al., 2016), SHAP (Lundberg

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

Proceedings of Recent Advances in Natural Language Processing,pages 748–757
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-086

and Lee, 2017) and Captum (Kokhlikyan et al., 2020). Second, we propose replacing those salient words with their closest contextual synonym. In this case, we compare the performance of six small language models (SLM), namely Llama-3.1-8B-Instruct, Llama-3.2-1B-Instruct, gemma-2-9b-it, Gemma-3-1b-it, Qwen2.5-7B-Instruct and Mistral-7B-Instruct-v0.3.

As far as we know, there is no standard evaluation measure that combines the success of an untargeted data-driven adversarial attack at inference time and its stealthiness, since an attack cannot be considered successful if it is evident. We thus propose a new evaluation measure called Attack Score Metric (ASM) that takes into account the amount of output perturbations of the victim LM and the amount of words changed in the input, since the quality of the attack weakens the higher the number of modifications and the greater the change from the original meaning. We also measure how the attack keeps the semantic meaning of the original text seeking to maintain it. We evaluate our untargeted adversarial attack in four text classification datasets of two different tasks (hate speech, and sentiment analysis) and of different text genres: reviews, comments on social media and dialogues. We perform the attack against three pre-trained LM: bert-base-cased, distilroberta-base and xlnet-base-cased. Although there are differences amongst the dataset, the results show that the joint leveraging of XAI methods and LLMs enables a more effective and stealthy adversarial attack

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., 2024) combining prompt engineering with automated adversarial optimization. The adversarial landscape extends to data poisoning scenarios, where techniques such as ModelSonar identify undetectable backdoors (Jia et al., 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 faces significant obstacles balancing computational efficiency and adversary resilience (Pontes et al., 2024).

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). In this work we show that LLMs can be used as a tool for generating poisoning data for attacking pre-trained LM.

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 contextual synonym. Figure 1 depicts the structure of our attack, which we explain as 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 work, we evaluate the following *a posteriori* XAI methods.

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 et al., 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 their synonym.[2] We leverage the text generation capacity of LLMs in order to generate contextual synonyms given an specific word and its sentence as context. Figure 2 shows the prompt designed to constrain an LLM to give a unique synonym word.

There are words that do not have synonyms or the LLM does not return any synonym words. In this case, we do not replace that 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 (see Section 4.4).

4.1 The Attacker Model

The attacker model is composed of two modules to select the salient words and then generate their corresponding synonyms. The first one corresponds to the *a posteriori* XAI method. As explained in Section 3.1, we compare 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-3.1-8B-Instruct (Grattafiori et al., 2024), Llama-3.2-1B-Instruct (Grattafiori et al., 2024), gemma-2-9b-it (Team et al., 2024), Gemma-3-1b-it (Team et al., 2025), Qwen2.5-7B-Instruct (Yang et al., 2024) (Team, 2024) and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023). All SLMs are smaller than 9B parameters. Likewise, all the SLMs used are instructed in order to assure a better realisation of a specific instruction prompt.

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

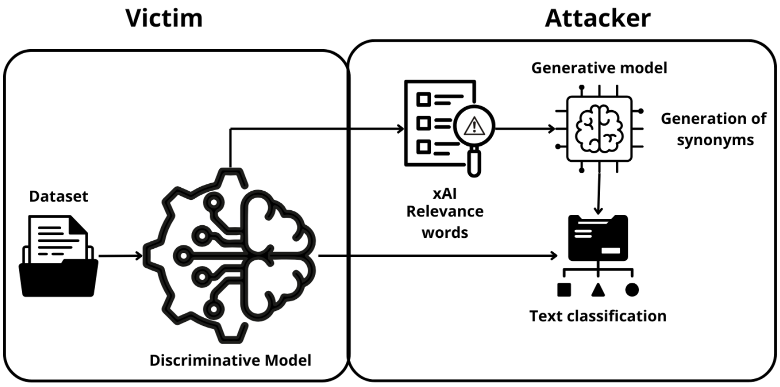

Figure 1: This is the outline of our untargeted adversarial attack.

> You are an assistant that replaces a single given word in a provided phrase with a contextually appropriate synonym.
> Instructions:
> 1. Carefully analyze the meaning and usage of the word within the context of the entire phrase provided.
> 2. Respond with a single replacement word that best fits the intended meaning in that specific context.
> 3. Your response must contain only the replacement word, with no extra text, punctuation, or explanation.
> 4. If no suitable contextual synonym exists, return a word that is as close as possible in meaning, even if not a perfect synonym.
> 5. Consider the phrase's tone, register, and intended meaning to ensure the replacement is natural and appropriate.
> Replace the word: WORD in this phrase: PHRASE

Figure 2: The prompt used for the synonym generation attack stage. WORD is the word we want to replace and PHRASE the sentence where is the word to replace.

4.2 The Victim Model

The victim models are pretrained language models (LM) for discriminative tasks. We evaluate our attack against four LMs, which means that we also assess their vulnerability against them. The LMs used are: bert-base-cased (Devlin et al., 2019), distilroberta-base (Sanh et al., 2019) and xlnet-base-cased (Yang et al., 2019). They represent a diversification in terms of their training scheme (Bert *vs.* Roberta *vs.* XLNet). We remark Captum could not be applied to xlnet-base-cased because of differences between their architectures and discrepancies among Captum tensor management and the tensor strcuture of xlnet-base-cased. All the LMs were fine-tuned in the three discriminative tasks.

4.3 Data

The attack evaluation was performed on four discriminative tasks. The LM were fine-tuned on the training set and evaluated on the test set of the following datasets.

rotten_tomatoes:[3] It is a sentiment analysis dataset of reviews from films. The size of the training set is 8,530 documents and of the test set 1,066 documents. The mean length of the test documents is 21.22 tokens (Pang and Lee, 2005).

FRENK-hate-en:[4] It is the English subset of the FRENK dataset (Ljubešić et al., 2021), a hate speech dataset. The size of the training set is 8,404 documents and of the test set of 2,301 documents. The mean length of the test documents is of 33.34 tokens (Ljubešić et al., 2019).

RECCON:[5] It is a dataset from conversations, indicating whether they contain emotions. The size of the training set is of 2,405 documents and of the test set of 816 documents. The mean length of the test documents is of 14.01 tokens (Poria et al., 2021).

sst2:[6] It is composed of sentences extracted from film reviews. The size of the training set is of 53,879 documents and of the test set of 872 documents. The mean length of the test documents is of 19.55 tokens (Socher et al., 2013).

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 then we assess the robustness of the attack according to the number of labels change, the number of words

[3] `https://huggingface.co/datasets/cornell-movie-review-data/rotten_tomatoes`

[4] `https://huggingface.co/datasets/classla/FRENK-hate-en`

[5] `https://huggingface.co/datasets/Deysi/sentences-and-emotions`

[6] `https://huggingface.co/datasets/stanfordnlp/sst2`

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. These evaluation measures shows the different performance of the non-attacked and attacked LMs, but they do not shed light regarding the effectiveness of the attack. In this context, the related literature lacks of a consensus on how to evaluate it. Hence, 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.

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

The ASR does not give any information about the stealth of the attack, hence we propose the ASM evaluation measure.

Attack Score Metric (ASM): It rectifies the ASR with respect to the amount of modifications carried out 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 ratio of the words changed with respect to the number of words in 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 computed this similarity by a vector representation of the original and manipulated text with the embedding model jina-embeddings-v3 model (Sturua et al., 2024). Then, we calculate the cosine value of the two vectors.

5 Results and Analysis

We analyse the results of our adversarial attack evaluation from different perspectives. First, we focus on XAI methods, assessing how effectively they identify salient words in order to increase the probability of success of the attack, and we examine the impact of these attacks on victim language models, measuring changes in their performance with the F1-Score measure (see Section 5.1). Finally, we study the role of the SLMs in the synonym generation stage by analysing errors related to incorrect or missing synonyms replacement (see Section 5.2), and conducting an ablation study to determine the importance of both identification of salient words and contextual synonym replacement in the overall attack process (see Section 5.3).

5.1 Adversarial Attack Performance Analysis

First, we analyse the performance of XAI methods under different experimental conditions, and subsequently evaluate how adversarial attacks substantially alter the performance of language models.

XAI method. The Figure 3 compares the XAI method using the ASM, where a higher value indicates that the XAI method is more effective at identifying key words whose substitution leads to successful attacks with less alterations of the input. In this context, LIME stands out with the highest ASM (0.1107), suggesting it is the most efficient at pinpointing impactful words for adversarial modifications, followed by Captum (0.0895), which is almost exactly at the general mean (0.0893), and Shap (0.0676), which is less effective by this metric. Thus, LIME provides the greatest leverage for adversarial interventions, while Shap is the least susceptible, with Captum occupying a middle ground.

Victim model. Since LIME performs better than the other two XAI methods in the selection of salient words, we show the effect of the attack on the LM using LIME as the XAI method. Figures 4, 5 and 6 show how the attack significantly harms the different LMs using the different LLMs.[7] We see that xlnet-base-cased consistently stands

[7] McNemar's test was used to calculate the performance difference between attacked and non-attacked LMs.

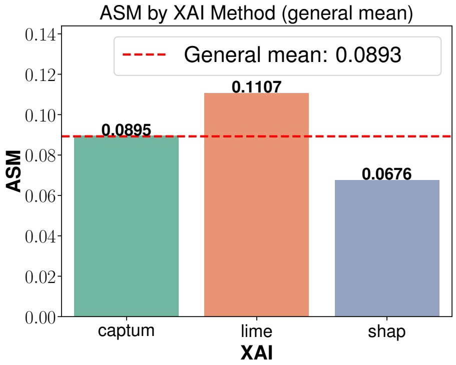

Figure 3: Comparison of ASM across xAI methods. The red dashed line sets the mean value.

out for its robust and stable performance, maintaining high effectiveness across tasks and showing resilience even in more challenging scenarios, bert-base-cased aexhibits slightly greater drops in F1 scores under attack, especially on datasets such as FRENK-hate-en and RECCON. In contrast, distilroberta-base, is more susceptible to performance degradation in demanding contexts, reflecting a trade-off between efficiency and robustness. These patterns highlight the distinct strengths and limitations of each architecture, with xlnet-base-cased excelling in stability, bert-base-cased offering a balanced profile but slightly less robust profile, and distilroberta-base prioritizing efficiency at the expense of some resilience.

Small Language Model. The figures 7 and 8 clearly illustrate the Attack Score Metric (ASM) and semantic preservation (cosine similarity) achieved by different generative models across all discriminative tasks. It is evident that Llama-3.2-1B-Instruct reached the highest mean ASM value (0.1072), significantly above the general mean of 0.0893, demonstrating its superior capability in generating synonyms that effectively alter the output of classification models. Interestingly, this comes at a cost to semantic preservation, as its cosine similarity (0.8283) falls below the general mean of 0.8461. By contrast, Qwen2.5-7B-Instruct shows the highest semantic preservation with a cosine similarity of 0.8717, but achieves the lowest ASM (0.0767) among all models. gemma-3-1b-it and Llama-3.1-8B-Instruct offer a more balanced performance, with ASM values of 0.0945 and 0.0853 respectively, and moderate cosine similarities. The results suggest that model size does not necessarily correlate with attack effectiveness, as the 1B parameter Llama-3.2-1B-Instruct outperforms larger models like Mistral-7B-Instruct-v0.3 and Llama-3.1-8B-Instruct in terms of ASM, challenging the assumption that larger models would generate more effective adversarial examples.

Table 1 shows some examples of how the SLM proposes synonyms that keep the original meaning and cause the alteration of the classification label. This also make evident 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, because we do not really know the real words that determine the classification of the LM, and we have to trust in the selection of the XAI method. Accordingly, we focus our error analysis on the generation of synonyms by the LLMs.

The errors observed in our adversarial attack primarily stem from the incorrect generation of synonyms or the absence of suitable replacements. Table 2 illustrates several examples of these issues. In the first case, the model switches the language of the target word. In the second, rather than replacing the word with a synonym, the model retains the original term and instead paraphrases the entire sentence, indicating a tendency to favor paraphrasing over direct substitution. In the third example, the model introduces additional words, thereby altering the original structure and expanding the content. These errors are attributable to the language model not strictly adhering to the prompt instructions, which may be due to limitations in the prompt design or inherent constraints of the model itself.

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 (RECCON and sst2) and two SLMs. Table 3 shows the results, and we see: (1) identifying salient words leads to a higher ASR compared to selecting words at random. However, this strategy may result in lower ASM values, since a LM might depend on more than two salient words to accurately classify an input text; and (2) the random replacement of the salient words tends to reach low

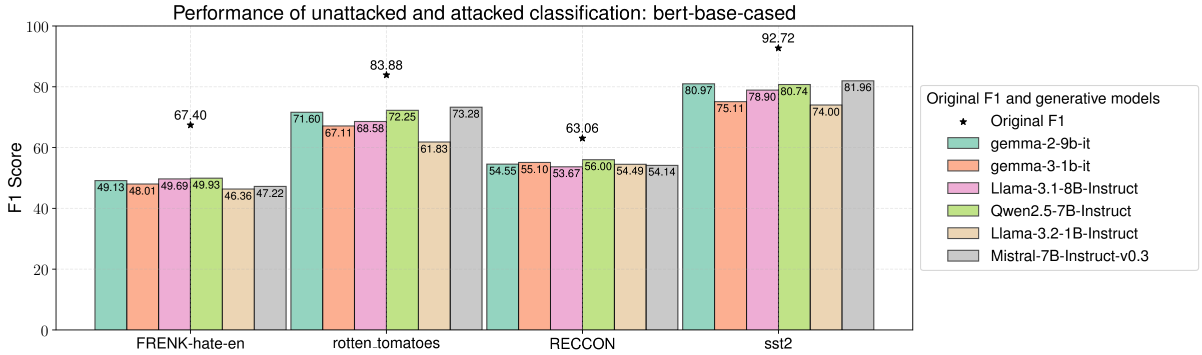

Figure 4: Performance comparison (F1 Score) of bert-base-cased on classification tasks without attacks and under generative model attacks across datasets.

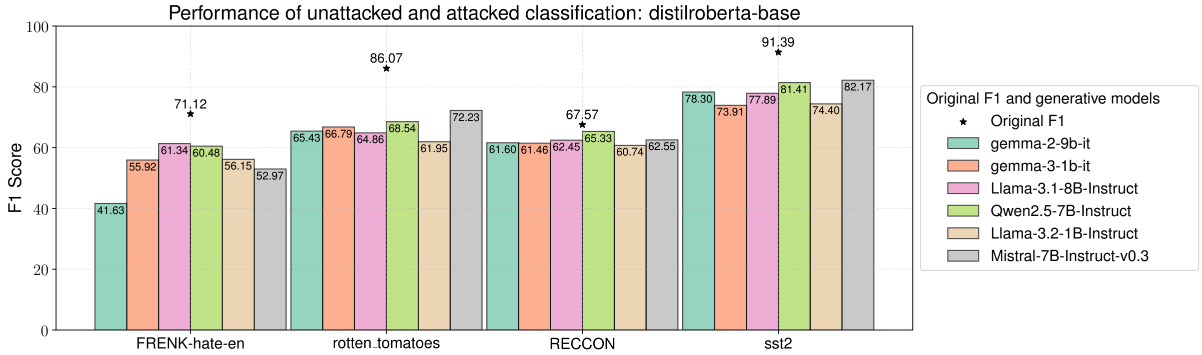

Figure 5: Performance comparison (F1 Score) of distilroberta-base on classification tasks without attacks and under generative model attacks across datasets.

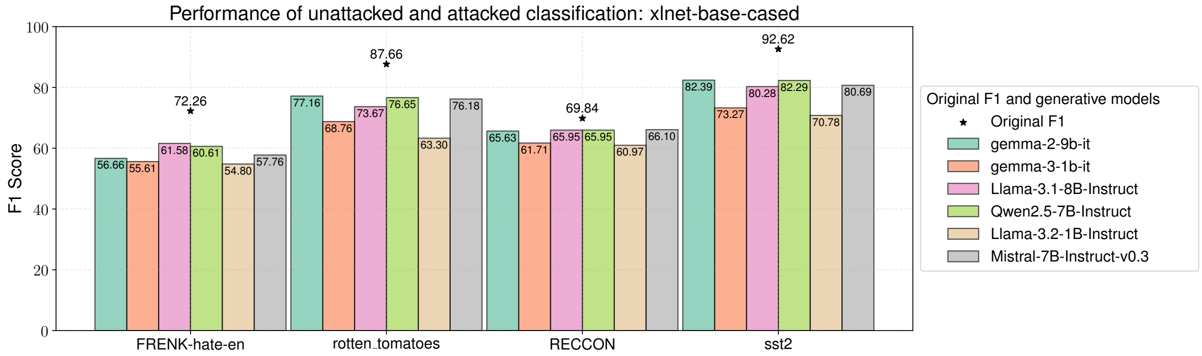

Figure 6: Performance comparison (F1 Score) of xlnet-base-cased on classification tasks without attacks and under generative model attacks across datasets.

Original text	**Modified text**	**Attacker model**
says the bigot	says the prejudiced	Llama-3.1-8B-Instruct
a moving and not infrequently breathtaking film.	an captivating and not occasionally stunning film.	Qwen2.5-7B-Instruct
Dad, I’m scared.	Dad, I’m terrified.	gemma-3-1b-it

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

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

Original text	Modified text	Attack model
how dare you!	how 胆敢 you!	Qwen2.5-7B-Instruct
The same to you .	The same Same (as) you. same to me	Mistral-7B-Instruct-v0.3
Good morning, Mary!	Good Good day, Hello!	gemma-2-9b-it

Table 2: Salient words (underlined ones) unfair replacement with their synonyms by SLMs.

SLM	Dataset	ASR	ASM	CS	Change
Gemma-2-9b-it	RECCON	0.0895	0.0598	0.8513	Our attack
		0.0539	0.0378	0.9107	2 random words with synonyms
		0.3775	0.2524	0.3235	XAI random word
	sst2	0.1525	0.1008	0.8168	Our attack
		0.0401	0.0284	0.9389	2 random words with synonyms
		0.4885	0.3229	0.2295	XAI random word
Llama-3.1-8B-Instruct	RECCON	0.1017	0.0680	0.8294	Our attack
		0.0539	0.0377	0.8997	2 random words with synonyms
		0.2500	0.1672	0.4092	XAI random word
	sst2	0.1732	0.1144	0.8029	Our attack
		0.0390	0.0276	0.9325	2 random words with synonyms
		0.6353	0.4198	0.3091	XAI random word

Table 3: Ablation analysis comparing the effectiveness of the proposed adversarial attack (targeted synonym replacement of salient words) with two baselines: (1) replacing two random words with synonyms, and (2) replacing salient words with random words.

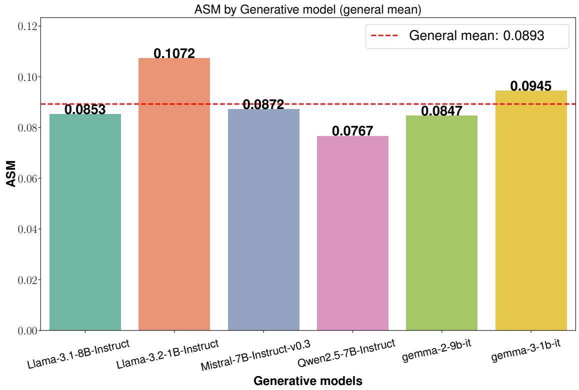

Figure 7: Comparison of ASM across SLMs. The red dashed line sets the mean value.

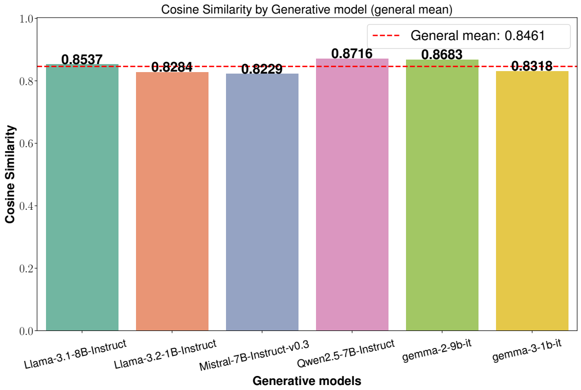

Figure 8: Comparison of CS across SLMs. The red dashed line sets the mean value.

their contextual synonyms.

6 Conclusions

This research introduces an untargeted adversarial attack against language models, leveraging the capabilities of LLMs and XAI methods. Our experimental results demonstrate that language models are vulnerable to subtle input modifications, particularly when these modifications target salient words.

The results allow us to conclude that our claim holds and contributes to: (1) a novel untargeted adversarial attack that preserves the semantic meaning of the input while achieving high success rates in altering model 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.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Shreya Goyal, Sumanth Doddapaneni, Mitesh M. Khapra, and Balaraman Ravindran. 2023. A Survey of Adversarial Defenses and Robustness in NLP. *ACM Comput. Surv.*, 55(14s).

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Muhammad Maaz Irfan, Sheraz Ali, Irfan Yaqoob, and Numan Zafar. 2021. Towards Deep Learning: A Review On Adversarial Attacks. In *2021 International Conference on Artificial Intelligence (ICAI)*, pages 91–96.

Jiabao Ji, Bairu Hou, Alexander Robey, George J Pappas, Hamed Hassani, Yang Zhang, Eric Wong, and Shiyu Chang. 2024. Defending Large Language Models against Jailbreak Attacks via Semantic Smoothing. *arXiv preprint arXiv:2402.16192*.

Jinyuan Jia, Yupei Liu, and Neil Zhenqiang 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.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B.

Narine Kokhlikyan, Vivek Miglani, Miguel Martin, Edward Wang, Bilal Alsallakh, Jonathan Reynolds, Alexander Melnikov, Natalia Kliushkina, Carlos Araya, Siqi Yan, et al. 2020. Captum: A unified and generic model interpretability library for PyTorch. *arXiv preprint arXiv:2009.07896*.

Xiaogeng Liu, Nan Xu, Muhao Chen, and Chaowei Xiao. 2024. AutoDAN: Generating Stealthy Jailbreak Prompts on Aligned Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Nikola Ljubešić, Darja Fišer, and Tomaž Erjavec. 2021. Offensive language dataset of Croatian, English and Slovenian comments FRENK 1.0. Slovenian language resource repository CLARIN.SI.

Nikola Ljubešić, Darja Fišer, and Tomaž Erjavec. 2019. The FRENK Datasets of Socially Unacceptable Discourse in Slovene and English.

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

Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales . In *Proceedings of the ACL*.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. 2022. Red Teaming Language Models with Language Models. *URL https://arxiv. org/abs/2202.03286*.

Marcos F. Pontes, Rodrigo C. Pedrosa, Pedro H. Lopes, and Eduardo 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)*.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, et al. 2021. Recognizing Emotion Cause in Conversations. *Cognitive Computation*, 13:1317–1332.

Marco Tulio Ribeiro, Sameer Singh, and Carlos 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.

Nuria Rodríguez-Barroso, Daniel Jiménez-López, M Victoria Luzón, Francisco Herrera, and Eugenio Martínez-Cámara. 2023. Survey on Federated Learning Threats: concepts, taxonomy on attacks and defences, experimental study and challenges. *Information Fusion*, 90:148–173.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Flavio Ambrosio da Silva. 2025. Navigating the dual-edged sword of generative AI in cybersecurity. *Brazilian Journal of Development*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Saba Sturua, Isabelle Mohr, Mohammad Kalim Akram, Michael Günther, Bo Wang, Markus Krimmel, Feng Wang, Georgios Mastrapas, Andreas Koukounas, Andreas Koukounas, Nan Wang, and Han Xiao. 2024. jina-embeddings-v3: Multilingual Embeddings With Task LoRA.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 Technical Report. *arXiv preprint arXiv:2503.19786*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving Open Language Models at a Practical Size. *arXiv preprint arXiv:2408.00118*.

Qwen Team. 2024. Qwen2.5: A Party of Foundation Models.

Zenan Wu, Liqin Tian, Yi Zhang, Yan Wang, and Yuquan Du. 2021. Network Attack and Defense Modeling and System Security Analysis: A Novel Approach Using Stochastic Evolutionary Game Petri Net. *Security and Communication Networks*.

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

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 Technical Report. *arXiv preprint arXiv:2407.10671*.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. *CoRR*, abs/1906.08237.

Yifan Yao, Jinhao Duan, Kaidi Xu, Yuanfang Cai, Zhibo Sun, and Yue Zhang. 2024. A Survey on Large Language Model (LLM) Security and Privacy: The Good, the Bad, and the Ugly. *High-Confidence Computing*, page 100211.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J Zico Kolter, and Matt Fredrikson. 2023. Universal and Transferable Adversarial Attacks on Aligned Language Models. *arXiv preprint arXiv:2307.15043*.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. 2024. PoisonedRAG: Knowledge Corruption Attacks to Retrieval-Augmented Generation of Large Language Models . *arXiv preprint arXiv:2402.07867*.