Heuristic Quality Assessment of Textual Adversarial Examples

Giuseppe Rizzo[1,†], Samuele Lo Cascio[1,†], Marco Morana[1,2,*,†] and Giuseppe Lo Re[1,2,†]

[1]Department of Engineering, University of Palermo, Italy
[2]Cybersecurity National Lab, CINI - Consorzio Interuniversitario Nazionale per l'Informatica, Italy

Abstract

In recent years, the field of Natural Language Processing (NLP) has experienced unprecedented growth, fueled by advances in machine learning and large-scale language models. Nevertheless, NLP systems remain highly susceptible to adversarial examples — carefully crafted inputs designed to trigger incorrect classifications. Due to the variability and complexity inherent in languages, defenses against such attacks are particularly challenging.In response to this scenario, some adversarial attack strategies alter the textual representation of inputs, thereby making them more likely to be detected by defence systems. Indeed, preserving the linguistic quality and effectiveness of adversarial examples requires substantial computing power and the simultaneous evaluation of multiple different factors. To address this issue, we present a heuristic approach for evaluating the quality of textual adversarial examples, with the aim of improving the resilience of the models under attack. Leveraging a data-driven design enables the quality assurance model to remain model-agnostic and highly modular, facilitating flexible integration with various cost-effective optimization processes. Furthermore, we present an extensive experimental analysis of the interplay between attackers and defenders, using game-theoretic frameworks to contextualize the impact of the quality of adversarial examples in the presence of anomaly detection mechanisms.

Keywords

Adversarial Machine Learning, NLP, Game Theory

1. Introduction

The increase in computational power has led to significant progress in Artificial Intelligence technologies. In particular, in the field of Natural Language Processing, the evolution from historical rule-based methodologies (such as Augmented Transition Networks [1]) and subsequent static statistical approaches, such as TF-IDF [2], reached a turning point with the introduction of Deep Learning architectures. Recurrent neural networks and, more recently, transformer architectures have enabled a more effective representation of the syntactic regularities inherent in natural languages, as well as allowing the modeling of latent semantic aspects in the textual data analyzed [3]. The use of such systems is not confined to a single task, as they operate as text feature extractors, allowing for a modular approach that can be adapted to different application requirements.

Nevertheless, these models share a common vulnerability: their exposure to so-called *adversarial examples*. These examples are instances specifically manipulated by a malicious agent aiming to induce the model to produce incorrect classifications or reject the input, thus generating malfunctions, service interruptions, and a consequent loss of trust on the part of the end user. This type of malicious attack, known as "evasion-attack", is one of the main threats to the robustness and reliability of machine learning models, as it aims to circumvent the system's decision-making mechanisms without apparently altering the input in detectable ways. Within NLP-tasks, this vulnerability manifests itself through the manipulation of texts at different levels of granularity [4], which can affect individual characters, words, or entire sentences. This condition is particularly advantageous for the attacker, as it

Joint National Conference on Cybersecurity (ITASEC & SERICS 2026), February 09-13, 2026, Cagliari, IT
*Corresponding author.
†These authors contributed equally.
giuseppe.rizzo24@unipa.it (G. Rizzo); samuele.locascio@unipa.it (S. L. Cascio); marco.morana@unipa.it (M. Morana); giuseppe.lo.re@unipa.it (G. L. Re)
0000-0002-5963-6236 (M. Morana); 0000-0002-8217-2230 (G. L. Re)

exploits the intrinsic variability and broad expressive capacity of natural languages, allowing for the introduction of imperceptible or contextually plausible alterations that can still lead to errors. However, such manipulations tend to deteriorate the syntactic and/or semantic coherence of the original texts, increasing the likelihood of detection by both automated systems and human inspection.

In order to ease this limitation, the literature frequently resorts to quality control mechanisms, the application of which, nonetheless, proves computationally onerous. To address this critical issue, we propose a data-centric approach to the qualitative evaluation of texts generated during the attack phase: the goal is to define efficient criteria and metrics that allow for the targeted optimization of adversarial generation procedures, thus improving the effectiveness of the hardening process of the models deployed without incurring the computational costs typical of traditional validation pipelines. In summary, our main contributions are as follows:

- The introduction of a novel heuristic, model-agnostic evaluation method designed to quantitatively and qualitatively assess the quality and validity of adversarial text samples. The proposed approach emphasizes computational efficiency and versatility across different architectures, thereby enabling consistent benchmarking of adversarial robustness in natural language processing systems.
- A comprehensive empirical evaluation of the proposed method was conducted across multiple benchmark datasets and diverse model architectures to rigorously investigate its performance and reliability. Further exploration of the method's limitations and boundary conditions through experimental analysis offers a nuanced understanding of its applicability in real-world adversarial scenarios.
- The usage of a game-theoretic validation framework grounded in the principles of game theory, integrated to systematically model and interpret the strategic interaction between the adversarial agent (attacker) and the defense mechanism (defender). This theoretical approach reinforces the empirical findings through formal analytical techniques.

The remainder of this paper is organized as follows. Section 2 reviews the current state of the art in adversarial attacks on NLP systems and relevant game-theoretical approaches. Section 3 introduces the formal threat model employed in this work. Section 4 outlines the experimental setup, while Section 5 presents and analyzes the results. Finally, conclusions are drawn in Section 6.

2. Related Works

The field of NLP Adversarial Machine Learning has evolved through several strands of research, with a particular emphasis on the development of attacks that generate adversarial examples with a satisfactory plausibility level in real-world contexts. Unlike adversarial examples in computer vision, where perturbations are often imperceptible pixel-level modifications, textual adversarial attacks must preserve both syntactic correctness and semantic coherence, making the construction of such examples considerably more challenging [4]. Nevertheless, several works have shown that even minimal textual modifications—such as synonym substitutions, paraphrasing, insertion or deletion of words, or reordering of sentence structure—can be sufficient to mislead state-of-the-art classifiers while remaining mostly undetected by human readers [5].

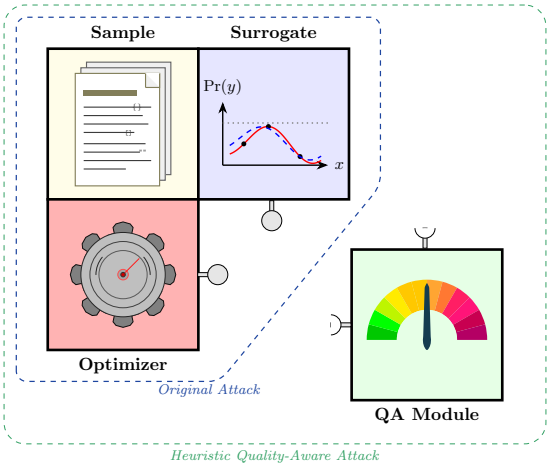
A wide range of approaches employ metaheuristic search methods, such as genetic algorithms, to generate samples. For instance, in [6] the authors propose a population-based method for creating adversarial examples and subsequently utilize these samples for adversarial training.

Building upon this work, the authors of [7] propose a certified robustness framework against adversarial word substitutions. To evaluate their method, they use a faster variant of the preceding attack which fixes neighbor sets, precludes repeated substitutions and employs a more efficient language model for scoring. This enables them to compute provable robustness guarantees and measure an empirical upper bound on attack success. Among the different approaches that leverage meta-heuristic techniques, one notable example is the method proposed in [8], which uses Particle Swarm Optimization (PSO) to

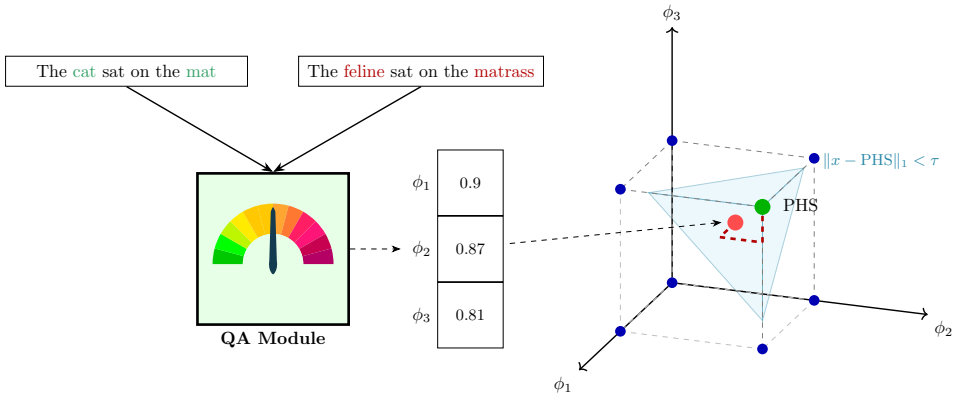
produce adversarial samples that preserve sememes. Carefully crafted perturbations that maintain the semantic integrity of the original text while misleading the models are used to challenge the robustness of state-of-the-art language models, including BERT and BiLSTM. On the other hand, Checklist [9] provides systematic, linguistically motivated transformations for constructing targeted adversarial examples that expose specific model failure modes and inform improvements in robustness. The framework implements hypothesis-driven tests, including invariance checks, directional expectations and minimal-pair perturbations, enabling the isolation of particular linguistic phenomena and the quantification of their effect on model behavior. Alongside these works is a class of algorithmic attacks that apply explicit selection criteria to determine which words to modify, thereby reducing the search space. These include TextFooler [10], which ranks words by their importance to the model's prediction, replacing the most influential ones to induce misclassification with minimal edits while preserving semantic similarity; PWWS (Probability Weighted Word Saliency) [11], which selects replacements based on the expected change in prediction probability; and BERT-ATTACK [12], which uses a pre-trained masked language model to suggest contextually appropriate replacements that maintain fluency and meaning more effectively. TextFooler and PWWS rely on surrogate signals (word importance scores and prediction probability changes) and constrained candidate pools (synonym lists or embedding nearest neighbors), making them effective in black-box settings. In contrast, BERT-ATTACK exploits a masked language model to generate more contextually plausible and grammatically coherent replacements, requiring fewer manual filtering steps. As has been described, effective adversarial testing is dependent on modularity and text quality. The proposed method combines these two aspects by systematically constraining the optimizer employed to generate high-quality adversarial examples.

Widening the perspective, it is evident that there is a clear parallel between the schematic frameworks of game theory and the adversarial dynamics between attacker and defender. Therefore, it is unsurprising that such formal methodologies have been employed in the literature to evaluate models in the field of AML. One notable example is the study presented in [13], in which the authors examine the Nash equilibrium between attack and defense strategies. A recurring theme, also explored in [14], in which the authors examine the use of adversarial training as an optimal response strategy within the framework of a two-player zero-sum game. Even when achieving a perfect equilibrium is not feasible, it is still possible to approximate it. This is demonstrated in [15], where the authors provide evidence for this claim through a framework that enables both participants to adopt randomized strategies. Although zero-sum games provide a fundamental theoretical basis, they often fail to capture more realistic interactions. Decision-making processes between players become more complex and asymmetric when players use more expressive formulations, such as Bayesian and Stackelberg games. These frameworks have proven to be of considerable value in the modeling of uncertainty and strategic behavior. For instance, in [16] the authors employ Stackelberg games to model the strategic interaction between a learner and multiple adversaries, capturing the asymmetry in their decision-making processes. In [17], Stackelberg games form the basis of an approach to adversarial regularization that aims to improve the model's generalization performance. The approach frames the learning process as a strategic interaction between the model and a virtual adversary, systematically exposing the model to challenging perturbations during training. On the other hand, in [18], the authors employ a game-theoretic framework to investigate the interaction between attackers and detection mechanisms. Using a Bayesian game, they model information asymmetry and derive the mixed-strategy Nash equilibrium to determine the optimal strategies for both parties, enabling a systematic evaluation of security and robustness against detection. Finally, in [19] the authors extend the game-theoretic approach by considering a pool of multiple models rather than relying on a single classifier. In a Repeated Bayesian Sequential Game framework, the learner interacts with the adversary and estimates the likelihood of an incoming query being adversarial. This information is then used to select the most appropriate model for prediction.

The literature review highlights the dual need to enhance the robustness of text processing systems against sophisticated perturbations and to make use of the strengths of formal, game-theoretic frameworks. Our work bridges these two areas by aiming to reduce the cost of generating adversarial samples while improving their effectiveness and demonstrating their applicability within a Bayesian game-theoretic context.



(a) System architecture



(b) Quality-Assurance flow

Figure 1: Overall framework: (a) Architecture and (b) QA mechanism.

3. Formulation

Suppose that an attacker operates under a strict black-box regime, meaning he has no knowledge of the model, parameters or pre-processing conducted by the system under attack: a threat model that is considered the most realistic in the literature [20]. Thus, given a target classifier $f(\cdot)$, he trains or otherwise obtains a surrogate model $\hat{f}(\cdot)$ that approximates its behavior to some extent. Therefore, he launches an attack $\mathcal{A}(\cdot)$ on the surrogate, leveraging a knowledge base Θ that encodes his objective $g(\cdot)$, the attacks' constraints Δ and the employed manipulations T [21], taking advantage of the transferability [22] of these samples to subsequently target the original model. Formally, the adversary searches the admissible perturbation space $S(\Delta, T)$ for an example x' that optimizes its objective with respect to the surrogate. This can be expressed universally as:

$$x' = \arg \max_{x \in \mathcal{S}(\Delta,T)} g\big(\hat{f}(x), y\big), \tag{1}$$

where g may be interpreted as a confidence score or a loss function computed with respect to the true label y. Conversely, in the case of a targeted attack, the adversary attempts to minimize the same objective in Eq. 1 while constraining the perturbation to induce a specific target label y_t different from the original label of the modified sample. Formally, a targeted formulation can be written as:

$$x' = \arg \min_{x \in \mathcal{S}(\Delta,T)} g\big(\hat{f}(x), y_t\big) \ \textbf{s.t} \ \ y_t \neq y, \tag{2}$$

Algorithm 1: Adversarial Quality Assurance

Data: x' (adversarial sample), x (original sample), τ (threshold)
Result: satisfying the constraint (bool)

$x'_s \leftarrow$ tokenize(x') ; /* Subset of words that constitutes x' */
$x_s \leftarrow$ tokenize(x);
features $\leftarrow$ [];
$S = (x'_s \cup x_s) \setminus (x'_s \cap x_s)$; /* Symmetric difference: modified words */
for *each metric* ϕ_i **do**
  if ϕ_i *is word-based* **then**
    scores $\leftarrow$ [];
    for $w \in S$ **do**
      score$_w \leftarrow \phi_i(w, x, x')$; /* Evaluate impact of word w */
      scores.append(score$_w$);
    features[i] $\leftarrow \frac{1}{|S|} \sum_{w \in S}$ score$_w$;
  else
    features[i] $\leftarrow \phi_i(x, x')$; /* Evaluate entire sentences */
return $\|\text{features} - \text{PHS}\|_1 \leq \tau$

where the constraint ($y_t \neq y$) indicates that the adversary aims to force the classifier toward the chosen target class. As for the optimization constraints Δ, if they admit a geometric description, the admissible perturbation set can be represented as a hypercube whose vertices correspond to the lower and upper bounds of each feature. Denoting by $\ell = (\ell_1, \dots, \ell_d)$ and $u = (u_1, \dots, u_d)$ the component-wise minimum and maximum limits determined, the feasible set may be written as the Cartesian product of closed intervals:

$$\Delta = \prod_{i=1}^{d} [\ell_i, u_i] = \{x \in \mathbb{R}^d \mid \ell_i \leq x_i \leq u_i, \ \forall i = 1, \dots, d\}. \tag{3}$$

Equivalently, when the bounds are specified relative to an original sample x_0 via a uniform perturbation budget ϵ, the constraint reduces to an L_p-ball around x_0:

$$\Delta = \{x \in \mathbb{R}^d \mid \|x - x_0\|_p \leq \varepsilon\}. \tag{4}$$

If the chosen norm is $p = 1$, this set is a convex polytope whose extreme vertices of the L_1-ball are given by the vectors $\pm\varepsilon e_i$ (i.e., perturbations that concentrate the entire budget on a single coordinate). From an operational standpoint, choosing the L_1 norm favors sparse perturbations (i.e., a few coordinates modified to a greater extent) over the "diffuse" perturbations induced by a generic $L_{p>1}$ constraint. The heuristic method proposed relies precisely on this type of constraint and is implemented in the *quality-assurance*(QA) module. This module, which is independent of the specific type of attack used (Fig. 1a), provides a data-centric assessment of the quality of the adversarial text samples generated by the attacker, based on a combination of appropriately selected metrics. Each of the metrics identified defines a coordinate within a d-dimensional space, whose values, by construction, are between 0 and 1 (Fig. 1b); these intervals constitute the boundaries of the domain previously described in Eq. 3. Formally, the module can be represented as a function that takes as input the pair of texts composed of the original sample x and its perturbed version x', returning a vector in $\mathbb{R}^d$:

$$f(x, x') \longrightarrow \{\phi_i(x, x')\}_{i=1}^{d}, \tag{5}$$

where $\phi_i(x, x')$ represents the value calculated by the i-th metric, normalized in the interval $[0, 1]$. Each metric could evaluate different specific aspects of the quality of the generated text, such as syntactic consistency, semantic fidelity, or lexical correctness. The assessment of the generated samples is carried out in the multidimensional space defined by the selected metrics, with respect to a fixed reference point: the unit edge, where all components assume the maximum value of 1, ideally corresponding to full agreement with the original text or, more generally, to maximum satisfaction of the quality criteria

considered. A perturbed sample x' is considered acceptable by the QA module if its distance L_1 from the *Point of Highest Similarity* (PHS) is less than a predefined threshold τ:

$$\|f(x, x') - \text{PHS}\|_1 < \tau. \tag{6}$$

The choice of norm L_1 plays a crucial role: it allows the evaluations of individual metrics to be decoupled, preventing a significant alteration in one dimension from being compensated for by high values in others. In this way, each metric contributes linearly and independently to the overall quality assessment, ensuring that any significant deviations are adequately penalized. Alg. 1 illustrates the implementation of the proposed methodology. Given an attack procedure defined by either Eq. 1 or Eq. 2, and a threshold value, τ, the attacker first tokenizes the adversarial and original examples to identify differing words (lines 1-4). For each of the selected d metrics, if the selected metric operates at the word level, the impact of each differing token is individually computed and then aggregated via mean-pooling to ensure consistency; this value is stored as the i-th feature (lines 5-11). Conversely, if the metric is defined at the sentence level, its corresponding value is adopted directly. Finally, after accumulating all the features associated with the sample under analysis, the L_1 distance from the PHS is computed and compliance with the threshold τ is verified. This approach facilitates the identification of scattered and controlled perturbations, preserving the syntactic consistency, semantic fidelity, and lexical correctness of the generated samples, and provides a rigorous basis for data-centric constraints in the generation of adversarial examples.

3.1. Game-Theoretic Evaluation Framework

In order to evaluate the effectiveness of the proposed module, we adopted a formalization in terms of strategic interaction between attacker and defender, using the principles of game theory applied to AML [23, 18]. We assume that the attacker generates a set of adversarial samples using the surrogate model, thus obtaining perturbations characterized by different levels of quality. These levels can be discretized by equibin subdivision and subsequently labeled, defining a set $Q = \{q_i\}_{i=1}^{|Q|}$, where each element q_i represents the label associated with the i-th quality interval. Before submitting any sample to the target model, the attacker can therefore define a tuple (s, q), where:

- s: indicates the nature of the sample, distinguishing between legitimate (0) and adversarial instances (1);
- q: represents the quality label associated with the sample, as derived from the quality assurance module.

Each quality bin is associated with a generation cost, c_q, which increasingly reflects the empirical observation that producing higher-quality adversarial samples requires greater computational and semantic effort. Each bin is also assigned an evasion score, ϵ_q, which is proportional to the success rate achieved against the surrogate model and indicative of the expected transferability of the attack to the target classifier. Finally, a reward term, r_q, is introduced to quantify the potential gain or utility obtained when an adversarial sample belonging to a given quality bin successfully evades detection. At the same time, the defender is equipped with an anomaly detection system, indicated by $d(\cdot)$, whose decision is based on a threshold mechanism κ applied to the score z associated with the observed sample. Starting from the observation of z, the defender can estimate the corresponding posterior distribution using a prior $\pi_0 = P(s = 1)$ that represents the probability that the sample is adversarial:

$$\pi(z) = \frac{f(z \mid s = 1)\pi_0}{f(z \mid s = 1)\pi_0 + f(z \mid s = 0)(1 - \pi_0)}, \tag{7}$$

where $f(z \mid s)$ is the conditional likelihood of observing z given type s. Note that in this formulation, the prior π_0 is fixed and not updated across rounds or observations; therefore, the defender selects a single optimal threshold τ once under a stationary belief about the environment. Accordingly, Table 1 presents the utilities $U(\cdot)$ for each player, which are determined by the possible game outcomes. The

a_A	a_D	$U_D(a_D)$	$U_A(a_A, s, q)$
Bluff	Accept	V	ρ
Bluff	Reject	$V - c_D$	$-\lambda$
Attack	Accept	$-\|L\|$	$r_q - c_q$
Attack	Reject	$V - \gamma$	$-c_q$

Table 1
Defender and Attacker Payoffs by Actions and Attack Quality

corresponding rewards and costs comprise the defender's benefit V from accepting a legitimate sample, the cost of a false rejection c_D, the loss incurred when an adversarial sample is accepted $|L|$, the defensive cost γ, the attacker's reward ρ for a legitimate sample being accepted and the rejection cost λ. The defender chooses to reject a sample if the expected utility of doing so is greater than acceptance [24]:

$$\mathbb{E}[U_D(\text{reject}) \mid z] \geq \mathbb{E}[U_D(\text{accept}) \mid z]. \tag{8}$$

Substituting the expectations with the posterior $\pi(z)$:

$$\pi(z)(V - \gamma) + (1 - \pi(z))(V - c_D) \geq \pi(z)(-|L|) + (1 - \pi(z))V, \tag{9}$$

which means that rejection is favorable when:

$$\pi(z) \geq \frac{c_D}{c_D - \gamma + V + |L|}. \tag{10}$$

Moreover, by leveraging Eq. 7 we could define the *likelihood ratio rule* for the optimal detection threshold from the previous equation:

$$\frac{f(z \mid s = 1)}{f(z \mid s = 0)} = \frac{(1 - \pi_0)}{\pi_0} \cdot \frac{c_D}{-\gamma + V + |L|} \quad \text{for } z = \tau^*. \tag{11}$$

By setting the optimal threshold, derived from Eq. 11, the defender can balance the trade-off between false positives (rejecting legitimate samples) and false negatives (accepting adversarial ones). In equilibrium, the detector acts as a rational filter that maximizes the defender's expected utility in an uncertain environment, encapsulating the game's fundamental strategic interplay between detection and evasion.

On the other hand, given the defender's fixed threshold τ, the attacker chooses a (possibly mixed) strategy, that maximizes its expected utility:

$$\max_{\sigma_A} \sum_{s,q} \sigma_A(s, q)\, u_A(s, q, d(\tau)), \tag{12}$$

A pair (σ_A^*, τ^*) constitutes a *Bayesian Nash Equilibrium (BNE)* if the following conditions hold:

1. Given the defender's strategy τ^*, the attacker's mixed strategy σ_A^* maximizes its expected utility:

$$\sigma_A^* = \arg\max_{\sigma_A}, \mathbb{E}[u_A \mid \tau^*], \tag{13}$$

2. Given the attacker's equilibrium strategy σ_A^*, the defender's decision threshold τ^* maximizes its expected utility under the fixed prior π_0:

$$\tau^* = \arg\max_{\tau}, \mathbb{E}[u_D \mid \sigma_A^*, \pi_0]. \tag{14}$$

No sequential updating or adaptive learning takes place; all expectations are evaluated with respect to the static prior π_0. This setting corresponds to a one-shot Bayesian game with incomplete information, modeling a strategic interaction between the attacker and the defender. The single-round formulation is motivated by two considerations: (i) adversarial examples are typically generated using surrogate models and (ii) the attacker may employ spoofing strategies to bypass tracking mechanisms.

Metric	Description
Bleu Score (B)	Measures precision of n-grams between generated and reference text, often used to evaluate machine translation quality.
GloVe-Similarity (GVS)	Computes semantic similarity between generated and reference text using cosine similarity of GloVe word embeddings.
Rouge Score (R)	Evaluates overlap between generated and reference text based on n-grams, word sequences, or word pairs.
Bert-Distance-to-Cluster (BDC)	Quantifies how close the embedding of a token (using BERT) is to a semantic cluster centroid (sentence's token mean pooling), indicating contextual alignment.
Information Gain (I)	Measures how much a token contributes to reducing uncertainty or entropy in the output space.

Table 2
Metrics used in the QA module during experimental evaluation. Word-based metrics are highlighted in gray.

4. Experimental Setup

To validate the proposed framework, we conducted a series of domain-specific classification experiments, employing a diverse set of models and adversarial attack strategies. All experiments were implemented using the TextAttack [25] framework, which offers a modular, extensible environment for adversarial NLP. This environment includes standardized attack recipes, datasets and pre-trained models. This ensured a consistent and reproducible experimental setup. We selected both Transformer-based architectures and a recurrent neural network as target models to cover different model paradigms. Specifically, we employed BERT [26], RoBERTa [27] and a word-level LSTM classifier, opportunely trained on three different datasets: AG-News [28], Movie Reviews[Rotten Tomatoes] [29], Yelp-Polarity [30]. For each of these datasets, the respective training, validation and test splits were employed in accordance with their standard set-up. AG-News comprises four balanced categories of English news articles (World, Sports, Business and Science/Technology), with the test partition containing $7,600$ examples evenly distributed across these categories. The Movie Reviews dataset consists of short movie review excerpts annotated for binary sentiment (positive or negative). It encompasses $10,662$ instances in total, of which approximately $1,066$ samples are conventionally assigned to the test set. Lastly, Yelp-Polarity comprises full-text user reviews from Yelp that have been labeled according to sentiment: positive or negative. It includes $560,000$ training samples and $38,000$ test examples, split evenly between the two categories. During the attack phase, 100 samples per dataset were extracted from the respective test sets. To assess robustness, we applied two distinct attack approaches available in the framework:

- **FasterAlzantot (FA)** [7]: a genetic algorithm-based adversarial attack that evolves perturbed text samples iteratively to mislead the target model;
- **PSO** [8]: which uses a particle swarm optimisation strategy to identify semantically preserving perturbations;

This combination of search-based, evolutionary, and behavioral methods enabled us to evaluate model robustness comprehensively, capturing vulnerability to lexical perturbations and inconsistencies. All experiments were performed using TextAttack's modular components (transformations, constraints, goal functions, and search methods) to ensure transparency, reproducibility, and comparability across attack types. In regards to the QA module, a selection of 5 metrics was chosen, the specifications of which are shown in Tab. 2. Surface-level metrics such as BLEU [31], ROUGE [32] and GloVe-Similarity [33] capture lexical and semantic overlap between the generated text and the adversarial one, ensuring that the model produces coherent and relevant outputs in context. In contrast, deeper, embedding-based metrics, such as BERT-Distance-to-Cluster [34] and Information Gain [35], assess the semantic consistency and informational contribution of tokens. The combined use of these metrics allowed for a multifaceted and efficient validation of the generation process. Regarding the detection mechanisms employed by the defender, three different defence strategies were adopted, each based on distinct hypotheses and operating principles:

Model Name	Dataset	Accuracy (%)
LSTM	AG News	91.4
	Movie Reviews	80.7
	Yelp Polarity	92.2
BERT	AG News	94.2
	Movie Reviews	87.6
	Yelp Polarity	96.3
RoBERTa	AG News	94.7
	Movie Reviews	89.9

Table 3
Performance of different models on various datasets. Accuracy is reported in percentage (%).

- **Local Outlier Factor** (LOF): applied to embeddings obtained downstream of a semantic extraction process using SBERT [36]. This allows the semantic consistency of the input samples to be evaluated with respect to the distribution of legitimate data, detecting any anomalies based on local density.
- **Perplexity-based detector**: which evaluates the probability of input samples in terms of perplexity associated with the reference linguistic model (GPT-2[1]), assuming that adversarial or out-of-distribution examples shows values that diverge significantly from those expected for legitimate data.

As for the definition of decision thresholds, both in the case of LOF in novelty detection mode and for the other two strategies, two complementary approaches were considered: on the one hand, the use of threshold (τ) values derived from the theoretical formulations reported in Eq. 11. On the other hand, the evaluation of a range of empirical thresholds in order to analyze the impact of these variations on the overall behavior of defense systems and, in particular, on the payoffs associated with both the attacker and the defender during the evaluation of the generated adversary samples.

Experiments were run on a workstation equipped with AMD Ryzen 5 5600G processor (6 cores/12 threads), 24GB of RAM, an NVIDIA GeForce RTX 4060 GAMING OC with 8GB GDDR6.

5. Results

Firstly, we report the accuracy values achieved by the models considered on their respective datasets, as shown in Table 3. For the LSTM model, we observe an accuracy of 91.4% on AG News, 80.7% on Movie Reviews, and 92.2% on Yelp Polarity. The BERT model shows superior performance, reaching 94.2% on AG News, 87.6% on Movie Reviews, and 96.3% on Yelp Polarity. Finally, RoBERTa achieves an accuracy of 94.7% on AG News and 89.9% on Movie Reviews. As can be noticed, the reported values are in general highly satisfactory, highlighting the superiority of Transformer-based models compared to LSTM. It should be noted that, consistent with preliminary empirical evaluations, the RoBERTa + Yelp Polarity combination was excluded from subsequent analyses.

Once these models were acquired, we evaluated their robustness in two separate phases. During the first one, we employed the aforementioned attacks in their original form, whereas in the second one, the attacks were deprived of their original constraints and supplemented by the QA module. As highlighted below, this allows us to consistently obtain higher quality samples, but also to geometrically approximate the final effect imposed by the original constraints, without compromising the success rate, thus balancing quality and strength effectively.

As reported in Fig. 2, we first observe that the proposed system does not significantly affect the effectiveness of the underlying attacks, as the performance remains comparable to that of the original methods. Specifically, the original FA achieves an average success rate of 64.3% across the three models,

[1]https://huggingface.co/docs/transformers/perplexity

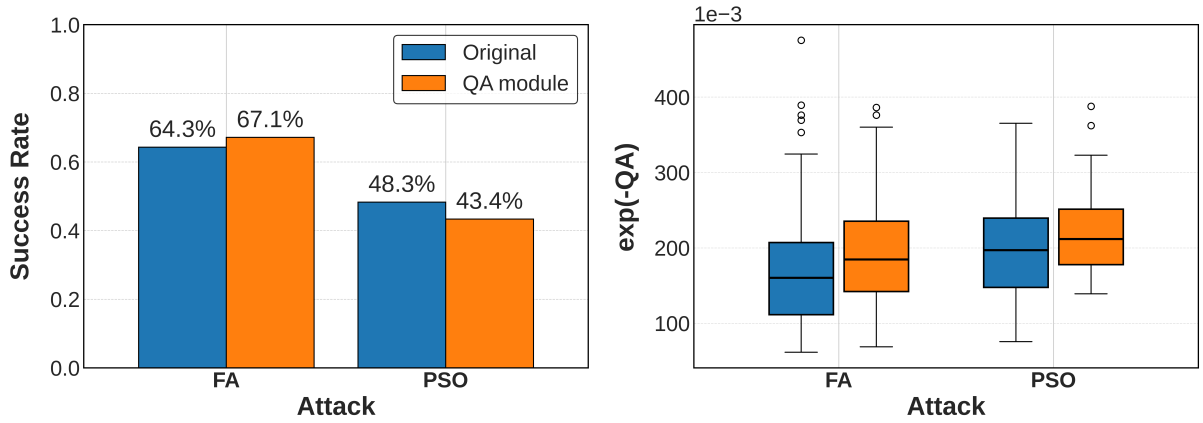


Figure 2: Success rate (left) and quality score distribution (right) for examples generated with and without the QA module.

while the same attack augmented with the QA module slightly improves this result, reaching 67.1% (+3%). Conversely, for PSO we observe an opposite trend: the original attack attains a success rate of 48.3%, surpassing the QA-augmented variant by approximately 5%, which achieves 43.4%. As for the quality of the corresponding adversarial examples, we observe that samples generated with the proposed QA module consistently achieve higher quality levels. The adopted visualization, based on the negative exponentiation of the quality score associated with each sample, is motivated by interpretability considerations: since quality is defined in terms of the minimum distance to the *PHS*, this transformation provides a more intuitive visual representation.

Referring to the metrics reported in Table 2, we can see that attacks without constraints lead to more pronounced variations than the reported metrics can capture. FA-based attacks in particular exhibit a median score of 189, while PSO-based ones reach a median of 206. These results suggest that, in the absence of additional semantic constraints, the generated adversarial examples tend to introduce stronger, and in some cases less controlled, perturbations. Including the QA module as an additional constraint significantly alters the behavior of the attacks. In this setting, adversarial examples display lower overall quality scores, with medians of 209 for FA-based attacks and 217 for PSO-based attacks. Furthermore, the dispersion of scores is notably reduced. Attacks generated with the QA module have a smaller interquartile range of 75 points for FA and 70 points for PSO compared to the larger ranges of 88 and 91 points observed for unconstrained attacks.

Overall, we can affirm that the proposed approach achieves results that are comparable to, and in some cases better than, those obtained with alternative strategies that rely on computationally expensive controls. These controls are instead effectively approximated by our modular geometric approach, yielding a more efficient solution without sacrificing performance.

To assess the impact of the proposed QA module from a differing perspective, we modeled the interaction between the attacker and defender as a game. In our evaluation model, the attacker employs a randomized strategy with a prior probability $\pi_0 = 0.5$ enabling an equal evaluation of the effects of adversarial versus original examples. In turn, the defender deployed anomaly-detection-based defense models, as described in the previous section. As can be seen in Fig. 3, the use of the QA module makes adversarial samples more difficult to detect for the anomaly detectors under study. In terms of true positive rate, the Perplexity-based anomaly detector identifies only 32% of adversarial examples when the QA module is used, compared to 37% for examples generated using the original attacks. Even when LOF is used, 40% of attacks are detected, versus 47% for samples obtained without the QA module. The use of the game theory-based formulation of the problem yields interesting results, as shown in the graph on the right in Figure 3. Using the above-defined parameters r_q, c_q, ρ, λ, V, L, c_D, and varying the prior probability of anomaly π_0, we analysed how the use of the QA module can give the attacker an advantage over the defender. For each of the prior probabilities π_0 studied, the introduction of the QA module resulted in an increase in the attacker's score, at the expense of the defender's score.

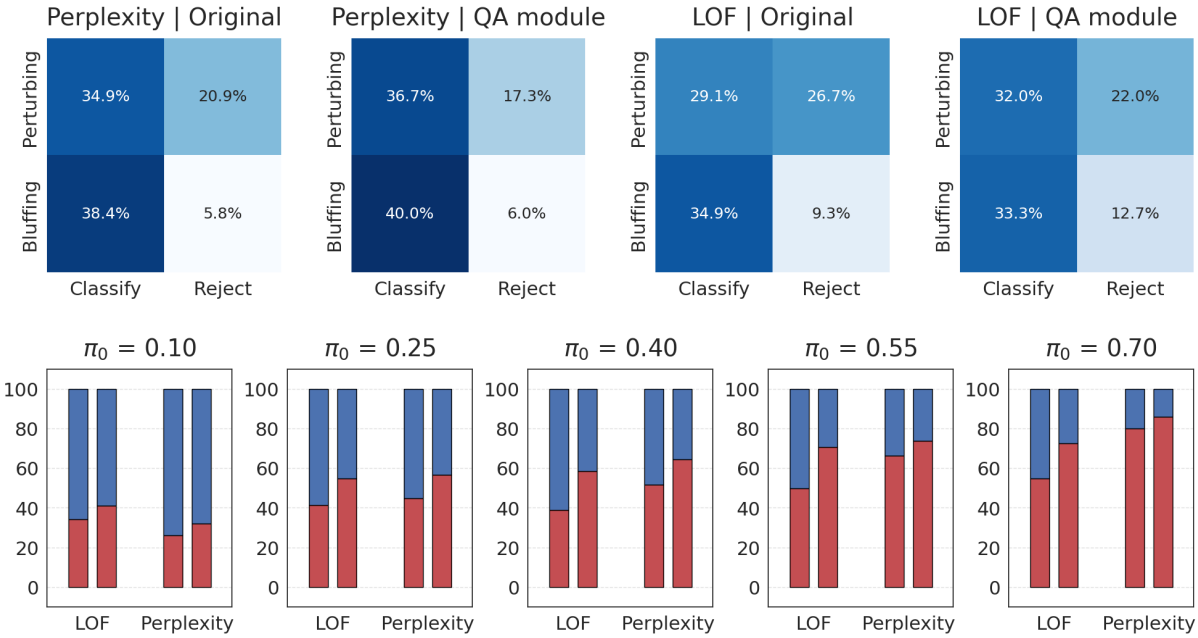


Figure 3: Up: performance of anomaly detectors under a game with prior probability $\pi_0 = 0.5$; Down: Game results for different anomaly detectors across varying values of π_0. For each anomaly detector two pairs of bars are shown: the first for Original attacks, the second for attacks using QA module. Attacker payoffs are in red, defender payoffs in blue

For low values of π_0, the attacker's payoff increases with the introduction of the QA module; however, this improvement is insufficient to outweigh the defender's advantage. Specifically, for $\pi_0 = 0.10$, when the defender employs LOF, the attacker's payoff increases from 34.12% to 41.13%, while under the perplexity-based anomaly detector it rises from 26.11% to 32.12%. For intermediate values of π_0, the QA module plays a more decisive role and can alter the outcome of the game. For instance, at $\pi_0 = 0.25$, the attacker's payoff increases from 41.33% to 54.76% against LOF, and from 44.65% to 56.52% when the perplexity-based detector is adopted. A similar trend is observed at $\pi_0 = 0.40$, where the attacker's payoff rises from 38.96% to 58.46% under LOF and from 51.55% to 64.29% under the perplexity-based detector. At higher values of π_0, the QA module substantially amplifies the attacker's advantage. In particular, for $\pi_0 = 0.55$, the attacker achieves a payoff of 70.68% against an LOF-based defender, compared to 49.69% in the absence of the QA module. When facing a perplexity-based detector, the attacker's payoff increases from 66.26% to 73.76%. The most pronounced advantage for the attacker is observed at $\pi_0 = 0.70$. In this setting, the attacker's payoff increases from 54.76% to 72.54% under LOF, while under the perplexity-based detector it rises from 80.12% to 85.81%, highlighting the strong impact of the QA module in regimes characterized by high prior probabilities.

6. Conclusions

In this work, we introduce a heuristic approach for generating high-quality textual adversarial samples. Owing to its highly modular and purely geometric design, our constraint approximation framework achieves a strong balance between efficiency and effectiveness, enabling the generation of adversarial examples that are both computationally efficient and highly impactfull. The experimental analysis demonstrates that attacks enhanced by our module achieve performance comparable to the original methods, incurring only a minor computational overhead, while producing samples that are more effective at bypassing the constraints imposed by common anomaly detection techniques. Additionally, we analyzed the competitive advantage gained by the attacker through the use of this technique using a game-theoretic framework.This analysis shows how the QA module can reverse the outcome of the attacker-defender interaction in favor of the attacker or amplify their advantage. Future research

could investigate integrating semantic- and context-aware constraints to enhance the naturalness and plausibility of generated samples. Another promising area for research is evaluating the proposed method against more advanced, adaptive anomaly detection systems, including those based on large language models. Finally, incorporating adaptive or learning-based constraint approximation strategies could strengthen the framework's robustness and versatility in dynamic attacker-defender scenarios.

Acknowledgments

Special thanks to Vincenzo Messina for his support in developing the preliminary evaluation code as part of his Master Thesis work.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] W. A. Woods, Transition network grammars for natural language analysis, Commun. ACM (1970).
[2] K. Sparck Jones, A statistical interpretation of term specificity and its application in retrieval, Taylor Graham Publishing, 1988.
[3] M. Ghaseminejad Raeini, The evolution of language models: From n-grams to llms, and beyond, Natural Language Processing Journal (2025).
[4] S. Qiu, Q. Liu, S. Zhou, W. Huang, Adversarial attack and defense technologies in natural language processing: A survey, Neurocomputing (2022).
[5] W. E. Zhang, Q. Z. Sheng, A. A. F. Alhazmi, C. Li, Generating textual adversarial examples for deep learning models: A survey, arXiv preprint arXiv:1901.06796 (2019).
[6] M. Alzantot, Y. Sharma, A. Elgohary, B.-J. Ho, M. Srivastava, K.-W. Chang, Generating natural language adversarial examples, arXiv preprint arXiv:1804.07998 (2018).
[7] R. Jia, A. Raghunathan, K. Göksel, P. Liang, Certified robustness to adversarial word substitutions, arXiv preprint arXiv:1909.00986 (2019).
[8] Y. Zang, F. Qi, C. Yang, Z. Liu, M. Zhang, Q. Liu, M. Sun, Word-level textual adversarial attacking as combinatorial optimization, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2020.
[9] M. T. Ribeiro, T. Wu, C. Guestrin, S. Singh, Beyond accuracy: Behavioral testing of NLP models with CheckList, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020.
[10] D. Jin, Z. Jin, J. T. Zhou, P. Szolovits, Is bert really robust? a strong baseline for natural language attack on text classification and entailment, in: Proceedings of the AAAI conference on artificial intelligence, 2020.
[11] S. Ren, Y. Deng, K. He, W. Che, Generating natural language adversarial examples through probability weighted word saliency, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2019.
[12] L. Li, R. Ma, Q. Guo, X. Xue, X. Qiu, BERT-ATTACK: Adversarial attack against BERT using BERT, in: B. Webber, T. Cohn, Y. He, Y. Liu (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2020.
[13] A. Pal, R. Vidal, A game theoretic analysis of additive adversarial attacks and defenses, in: Proceedings of the 34th International Conference on Neural Information Processing Systems, 2020.
[14] M.-F. Balcan, R. Pukdee, P. Ravikumar, H. Zhang, Nash equilibria and pitfalls of adversarial training in adversarial robustness games, in: International Conference on Artificial Intelligence and Statistics, 2023.

[15] L. Meunier, M. Scetbon, R. B. Pinot, J. Atif, Y. Chevaleyre, Mixed nash equilibria in the adversarial examples game, in: International Conference on Machine Learning, 2021.
[16] A. S. Chivukula, W. Liu, Adversarial deep learning models with multiple adversaries, IEEE Transactions on Knowledge and Data Engineering (2019).
[17] S. Zuo, C. Liang, H. Jiang, X. Liu, P. He, J. Gao, W. Chen, T. Zhao, Adversarial regularization as stackelberg game: An unrolled optimization approach, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2021.
[18] H. Zeng, B. Chen, K. Deng, A. Peng, Adversarial example detection bayesian game, in: 2023 IEEE International Conference on Image Processing (ICIP), 2023.
[19] P. Dasgupta, J. B. Collins, M. McCarrick, Playing to learn better: Repeated games for adversarial learning with multiple classifiers, 2020.
[20] B. Biggio, F. Roli, Wild patterns: Ten years after the rise of adversarial machine learning, in: Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, 2018.
[21] F. Pierazzi, F. Pendlebury, J. Cortellazzi, L. Cavallaro, Intriguing properties of adversarial ml attacks in the problem space, in: 2020 IEEE symposium on security and privacy (SP), 2020.
[22] A. Demontis, M. Melis, M. Pintor, M. Jagielski, B. Biggio, A. Oprea, C. Nita-Rotaru, F. Roli, Why do adversarial attacks transfer? explaining transferability of evasion and poisoning attacks, in: 28th USENIX security symposium (USENIX security 19), 2019.
[23] C. T. Do, N. H. Tran, C. Hong, C. A. Kamhoua, K. A. Kwiat, E. Blasch, S. Ren, N. Pissinou, S. S. Iyengar, Game theory for cyber security and privacy, ACM Computing Surveys (CSUR) (2017).
[24] R. Gibbons, Game theory for applied economists, Princeton University Press, 1992.
[25] J. Morris, E. Lifland, J. Y. Yoo, J. Grigsby, D. Jin, Y. Qi, Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, 2020.
[26] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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), 2019.
[27] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, V. Stoyanov, Roberta: A robustly optimized bert pretraining approach, arXiv preprint arXiv:1907.11692 (2019).
[28] X. Zhang, J. J. Zhao, Y. LeCun, Character-level convolutional networks for text classification, in: NIPS, 2015.
[29] B. Pang, L. Lee, Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales, in: Proceedings of the ACL, 2005.
[30] X. Zhang, J. Zhao, Y. LeCun, Character-level convolutional networks for text classification, Advances in neural information processing systems (2015).
[31] K. Papineni, S. Roukos, T. Ward, W.-J. Zhu, Bleu: a method for automatic evaluation of machine translation, in: Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002.
[32] C.-Y. Lin, Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, 2004.
[33] J. Pennington, R. Socher, C. D. Manning, Glove: Global vectors for word representation, in: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014.
[34] E. Reif, A. Yuan, M. Wattenberg, F. B. Viegas, A. Coenen, A. Pearce, B. Kim, Visualizing and measuring the geometry of bert, Advances in neural information processing systems (2019).
[35] Y. Li, D. McLean, Z. A. Bandar, J. D. O'shea, K. Crockett, Sentence similarity based on semantic nets and corpus statistics, IEEE transactions on knowledge and data engineering (2006).
[36] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing

and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, 2019.