Enhancing Robustness in Medical Question Answering Systems with Novel Defense Models against Adversarial Attacks

Atrab A. Abd El-Aziz[1], Reda A El-Khoribi[2], and Nour Eldeen Khalifa[2]*

[1]Department of Information Technology, Faculty of Computers and Information, KafrelSheikh University, Egypt
[2]Department of Information Technology, Faculty of Computers and Artificial Intelligence, Cairo University, Egypt

Abstract
Medical Question Answering (MQA) systems are essential for accurate medical diagnoses but face significant threats from adversarial attacks that manipulate input text and lead to potential misinterpretations and erroneous recommendations. Although extensive research has focused on defenses for medical images, there is a notable gap in protections for MQA systems. To the best of our knowledge, this paper is the first to address this gap by introducing three advanced defense models specifically designed for MQA systems. The proposed models target both word-level (synonym substitution, word deletion) and character-level (random character insertion) attacks on the pre-trained BERT model. The Synonym Substitution Embedding (SSE) Defense Framework is designed to counter word synonym substitution attacks using Term Frequency-Inverse Document Frequency (TFIDF) and pre-trained transformers for enhanced synonym embedding. The second model, (CosineDefender) utilizes cosine similarity to mitigate these adversarial perturbations. While the third model (JaccardDefender) employs Jaccard similarity for defense against the same attacks. Evaluation of these models is conducted on three datasets: two medical datasets (Symptom2Disease and Medical Symptoms Text and Audio Classification) and one natural language dataset (AGs News) for comparative purposes. Results show that the SSE model reduces the attack success rate from 8.7% to 0.4% on the AGs News dataset. For Symptom2Disease, attack success rates are high (10.2%, 12.8%, and 62%) for word synonym substitution, word deletion, and random character insertion, but CosineDefender lowers these rates to 3.4%, 4.3%, and 12.8%. JaccardDefender performs best, achieving the lowest attack success rates (3.4%, 3.5%, and 3.4%) and highest accuracy across datasets. These findings highlight the effectiveness of these models in improving MQA system resilience against adversarial threats.

Keywords:
Adversarial Attacks, BERT, Medical Question Answer (MQA), Term Frequency-Inverse Document Frequency (TFIDF).

1. Introduction

The integration of natural language processing (NLP) systems in healthcare has revolutionized MQA and diagnostic support. Nevertheless, the increasing reliance on these systems has rendered them susceptible to adversarial attacks. These attacks, previously explored in computer vision and now applicable to NLP, manipulate input data subtly to deceive deep learning models. Within the context of MQA, such attacks can introduce incorrect diagnoses, and raise ethical concerns [1]. Textual data must adhere to multiple properties including grammatical, lexical, and semantic constraints. As a result, numerous efficient adversarial image attack techniques such as gradient-based methods cannot be easily transferred to text data. This is due to the risk of generating incorrect characters and non-existent terms.

Recent investigations have illuminated the susceptibility of medical NLP models to adversarial attacks. Tactics like synonym substitution, character insertions, and deletions can significantly alter model outputs while maintaining linguistic coherence. Consequently, a demand for robust defense mechanisms has arisen to ensure the precision and security of medical question-answering systems [2].

This paper highlights the adversarial attacks targeting DL-based MQA and examines their potential ramifications on patient care and clinical decision-making. Drawing inspiration from the broader realm of adversarial DL and the distinctive challenges presented by healthcare applications. This paper introduces three new defense frameworks tailored to mitigate the impact of adversarial attacks. The SSE defense model against word substitution attacks is designed with a focus on applications such as news and MQA Bert-based text classification. This model dynamically enhancing the robustness of deep learning-powered text classifiers.

Our method's effectiveness is demonstrated through experiments utilizing a BERT pre-trained classification model and the widely recognized AG's news dataset, a common benchmark for text classification. The results indicate that our approach surpasses existing word substitution adversarial defense methods in terms of both attack success rate and model accuracy. Furthermore, we assess the SSE model's performance using two different MQA datasets. The outcomes of the experiments reveal not only its efficacy on news-related data but also its transferability to medical datasets.

In the alternate two defense models, protection against three variations of character and word-level attacks is accomplished through the utilization of Cosine similarity and Jaccard similarity techniques. These metrics facilitate the identification of the most comparable attributes between the initial and adversarial instances. Similarly, these models

Manuscript received April 5, 2026
Manuscript revised April 20, 2026
https://doi.org/10.22937/IJCSNS.2026.26.4.2

are implemented on the previously mentioned three datasets, encompassing three distinct types of attacks. The study also delves into the limitations of the proposed models and their impact on transformer models.

Novel Contributions:

1. **Development of Defense Models for MQA Systems:** This paper introduces three innovative defense models specifically designed to counter adversarial attacks in MQA systems. These models address both word-level and character-level perturbations, filling a significant gap in existing research on MQA system defenses.

2. **Synonym Substitution Embedding (SSE) Defense Framework:** The SSE model employs Term Frequency-Inverse Document Frequency (TFIDF) and pre-trained transformers to refine synonym embeddings by effectively defending against word synonym substitution attacks.

3. **CosineDefender and JaccardDefender Models:** These models utilize cosine and Jaccard similarities, respectively, to enhance robustness against the three mentioned adversarial attacks. while JaccardDefender demonstrates superior performance with the lowest attack success rates across datasets.

4. **Comprehensive Evaluation on Diverse Datasets:** The proposed models are rigorously evaluated on three diverse datasets including medical and natural language datasets and provided a comparative analysis of their effectiveness.

5. **Highlighting a Research Gap:** This paper addresses a critical gap in the current literature by focusing on MQA systems which have been overlooked in previous adversarial defense research. To our knowledge, this is the first work to tackle adversarial attacks specifically in the context of MQA systems.

This paper is organized as follows: Section 2 presents a comprehensive literature review and summarizes relevant research supporting the proposed approach. Section 3 introduces the proposed models and outlines the methodology in detail. Section 4 describes the experimental setup and evaluates the results, followed by an in-depth discussion of the findings in Section 5. Finally, Section 6 concludes the paper with key insights and potential future work.

2. Literature Reviews

Adversarial NLP text-based attacks and defenses have emerged as dynamic areas of research in recent times. Within the domain of medical text, numerous tasks are encountering the risk of adversarial attacks. For instance, tasks like machine translation, text classification, medical question and answer (MQA), and others. These models are particularly susceptible to malicious adversaries. The initial focus of this section is on addressing the issue of adversarial attacks and their corresponding defense mechanisms in the context of text classification tasks. Subsequently, an initial overview of attack models is explained, specifically delving into the realm of several commonplace word-level synonym adversarial attack strategies. Figure 1 shows a medical scenario for a text adversarial attack.

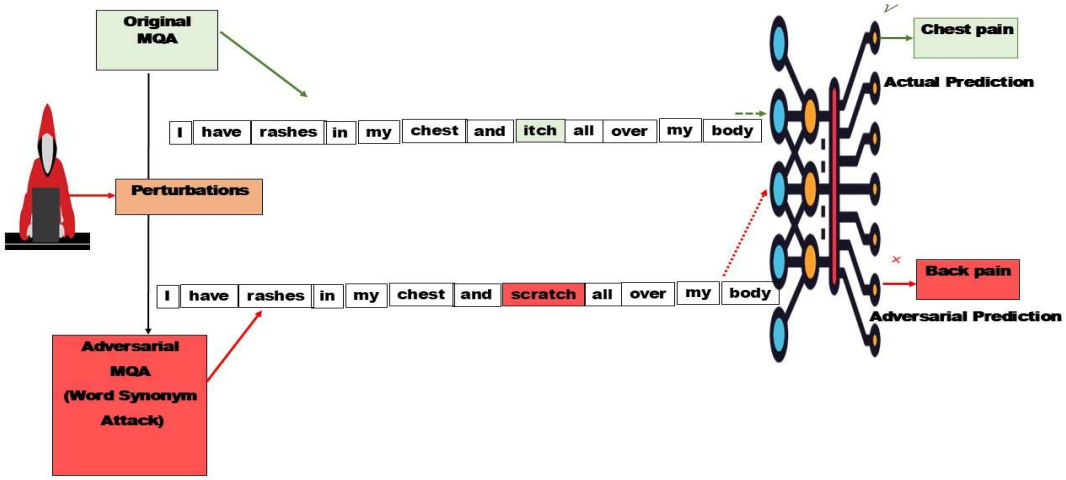


Figure 1. Medical Question-Answer Adversarial Attack Scenario

Definition of the Problem: Considering a text classifier denoted as C: X → Y, where X represents the input space and Y signifies the output space, let's assume there exists an input text denoted as x ∈ X. With this setup, the classifier is capable of generating a predicted true label y_{true} through a posterior probability denoted as P [3] as illustrated in Equation (1).

$$argmax_{y_i \in y} P((y_i|x) = y_{true} \quad (1)$$

Adversarial attack Definition: is the adversary's capability to generate an adversarial example x' by incorporating a perturbation $\Delta\chi$ that is imperceptible to human observation on the classifier C as illustrated in Equation (2).

$$argmax_{y_i \in y} P((y_i|x') \neq y_{true}$$
$$x' = x + \Delta x, \|\Delta x\| p < \in \quad (2)$$

Here, ∈ is a parameter that regulates the magnitude of the small perturbation in a way that ensures the crafted example remains imperceptible to human senses. The notation $\|\Delta x\| p$ represents the p-norm.

1.1 2.1 Classification of Text Adversarial Attacks

Given that textual data varies from data in image or audio domains, attack types also differ. Depending on the components altered within the text, adversarial attack techniques can be categorized into four distinct types: character-level, word-level, sentence-level, and multi-level attacks. In these attack categories, manipulations typically involve the insertion, removal, swapping/replacement, or flipping of text data. However, it's important to note that not all of these options are necessarily employed across different levels of attacks [4]. Current advanced adversarial attacks on text classification can be classified into the following categories:

A. **Character-level Attack:** In this attack, modifications are made to individual characters within the text. This can involve altering characters by replacing them with new characters, special symbols, or numbers. The attack techniques include adding new characters to the text, swapping characters with neighboring ones, removing characters from words, or flipping them. These manipulations aim to create subtle changes in the text while attempting to maintain its overall structure and coherence. Gao et al. [5] also introduced Deep-WordBug, a technique that introduces minor character perturbations to create adversarial examples by creating typos and grammatical inconsistencies in the sentence against DNN classifiers. Ebrahimi et al. [6] proposed an efficient method named Hotflip, which generates white-box adversarial texts to deceive character-level neural networks. Additionally, text adversarial samples were generated in both white-box and black-box scenarios. However, these approaches are susceptible to defense by incorporating a word recognition model before inputting data into the neural network.

- **Random Character Insertion (RCI) Attack:** this technique involves introducing random characters into an input text. The objective here is to disturb the model's comprehension of the text's meaning while keeping the text's overall structure and coherence intact. The attacker initiates by selecting a clean input text for modification. The attacker identifies random positions within the input text and inserts arbitrary characters. These characters can include letters, digits, symbols, or a combination of these elements. The inserted characters are intended to interfere with the semantic coherence and contextual flow of the text. Nonetheless, the attacker takes care to ensure that the inserted characters do not render the text conspicuously altered or nonsensical. the altered text is fed into the NLP model. The anticipation is that the model's interpretation of the text is altered sufficiently to lead to misclassification or inaccurate outcomes.

Defenses against this type of attack have included methods such as spell checkers [7,8]. However, these same defenses prove to be particularly susceptible to word-level attacks that maintain language coherence. Against syntactically accurate attacks, Dirichlet Neighborhood Ensemble (DNE) [9], effective strategies encompass adversarial training (AT) [10], Adversarial Sparse Convex Combination (ASCC) [11], and Synonym Encoding Method (SEM) [12]. The first three methods utilize some form of data augmentation by training the model on perturbed samples. Conversely, the last approach introduces an encoding step before the input layer of the target model and trains it to eliminate potential perturbations. Moreover, there are methods for detecting adversarial inputs. In contrast to other defense strategies, these methods possess the capability to explicitly identify manipulated inputs and

generate alert signals. there are two available methods: learning to DIScriminate Perturbation (DISP) [13] and Frequency-Guided Word Substitution (FGWS) [14]. The former approach leverages the frequency characteristics of adversarial words and represents the latest and most accurate technique in this regard. Beyond the security concerns highlighted earlier and the methods for text adversarial attacks, it's important to acknowledge that the field of healthcare security encompasses a broader range of issues. Numerous challenges to healthcare security, adversarial attacks, and defense strategies have been explored and put forth as well.

The integration of NLP systems in healthcare has brought about transformative advancements, yet the regulatory landscape and industry standards governing their security and robustness remain largely undefined. This lack of specificity poses challenges in ensuring the consistent and reliable performance of these systems, particularly in the face of adversarial attacks. Establishing clear guidelines and benchmarks for security measures, testing protocols, and model validation would be instrumental in fostering trust and mitigating potential risks associated with the deployment of NLP technologies in critical healthcare applications.

Traditional static defense mechanisms often rely on fixed rules or predetermined patterns to identify and mitigate attacks, making them vulnerable to adaptive adversaries who can easily circumvent these defenses. In contrast, the SSE defense model employs a dynamic approach by leveraging contextual information and semantic relationships between words to detect and neutralize adversarial perturbations. This adaptability allows the SSE model to effectively counter a wider range of attacks, including those that may not conform to predefined patterns.

B. **Word-Level Attack:** involve the substitution of words within original texts using synonyms, antonyms, or by simulating typing errors. Alternatively, words might be entirely removed from the text to create variations. This is a strategy employed to maintain semantic coherence while altering the content. Liang et al. [15] present a technique that involves identifying suitable terms for insertion, replacement, and deletion based on the calculation of the most substantial gradient magnitude of the cost function and the word frequency. However, their approach necessitates a notable degree of human involvement in the creation of adversarial instances. In order to sustain semantic consistency and minimize the likelihood of human detection, their method demands manual efforts. Word-level attacks pose a greater challenge to detect due to their ability to preserve the semantic meaning and grammatical correctness of the original text. Synonym substitutions, in particular, can seamlessly replace words while maintaining contextual coherence, making the attack less conspicuous. The vast number of potential synonyms further expands the attack space, making it difficult to anticipate and defend against all possible variations. Additionally, these attacks can be effectively executed in black-box scenarios, where the attacker lacks knowledge of the target model's internal workings, enhancing their versatility and applicability. For examples,

- **Word Synonym Substitution (WSS) Attack:** this involves the replacement of words in a sentence with their synonyms, to retain the overall meaning and context of the text. Initially, a system must identify suitable synonyms for the words present in the input text. This task can be accomplished using resources such as WordNet or pre-trained word embeddings. The selected synonyms should ensure that the intended meaning and syntactic structure of the sentence are preserved. Furthermore, the substituted word should seamlessly integrate into the sentence, maintaining coherence and readability. Adversarial examples generated through word synonym substitution are designed to deceive machine learning models. To assess the efficacy of this technique, one can measure its impact on model performance and evaluate how well the substituted text maintains the original meaning and context.

- **Random Word Deletion (RWD) Attack**: is a method that involves the removal of words from an input text in a randomized manner, aiming to disrupt the model's comprehension of the text's meaning while striving to uphold its grammatical structure. The primary aim of this attack is to introduce alterations to the input while maintaining an appearance of innocence. The process initiates with a clean input text that they intend to manipulate. This input could encompass a sentence, paragraph, or an extended piece of text. The attacker randomly eliminates words from the input. This random selection contributes to unpredictability and minimizes the chances of detection. The challenge for the attacker lies in perturbing the semantic flow of the text while adhering to grammatical correctness, thereby reducing suspicion. Following the deletion of

words, the modified text is inputted into the NLP model. The expectation is that the model's interpretation of the text is distorted sufficiently to lead to misclassification or inaccurate results. The success of the attack is determined by whether the model produces an output that deviates from the desired outcome.

As a result, other papers primarily focus on the approach of word substitution to achieve automated generation. The pivotal distinction among these subsequent methodologies lies in their methods of generating alternative words. Samanta et al. [16] proposed the construction of a candidate pool containing synonyms employing the Fast Gradient Sign Method (FGSM), genre-specific keywords, and typographical errors. In contrast, Papernot et al. [17] perturb a word vector by computing its forward derivative and subsequently mapping this perturbed word vector to the nearest word within the word embedding space.

The adversarial attack technique "DeepWordBug" introduced in [5] also tested word-level attacks for crafting adversarial instances. It relies on a scoring strategy to identify and modify the most significant words, leading to substantial changes in the classification outcome. This method successfully manipulated classification results to a considerable degree. Building upon the synonym substitution technique, Ren, et al. [18] introduced a greedy algorithm called PWWS, designed specifically for text adversarial attacks. The method is devised to perturb an initial text example into an adversarial one. They introduced a novel word replacement sequence that takes into account both the saliency of words and the classification probability. By ensuring that the altered example retains a semantic similarity to the original, it becomes challenging for humans to detect any anomalies in the modified text. This algorithm focuses on word-level adversarial examples, which tend to be less noticeable to humans and present greater challenges for deep neural networks (DNNs) to counteract or defend against.

Yang et al. [19] introduce two distinct techniques: the Greedy Attack, which relies on perturbation, and the Gumbel Attack, which is built upon scalable learning. To reinstate the interpretability of adversarial attacks utilizing the word substitution approach, Sato et al. [20] confine the direction of perturbations to existing words within the input embedding space. In [21], the susceptibility of Deep Learning-based Text Understanding techniques to adversarial text attacks is thoroughly examined. The authors devised a comprehensive attack framework, TextBugger, to create adversarial texts. Within this framework, they adopted distinct approaches for character-level and word-level perturbations. For character-level perturbation, their method involves introducing misspelled versions of significant words, leading to efficient misclassification of models; however, these misspellings are easily detectable. Conversely, for word-level perturbation, they opted to select substitute words from the word embedding space, utilizing the GloVe model [22] as the basis for their choice.

Defenses against word-level text adversarial attacks have seen limited research activity. As far as our knowledge extends, the work of [23] stands out as the sole attempt at countering attacks based on synonym substitution. They introduced the Synonym Encoding Method (SEM), which involves encoding synonyms using identical word embeddings to counteract adversarial perturbations. Nevertheless, SEM requires an additional encoding step before regular training and is constrained by fixed synonym substitution. Our framework, in contrast, employs a unified training approach and offers a versatile synonym substitution encoding strategy.

The distinction between the random word substitution attack and other adversarial attacks lies in its specific method of manipulating text input. While other attacks might involve character-level changes, sentence insertions, or combinations of different techniques, random word substitution focuses solely on replacing words within the original text with their synonyms. This targeted approach aims to maintain the overall meaning and grammatical structure of the sentence while subtly altering the content to mislead machine learning models. The random nature of the substitution adds an element of unpredictability, making it harder for defense mechanisms to anticipate and counteract the attack.

C. **Sentence-Level Attack:** adversarial examples are created by inserting entirely new sentences. While other approaches have been less explored, this method focuses on introducing new sentence structures to manipulate the model's predictions.

Cao et al. [24] introduced the black-box adversarial attack named Twin Answer Sentences Attack (TASA). This attack designed to alter the context of a question without compromising its fluency or the accuracy of the correct answer. TASA identifies the relevant answer sentence in the context and generates two modified sentences by replacing the key terms with synonyms to exploit biases in question-answering models. This lead to produce a misleading answer that

directs the model toward an incorrect response by introducing irrelevant entities. In [25], two techniques are proposed for generating adversarial examples targeting Math Word Problem (MWP) solvers. The "Question Reordering" is the first approach involves rearranging the question portion to appear at the beginning of the problem text. The "Sentence Paraphrasing" is the second technique focuses on rephrasing each sentence in the problem while preserving both its semantic meaning and numerical information.

D. **Multi-Level Attack:** encompasses a combination of character, word, and sentence-level techniques. These attacks leverage various levels of manipulation to create more complex and impactful adversarial examples.

Authors in [26] introduced the Visuo-Adaptive DualStrike (VADS) attack which is a novel method that combines transfer-based and query-based strategies to exploit vulnerabilities in Visual Question Answering (VQA) systems. VADS employs a momentum-like ensemble method to identify potential attack targets and compress perturbations. It then uses a query-based strategy to dynamically adjust the perturbation weights for each surrogate model. Evaluation across two datasets demonstrated that VADS surpasses existing adversarial techniques in both efficiency and success rate. In another study, authors in [27] are the first to investigate and successfully execute attacks on a multilingual Question Answering (MLQA) system pre-trained with multilingual BERT using various adversarial strategies. Their approach demonstrates that these attacks can degrade system performance by up to 85%. They reveal that the model exhibits a preference for English and the language of the question, often neglecting other languages present in the QA pair. Additionally, the authors show that incorporating these attack strategies during the training phase can help mitigate the impact of such adversarial attacks.

2.2 Defense Techniques against Adversarial Text

Research on text adversarial defense techniques has predominantly explored three primary approaches [28]:

A. **Adversarial Training:** In these strategies, the model's training process is altered to acquire robust features and heightened resilience against adversarial attacks. During testing, inputs are also adjusted to prevent the introduction of adversarial perturbations. A significant challenge in adversarial training arises from the requirement to be aware of various attack strategies during the training process. The limitation stems from the fact that adversaries typically do not disclose their attack techniques, rendering adversarial training constrained by the user's awareness. If a user attempts to incorporate adversarial training to counteract all known attacks within their knowledge, the resulting model's capability to carry out accurate classification could be severely compromised. This is due to the model acquiring minimal information about the genuine data, which ultimately hampers its classification performance. Adversarial training, while enhancing robustness against attacks, can sometimes negatively impact the model's ability to classify clean, authentic data. This is because the model learns to focus on features that distinguish adversarial examples from clean ones, potentially overlooking subtle nuances important for accurate classification of real-world data.

B. **Modifying Networks:** This approach involves enhancing the model's architecture by incorporating additional layers, and sub-networks, or modifying loss and activation functions to bolster its defensive capabilities.

C. **Network Add-on:** External networks are integrated into the system as supplementary components for classifying previously unseen data, thus augmenting the defense mechanisms against adversarial inputs.

A defense model is considered successful when the generated adversarial example x' fails to deceive the classifier C∗ or given an input text example x, the attacker is unable to create an adversarial example x' . In this paper, we introduce many defense models against three types of natural and medical text attacks depending on the modifying network's direction which involves enhancing the model's architecture by adding some preprocessing steps in the test phase to reduce the attack impact on the clean model.

The remainder of the paper will introduce the proposed models and provide a detailed explanation of the methodology in Section 3. Additionally, the experimental setup and results will be presented in Section 4. Section 5 offers an in-depth discussion of the findings. Finally,

Section 6 concludes the paper with key insights and suggestions for future work.

3. Proposed Frameworks

The proposed models defend against text adversarial attacks by examining changes in the classification outcome once the text modification module is applied. This method operates under the assumption that the intended target of potential attacks is a BERT model which is the cutting-edge model for text recognition.

1.1 First: Synonym Substitution Embedding (SSE) Defense Framework

Synonym Replacement is a defense mechanism that entails the identification of words within the input text that can be substituted with synonyms, maintaining the text's overall meaning. Synonyms are words that possess similar meanings but may exhibit distinct linguistic forms. For instance, in the sentence "The weather is nice," the word "nice" could be substituted with "pleasant," all the while preserving the sentence's intended meaning. Figure 2 shows the overall processes of the SSE defense model. This defense mechanism is implemented through a structured seven-steps process, as outlined below:

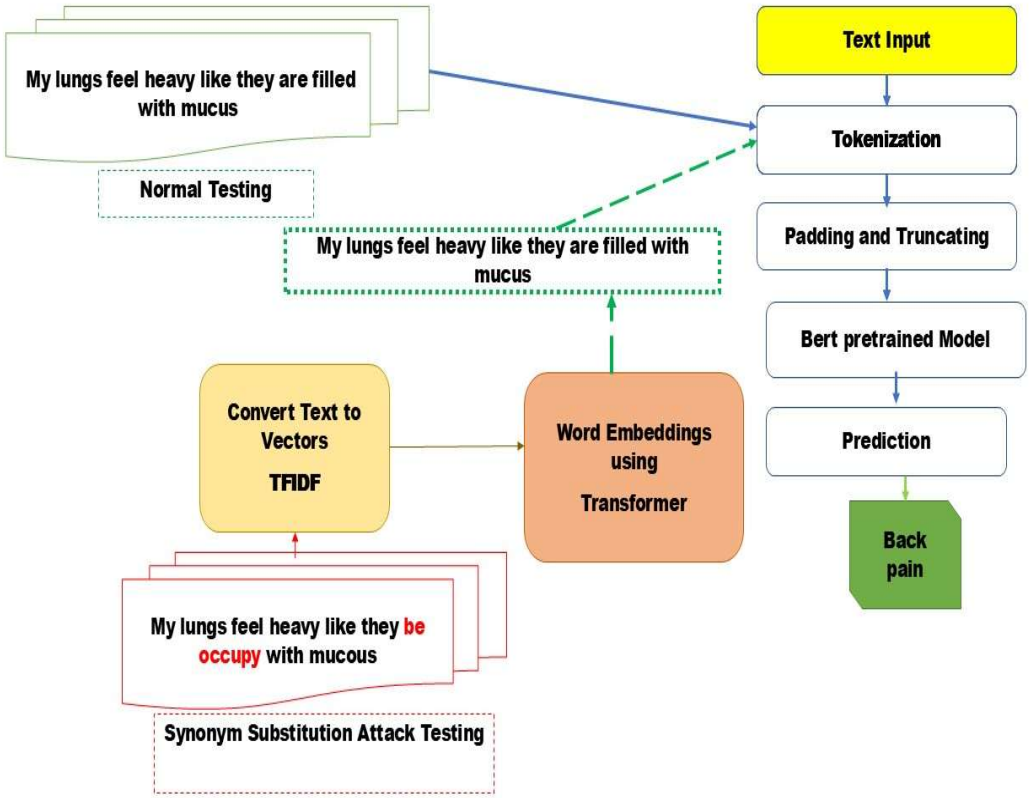


Figure 2. SSE proposed model structure against word level attack

1. **Preprocessing:** The three datasets undergo a sequence of preprocessing steps designed to eliminate inconsistent, missing, and redundant values. The text in all datasets is converted to lowercase. Next, tokenization is carried out, involving the segmentation of the provided text into distinct linguistic units known as tokens. Following this, padding is applied by inserting specific tokens, often represented by a PAD token, at the end of sequences to ensure consistent lengths. Lastly, text truncation is applied, which involves removing tokens from sequences that exceed a predetermined maximum length. To ensure uniformity, the maximum sequence length is determined based on the longest text in the dataset. Questions with shorter lengths are padded with zeros to align them appropriately. In the realm of deep neural models, we investigate BERT as a state-of-the-art model which is an attention mechanism that captures contextual relationships between words in a sentence. It is designed for text classification, accommodating both word-level and character-level data processing.
2. **TF-IDF** is an acronym for Term Frequency-Inverse Document Frequency, encompassing two interconnected metrics for gauging the relevance of a word within a document. Each word is assigned distinct Term Frequency and Inverse Document Frequency values. The TF*IDF score results from the multiplication of these individual weights. A higher TF*IDF score signifies infrequent occurrence. Specifically, TF corresponds to Term Frequency, reflecting how often a term appears in a document, while IDF stands for Inverse Document Frequency, indicating the significance of the term across the entire collection of documents [29].
3.

$$TF = \frac{Number\ of\ times\ this\ word\ ocures}{Number\ of\ words\ in\ sentence}$$

$$IDF = \log \frac{Number\ of\ Sentences}{Number\ of\ sentences\ where\ this\ word\ occured}$$

4. **Sentence Transformer:** the pre-trained MiniLM-L6-v2 [30] sentence transformer model is employed to extract N-gram key-phrases from real-time datasets. It is designed to encode sentences or text passages into fixed-dimensional vectors that capture semantic information. It has a size of 80MB and consists of 384 hidden layers. This model is known for its efficient encoding speed, capable of processing 14,200 sentences per second on a V100 Graphics Processing Unit (GPU).
5. **Fusion of Sentence Transformers and TF-IDF**: The combined use of sentence transformers and TF-IDF enhanced the embeddings' ability to capture semantic meaning and word importance. This integrated approach results in more accurate semantic similarity assessments and improved defense against adversarial attacks.
6. **Deep Neural BERT Model:** The BERT-Base-Uncased model is fine-tuned using the Adam optimizer, and different learning rates are tested over 3 epochs. Categorical cross-entropy is employed as the loss function to minimize during the training process. The parameters of our BERT-Base-Uncased model are evaluated throughout the training phase. The optimal model is identified when the validation loss is minimized, achieved through adjusting hyperparameters. Notably, altering hyperparameters significantly influences the model's performance. A learning rate of 2e-5 yields superior results in comparison to other learning rates [31].
7. **Adversarial Text Generation:** This step involves the creation of adversarial text examples using the word synonym substitution attack. This process modifies the original text to introduce perturbations aimed

at challenging the model's robustness. The objective is to generate adversarial examples that test the resilience of the model against various forms of textual manipulation. Once the adversarial text is generated, it undergoes transformation into numerical vectors using the Term Frequency-Inverse Document Frequency (TF-IDF) method. TF-IDF serves to identify the most salient words within the text by quantifying their importance based on their frequency of occurrence.

8. **Similarity Calculation Using Sentence Transformer:** In the final step, a sentence transformer model is employed to compute semantic similarity between words using Euclidean distance. The pre-trained MiniLM-L6-v2 model is used to encode text into fixed-dimensional vectors that capture the semantic meaning of sentences. The model's efficiency in processing, with a capacity to handle 14,200 sentences per second on a V100 GPU, allows for rapid encoding and comparison. The MiniLM-L6-v2 model with 384 hidden layers and a size of 80MB generates dense vector embeddings that encapsulate the semantic content of the sentences. This approach enhances the model's ability to detect and counteract word substitution attacks by evaluating the contextual similarity of words.

The fusion of Sentence Transformers and TF-IDF results in enhanced embeddings and improved performance when assessing semantic similarity. Transformers are employed to create dense vector representations, or embeddings, for sentences. These embeddings effectively encapsulate the semantic meaning of sentences within a continuous vector space. This integrated approach yields enhanced embeddings that not only capture the semantic intricacies of sentences but also the importance of individual words within those sentences.

1.2 Second: CosineDefender and JaccardDefinder Defense Frameworks

The objectives of both CosineDefender and JaccardDefinder defense techniques are to introduce controlled alterations to input text, creating difficulties for adversarial attacks to influence the model's predictions. Nevertheless, maintaining a careful equilibrium in the amount of noise added is crucial to prevent adverse effects on the model's accuracy with clean data. Furthermore, the efficiency of these defense models can differ depending on the particular NLP architecture, the characteristics of the attacks, and the extent of robustness testing they undergo. These two models protect against three distinct types of attacks. Among these, two are at the word level which involving attacks such as word substitution and random word deletion. The third attack is conducted at the character level and is known as Noise Injection. As shown in Figure 3, This defense model shares similarities with the first model in terms of its clean model training structure. However, there are two notable distinctions: firstly, this framework is designed to defend against three distinct types of attacks. Secondly, the defense techniques employed have been altered, as two additional methods cosine similarity and Jaccard similarity are tested as defense mechanisms against these three attack types. The CosineDefender and JaccardDefender frameworks share the same initial preprocessing and BERT model steps as the SSE Defense model. However, instead of using TF-IDF and Sentence Transformers, these frameworks utilize Cosine Similarity and Jaccard Similarity to detect adversarial perturbations and defend against adversarial attacks. Below is a detailed description of each step involved: As with the SSE Defense model, the preprocessing and BERT stages remain consistent:

1. **Preprocessing** includes converting text to lowercase, tokenizing it, applying padding, and truncating sequences. These steps standardize the input for model processing.
2. **BERT Model:** The tokenized text is passed through the BERT model which generates contextual embeddings for each token in the input sequence. BERT's bidirectional attention mechanism helps capture semantic relationships between words and outputs dense vector representations of the text.

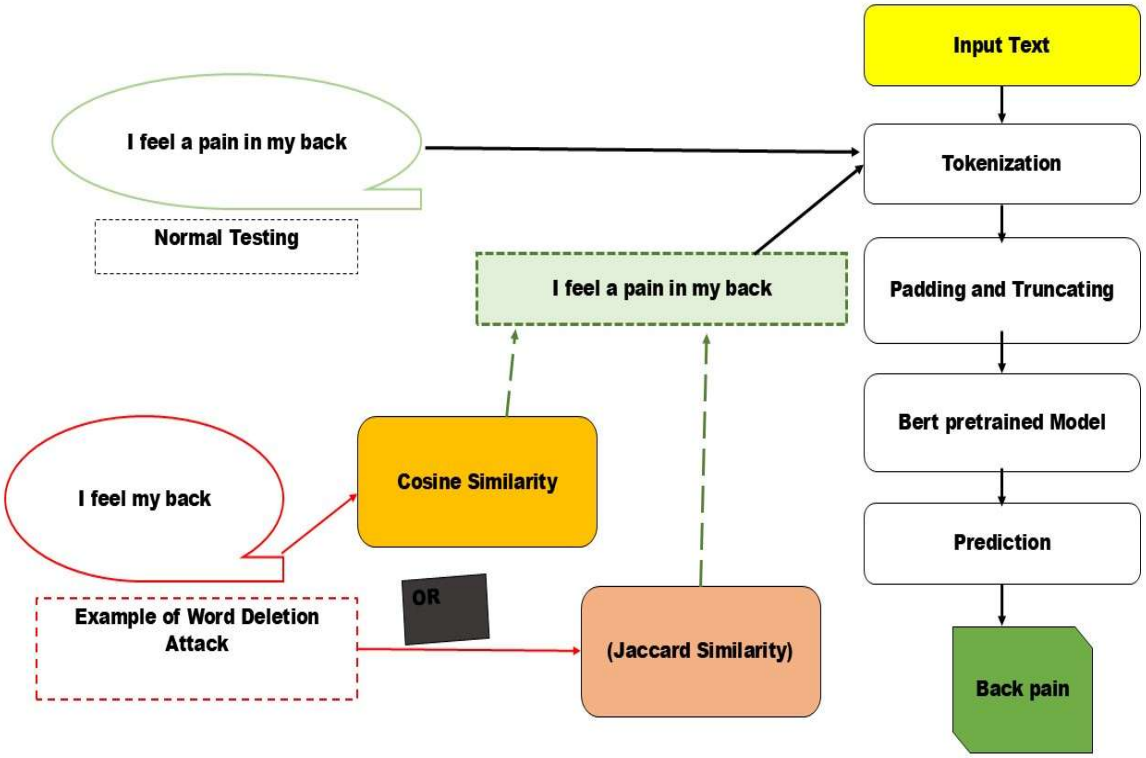


Figure 3. The proposed JaccardDefender and CosineDefender models against word deletion, substituation and character insertion attacks

3. **Cosine similarity**: In this framework, we replace TF-IDF and Sentence Transformers with Cosine Similarity to measure the semantic similarity between the original input and its adversarially perturbed version. Cosine Similarity is a prevalent measurement used in the field of NLP to gauge the similarity between two vectors. Its applicability extends to the realm of adversarial defense, particularly in assessing semantic similarity. This metric operates by computing the cosine of the angle formed between vectors, which can represent embeddings of words, phrases, or even entire documents within NLP. When the cosine similarity between two vectors is high, it signifies that these vectors align closely in direction, implying shared semantic meanings. The importance of cosine similarity stems from its capability to reveal the semantic connections present in textual elements. Particularly in the context of adversarial attacks, where alterations are frequently inconspicuous yet uphold semantic consistency, cosine similarity assumes a pivotal function. Through the computation of cosine similarity between the initial input and its perturbed version, it becomes viable to measure the extent of semantic diversion between these instances. A significant reduction in cosine similarity might signal the potential occurrence of an adversarial attack [29].

$$cosine\_similarity(A,B) = \frac{A.B}{\|A\|*\|B\|} \qquad (3)$$

Where A · B represents the dot product of vectors A and B. ||A|| and ||B|| represent the magnitudes

(Euclidean norms) of vectors A and B. The values of cosine similarity fall within the range of -1 to 1. A score of 1 indicates perfect similarity, 0 indicates no similarity, and -1 signifies perfect dissimilarity.

In summary, cosine similarity proves to be a versatile tool, proficient in quantifying semantic similarity and adept at detecting semantic alterations induced by adversarial perturbations.

4. **Jaccard similarity:** The JaccardDefender defense model replaces the TF-IDF and Sentence Transformers with Jaccard Similarity. Jaccard Similarity is another widely used metric in NLP, for quantifying the similarity between sets. Jaccard similarity can play a role in defending against adversarial attacks. Like cosine similarity, Jaccard similarity can be harnessed to detect changes in the semantic distance between the original input and its perturbed counterpart caused by adversarial attacks. A significant drop in Jaccard similarity between the perturbed gradient and the original gradient could indicate the presence of adversarial modifications. By setting a Jaccard similarity threshold, it becomes feasible to identify highly similar text between adversarial input and the testing data. Text inputs exceeding the threshold might be potentially recoverable and subjected to additional analysis. The Jaccard similarity coefficient is defined as the size of the intersection of the sets divided by the size of the union of the sets [33]. Mathematically, it can be expressed as:

$$J(A,B) = |A \cap B| / |A \cup B| \quad (4)$$

The Jaccard similarity coefficient, denoted as J (A, B), quantifies the similarity between sets A and B. The size of the intersection of sets A and B is represented by |A ∩ B|, while the size of their union is represented by |A ∪ B|. To summarize, Jaccard similarity plays a role in identifying and assessing adversarial examples. Its effective application necessitates its incorporation into a comprehensive defense strategy.

5. Fusion of BERT with Cosine and Jaccard Similarity

Both the CosineDefender and JaccardDefender defense models combine the BERT-generated embeddings with the respective similarity metrics (Cosine or Jaccard) to assess and counter adversarial attacks. The fusion of deep embeddings from BERT and similarity-based approaches allows the system to capture both contextual and token-level nuances to improve its resilience against a wide range of adversarial manipulations.

4. Experiment Results

In this section, we conduct experimental assessments to gauge the efficacy of the proposed models. We initiate by outlining the experiment's configuration and subsequently provide an account of the outcomes achieved across three distinct real-world datasets. The results illustrate that models operating within the proposed defense models exhibit significantly improved performance in defending against adversarial examples.

4.1 Datasets

We evaluate the performance of our proposed defense models using three established benchmark datasets: AG's News, Symptom2Disease, and Medical Symptoms Text and Audio Classification (MSTAC).

AG's News dataset [34] is collected from news articles. It utilizes solely the title and description fields. It includes four classes of samples world, sports, business, and Sci/Tec. Each class comprises 30,000 training samples and 1,900 testing examples. For our subsequent experiments, we conducted many experiments such as using all training and test data as provided by the dataset. Then, we selected a subset of the original data. Specifically, we sampled 1,000 and 2,000 testing examples from the original pool of data. The AG's News dataset was chosen as a benchmark due to its widespread recognition and established use in evaluating text classification models. Its balanced distribution of classes and substantial size make it a reliable and comprehensive dataset for assessing the performance and robustness of our proposed defense models in a general text classification setting.

Symptom2Disease medical dataset [35] The dataset consists of information about 24 distinct diseases, with each disease being associated with 50 symptom descriptions. This accumulates to a total of 1,200 data points. For our experiments, we split the dataset into 70% of 840 training samples,120 validation samples, and 240 testing samples.

Medical Symptoms Text and Audio Classification [36]: this dataset encompasses numerous audio transcriptions of

prevalent medical symptoms such as "knee pain" or "headache". It consists of information about 25 distinct diseases. This dataset contains a total of 6,661 data points. This dataset holds the potential for training agents within the medical domain. For our experiments, we split the dataset into 70% of 4,662 training samples,666 validation samples, and 1,333 testing samples.

4.2 Performance metrics

The primary metrics utilized to assess the effectiveness of various defense frameworks in this study include [37]:

Accuracy: the ratio of correctly predicted samples to the total number of testing samples.

Accuracy Shift (AS): the reduction in accuracy observed before and after an adversarial attack.

Attack-Success Rate (ASR): dividing the number of examples successfully manipulated by attack models against the number of examples that were initially correctly predicted with no attack. Attack-Success Rate is calculated using the following formula:

$$\text{ASR}= \frac{Accuracy\ Shift}{Number\ of\ Correctly\ Predicted\ Examples\ with\ No\ Attack} \quad (5)$$

Indeed, a stronger defense performance by the target model leads to a lower Attack-Success Rate for the attacker. In other words, as the defense becomes more effective, the attacker's success in manipulating examples decreases.

4.3 Results

In order to showcase the efficacy of the proposed models, we utilized the AG's News dataset as a benchmark for validation and comparative analysis.

A. Results of SSE Model

Table 1 outlines the performance metrics for the SSE Defense Model. The table shows the accuracy of the baseline pretrained BERT model without any attack, the accuracy after a random word substitution attack, the accuracy_shift due to the attack, the attack success rate and the accuracy of the model after applying the SSE defense model. The model is evaluated across three datasets: AG's News, MSTI, and Symptom2Disease. From the results, we can observe that the SSE Defense Model demonstrates a significant improvement in mitigating the impact of adversarial attacks. Below are key observations based on the metrics:

AG's News Dataset :The attack caused a notable accuracy shift from 92.8% to 84.7%, resulting in an attack success rate of 8.1% .After applying the SSE defense, the accuracy rebounded to 92.4%, showing the model's ability to recover from adversarial attacks effectively.

MSTI Dataset :The attack success rate was higher for the MSTI dataset with a 17.6% drop in accuracy (from 99.6% to 82%) .The SSE model successfully reduced this drop, restoring the accuracy to 98.1% which demonstrating strong defense against adversarial attacks in medical datasets.

Symptom2Disease Dataset :A similar trend is observed in this dataset, with the attack lowering accuracy by 10%. The SSE defense restored accuracy to 97.2% closely matching the no-attack accuracy of 97.5%.

Figure 4 visualizes the performance of the AG's News dataset under attack and after applying the SSE defense model. It highlights the dramatic reduction in attack success rate from 8.1% to 0.4%, showcasing the robustness of the SSE defense in restoring model performance post-attack.

These results validate that the SSE model is highly effective in mitigating adversarial attacks across all tested datasets. The model consistently reduces the impact of the attack by restoring accuracy to near-original levels and significantly lowering the attack success rate.

Table 1. Results of the SSE Defense Model against Random Word Substitution Attack

Dataset	No-Attack Accuracy	After Attack Accuracy	Accuracy Shift	Attack Success Rate	SSE Defense model
AG's News	92.8%	84.7%	8.1	8.7%	92.4%
MSTI	99.6%	82%	17.6	17.6%	98.1%
Symptom2Disease	97.5%	87.5%	10	10.2%	97.2%

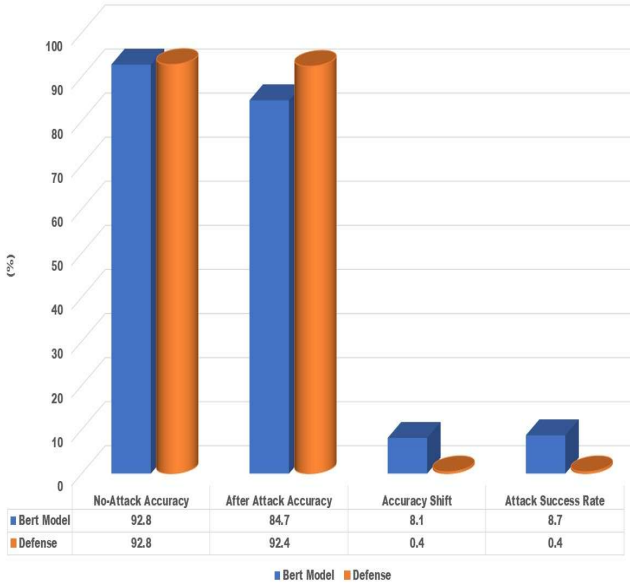


Figure 4 Evaluation of AG's news classification before and after SSE Technique.

B. Results of CosineDefender and JaccardDefinder Defense Frameworks

In this section, we evaluate the performance of the CosineDefender and JaccardDefender defense models across the AG's News, MSTI, and Symptom2Disease datasets. These models evaluated against three different adversarial attacks: random word substitution, random word deletion, and random character insertion.

AG's News Dataset : Table 2 outlines the accuracy results of the pretrained BERT base model when exposed to random word substitution, random word deletion, and random character insertion attacks on the AG's News dataset. Table 3 shows the proposed (Jaccard and Cosine) similarity defense techniques against those three attacks under the same attack condition. The defense framework with the highest classification accuracy is highlighted in bold which signifying its superior performance in mitigating attacks. Analyzing Table 2 and Table 3 reveals several significant observations from the accuracy results across different scenarios:

a) When no attack is present, the baseline model typically exhibits the highest accuracy. However, in the presence of various attacks, the framework consistently achieves lower accuracy.

b) Attack impact of the random character insertion attack emerges as the most potent adversarial method across all attack types by causing the largest attack success rate of 9.4. The baseline model suffers a significant decline, underscoring the destructive nature of this attack .The random word deletion attack also results in a substantial high attack success rate of 8.7 but is less severe than character insertion. While, random word substitution has a moderate effect by causing the smallest ASR of 5.17.

c) Defense Performance of the JaccardDefender framework over AG's news dataset consistently outperforms CosineDefender in mitigating the impact of attacks. For example, under random character insertion, the Jaccard similarity effectively limits the accuracy loss by maintaining a minimal reduction in accuracy compared to the 'No attack' baseline 92.2 .However, CosineDefender also exhibits solid performance particularly in defending against random word substitution attacks. It is slightly less effective than JaccardDefender when confronting more aggressive attacks like character insertion.

Table 2. Impact of Attacks on AG's News Dataset using BERT Model

Attack Type	No-Attack Accuracy (%)	After Attack Accuracy (%)	Accuracy Shift	Attack Success Rate (%)
Random word substitution	92.8	84.7	8.1	8.7
Random word Deletion	92.8	88	4.8	5.17
Random character Insertion	92.8	84	8.8	9.4

Manuscript received April 5, 2026
Manuscript revised April 20, 2026
https://doi.org/10.22937/IJCSNS.2026.26.4.2

Table 3. Evaluation of JaccardDefender and CosineDefender models on AG's news

Attack	No-Attack (%)	After Attack (%)	JaccardDefender Accuracy (%)	JaccardDefender ASR (%)	CosineDefender Accuracy (%)	CosineDefender ASR (%)
Random word substitution	92.8	84.7	**92**	.86	91.4	1.5
Random word Deletion	92.8	88	**92.2**	.6	92.2	.6
Random character Insertion	92.8	84	**92.29**	.54	92.2	.6

2- MSTI Dataset

The selection of the MSTI (Medical Symptoms Text and Audio Classification) dataset for validating the defense methods in the medical domain was primarily influenced by its substantial size, encompassing a wide array of medical symptoms and corresponding transcriptions. This extensive dataset provided a robust foundation for training and evaluating the models' effectiveness in handling diverse medical inquiries and potential adversarial attacks. Additionally, the dataset's inclusion of both text and audio transcriptions presented a unique opportunity to assess the models' resilience across different modalities of medical data, ensuring a comprehensive evaluation of their defensive capabilities in real-world medical scenarios.

To validate the efficacy of our defense methods in the medical domain, we selected the MQA dataset referred to as " MSTI". For testing purposes, we maintained an equal number of original test examples to use in the attacks. As a baseline, we also provided the classification accuracy of the original examples. Additionally, we included a comparison of different adversarial attack methods, encompassing Random Replacement, Random Deletion, and Random Character Insertion. As depicted in Table 4, the Bert-based classification model achieve high accuracy (99.6%) when classifying original examples. However, the introduction of adversarial attacks significantly reduced the model's accuracy. For example, the random word substitution decreased the accuracy to 82%, random word deletion dropped the accuracy further to 75% and random character insertion resulted in the most drastic reduction with the accuracy plummeting to 44.4%. This illustrates the model's vulnerability to these perturbations particularly character-level attacks. Table 5 presents the results of the two proposed defense models, CosineDefender and JaccardDefender, when subjected to these adversarial attacks on the MSTI dataset. Both defense frameworks proved to be highly effective but the Jaccard similarity model consistently achieved higher accuracy than the Cosine similarity model. This was especially evident under the more challenging attack scenarios like random character insertion where JaccardDefender maintained a relatively high accuracy of 92.2%, compared to 60.1% for CosineDefender.

Table 4.Impact of Attacks on *MSTI* Dataset using BERT Model

Attack	No-Attack Accuracy (%)	After Attack Accuracy (%)	Accuracy Shift	Attack Success Rate (%)
Random word substitution	99.6	82	17.6	17.6
Random word Deletion	99.6	75	24.6	26.5
Random character Insertion	99.6	44.4	55.2	59.4

Manuscript received April 5, 2026
Manuscript revised April 20, 2026
https://doi.org/10.22937/IJCSNS.2026.26.4.2

Table 5 Evaluation of JaccardDefender and CosineDefender Defense Models on MSTI dataset

Attack	No-Attack (%)	After Attack (%)	JaccardDefender Accuracy (%)	JaccardDefender ASR (%)	CosineDefender Accuracy (%)	CosineDefender ASR (%)
Random word substitution	99.6	82	**97**	2.2	93.9	6
Random word Deletion	99.6	75	**94.9**	4.7	89.2	10.4
Random character Insertion	99.6	44.4	**92.2**	7.4	60.1	39.6

3. Symptom2Disease Dataset

To further validate the proposed defense models, their performance is evaluated using another MQA dataset named Symptom2Disease. It encompasses a diverse set of medical symptoms and disease relationships. The goal was to showcase the robustness of the JaccardDefender and CosineDefender frameworks against the same three adversarial attacks. As illustrated in Table 6, the attacks resulted in a noticeable drop in the BERT-based model's classification accuracy. It shown that with No-Attack Accuracy of the baseline BERT model was 97.5. While , under random word substitution caused the accuracy to drop to 87.5% with an attack success rate of 10.2% . Random word deletion resulted in an accuracy of 85% with and an attack success rate of 12.8%. Random character insertion had the most pronounced effect by decreasing the accuracy drastically to 37%, with an accuracy shift of 60.5% and an attack success rate of 62%. Table 7 shows that under random word substitution both JaccardDefender and CosineDefender restored accuracy to 94.1% with an ASR of 3.4% .For random word deletion, JaccardDefender achieved 94% accuracy with a 3.5% ASR, while CosineDefender achieved 93.3% with a 4.3% ASR. The most significant difference was observed in the random character insertion attack where JaccardDefender reached 94.1% accuracy with a 3.4% ASR, while CosineDefender only restored accuracy to 85% with a 12.8% ASR.

Table 6. Impact of Attacks on Symptom2Disease Dataset against BERT Model

Attack	No-Attack Accuracy (%)	After Attack Accuracy (%)	Accuracy Shift	Attack Success Rate (%)
Random word substitution	97.5	87.5	10	10.2
Random word Deletion	97.5	85	12.5	12.8
Random character Insertion	97.5	37	60.5	62

Table 7. Evaluation of JaccardDefender and CosineDefender defense models on Symptom2Disease dataset

Attack	No-Attack (%)	After Attack (%)	JaccardDefender Accuracy (%)	JaccardDefender ASR (%)	CosineDefender Accuracy (%)	CosineDefender ASR (%)
Random word substitution	97.5	87.5	**94.1**	3.4	94.1	3.4
Random word Deletion	97.5	85	**94**	3.5	93.3	4.3
Random character Insertion	97.5	37	**94.1**	3.4	85	12.8

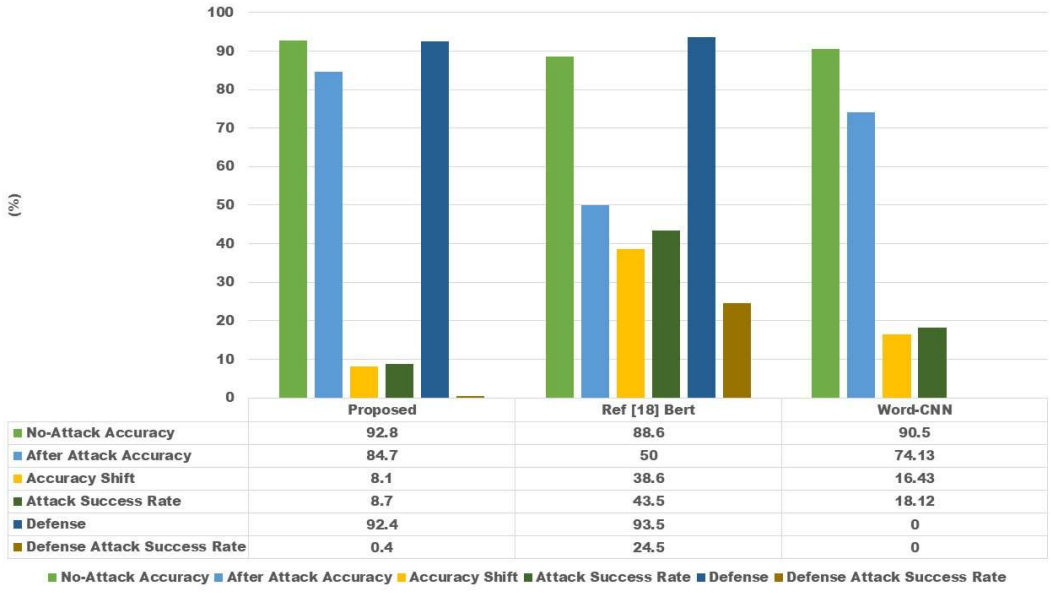
Across different dataset configurations and attack methods, the JaccardDefender consistently offers superior generalization and defense capabilities especially for synonym-based adversarial attacks. Across varying configurations (including different test sizes datasets, and attack methods) the JaccardDefender framework consistently delivers improved performance with minimal accuracy degradation. This highlights the generalizability of the Jaccard similarity defense framework in fortifying deep neural networks against synonym-based adversarial attacks.

4.4 Comparative Analysis

We further conducted a comparison between our SSE Model framework and previous research studies on the AG's News dataset. The comparison was performed under the random word substitution attack model as depicted in Figure 5. This evaluation was performed for several reasons. For example, the random word substitution represents one of the most prevalent word-level attack techniques in adversarial settings .We directly reference results from prior works [18, 19] which also utilized the same attack conditions for comparative analysis.

Upon examining Figure 5, we observe that the SSE Model achieves superior After-Attack Accuracy compared to the adversarial training defense technique (as detailed in Ref. [18]). Specifically, the SSE model reduces the attack success rate by approximately 23%–24%, despite potential differences in base model parameters between the studies. Our analysis indicates that, in contrast to the adversarial training technique in Ref. [18], the models under the SSE framework experience a smaller accuracy shift with an average decline of just 0.4%. By comparison, adversarial training shows an average accuracy decrease of 24% under similar attack conditions. Moreover, in cases where a Word-CNN model was employed, the accuracy shift was as high as 16.4%, leading to unsatisfactory performance. It is worth noting that the authors of Ref. [18] only assessed the impact of random word substitution attacks on the Word-CNN model without proposing any defense mechanisms.

In conclusion, our comparative analysis underscores the effectiveness of the SSE Model in mitigating adversarial attacks particularly when compared to prior approaches. These models not only maintain higher accuracy post-attack but also exhibit superior defense performance by rendering adversarial attacks significantly less successful.



	Proposed	Ref [18] Bert	Word-CNN
No-Attack Accuracy	92.8	88.6	90.5
After Attack Accuracy	84.7	50	74.13
Accuracy Shift	8.1	38.6	16.43
Attack Success Rate	8.7	43.5	18.12
Defense	92.4	93.5	0
Defense Attack Success Rate	0.4	24.5	0

Figure 5. Comparison between the proposed models and previous works on AG's News

5. Discussion

Dataset: The performance of the proposed model varies across different datasets. When considering adversarial sentences, the impact of using the proposed method for defending against adversarial examples is minimal for the AG's dataset, whereas its effect is significant for the Symptom2Disease and MSTI medical datasets. This disparity can be attributed to the nature of the medical question-answer dataset, which contains a few numbers of instances. Conversely, the AG's dataset consists of four

classes, differing from other datasets that typically have 24 or 25 classes. In a broader sense, the proposed model for each dataset necessitates more word modifications and repetitions, leading to reduced attack effectiveness while maintaining accuracy in the original sentences.

Applications: The proposed models extend to medical domains. Particularly in cases involving medical question-answer scenarios, the proposed method can mitigate the risk of misdiagnosis. Even when attackers present these medical questions as adversarial sentences, the application of our method can identify them accurately as original sentences. This capability is especially significant in medical scenarios, where the misrecognition of a classified document could lead to substantial misdiagnosis. Furthermore, the proposed method can be employed to enhance the safety of text recognition models in domains such as military data or public policy-based projects. This usage ensures the reliability and accuracy of text recognition of critical applications.

Limitation:
The proposed methods have some limitations such as: It necessitates a distinct process for defending other types of text adversarial attacks. Nevertheless, it's important to note that the proposed method still provides a degree of effectiveness even in different environments. Indeed, there is no universal defense mechanism that can ensure complete protection against all forms of adversarial attacks. Attackers consistently modify their strategies, posing a challenge for defenders to foresee and counteract every potential attack vector. Consequently, a defense approach that proves effective against a specific type of attack might not be equally successful against other types. The scope of this study is constrained by the use of three datasets due to time and resource limitations. Expanding the research to include additional datasets could provide a more comprehensive understanding of adversarial attacks across various domains. It is also worth noting that no existing paper comprehensively covers adversarial attacks across all types of datasets. Therefore, while our work provides significant insights, future research is planned to incorporate additional datasets to further validate and generalize our findings. Therefore, exploring alternative universal techniques for universal text adversarial examples is required.

6. Conclusion

The three defense models introduced in this paper include Synonym Substitution Embedding (SSE) Defense Framework, CosineDefender, and JaccardDefender demonstrate significant efficacy in enhancing the resilience of Medical Question Answering (MQA) systems against adversarial attacks. Evaluations across diverse datasets including Symptom2Disease, Medical Symptoms Text and Audio Classification, and AG's News validate the robustness of these models against both word-level (synonym substitution, word deletion) and character-level (random character insertion) attacks. The SSE Defense Framework effectively reduces attack success rates from 8.7% to 0.4% on AG's News and mitigates adversarial impact on medical datasets. Notably, CosineDefender lowers attack success rates for word synonym substitution, word deletion and character insertion to 3.4%, 4.3%, and 12.8%, respectively. On the other hand, JaccardDefender achieves the highest overall performance with success rates of 3.4%, 3.5%, and 3.4% coupled with superior accuracy across all datasets. These results underscore the effectiveness of the proposed models in bolstering the reliability of MQA systems in the face of evolving adversarial threats .

Despite these promising results, there are several limitations that should be addressed. For examples, the models were evaluated on a limited set of datasets which may not fully represent the diverse nature of medical question answering tasks. Additionally, the models primarily address specific types of adversarial attacks and may not be effective against all possible adversarial strategies .

Future work should focus on expanding the evaluation to include a broader range of datasets and adversarial attack types. Integrating additional defense mechanisms and exploring hybrid approaches could enhance the models' ability to handle a wider spectrum of adversarial threats. Further research into optimizing computational efficiency and scalability will also be crucial for practical deployment in diverse MQA systems. Addressing these aspects will help ensure comprehensive protection and maintain the reliability of MQA systems in the face of evolving adversarial challenges.

References

[1] Carlini, N. & Wagner, D. Audio adversarial examples: Targeted attacks on speech-to-text. In 2018 IEEE security and privacy workshops (SPW), 1–7 (IEEE, 2018).

[2] Samanta,S.,Mehta,S.,&Singh,S.(2020).Asurveyonsecurityth reatsanddefensetechniquesofmachinelearninginadversarialen vironment.JournalofAmbientIntelligenceandHumanizedCo mputing,11(6),2617-2632. [Accessed 22-07-2024].

[3] Goyal, S., Doddapaneni, S., Khapra, M. M. & Ravindran, B. A survey of adversarial defenses and robustness in nlp. ACM Comput. Surv. 55, 1–39.(2023)

[4] Han, X., Zhang, Y., Wang, W. & Wang, B. Text adversarial attacks and defenses: Issues, taxonomy, and perspectives. Secur. Commun. Networks 2022, 6458488.(2022)

[5] Gao, J., Lanchantin, J., Soffa, M. L. & Qi, Y. Black-box generation of adversarial text sequences to evade deep learning classifiers. In 2018 IEEE Security and Privacy Workshops (SPW), 50–56 (IEEE, 2018).

[6] Ebrahimi, J., Rao, A., Lowd, D. & Dou, D. Hotflip: White-box adversarial examples for text classification. arXiv preprint arXiv:1712.06751.(2017) 16/18
[7] Pruthi, D., Dhingra, B. & Lipton, Z. C. Combating adversarial misspellings with robust word recognition. arXiv preprint arXiv:1905.11268.(2019)
[8] Huang, P.-S. et al. Achieving verified robustness to symbol substitutions via interval bound propagation. arXiv preprint arXiv:1909.01492.(2019)
[9] Zhou, Y., Zheng, X., Hsieh, C.-J., Chang, K.-w. & Huang, X. Defense against adversarial attacks in nlp via Dirichlet neighborhood ensemble. arXiv preprint arXiv:2006.11627 .(2020)
[10] Goodfellow, I. J., Shlens, J. & Szegedy, C. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.(2014)
[11] Dong, X., Luu, A. T., Ji, R. & Liu, H. Towards robustness against natural language word substitutions. arXiv preprint arXiv:2107.13541.(2021)
[12] Chen, Y. et al. Why should adversarial perturbations be imperceptible? rethink the research paradigm in adversarial nlp. arXiv preprint arXiv:2210.10683.(2022)
[13] Zhou, Y., Jiang, J.-Y., Chang, K.-W. & Wang, W. Learning to discriminate perturbations for blocking adversarial attacks in text classification. arXiv preprint arXiv:1909.03084 .(2019)
[14] Mozes, M., Stenetorp, P., Kleinberg, B. & Griffin, L. D. Frequency-guided word substitutions for detecting textual adversarial examples. arXiv preprint arXiv:2004.05887 .(2020)
[15] Liang, B. et al. Deep text classification can be fooled. arXiv preprint arXiv:1704.08006.(2017)
[16] Samanta, S. & Mehta, S. Towards crafting text adversarial samples. arXiv preprint arXiv:1707.02812.(2017)
[17] Papernot, N., McDaniel, P., Swami, A. & Harang, R. Crafting adversarial input sequences for recurrent neural networks. In MILCOM 2016-2016 IEEE Military Communications Conference, 49–54 (IEEE, 2016).
[18] Ren, S., Deng, Y., He, K. & Che, W. Generating natural language adversarial examples through probability weighted word saliency. In Proceedings of the 57th annual meeting of the association for computational linguistics, 1085–1097 .(2019)
[19] Yang, P., Chen, J., Hsieh, C.-J., Wang, J.-L. & Jordan, M. I. Greedy attack and gumbel attack: Generating adversarial examples for discrete data. J. Mach. Learn. Res. 21, 1–36 .(2020)
[20] Sato, M., Suzuki, J., Shindo, H. & Matsumoto, Y. Interpretable adversarial perturbation in input embedding space for text.arXiv preprint arXiv:1805.02917.(2018)
[21] Li, J., Ji, S., Du, T., Li, B. & Wang, T. Textbugger: Generating adversarial text against real-world applications. arXiv preprint arXiv:1812.05271.(2018)
[22] Pennington, J., Socher, R. & Manning, C. D. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 1532–1543.(2014)
[23] Wang, X., Hao, J., Yang, Y. & He, K. Natural language adversarial defense through synonym encoding. In Uncertainty in Artificial Intelligence, 823–833 (PMLR, 2021).
[24] Cao, Y. et al. Tasa: Deceiving question answering models by twin answer sentences attack. arXiv preprint arXiv:2210.15221 (2022).
[25] Kumar, V., Maheshwary, R. & Pudi, V. Adversarial examples for evaluating math word problem solvers. arXiv preprint arXiv:2109.05925 (2021).
[26] Zhang, B., Li, J., Shi, Y., Han, Y. & Hu, Q. Vads: Visuo-adaptive dualstrike attack on visual question answer. Comput. Vis.Image Underst. 104137 (2024).
[27] Karra, R. & Lasfar, A. Analysis of qa system behavior against context and question changes. Int. Arab. J. Inf. Technol. 21,191–200 (2024).
[28] Wang, W., Wang, R., Wang, L., Wang, Z. & Ye, A. Towards a robust deep neural network against adversarial texts: A survey. ieee transactions on knowledge data engineering 35, 3159–3179.(2021)
[29] Kumar, V. & Subba, B. A tfidfvectorizer and svm based sentiment analysis framework for text data corpus. In 2020 national conference on communications (NCC), 1–6 (IEEE, 2020).
[30] Wilianto, D. & Girsang, A. S. Automatic short answer grading on high school's e-learning using semantic similarity methods. TEM J. 12.(2023)
[31] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805.(2018)
[32] Lahitani, A. R., Permanasari, A. E. & Setiawan, N. A. Cosine similarity to determine similarity measure: Study case in online essay assessment. In 2016 4th International conference on cyber and IT service management, 1–6 (IEEE, 2016).
[33] Soto, J. E., Hernández, C. & Figueroa, M. Jacc-fpga: A hardware accelerator for jaccard similarity estimation using fpgas in the cloud. Futur. Gener. Comput. Syst. 138, 26–42 .(2023)
[34] AGâˇZsnews,https://www.kaggle.com/datasets/amananandrai/ag-news-classification-dataset, availableonline(27Aug2023). [Accessed 22-07-2024].
[35] Symptom2Disease,https://www.kaggle.com/datasets/niyarrbarman/symptom2disease,availableonline(27Aug2023). [Accessed 22-07-2024 .
[36] MedicalSymptomsTextandAudioClassification,https://www.kaggle.com/code/paultimothymooney/ medical-symptoms-text-and-audio-classification,availableonline(27Aug2023). [Accessed 22-07-2024].
[37] Cresswell, W. & Quinn, J. L. Attack frequency, attack success and choice of prey group size for two predators with contrasting hunting strategies. Animal Behav. 80, 643–648 (2010).

Atrab A. Abd El-Aziz conducted the experiment practically through Google Colab. Also, she collected and processed the data from Kaggle website, tested the proposed models, showed the results, and then wrote the method. Prof.Dr. Nour Eldeen Khalifa presented the results, represented them in the figures shown, and then made comparisons. Prof.Dr. Reda A El-Khoribi worked on the previous studies and then wrote the introduction and previous studies to be a strong reference for us to continue our work.