Journal of Theoretical and Applied Information Technology
15th February 2024. Vol.102. No 3

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

STRENGTHENING QA SYSTEM ROBUSTNESS: BERT AND CHARACTER EMBEDDING SYNERGY

RACHID KARRA[1] , ABDELALI LASFAR[2]
[1,2]LASTIMI Laboratory, Mohammadia School Of Engineers,
Mohammed V University in Rabat, Morocco
E-mail: [1]rachid.karra@est.um5.ac.ma, [2]abdelali.lasfar@est.um5.ac.ma

ABSTRACT

The inherent nonlinearity of neural networks renders them vulnerable to adversarial attacks. As artificial intelligence-based question-answering (QA) systems continue to proliferate across sectors like education and health, ensuring their security and predictable behavior becomes imperative for widespread adoption among the general public. Robustness measures the resilience of Natural Language Processing (NLP) models against adversarial attacks. A robust QA system exhibits resilient behavior when encountering maliciously crafted questions. Previous experiences have indicated that utilizing character embeddings enhances resistance against contradictory misspelled questions. To validate this, we fine-tuned BERT with character embeddings on SQuAD dataset for question-answering tasks and assessed both models using {question, context, answer} tuples. The results unequivocally demonstrate that BERT with character embeddings yields superior performance. Finally, we propose a comprehensive framework aimed at safeguarding question-answer type dialogue systems.

Keywords: *Adversarial Attacks, BERT, Character Embedding, QA Systems, SQuAd*

1. INTRODUCTION

Dialogue systems have taken a key place in our daily lives. With the advent of LLMs (Large Language Models), the reflex of seeking an answer in a dialogue system will become as ingrained as searching from a search engine. In the realm of education, dialog systems have reshaped the learning landscape by providing personalized learning experiences and answering various student questions [1]. These systems offer customizable explanations, exercises, and resources, promoting an inclusive learning environment. Dialogue systems have become integral in transforming the health sector by offering unparalleled support and accessibility. These systems, encompassing chatbots and virtual assistants, have significantly enhanced patient engagement by providing immediate responses to inquiries, enabling 24/7 support, and efficiently triaging symptoms for accurate diagnoses [2]. Moreover, their integration with telemedicine platforms has revolutionized remote healthcare, facilitating consultations and real-time monitoring. Beyond patient care, these systems also serve as educational tools, disseminating health-related information to promote awareness and healthy practices among individuals and communities.

The majority of assessments for NLP (Natural Language Processing) models have typically been conducted during the testing phase, primarily focusing on their performance indicators. In these evaluations, researchers and practitioners scrutinize the models' capabilities and efficiencies, analyzing their behavior and outcomes in relation to predefined benchmarks and metrics [3]. This testing-centric approach provides insights into how well the NLP models perform across various tasks, contributing to a comprehensive understanding of their overall effectiveness and suitability for real-world applications.

Adversarial attacks on Natural Language Processing (NLP) models represent a critical area of research that explores vulnerabilities within these models [3]. These attacks involve the deliberate manipulation of input data to deceive or mislead NLP systems, causing them to produce incorrect outputs or classifications. Adversarial attacks pose a significant challenge to the reliability and security of NLP models, raising concerns about their robustness in real-world applications [4]. These attacks can take various forms, such as adding imperceptible perturbations to the input text, crafting specially designed examples, or utilizing sophisticated algorithms to exploit weaknesses in the model's decision-making process. The objective

of these attacks is to exploit the model's vulnerabilities, leading it to make erroneous predictions or classifications, even with minimal alterations to the input text.

Research in this field aims to understand the underlying mechanisms behind these vulnerabilities and develop defense mechanisms to enhance the robustness and resilience of NLP models against such attacks [5]. Various techniques, including adversarial training, robust optimization, and model modification, are being explored to mitigate the impact of adversarial attacks on NLP models. Goodfellow et al. [6] laid the groundwork for understanding adversarial attacks across machine learning models. Further research has expanded this domain, exploring adversarial attacks specifically tailored for NLP tasks, such as text classification, sentiment analysis, question-answering systems.

Adversarial attacks against question-answering (QA) systems are a significant area of concern in the field of natural language processing (NLP). These attacks involve creating input to mislead QA systems, causing them to provide incorrect or misleading answers [7]. The QA system's vulnerability to adversarial attacks poses a crucial challenge, potentially leading to erroneous results that could have grave consequences in real-world applications. These disruptions aim to exploit vulnerabilities in the neural model's inference, causing it to provide inaccurate responses or generate misleading explanations.

Adversarial attacks in QA systems are based on the intrinsic vulnerabilities of neural networks. They can take several forms. In the case of vision, attackers can introduce minimal modifications to alter recognition or classification. Subtle changes like homoglyphs in questions or entered passages, which are almost imperceptible to humans, but significantly alter the output of the system. We can classify NLP adversarial attacks by the information collected on the model (black-box or white-box), impact (targeted or non-targeted) and text granularity (character, word, or sentence level), and attack strategy. Adversarial attacks are divided into two types as either targeting training data or neural network models [8].

Researchers have studied various techniques to understand, detect, and stop adversarial attacks against QA systems. Adversarial learning, where models are trained on examples developed in an adversarial manner, is an approach aimed at improving the robustness of QA systems [9]. Additionally, adversarial defenses such as input disruption constraints, defensive distillation, and adversary fine-tuning are explored to strengthen these systems against attacks. Several scientific studies have looked at adversarial attacks specifically targeted at QA systems. Jia & Liang [10] highlighted how adversarial attacks can significantly degrade the performance of reading comprehension models. Other works, such as Zügner & Günnemann [11], have examined adversarial attacks in the context of graph-based QA system.

The results of Karra and Lasfar [9] show that R-NET handles spelling errors better than BERT. Indeed, BERT and R-NET give quite similar results in the case of a question with one error, and R-NET obtains better results (more correct answers) in the case of questions with two and three errors. These results go against what we expected because BERT uses more parameters than R-NET, based on the transformer architecture and it is trained on more data. In addition, BERT is better in F1 and EM scores than R-NET.

However, analysis of the architecture of each of the two models shows that these results can be explained by the layers of word embedding adopted by each of the models. BERT adopts token-based, segment-based, and positional embeddings, but not any character-based embedding like R-NET, which has in addition to word-based embedding, it also contains word-based embedding on characters which is useful for processing words outside of vocabulary.

The findings align with [12] which utilized a modified BERT variant, Character-BERT-medical, yielding improved outcomes within the medical domain. This involved substituting subword embeddings with character embeddings (utilizing Character-CNN, integrated within the ELMo architecture), notably enhancing performance, especially when confronted with spelling errors. Additionally, [13] demonstrated BERT's limitations against various adversarial attacks, particularly those involving word modifications. Their research highlighted the significant impact of mistyping, as it generates uncommon samples for subword embeddings, consequently undermining BERT's performance.

In our study, we evaluated the robustness of two variants of the BERT model (subwords and character embedding) using a question-answering task. While the performance of dialog systems is crucial in a production environment, the robustness and security of question processing are also important. Our study also sheds light on the degrees of degradation of the dialogue system.

In the following sections, we aim to justify our selection of character embedding as a defense mechanism against adversarial misspelling attacks. We will demonstrate the effectiveness of fine-tuning the BERT-based model using character embedding for the QA system task. Subsequently, we plan to evaluate the robustness of two models, namely BERT with character embedding (BERT-characters) and BERT-base, by subjecting them to 790 {question, context} pairs. Lastly, we intend to propose a comprehensive framework designed to enhance the security of the QA system against adversarial attacks while also improving their management.

2. MATERIALS AND METHODS

2.1 Experiments

Adversarial attacks can be categorized into several types. In the case of white-box attacks, the tester possesses knowledge about the neural architecture and hyperparameters of the model being targeted. Conversely, black-box attacks occur when the tester lacks direct access to the model's internal details. In this scenario, the only available means to gather information about the QA system is through the posed questions. Research indicates that targeted questions can unveil certain characteristics of the model [14]. These questions might include specific symbols or special characters encoded differently to probe system vulnerabilities. For our experiment, we chose substitution attacks at the character level, and we categorized them into 3 levels: substitution of one character, two characters and three characters (see Table 1).

Alterations in questions vary based on the token impacted by the change and its position within the sequence. In the illustrative instance presented in Table 1, switching from 'P' (representing polynomial) to 'O' (denoting Landau or Big O) results in a markedly distinct interpretation of the question.

Table 1: questions with misspelling changes

Original	What does the P mean in complexity theory?
1sp-Error	What does the P mean in comflexity theory?
2sp-Error	What dods the P mean il complexity theory?
3sp-Error	ghat does the O meap in complexity theory?

For fine-tuning the model, we used the dataset SQuAD 1.1 (Stanford Question Answering Dataset). It is a benchmark dataset for context-based answer retrieval tasks. It is composed of question-answer pairs derived from Wikipedia articles. It contains contextual passages (C), questions (Q) and paragraphs (A) such as tuples (C, Q, A). The dataset is designed to evaluate the models' ability to accurately understand, and answer questions based on associated contexts, formulated as follows:

$$\text{SQuAD 1.1} = \{(C, Q, A)\} \quad (1)$$

This widely used dataset facilitates model development and evaluation in natural language understanding and question answering tasks. Version 2.0 also contains questions that are not answered from context.

Our innovative strategy aimed at enhancing BERT's understanding of linguistic nuances led us to use a model who integrates character-level embeddings alongside the word-level ones, using convolutional neural networks (CNN) and long short-term memory (LSTM) networks. Subsequently, we fine-tuned the BERT architecture, customizing it to seamlessly integrate character embeddings into the model's input layers. This modification empowered BERT to comprehend both word-level semantics and nuanced character-level information concurrently.

2.2 Model

Transformer-based models commonly utilize word or subword embeddings [15], [16] employing a standardized processing approach. This involves utilizing pre-trained models and selecting layers based on the specific requirements of the datasets, rather than starting the model training from the ground up. Despite BERT displaying superior F1 and EM scores compared to R-NET, as discussed earlier, our exploration in the preceding section revealed R-NET's heightened resilience against adversarial attacks. Notably, embedding at the character level substantially fortifies the model's robustness [13].

BERT exhibits increased vulnerability to character change attacks, particularly in words that draw more focus within the query [13]. To counter this susceptibility, a strategy was employed involving two RNN-based layers: the initial layer integrated GloVe embeddings, while the subsequent layer leveraged character n-gram embedding. Notably, previous attempts by other researchers to incorporate character embedding into

the R-NET model, utilizing a convolutional neural network, proved unsuccessful [17]. ELMO (Embeddings from Language Models), a pre-trained model rooted in bidirectional LSTMs and character-level integration [18], employs a methodology where words undergo segmentation into characters, entering a convolutional neural network, and are then optimized through global max-pooling. When contextualized word embeddings and character embedding are combined, the result is a hybrid representation that benefits from both the contextual understanding provided by ELMO and the ability of character embeddings to handle out-of-vocabulary words and spelling variations [18].

In this process, the information flows through two Highway layers that incorporate a residual connection. These layers encompass two crucial elements: the transformation gate (referred to as T) and the carry gate (known as C). Their role lies in quantifying the output y generated by altering the input x through non-linear transformations. To illustrate, when T equals 0, the input x traverses directly without any modification, giving rise to the term "Highway" as it allows direct passage (see Figure 1).

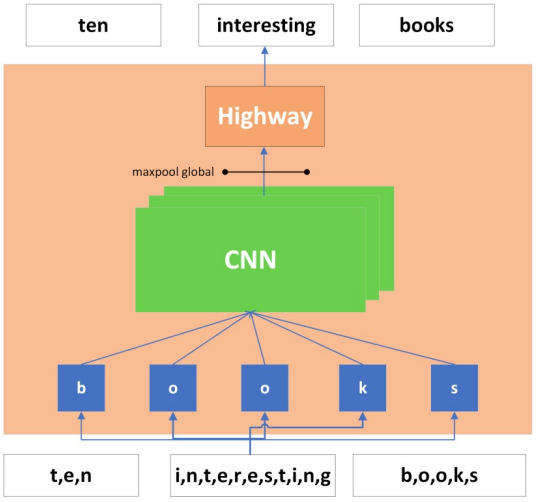

Figure 1: The CNN corporation at the character level

For example, the specific subword units that "looking" could be segmented into by BERT's tokenizer may include "look" and "##ing". For example, BERT's tokenizer might segment the word "looking" into specific subword units such as "look##" and "##ing". However, an error in a token like "louk##" fails to rectify the misspelling, leading to an incorrect reconstruction of the word as "louking". CNN representation exhibits an understanding of character distributions.

For instance, in the event of an error within a word like "raim," this framework maintains the capability to establish a connection with the accurately spelled word "rain". Building upon the approach outlined by ELMO [18], [12] opted to employ embedding within the CNN architecture, replacing the BERT subword embedding. We relied on the CharacterBERT model [12], which not only matches the performance of BERT-base or uncased but also demonstrates superior performance by an average of 2 points across various datasets (MEDNLI, ChemPro).

This adaptation enhances the model's resilience when encountering data with degradation. Throughout the training phase, all fine-tuning experiments are run on a single Tesla P100/16GB and implemented the AdamW optimizer.

To fine-tune Character-BERTgeneral for good performance in question answering (QA) tasks, our first crucial step involved selecting a suitable dataset. We opted for the SQuAD v1.1 (Stanford Question Answering Dataset) [19], a rich resource containing context paragraphs and corresponding questions along with answer spans.

Leveraging BERT's tokenizer, we intricately processed the text data, tokenizing it into word-level tokens and segmenting words into subword tokens using the WordPiece tokenizer. Simultaneously, we embarked on a novel approach, engineering character-level representations by disassembling words into individual characters.

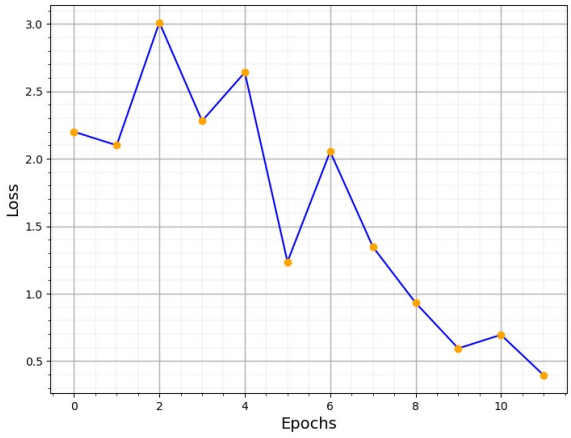

Figure 2: Model loss in 12 epochs

For the fine-tuning process, we initialized the modified BERT (Character-BERTgeneral) model with pre-trained weights and embarked on refining it using the SQuAD dataset. Implementing the AdamW optimizer, we orchestrated 12 epochs of training, optimizing the model's parameters and

adjusting hyperparameters with a learning rate of 1.1e-5. This refinement over multiple epochs ensured that the model effectively learned from the dataset, improving its ability to provide accurate answers in diverse question answering scenarios.

3. RESULTS AND DISCUSSION

3.1 One misspelling error questions

The results of single-error questions (Figure 3) show that the percentage of correct answers of BERT-CAR is 84.1% compared to 78.35% for the normal BERT model. The two models do not record any partial classified response and record respectively 126 and 171 incorrect responses, a reduction of around 26.3%.

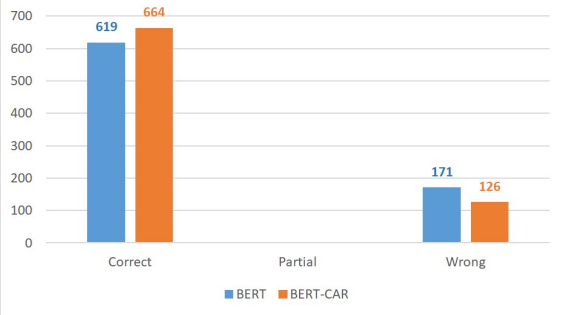

Figure 3: Comparative Analysis of BERT and BERT-CAR Responses to Adversarial Examples with a Single Error

Figure 4 shows that out of 790 questions the BERT-CAR model obtains 23 sentences with no wrong answers, 17 for the category of one and two wrong answers. BERT-CAR does not have error categories with 7 or more wrong answers.

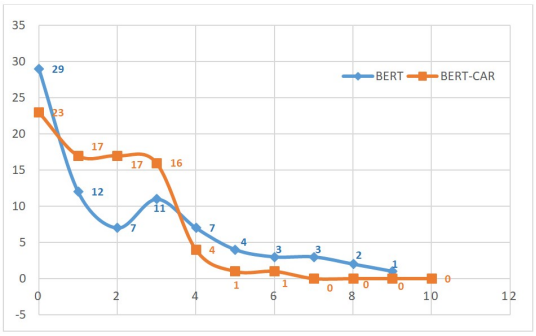

Figure 4: Categorization of questions according to the number of wrong answers (1 error).

The comparison between BERT-CAR and BERT shows that there is a small improvement compared to the latter with an increase in the number of sentences containing a low number of errors except for the category of sentences with no errors where BERT achieves a best score. BERT-CAR performs better compared to BERT in handling a single error.

3.2 Two misspelling errors questions

In situations where two errors were detected, the count of accurate responses generated by BERT-CAR dropped notably to 521 from the previous 664, marking a significant 21.5% decrease. Conversely, the tally of incorrect answers surged from 126 to 265, showcasing a substantial increase of 110.3%. Remarkably, partial responses remained nearly unchanged, marginally declining from 5 to 4. Comparatively, when measured against BERT, BERT-CAR exhibited a notable 15.5% enhancement in delivering accurate answers while concurrently demonstrating a reduction of -20.6% in incorrect responses (see Figure 5). These metrics underscore the promising performance advancements of BERT-CAR over BERT, particularly in the realm of robustness.

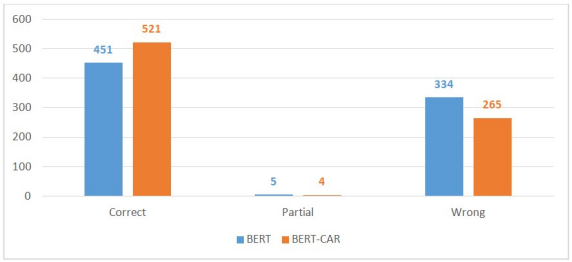

Figure 5: Comparative Analysis of BERT and BERT-CAR Responses to Adversarial Examples with a 2 Errors

Both models exhibit a similar curve trend in the scenario of two errors. In the 8-error category, BERT encompasses 8 questions, whereas BERT-CAR only comprises 4 questions. Notably, BERT-CAR outperforms BERT in the initial range of error counts, demonstrating a superior performance, albeit with a gradual increase in incorrect responses towards the latter part.

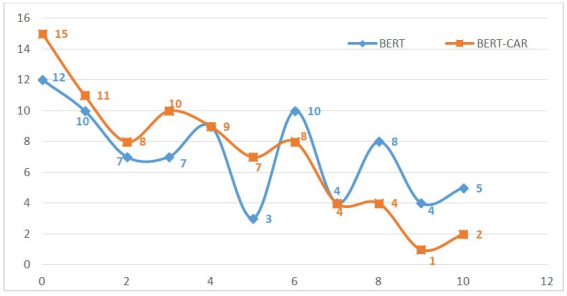

Figure 6: Categorization of questions according to the number of wrong answers (2 errors).

3.3 Three misspelling errors questions

As expected, the outcomes for sentences with three errors show a decline even in the case of the BERT-CAR model. Consequently, it achieved a total of 491 correct answers, translating to a percentage of 62.2%, while yielding 298 incorrect answers in comparison.

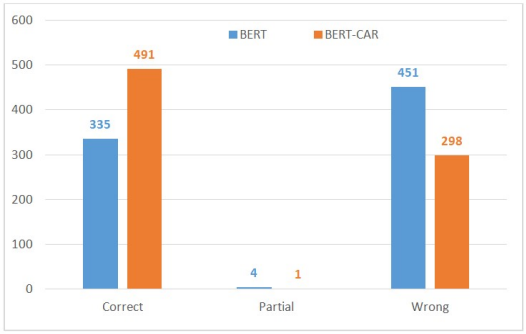

Figure 7: Distribution of answers to contradictory examples with 3 errors

Figure 8 shows that BERT-CAR obtains 7 sentences with no incorrect answers, 9 for the single error category. However, it should be noted that the greatest number of answers concern those with 3, 4 and 5 errors.

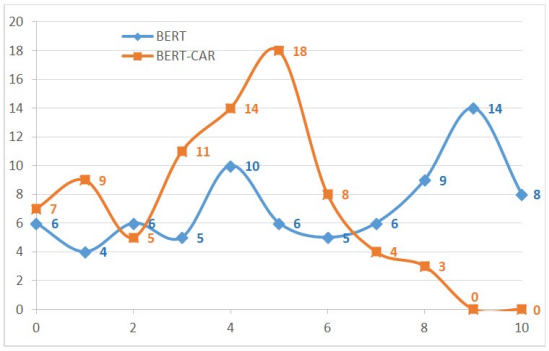

Figure 8: Categorization of questions according to the number of wrong answers (2 errors)

In a comprehensive evaluation, BERT-CAR demonstrates superior performance compared to BERT. This enhancement is particularly noteworthy in the mid-range spectrum, specifically pertaining to sentences containing three, four, five, and six errors. Consequently, BERT-CAR exhibits greater resilience than BERT, notably excelling in inferring three-error questions

Utilizing character-level embedding, the model achieves an impressive protection rate of 84.1% for questions with a single error, 65% for those with two errors, and 62.2% for three-error questions. This amalgamation leverages the performance prowess of BERT, harnessed through Transformers, in conjunction with the resilience of R-NET against spelling errors, employing character embedding within a CNN architecture [14].

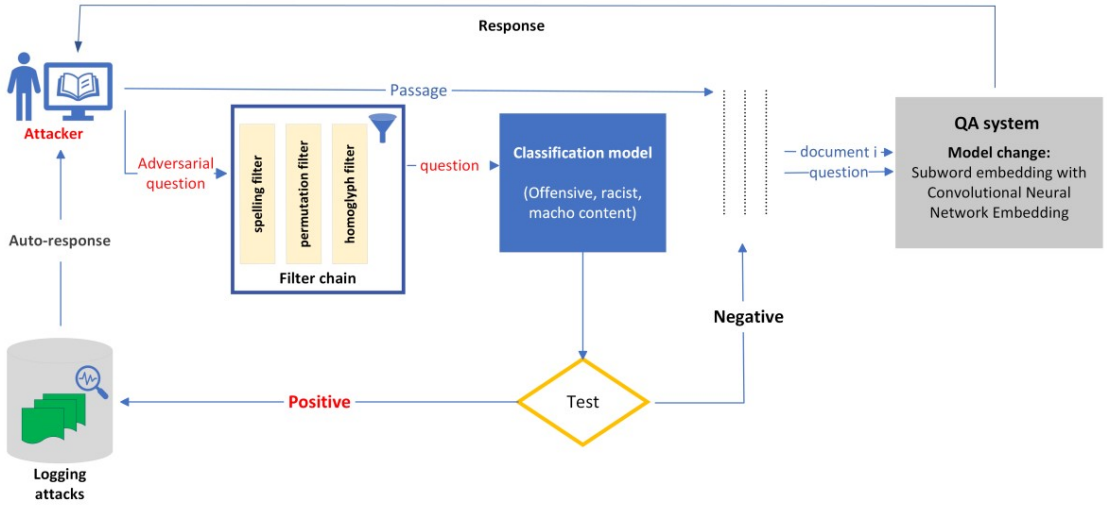

Figure 9: The schematic representation outlining the safeguarding measures for the QA system against adversarial attacks

3.4 Secure QA-Systems

Natural Language Processing (NLP) models were initially designed for inference and prediction rather than security. To bolster their resilience, a pioneering approach [20], [21] suggests training the model from inception using contradictory examples. This strategy involves augmenting the training data by incorporating responses such as 'I don't know' or 'no answer'. This proactive measure aims to mitigate vulnerabilities within the model architecture, fortifying its robustness against potential threats or adversarial inputs.

The model's architecture prioritizes security considerations and defenses against adversarial attacks. During the training phase, the inclusion of contradictory data serves to inoculate the model against potential attacks, while additional measures involve integrating security components to forestall inappropriate behaviors. An alternative approach involves implementing upstream detection mechanisms to discern adversarial examples, enabling redirection without necessitating alterations to existing NLP models [22]. These proactive strategies aim to fortify the model's resilience, minimizing susceptibility to malicious inputs and enhancing its overall security posture.

In our efforts to mitigate adversarial attacks, we employ an initial mechanical shield, akin to non-neural methods [23], that rectifies students' linguistic nuances and spelling inaccuracies. Language Learning Models (LLMs) tailored for public consumption leverage filters and proxies to moderate user-generated content, employing automatic filters to flag inappropriate queries or relying on human moderators for oversight [24], [25].

These strategies act as protective measures, aiming to maintain the integrity of interactions and content while reducing the potential impact of adversarial inputs on the model. A text error is defined as any word within a predetermined dictionary that comprises characters outside the specified register [a-zA-Z]. The implemented shield effectively nullifies an entire category of "Character Manipulation" attacks, encompassing actions such as adding tokens, deleting characters, substituting homoglyphs (such as replacing '0' with the number, or 'O' with the letter), permutation, and substitution based on chance or proximity to neighboring keys on the keyboard (see Figure 9).

Employing a model trained on contradictory data serves as an ultimate line of defense. In our scenario, we substituted BERT-base with Character-BERTgeneral, refined through fine-tuning with character-level embedding using CNN networks. This particular model exhibited significantly heightened resilience when subjected to adversarial "character manipulation" attacks.

4. CONCLUSION AND FUTUR WORKS

The continuous evolution of dialogue systems and their integration with various platforms continue to revolutionize various industries, making their utilization more accessible, adaptable, and secure for individuals. The fine-tuned BERT-CAR model, equipped with character embeddings, demonstrated adaptability, and proved adept at handling misspelled words, proving its efficacy in addressing adversarial attacks with enhanced precision and reliability. Character-level embedding inherently fortifies the model, providing inherent defense mechanisms against adversarial attacks, including character substitution or permutation. Defending question answering (QA) systems against adversarial attacks requires comprehensive strategies aimed at enhancing their robustness and resilience. Apart from adversarial training, several other approaches and defenses have been explored in the paper to mitigate the impact of adversarial attacks on QA systems. Designing QA models with robust feature representations can aid in reducing their vulnerability to adversarial attacks. Utilizing methods like feature denoising, robust feature extraction, or incorporating domain-specific knowledge into feature representations can bolster the model's robustness.

Securing a QA (question-answering) system requires implementing multiple system enhancement and immunization strategies due to the diverse nature of attacks. These measures encompass a range of defenses, given the wide array of potential attack vectors targeting QA systems. Protecting a QA system necessitates a comprehensive approach, employing diverse strategies to fortify its defenses against many potential threats and attacks.

In the ever-evolving landscape of technology, the integration of multimodal dialogue systems has become a pivotal aspect, ushering in a new era of diverse data processing. This paradigm shift introduces heightened challenges in securing the intricate interplay of image, text, and structured data. As we delve into the complexities of this multifactorial landscape, the need for robust security measures becomes paramount. Understanding the nuances of each modality is essential in fortifying against potential contradictory attacks. The dynamic nature of these systems necessitates continuous research and

adaptation to ensure their resilience in real-world applications. Amidst these challenges, the pursuit of effective security solutions remains at the forefront, promising a safeguarded environment for the seamless functioning of multimodal dialogue systems.

Our study exclusively focused on adversarial attacks of the substitution type, recognizing that various other types, including permutation, the addition of special characters, and changes to encoding, exist. The rationale behind our choice lies in the prevalence of substitution attacks, which stands as the most common form encountered by users of dialogue systems. The implementation of special character attacks is notably challenging and demands a substantial number of attempts. Looking ahead, our forthcoming research endeavors will encompass a broader spectrum of adversarial attacks, aiming to develop a comprehensive framework for automating robustness tests against dialogue systems.

REFERENCES:

[1] R. Karra and A. Lasfar, "Enhancing Education System with a Q&A Chatbot: A Case Based on Open edX Platform," in *Digital Technologies and Applications*, S. Motahhir and B. Bossoufi, Eds., Cham: Springer International Publishing, 2021, pp. 655–662.

[2] E. Adamopoulou and L. Moussiades, "Chatbots: History, technology, and applications," *Machine Learning Application*, vol. 2, p. 100006, Dec. 2020.

[3] Z. Kong, J. Xue, Y. Wang, L. Huang, Z. Niu, and F. Li, "A Survey on Adversarial Attack in the Age of Artificial Intelligence," *Wireless Communications and Mobile Computing*, vol. 2021, pp. 1–22, Jun. 2021.

[4] C. Szegedy *et al.*, "Intriguing properties of neural networks," in *International Conference on Learning Representations,* Banff, Canada, 2014.

[5] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The Limitations of Deep Learning in Adversarial Settings," in *IEEE European Symposium on Security and Privacy (EuroS&P)*, Saarbrucken: IEEE, Mar. 2016, pp. 372–387.

[6] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples." in *International Conference on Machine Learning*, vol. 37, 2015

[7] M. Bartolo, T. Thrush, R. Jia, S. Riedel, P. Stenetorp, and D. Kiela, "Improving Question Answering Model Robustness with Synthetic Adversarial Data Generation," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, 2021, pp. 8830–8848.

[8] I. Alsmadi *et al.*, "Adversarial Machine Learning in Text Processing: A Literature Survey," *IEEE Access*, vol. 10, pp. 17043–17077, 2022.

[9] R. Karra and A. Lasfar, "Impact of Data Quality on Question Answering System Performances," *Intelligent Automation & Soft Computing*, vol. 35, no. 1, pp. 335–349, 2023.

[10] R. Jia and P. Liang, "Adversarial Examples for Evaluating Reading Comprehension Systems," *in Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing,* Copenhagen, Denmark, pp. 2021–2031, 2017.

[11] D. Zügner, A. Akbarnejad, and S. Günnemann, "Adversarial Attacks on Neural Networks for Graph Data," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Jul. 2018, pp. 2847–2856.

[12] Boukkouri H. E., Ferret O., Lavergne T., Noji H., Zweigenbaum P., and Tsujii J., "CharacterBERT: Reconciling ELMo and BERT for Word-Level Open-Vocabulary Representations From Characters," in *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 6903–6915, 2020.

[13] L. Sun *et al.*, "Adv-BERT: BERT is not robust on misspellings! Generating nature adversarial samples on BERT," *ArXiv200304985 Cs*, DeepAI publication, 2020.

[14] W. E. Zhang, Q. Z. Sheng, A. Alhazmi, and C. Li, "Adversarial Attacks on Deep Learning Models in Natural Language Processing: A Survey," *ACM Transactions on Intelligent Systems and Technology,* Volume 11, Issue 3, No: 24, pp 1–41, 2020

[15] A. Vaswani *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pp. 5998–6008, 2017.

[16] Devlin J., Chang M.-W., Lee K., and Toutanova K., "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in *North American Chapter of the Association for Computational Linguistics*, pp. 4171–4186, 2019.

[17] "R-NET: Machine Reading Comprehension With Self-Matching Networks." Natural Language Computing Group, Microsoft Research Asia, 2017.

[18] M. Peters *et al.*, "Deep Contextualized Word Representations," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, New Orleans, Louisiana: Association for Computational Linguistics, 2018, pp. 2227–2237.

[19] Rajpurkar P., Zhang J., Lopyrev K., and Liang P., "SQuAD: 100,000+ Questions for Machine Comprehension of Text," in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, 2016.

[20] J. X. Morris, E. Lifland, J. Y. Yoo, J. Grigsby, D. Jin, and 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,* pp. 119–126, 2020.

[21] H. Zhang and J. Wang, "Defense Against Adversarial Attacks Using Feature Scattering-based Adversarial Training", in *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, no. 164, pp. 1831–1841, 2019.

[22] V. Raina and M. Gales, "Residue-Based Natural Language Adversarial Attack Detection," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Seattle, United States: Association for Computational Linguistics, 2022, pp. 3836–3848.

[23] M. das G. Bruno Marietto *et al.*, "Artificial Intelligence Markup Language: A Brief Tutorial," *International Journal of Computer Science and Engineering Survey*, vol. 4, no. 3, pp. 1–20, Jun. 2013.

[24] S. Eger *et al.*, "Text Processing Like Humans Do: Visually Attacking and Shielding NLP Systems", 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),* Minneapolis, Minnesota, pp. 1634–1647, 2019.

[25] T. Le, N. Park, and D. Lee, "SHIELD: Defending Textual Neural Networks against Multiple Black-Box Adversarial Attacks with Stochastic Multi-Expert Patcher," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 6661–6674.