Indonesian Journal of Electrical Engineering and Computer Science
Vol. 38, No. 3, June 2025, pp. 1735~1744
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v38.i3.pp1735-1744

TextBugger: an extended adversarial text attack on NLP-based text classification model

Sanjaikanth E. Vadakkethil Somanathan Pillai[1], Srinivas A. Vaddadi[2], Rohith Vallabhaneni[2], Santosh Reddy Addula[2], Bhuvanesh Ananthan[3]
[1]School of Electrical Engineering and Computer Science, University of North Dakota, Grand Forks, United States
[2]Department of Information Technology, University of the Cumberlands, Williamsburg, United States
[3]Department of Electrical and Electronics Engineering, PSN College of Engineering and Technology, Tirunelveli, India

Article Info

Article history:

Received Mar 19, 2024
Revised Nov 19, 2024
Accepted Nov 24, 2024

Keywords:

Attack detection
BERT
Natural language processing
Robustly optimized BERT
Text adversaries
XLNet

ABSTRACT

Recently, adversarial input highly negotiates the security concerns in deep learning (DL) techniques. The main motive to enhance the natural language processing (NLP) models is to learn attacks and secure against adversarial text. Presently, the antagonistic attack techniques face some issues like high error and traditional prevention approaches accurately secure data against harmful attacks. Hence, some attacks unable to increase more flaws of NLP models thereby introducing enhanced antagonistic mechanisms. The proposed article introduced an extended text adversarial generation method, TextBugger. Initially, preprocessing steps such as stop word (SR) removal, and tokenization are performed to remove noises from the text data. Then, various NLP models like Bi-directional encoder representations from transformers (BERT), robustly optimized BERT (ROBERTa), and extreme learning machine neural network (XLNet) models are analyzed for outputting hostile texts. The simulation process is carried out in the Python platform and a publicly available text classification attack database is utilized for the training process. Various assessing measures like success rate, time consumption, positive predictive value (PPV), Kappa coefficient (KC), and F-measure are analyzed with different TextBugger models. The overall success rate achieved by BERT, ROBERTa, and XLNet is about 98.6%, 99.7%, and 96.8% respectively.

Corresponding Author:

Sanjaikanth E. Vadakkethil Somanathan Pillai
School of Electrical Engineering and Computer Science, University of North Dakota
Grand Forks, ND 58202, United States
Email: s.evadakkethil@und.edu

1. INTRODUCTION

In today's scenario, the use of the deep learning (DL) approach keeps on increasing results in the introduction of natural language processing (NLP) models. It is noted that fascinating results are obtained while processing the NLP models in various fields like question answering, sentimental analysis (SA), language translation, and text manipulation. Astudillo *et al.* [1], it is noted that integrating suitable perturbations cannot be easily identified to text data that deliberates the DL models to produce errors resulting in adversarial attacks mainly encompassed in computer vision applications. Recently, studies on adversarial attacks made outstanding intimidation in NLP, image processing, face identification, and intrusion detection processes [2], [3]. It is analyzed that particular NLP processes like spam identification, and sensitive data detection are playing an integral role in data processing and security on networks. Hence, it is necessary to enhance the performance of NLP models based on DL techniques.

However, creating adversarial inputs for texts is highly challenging compared to creating adversarial inputs in images [4], [5]. The texts are highly random, conquering the persistent concept of an image. Moreover, the hostile text inputs are obtained via a disturbing character-level process that causes vulnerability during word correction and readable processes [6], [7]. This process can create high security to some extent regarding character-level attacks. But this alteration subjects to increased gradient attacks that are not directly implemented on the text. In addition to this, integrating sub-word perturbation may change the text into out-of-vocabulary (OOV) words. The textual perturbation can create an enhanced impact on semantics than on images. As a result, it is difficult to enhance the models to generate adversarial textual examples [8], [9]. To overcome the cons of existing methodologies, this article integrates the textual features and model features to develop a multiple attack technique named, TextBugger.

Motivation: nowadays, the DL models are becoming more popular in classifying adversarial text based on original texts. However, generating adversarial data is highly challenging and it is not as image adversaries. To overcome this issue, NLP-based DL models are introduced that automatically learn meaningful sentences and classify the hostile text effectively. Some of the commonly used NLP schemes are Bi-directional encoder representations from transformers (BERT), robustly optimized BERT (ROBERTa), and extreme learning machine neural network (XLNet) models that use contextual embedding property and prevent long-term dependency problems. Motivated by this, the developed framework investigated several NLP models in classifying adversarial texts using original texts. The key contributions of the developed framework are described as follows:

- To introduce an extended text attack NLP scheme to analyze its performance in classifying adversarial outcomes.
- To analyze various natural language models like BERT, ROBERTa, and XlNet in classifying adversarial text based on textual output.
- To validate the existing BERT, ROBERTa, and XLNet-based NLP models by assessing different performance measures like accuracy, Kappa coefficient (KC), positive predictive value (PPV), and F-measure metrics.

The upcoming sections are organized as follows: section 2 outlays the section about related work, section 3 deliberates over the suggested methodology, section 4 presents the results and discussion, and section 5 represents the conclusion of the proposed framework.

2. LITERATURE SURVEY

Seyyar *et al.* [10] defined the BERT model for classifying text attacks to assist various text-related applications. In this study, HTTP requests were considered to detect genuine and malicious texts effectively. Moreover, six fully connected (FC) layers of multilayer perceptron (MLP) were utilized to classify the adversarial texts. In the experimental part, accuracy and F-measure were analyzed and distinguished from other studies. However, the long-term dependency problems were unsolved for larger documents.

Liu *et al.* [11], put forth a secure text similarity protocol for malicious text classification attacks in the DL model. Here, the elliptic-curve cryptography (ECC) technique was introduced to enhance the model execution efficiency. Then, the malicious behavior of the semi-honest protocol was examined and combined with zero-knowledge-proof and cut-choose schemes. In the experimental part, accuracy and execution time were analyzed and distinguished from other studies. However, this method was highly sensitive to word length and increased error during the training process.

Zhang *et al.* [12], established the DL-based adversarial text classification technique using a virtual training process. For word embedding, bag-of-words (BoW) was utilized, performing vectorization over each database. The Elec, IMDB, and Rotten-based third benchmark datasets were used for the training process. In the experimental part, accuracy, and loss were analyzed and distinguished from other studies. However, this method causes high black-box issues and lacks its interpretability over unstructured text data.

Bajaj and Vishwakarma [13], a hostile attack protocol for outputting text vulnerabilities over DL-based sentiment classifiers. Various popular NLP-based DL models like convolutional neural network (CNN) and long short-term memory (LSTM) along with five different transformer methods were utilized. Moreover, the MR and IMDB-based two benchmark datasets were used for the training process. In the experimental part, accuracy, sensitivity, and run-time were analyzed and distinguished from other studies. However, recent NLP models like ROBERTa and XLNet failed to consider for classifying adversarial texts.

Bao *et al.* [14], introduced a score level network for detecting hostile texts accurately. Here, the class-aware score network (CASN) model was emphasized to identify the text over adversarial training. Moreover, the cosine similarity was performed to denoise the unwanted text data. The SST-1, SST-2, IMDB, and AGNEWS-based four benchmark datasets were used for the training process. In the experimental part,

area under the receiver operating characteristic curve (AUROC), and F-measue were analyzed and distinguished from other studies. However, the time complexity was highly likely to cause high overfitting issues.

Meanwhile, deep neural networks (DNNs)-based text classification is becoming increasingly significant in today's information analysis and comprehension. For example, sentiment analysis of user reviews and comments is a key component of many online recommendation systems [15]. These kinds of algorithms would often divide the reviews and comments into two or three groups, then rank the movies or products based on the results. Text classification plays a crucial role in improving the safety of online discussion spaces. For example, it can be used to automatically identify online toxic content [16], which includes insults, sarcasm, abuse, harassment, and irony. Numerous research works have examined the security of existing machine learning models and have put forth several attack techniques, such as exploratory and causal attacks [17]–[19]. Exploratory attacks create hostile testing cases (adversarial examples) in order to elude a particular classifier, while causative attacks try to modify the training data in order to trick the classifier itself. Numerous methods have been put forth to produce robust classifiers in order to fend off these attacks [20], [21]. Adversarial assaults have demonstrated a high attack success rate in image classification tasks recently [22], which has put many intelligent devices such as self-driving cars in grave danger [23], [24].

Research gaps in TextBugger: an extended adversarial text attack on NLP-based text classification models present several opportunities for exploration. One key area is the robustness of models against more sophisticated adversarial attacks. TextBugger has demonstrated vulnerabilities in text classification models, but further research is needed to explore more complex and context-aware perturbations. Such advanced attacks could exploit deeper linguistic features, requiring models to be equipped with stronger defenses capable of recognizing subtle changes in adversarial inputs.

Another significant research gap lies in developing defense mechanisms specifically tailored to textual data. While TextBugger exposes weaknesses in existing NLP models, the study of effective defense strategies remains underdeveloped. Techniques like adversarial training, noise-injection, and certified robustness have been explored in vision models but need further refinement and testing in the NLP domain, particularly in handling diverse text structures and meanings.

The cross-lingual and multi-task vulnerabilities of NLP models under adversarial attacks also warrant further investigation. TextBugger primarily focuses on English text, leaving open questions about how adversarial attacks impact models that operate in multiple languages or perform various tasks like sentiment analysis and named entity recognition. Research in this area can provide insights into the generalization and transferability of adversarial vulnerabilities across linguistic boundaries.

Another gap relates to the transferability of adversarial examples. While TextBugger showcases vulnerabilities in specific models, it remains unclear how transferable these adversarial attacks are across different architectures, particularly in modern transformer-based models like BERT and GPT. Exploring the cross-model transferability of adversarial attacks can help understand how to build more robust architectures that can defend against a wider array of threats.

Furthermore, human perceptibility and semantic preservation is another important area for future research. Although TextBugger aims to create adversarial examples that remain imperceptible to humans, the extent to which these attacks preserve the original meaning and coherence of the text requires further evaluation. Studies are needed to assess the balance between attack success and the preservation of semantics, especially for more complex NLP tasks where maintaining meaning is crucial.

The real-world applicability of TextBugger-style attacks also requires further research. Evaluating how adversarial text manipulations impact real-time applications, such as spam detection, fake news moderation, and content filtering systems, is essential. Understanding the behavior of these attacks in practical settings, especially those with human-in-the-loop systems or multiple layers of filtering, can shed light on potential defense mechanisms and system vulnerabilities.

Moreover, adversarial attacks on transformer-based models remain relatively unexplored. TextBugger's analysis focuses primarily on traditional NLP models, but with the increasing adoption of transformers, there is a pressing need to understand how resilient these newer models are to adversarial attacks. Research into extending TextBugger's methodology to transformer-based architectures like BERT, GPT, and T5 will provide insights into the robustness of state-of-the-art models.

Lastly, there is a gap in understanding attack generalization across various NLP tasks beyond text classification. The versatility of adversarial attacks, such as TextBugger, in other domains like machine translation, text summarization, or question-answering systems remains largely unexplored. Investigating how these attacks generalize to more complex and diverse NLP tasks will help identify more comprehensive defense strategies. By addressing these research gaps, advancements can be made in building more secure and resilient NLP models, which are crucial for the reliable deployment of AI-driven systems in real-world applications.

Problem statement: From the deep analysis of the conventional studies, it is noted that the vulnerability of these models has failed to mitigate in terms of harmful hostile attacks. The minor changes in input texts can lead to inaccurate classifications. These adversarial attacks weaken the model and cause serious consequences like the transmission of manipulated data or vulnerability to automated techniques. Recently, several challenges have been faced to identify effective techniques that can accurately secure text attacks and enhance the reliability of the text classification process. Nowadays, NLP models are playing an integral role in several text-related applications that maintain their popularity even though larger samples are processed. Hence, this article investigated various NLP models in text classification attacks over original examples.

3. PROPOSED METHOD

The proposed article introduced an extended text adversarial generation method, TextBugger. Initially, preprocessing steps such as keyword selection (KS) are performed to remove noises from the text data. Then, various NLP models like BERT, ROBERTa, and XLNet models are analyzed for outputting hostile texts. Figure 1 indicates the workflow of the developed framework.

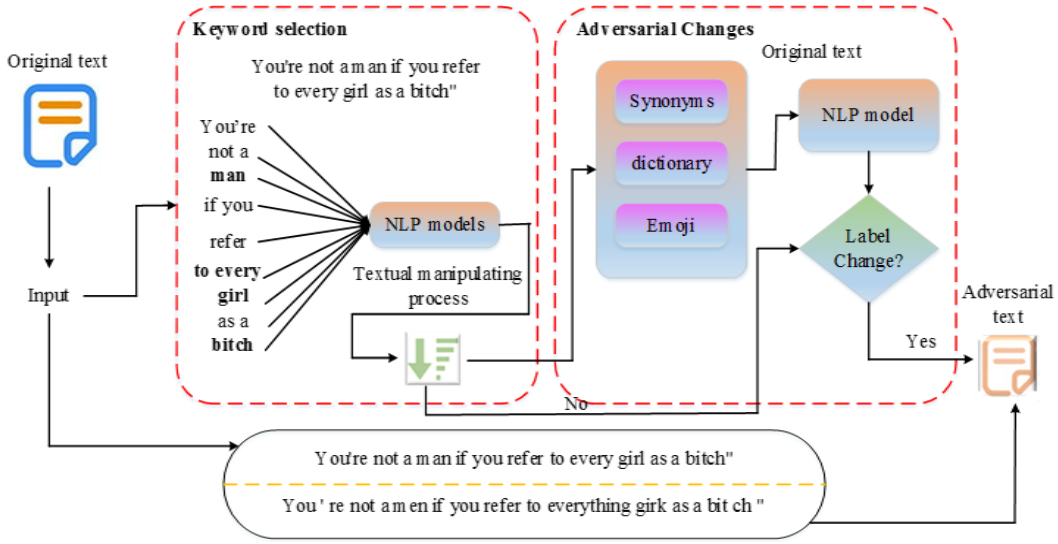

Figure 1. Workflow of the developed framework

3.1. Preprocessing stage

Initially, the raw text data collected from public sources are preprocessed by performing the Tokenization process. The detailed analysis of each stage is depicted below.

3.1.1. Keyword selection

It is the process of separating the textual data into minute units (keywords) that can easily recognize the text attacks accurately. An example of the KS process is conquered below in Table 1.

Table 1. A few examples of the KS process

Input	KS process
On a mission to find some zebra cakes	On a mission, mission to find, to find, some zebra, zebra cakes
This bitch had horseradish ponytail today she. dyed her bald head ass hair red and put that bitch in a ponytail smh	This bitch had, had horseradish ponytail, pony tail she. dyed her, bald head ass hair red, and put that, bitch in a ponytail smh

3.2. Text classification attacks on different NLP models

The selected keywords are then fed into the different NLP models like BERT, ROBERTa, and XLNet models to analyze their performance on the text classification attack process. The detailed analysis of the different NLP models is depicted below.

3.2.1. BERT-based NLP model

In the BERT technique, the synonyms of the word in a given sentence are manipulated based on other words adjacent to it. The BERT model provides all the input in a single duration to solve the long-term dependencies between words and it is of two types: BERT base and BERT large model. In the BERT base technique, a total of twelve transformer encoders are present for the training process. In the BERT large technique, a total of twenty-four transformer encoders are present. Here, the tuning process is very easy and provides outstanding classification performance. The following steps are performed in the BERT-based NLP model for the text classification attack process:

- Separate the collected text data based on training and testing sets using the train-test split process.
- Transform the training set based on corresponding Python tensors for the NLP technique.
- Determine the batch size to generate tensors repeatedly to enhance the BERT technique.
- Train the BERT using the network parameters and analyze the success rate performance. The outcome of BERT-NLP model is depicted in Table 2. Figure 2 indicates the architecture of BERT model.

Table 2. Adversarial text outcome from the BERT-NLP model

Original input	Adversarial outcome
Modern day singers talk about the same shit rappers talk about lol....hoes	Modern day singers talk about the same sit rappers talk about lol....hKes
You had to throw in the faggot word smh smh smh	You had to toss in tte faoggt wod snh snh snh

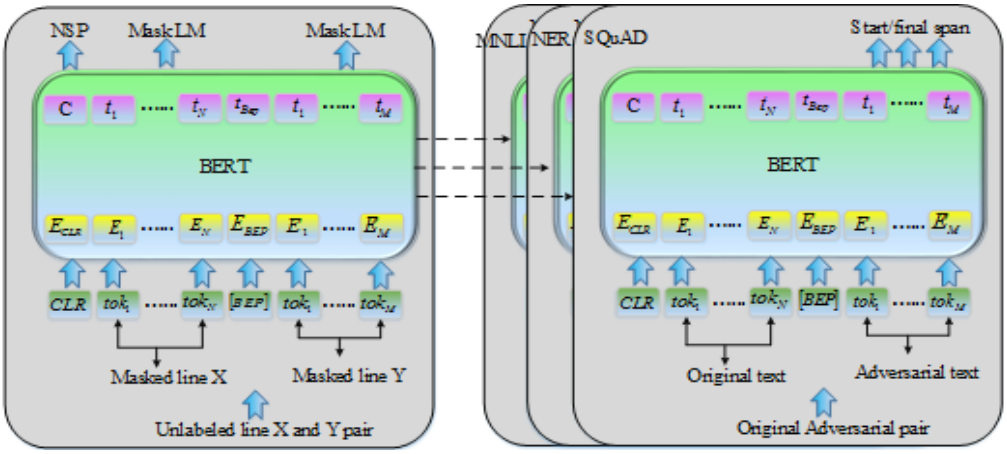

Figure 2. Architecture of BERT model

3.2.2. ROBERTa-based NLP model

The ROBERTa technique is the improved version of the BERT scheme and aids in solving long-term dependency problems during the training process. As like BERT model, the ROBERTa model also uses the transformers thta consist of three elements: heads, transformers, and tokenizer. The transformers convert the sparse data into contextual embedding's for depth-level training. The head covers the transformer that assists the contextual embedding for upcoming training process. The tokenizer assists in altering original text into index sparse encodings. The ROBERTa uses the byte-pair character-level encodings capable of training larger text data over 50,000 subset units. Apart from this, the ROBERTa model fine-tunes more effectively compared to BERT models. The following steps are performed in the ROBERTa-based NLP model for the text classification attack process:

- Initially, the actual text data is tokenized into sub-words so the word embedding are encoded easily. A specialized token such as <s> and </s> to represent the starting and ending word sequence. Moreover, <pad> token assisting text padding to increase the length of word vector.
- For text learning, the words are converted into useful numerical interpretation. The tokenizer encodes the actual text into an attention mask (deliberates the presents and absence of tokens for the training process) and text IDs (contains token index and token numerical interpretation).
- The text IDs and attention masks are then fed into the ROBERTa scheme that consist of 12 base layers, more than 120 million parameters and 768 hidden vectors that creates useful word embedding as the feature engineering. The outcome of ROBERTa-NLP model is depicted in Table 3.

Table 3. Adversarial text outcome from the ROBERT-NLP model

Original input	Adversarial outcome
Modern day singers talk about the same shit rappers talk about lol....hoes	Modern day singers talk about the same siht rappers talk about lol....hookers
Wtf was drake asking us to pull over so he can get my autograph bitch	Wtf was drake asking us to pull over so he can get my autograph bithc

3.2.3. XLNet-based NLP model

The XLNet utilizes the property of permutation language model (PLM) to integrate the pros of autoregressive (AR), and autoencoder (AE). The AR model acts as a decoder of transformer and process the present data to classify the corresponding outcome. In the AE technique, BERT model is utilized where the particular words of the input text are masked and the outcome is retained. The tokens are arranged dynamically in PLM in a sentence format and utilize AE to detect final few tokens. While detecting the token, dual token information is utilized and understand the dependency among the tokens.

Moreover, XLNet implements the recursive mechanism and authorized position encoding in the transformer. XLNet store the hidden unit sequence during every permutation and the authorized position encoding is balanced between various permutations. Due to the use of transformers, it can enhance the extracted features by utilizing the pros of NLP over larger texts. Because of the aforementioned property of XLNet, it can completely indicate every token based on semantic representationin Table 4.

Table 4. Adversarial text outcome from the XLNet-NLP model

Original input	Adversarial outcome
Modern day singers talk about the same shit rappers talk about lol....hoes	Modern day singers talk about the same poop rappers talk about lol....ducklings
Wtf was drake asking us to pull over so he can get my autograph bitch	Wtf was drake asking us to pull over so he can egt my autograph bicth

4. RESULTS AND DISCUSSIONS

The developed method is processed and analyzed via the Python platform. For the simulation process, a text classification attack benchmark database (TCAB) [25] is utilized which consists of different adversarial attacks on traditional text classification models trained on various sentiments and abusive domain contents. In the training part, 552,364 samples are considered clean, and the remaining as unperturbed data. For the testing process, 178,607 samples are considered clean, and the remaining as unperturbed texts. Various performance analyses like accuracy, KC, F-measure, and PPV are computed and compared with different NLP models.

4.1. Assessment metrics

$$Success\_rate = \frac{w+x}{w+x+y+z} \quad (1)$$

$$F_1 - score = 2 \times \left(\frac{Pr\,ecision \times Re\,call}{Pr\,ecision + Re\,call}\right) \quad (2)$$

$$Kappa_{coeffective} = \frac{2\times(x\times w - y\times z)}{(x+z)(z+w)+(x+y)(y+w)} \quad (3)$$

$$PPV(\%) = \frac{x}{x+z} \quad (4)$$

Here, w, x, y, z indicates the true negative (TN), true positive (TP), false negative (FN), and false positive (FP) respectively.

4.2. Comparative analysis of developed method over conventional techniques

In this section, the outcomes achieved by various NLP models in producing hostile texts are analyzed by assessing success rate, F-measure, KC, time consumption, and PPV metrics. The detailed analysis of the obtained outcomes is conquered below. Figure 3 depicts the overall accuracy and loss analysis of the NLP models. The NLP models like BERT, ROBERTa, and XLNet models are trained and tested for classifying hostile attacks on input texts. From the graphical interpretation, it is noted that the NLP models outperform well by minimizing losses during the training, and testing process. Table 5 tabulates the

comparative analysis of different NLP models. While analyzing the performance of different NLP models, ROBERTa model outperforms better in terms of success rate, and time consumption.

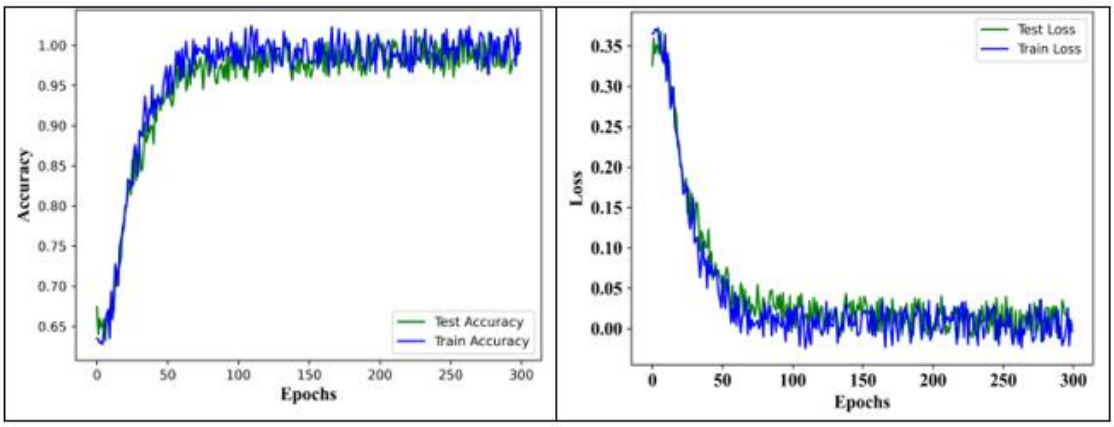

Figure 3. Overall accuracy and loss analysis

Table 5. Comparative analysis of different NLP models

Methods	Success rate (%)	F-measure (%)	KC (%)	PPV (%)	Time consumption (s)
ROBERTa	99.7	99.65	98.90	99.68	106.28
BERT	98.6	98.59	97.87	98.6	2184.08
XLNet	96.8	96.66	95.45	95.92	7691.018

4.3. Practical impacts of TextBugger

An extended adversarial text attack on NLP-based text classification models are substantial and multifaceted. Firstly, TextBugger highlights the vulnerabilities of NLP models to adversarial attacks, significantly raising awareness about the need for enhanced security measures. This newfound awareness drives researchers and practitioners to address these weaknesses and develop more robust models that can resist such manipulative inputs. TextBugger also presents challenges related to maintaining semantic integrity in the face of adversarial examples. The attack's ability to generate text modifications that remain semantically similar to the original content highlights the need for methods that can detect and mitigate such subtle manipulations without compromising the model's performance or understanding.

Finally, the exploration of adversarial text attacks by TextBugger may drive cross-disciplinary research efforts. By integrating insights from NLP, cybersecurity, and artificial intelligence, it fosters a comprehensive approach to developing solutions that enhance the overall security framework for text classification systems. This interdisciplinary collaboration can lead to more effective and resilient security measures in NLP applications. Overall, TextBugger's practical impacts are significant, leading to improved model security, refined evaluation metrics, better model design, and enhanced defenses in real-world applications. Its findings drive advancements in creating robust and secure NLP systems, addressing key challenges and fostering cross-disciplinary research.

5. CONCLUSION

The developed method investigated various existing NLP models to classify adversarial texts on original examples. Common textBuggers like BERT, ROBERTa, and XLNet models are analyzed by inputting actual texts for the training process. The extensive simulation is carried out in the publicly available TCAB dataset to analyze these models. The outcomes of this simulation proved that the ROBERTa-based textbugger model is highly effective and fast. To prove the robustness of developed scheme, other existing approaches are also experimented with in terms of success rate, time consumption, PPV, F-measure, and KC. The simulation process is carried out in the Python platform and the overall success rate achieved by BERT, ROBERTa, and XLNet is about 98.6%, 99.7%, and 96.8% respectively. However, the developed scheme failed to consider other DL models like Bi-LSTM, CNN, and LSTM models to classify adversarial texts based on input texts. In future studies, other DL models are also considered and their performance will be analyzed by inputting various text examples. An extended adversarial text attack on NLP-based text

classification models should focus on several key areas. Advancing attack strategies to exploit deeper NLP model features, developing more effective and tailored defense mechanisms, and understanding the impact of adversarial attacks on cross-lingual and multi-task models are crucial. Research should also explore the transferability of adversarial examples across different architectures, examine how these attacks affect text readability and semantic integrity, and assess their real-world applicability in systems like automated content moderation and sentiment analysis.

ACKNOWLEDGEMENTS

The author with a deep sense of gratitude would thank the supervisor for his guidance and constant support rendered during this research.

FUNDING INFORMATION

No funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Sanjaikanth E. Vadakkethil Somanathan Pillai		✓	✓		✓		✓	✓		✓			✓	
Srinivas A. Vaddadi		✓	✓	✓	✓	✓		✓	✓	✓		✓	✓	
Rohith Vallabhaneni	✓			✓	✓	✓	✓			✓	✓		✓	
Santosh Reddy Addula	✓	✓	✓	✓			✓	✓	✓	✓				
Bhuvanesh Ananthan	✓		✓		✓	✓			✓		✓	✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author, [S.E.V.S.P], upon reasonable request.

REFERENCES

[1] E. B.-Astudillo, W. Fuertes, S. S.-Gordon, D. N.-Agurto, and G. R.-Galán, "A phishing-attack-detection model using natural language processing and deep learning," *Applied Sciences (Switzerland)*, vol. 13, no. 9, 2023, doi: 10.3390/app13095275.

[2] B. He, M. Ahamad, and S. Kumar, "Petgen: personalized text generation attack on deep sequence embedding-based classification models," in *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, Aug. 2021, pp. 575–584, doi: 10.1145/3447548.3467390.

[3] I. Alsmadi *et al.*, "Adversarial attacks and defenses for social network text processing applications: techniques, challenges and future research directions," *Arxiv*, 2021, [Online]. Available: http://arxiv.org/abs/2110.13980.

[4] F. Marulli, L. Verde, and L. Campanile, "Exploring data and model poisoning attacks to deep learning-based NLP systems," *Procedia Computer Science*, vol. 192, pp. 3570–3579, 2021, doi: 10.1016/j.procs.2021.09.130.

[5] S. Uplenchwar, V. Sawant, P. Surve, S. Deshpande, and S. Kelkar, "Phishing attack detection on text messages using machine learning techniques," in *2022 IEEE Pune Section International Conference (PuneCon)*, Dec. 2022, pp. 1–5, doi: 10.1109/PuneCon55413.2022.10014876.

[6] Z. Zhou, H. Guan, M. Bhat, and J. Hsu, "Fake news detection via NLP is vulnerable to adversarial attacks," in *Proceedings of the 11th International Conference on Agents and Artificial Intelligence*, 2019, vol. 2, pp. 794–800, doi: 10.5220/0007566307940800.

[7] H. Ali *et al.*, "All your fake detector are belong to us: evaluating adversarial robustness of fake-news detectors under black-box settings," *IEEE Access*, vol. 9, pp. 81678–81692, 2021, doi: 10.1109/ACCESS.2021.3085875.

[8] X. Li, L. Chen, and D. Wu, "Turning attacks into protection: social media privacy protection using adversarial attacks," in *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*, Philadelphia, PA: Society for Industrial and Applied Mathematics, 2021, pp. 208–216.
[9] J. R. Asl, M. H. Rafiei, M. Alohaly, and D. Takabi, "A semantic, syntactic, and context-aware natural language adversarial example generator," *IEEE Transactions on Dependable and Secure Computing*, vol. 21, no. 5, pp. 4754–4769, Sep. 2024, doi: 10.1109/TDSC.2024.3359817.
[10] Y. E. Seyyar, A. G. Yavuz, and H. M. Unver, "An attack detection framework based on BERT and deep learning," *IEEE Access*, vol. 10, pp. 68633–68644, 2022, doi: 10.1109/ACCESS.2022.3185748.
[11] X. Liu *et al.*, "Secure computation protocol of text similarity against malicious attacks for text classification in deep-learning technology," *Electronics*, vol. 12, no. 16, p. 3491, Aug. 2023, doi: 10.3390/electronics12163491.
[12] W. Zhang, Q. Chen, and Y. Chen, "Deep learning based robust text classification method via virtual adversarial training," *IEEE Access*, vol. 8, pp. 61174–61182, 2020, doi: 10.1109/ACCESS.2020.2981616.
[13] A. Bajaj and D. K. Vishwakarma, "HOMOCHAR: a novel adversarial attack framework for exposing the vulnerability of text based neural sentiment classifiers," *Engineering Applications of Artificial Intelligence*, vol. 126, p. 106815, Nov. 2023, doi: 10.1016/j.engappai.2023.106815.
[14] R. Bao, R. Zheng, L. Ding, Q. Zhang, and D. Tao, "CASN: class-aware score network for textual adversarial detection," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2023, vol. 1, pp. 671–687, doi: 10.18653/v1/2023.acl-long.40.
[15] W. Medhat, A. Hassan, and H. Korashy, "Sentiment analysis algorithms and applications: a survey," *Ain Shams Engineering Journal*, vol. 5, no. 4, pp. 1093–1113, Dec. 2014, doi: 10.1016/j.asej.2014.04.011.
[16] C. Nobata, J. Tetreault, A. Thomas, Y. Mehdad, and Y. Chang, "Abusive language detection in online user content," in *Proceedings of the 25th International Conference on World Wide Web*, Apr. 2016, pp. 145–153, doi: 10.1145/2872427.2883062.
[17] M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar, "The security of machine learning," *Machine Learning*, vol. 81, no. 2, pp. 121–148, Nov. 2010, doi: 10.1007/s10994-010-5188-5.
[18] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar, "Can machine learning be secure?," in *Proceedings of the 2006 ACM Symposium on Information, computer and communications security*, Mar. 2006, vol. 2006, pp. 16–25, doi: 10.1145/1128817.1128824.
[19] L. Huang, A. D. Joseph, B. Nelson, B. I. P. Rubinstein, and J. D. Tygar, "Adversarial machine learning," in *Proceedings of the 4th ACM workshop on Security and artificial intelligence*, Oct. 2011, pp. 43–58, doi: 10.1145/2046684.2046692.
[20] B. Biggio, G. Fumera, and F. Roli, "Design of robust classifiers for adversarial environments," in *2011 IEEE International Conference on Systems, Man, and Cybernetics*, Oct. 2011, pp. 977–982, doi: 10.1109/ICSMC.2011.6083796.
[21] D. Sculley, G. M. Wachman, and C. E. Brodley, "Spam filtering using inexact string matching in explicit feature space with on-line linear classifiers," *NIST Special Publication*, 2006.
[22] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (SP)*, May 2017, pp. 39–57, doi: 10.1109/SP.2017.49.
[23] I. Evtimov *et al.*, "Robust physical-world attacks on machine learning models," *Computer Vision and Pattern Recognition*, 2017, [Online]. Available: http://arxiv.org/abs/1707.08945.
[24] J. Gao, J. Lanchantin, M. Lou Soffa, and Y. Qi, "Black-box generation of adversarial text sequences to evade deep learning classifiers," in *2018 IEEE Security and Privacy Workshops (SPW)*, May 2018, pp. 50–56, doi: 10.1109/SPW.2018.00016.
[25] A. Kalyani *et al.*, "TCAB: text classification attack benchmark dataset," *Zenodo*, 2022. https://zenodo.org/records/7226519.

BIOGRAPHIES OF AUTHORS

Sanjaikanth E. Vadakkethil Somanathan Pillai (Senior Member, IEEE) holds an MS in software engineering from The University of Texas at Austin, Texas, USA, and a BE from the University of Calicut, Kerala, India. Currently pursuing a Ph.D. in computer science at the University of North Dakota, Grand Forks, North Dakota, USA, his research spans diverse areas such as mobile networks, network security, privacy, location-based services, and misinformation detection. He is a proud member of Sigma Xi, The Scientific Research Honor Society, underlining his commitment to advancing scientific knowledge and research excellence. He can be contacted at email: s.evadakkethil@und.edu.

Srinivas A. Vaddadi is a dynamic and forward-thinking professional in the field of Cloud and DevSecOps. With a solid educational foundation in computer science, he embarked on a journey of continuous learning and professional growth. Their relentless pursuit of knowledge and commitment to staying at the forefront of industry advancements has earned them recognition as a thought leader in the Cloud and DevSecOps space. He can be contacted at email: vsad93@gmail.com.

Dr. Rohith Vallabhaneni is a dedicated worker with a strong work ethic in leading teams to solve organizational issues. He is capable of learning all aspects of information within a company and using the technical knowledge and business background to effectively analyze security measures to determine their effectiveness in order to strengthen the overall security posture. He has great work ethic and outstanding team leadership skills and seek to accomplish organizational goals, while growing in knowledge and experience. He can be contacted at email: rohit.vallabhaneni.2222@gmail.com.

Santosh Reddy Addula a senior member of IEEE, is a research scholar at the University of the Cumberlands. His educational qualifications include a Ph.D. and a Master of Science in information technology. With extensive experience in the IT industry, he has demonstrated expertise across multiple domains. He is an innovator who has made significant contributions to academic research through his articles as an author and co-author. Additionally, he serves as a reviewer for esteemed journals, demonstrating his commitment to advancing knowledge and upholding high standards in scholarly publications within his field. He can be contacted at email: santoshaddulait@gmail.com.

Dr. Bhuvanesh Ananthan received the B.E. degree in electrical and electronics engineering from Anna University in 2012, M.Tech. in power system engineering from Kalasalingam University in 2014 and Ph.D. degree from Faculty of Electrical Engineering of Anna University in 2019. He has published more than 100 papers in reputed international journals, 75 papers in international conferences and 20 books. He is a life time member of International Society for Research and Development, International Association of Engineers. He can be contacted at email: bhuvanesh.ananthan@gmail.com.