Proceedings of CONF-SPML 2026 Symposium: The 2nd Neural Computing and Applications Workshop 2025
DOI: 10.54254/2755-2721/2026.TJ30298

Analysis of the Effectiveness of Deep Learning Spam Email Classifiers against Text-Based Attacks

Jianchao Li

School of Electrical Engineering, Guangzhou Nanfang College, Guangzhou, China
outlook_165D168DEE309A11@outlook.com

Abstract. With the widespread application of machine learning, particularly deep learning models, in the field of cybersecurity, the intelligence of spam filtering systems has been continuously enhanced. Deep learning classifiers, with their advantages such as character-level feature learning and semantic invariance, have become the preferred choice for deployment. However, these models rely on surface text features, making them vulnerable to adversarial attacks. As a result, they exhibit significant vulnerability when facing carefully constructed text adversarial attacks. Text adversarial attacks, through covert modifications such as synonym substitution and character perturbation, can mislead the model to misjudge malicious emails, leading to risky spam emails such as phishing and fraud passing through the defense system. This study first elaborates on three attack methods, namely character-level attack, word-level attack, and sentence-level attack. Secondly, it introduces the existing limitations of spam email attacks and then this study comprehensively reviews the key findings in the existing research results: deep learning models generally have a high attack success rate (ASR). The aim is to provide a theoretical basis for building a more robust next-generation spam email filtering system.

Keywords: Deep Learning, Spam Emails, Character Set Attack, Word-level Attack

1. Introduction

Nowadays, with the development of technology, traditional paper-based communication and interaction have gradually been replaced by email, which is convenient to carry and takes less time to convey information. Due to the soaring usage rate of email, some criminals send a large number of emails (spam emails) to users without their permission. Spam emails usually consume users' time and energy, and sometimes are used for fraudulent activities. The senders of spam emails package the emails as reputable companies in order to take advantage of the recipient's trust and disclose personal privacy information, such as personal bank account numbers and payment passwords, thereby causing serious economic losses. Due to the serious social harm caused by spam emails, spam email filters have emerged. With the advent of machine learning, especially deep learning models, not only being used in the field of image processing but also widely applied in natural language processing and cybersecurity fields, especially in the detection field [1], the intelligence of spam email filtering systems has been increasingly enhanced. Deep learning classifiers, with their advantages such as character-level feature learning and semantic invariance, have become the

DOI: 10.54254/2755-2721/2026.TJ30298

preferred choice for deployment. However, such models show significant vulnerability when facing carefully constructed text adversarial attacks. This paper selects an evaluation metric, namely the attack success rate (ASR), as the core metric to comprehensively quantify the attack effect and concealment. Through experimental verification of three mainstream attack techniques at the character level, word level, and sentence level, combined with the test results of typical deep learning classifiers such as shallow CNN and dense.

2. Spam classification based on character-level attacks

Character-level attacks, as a relatively easy-to-implement attack method, do not require users to understand the underlying working principle of the training data. They merely require users to make low-level modifications to characters, such as inserting placeholder symbols (like underscores), spelling mistakes, or using words with similar pronunciations to deceive online detection models [2], and even by reversing characters [3]. Current defenses against character-level attacks mainly include recognition defenses relying on powerful word recognition modules, where this module assumes that attackers follow unrealistic rules and does not allow simple operations such as inserting or deleting spaces [4]. Secondly, character frequency statistics, by calculating the occurrence frequency of special characters in the text, trigger warnings when the threshold is exceeded. Next is spelling correction filtering, by integrating spelling check tools such as BERT-based correction models. Yonatan Belinkov and Yonatan Bisk demonstrated that character-level machine translation models are sensitive to natural character-level perturbations (spelling mistakes) and adversarial selection perturbations [5]. Both the email model and the machine translation model belong to text models, so this study first used character-level attacks to attack the email model. The mechanism of character-level attacks modifies at the character level, simulating typing errors. Boucher et al. proposed character set attacks, which can bypass human eye detection with their minimal perturbation, thereby manipulating the output of various natural language processing (NLP) systems. They obtained that character-level attacks pose a significant threat to text NLP systems, reducing the performance of vulnerable models by a significant margin with just one attack, and the performance of most models would be severely damaged after three attacks [6]. Gao et al. used the Deepwordbug method in a black-box situation to attack two models with Enron spam data sets, namely the word-based LSTM and the character-based CNN (Char-CNN). This method used the scoring function created in this study to mark important word elements, and then used four methods of exchanging characters, replacing characters, deleting characters, and inserting characters to attempt to create minor perturbations to affect classification. It respectively reduced the performance of the word-based LSTM model by 68% and the performance of the Char-CNN model by 48% [7]. This indicates that minor perturbations on characters can bypass the recognition of shallow neural networks. This precisely highlights the vulnerability of lightweight classifiers.

3. Based on word-level attacks

The attack methods of word-level attacks include word form transformation, replacing the keyword elements (which refer to those words that, when modified, will cause the classifier to make incorrect classifications) with their synonyms [8]. In the research by Ding et al., they argued that character set attacks would generate unnatural sentences [9]. Such sentences might be more easily detected, so this paper analyzed word-level attacks to detect the attack success rate of word-level attacks on deep learning classifiers. Word-level attacks rely on the method of replacing words with semantic roots. Such an attack method can find more suitable candidate replacement words, and has significant

DOI: 10.54254/2755-2721/2026.TJ30298

superiority compared to word replacement methods based on synonyms or word vectors [10]. In addition, there is also the attack method of word embedding perturbation. Now, the defense against word-level attacks includes identifying the synonyms of high-threat words, and the next step is the mixed module defense. Liao et al. constructed a word-level attack method based on gradient adaptive word embedding perturbation (AG WEP), which increased the misclassification rate of the CNN classifier by 42%. This study used a GloVe word embedding model with context awareness to convert the words in spam emails into a machine-readable format. Although this method would disrupt the grammar of the sentence, the core original information of the spam email was still retained [11]. Such attacks would have smaller perturbations than character set attacks, thereby increasing their concealment. It can more effectively verify the significant vulnerability of deep learning classifiers when facing carefully constructed text adversarial attacks. This phenomenon indicates that word-level attacks are more concealed and less recognizable by deep learning classifiers than character-level attacks. The detection of deep learning classifiers may be comparative semantics, and it does not focus on the keywords, but it will consider this as a benign email and send it to the user.

4. Based on sentence-level attacks

Sentence-level attacks are basically created by inserting or replacing clauses, and even modifying the grammar while maintaining the original meaning of the sentence [3]. The changed sentences may have incorrect meanings or grammar, which makes such attacks easily detectable. Although sentence-level attacks are less covert than the above two attack methods, they mainly manifest in generating unnatural sentences with low sentence fluency and sometimes causing incorrect representations. To verify the vulnerability of deep learning classifiers, in addition to the above two attack methods, this paper also employs sentence-level attacks. Sentence-level attacks mainly focus on operations on entire sentences or text fragments. Attackers may introduce grammatical errors, syntactic anomalies, or semantic inconsistencies to affect the model's detection or analysis. Sentence-level attacks include grammatical errors, syntactic anomalies, or semantic inconsistencies, etc. [12]. Hotoˇglu et al. increased the false detection rate of the dense classifier by 30% for real datasets such as Enron and SpamAssassin by adding new sentences and optimizing the scoring function [13]. The success rate of sentence-level attacks is significantly lower than that of the above two attacks, indicating that the perturbations applied by sentence-level attacks are higher than those at the character and word levels.

5. Current limitations and future prospects

Most of the experiments on text adversarial attacks regarding spam emails rely on publicly available datasets for verification. Although these datasets are well-labeled, they suffer from common scenarios and lack significant content differences, all being in the same language. Nowadays, spam emails have exhibited characteristics such as multi-language mixture, polymorphic integration (text combined with malicious attachments), and dynamic changes in themes (AI generating a large number of images, texts, etc., highly realistic content, and dynamically adjusting the email structure to bypass static feature-based detection rules). Secondly, the attack methods discussed in this paper are overly idealized. Most of the cited experiments are conducted in a white-box environment, without simulating the real-world situation where attackers have no knowledge of the model's information. Moreover, they do not simulate the continuous adjustment of attack methods by attackers based on different defense systems in the real world. Additionally, although semantic

DOI: 10.54254/2755-2721/2026.TJ30298

similarity indicators are introduced in the evaluation system, there is a lack of human subjective assessment verification. Machines may have a high semantic similarity for some synonyms, but they ignore human individual habits and thinking patterns, and do not test the recognition rate of human users for these samples. However, real users may detect anomalies due to differences in word usage habits, resulting in the actual success rate of the attack being lower than the model's assessment results. This disconnection between machine assessment and human perception makes the conclusion on the concealment of the attack lack practical significance.

To address these issues, it is urgent to establish a dataset that better matches the current attack scenarios. Such a dataset should meet three requirements: Firstly, it should have multiple language mixed samples, including more than one language and the minimum language proportion should not be lower than 10%. Additionally, a sample of multiple types of integration should be designed, such as text combined with OCR text or QR codes. Moreover, dynamic attack trajectories should be recorded and labeled, and the process of the attacker adjusting their attack methods based on different defense models should be simulated, such as switching from character perturbation to word embedding replacement. The real-time threat intelligence of international anti-spam organizations can be relied on to update the sample library in real time to ensure the data is real-time and effective. In addition to building a new dataset, a human subjective indicator should also be added. That is, a human subjective assessment of the concealment of adversarial samples. Besides these two points, it is necessary to focus on breaking through lightweight model-specific defense technologies, including: (1) lightweight adversarial training, reducing training costs by freezing the bottom-level parameters and using virtual adversarial samples (without the need for real labels); (2) dynamic feature enhancement, integrating text statistical features (such as the frequency of spam words) and semantic features (such as sentence vectors) in real-time during inference to improve the discrimination of adversarial samples; (3) federated robust learning, sharing adversarial sample gradients between user terminal devices (without revealing the original data), reducing the ASR of edge device models.

6. Conclusions

This paper, through discussion and theoretical analysis, concludes that text adversarial attacks pose a serious threat to deep learning spam email classifiers. The performance of deep learning models, specifically the word-based LSTM model and the Char-CNN model, decreased by 68% and 48% respectively under character-level attacks. Word-level attacks could increase the misclassification rate of the CNN classifier by 42%. The evaluation indicators constructed and the vulnerability mechanisms revealed in this study not only fill the gap in the research on the adversarial robustness of deep learning text classifiers, but also provide theoretical support and practical paths for the security design of the next-generation spam email filtering system. In the future, it is necessary to further break through the limitations of datasets, model coverage, and evaluation systems, and promote deep learning classifiers to move from laboratory datasets to practical, efficient, and deployable solutions in real network environments, to build the first line of defense in network security.

References

[1] Lin, Z., Liu, Z., & Fan, H. (2025). Improving Phishing Email Detection Performance of Small Large Language Models. arXiv preprint arXiv: 2505.00034.

[2] Eger, S., & Benz, Y. (2020). From Hero to Z\'eroe: A Benchmark of Low-Level Adversarial Attacks. arXiv preprint arXiv: 2010.05648.

DOI: 10.54254/2755-2721/2026.TJ30298

[3] Chen, X., Salem, A., Chen, D., Backes, M., Ma, S., Shen, Q., ... & Zhang, Y. (2021, December). Badnl: Backdoor attacks against nlp models with semantic-preserving improvements. In Proceedings of the 37th Annual Computer Security Applications Conference (pp. 554-569).
[4] Li, L., Ma, R., Guo, Q., Xue, X., & Qiu, X. (2020). Bert-attack: Adversarial attack against bert using bert. arXiv preprint arXiv: 2004.09984.
[5] Belinkov, Y., & Bisk, Y. (2017). Synthetic and natural noise both break neural machine translation. arXiv preprint arXiv: 1711.02173.
[6] Boucher, N., Shumailov, I., Anderson, R., & Papernot, N. (2022, May). Bad characters: Imperceptible nlp attacks. In 2022 IEEE Symposium on Security and Privacy (SP) (pp. 1987-2004). IEEE.
[7] Gao, J., Lanchantin, J., Soffa, M. L., & Qi, Y. (2018, May). Black-box generation of adversarial text sequences to evade deep learning classifiers. In 2018 IEEE Security and Privacy Workshops (SPW) (pp. 50-56). IEEE.
[8] Gao, J., Lanchantin, J., Soffa, M. L., & Qi, Y. (2018, May). Black-box generation of adversarial text sequences to evade deep learning classifiers. In 2018 IEEE Security and Privacy Workshops (SPW) (pp. 50-56). IEEE.
[9] Jin, D., Jin, Z., Zhou, J. T., & Szolovits, P. (2019). Is bert really robust? natural language attack on text classification and entailment. arXiv preprint arXiv: 1907.11932, 2(10).
[10] Zang, Y., Qi, F., Yang, C., Liu, Z., Zhang, M., Liu, Q., & Sun, M. (2019). Word-level textual adversarial attacking as combinatorial optimization. arXiv preprint arXiv: 1910.12196.
[11] Gregory, J., & Liao, Q. (2023, September). Adversarial spam generation using adaptive gradient-based word embedding perturbations. In 2023 IEEE International Conference on Artificial Intelligence, Blockchain, and Internet of Things (AIBThings) (pp. 1-5). IEEE.
[12] Huq, A., & Pervin, M. (2020). Adversarial attacks and defense on texts: A survey. arXiv preprint arXiv: 2005.14108.
[13] Hotoğlu, E., Sen, S., & Can, B. (2025). A Comprehensive Analysis of Adversarial Attacks against Spam Filters. arXiv preprint arXiv: 2505.03831.