[4] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, “Badnets: Evaluating back-

dooring attacks on deep neural networks,” IEEE Access, vol. 7, pp. 47230–

47244, 2019, doi: 10.1109/ACCESS.2019.2909068.

[5] Y. Liu et al., “Trojaning attack on neural networks,” in Proc. 25th

Annu. Netw. Distrib. Syst. Secur. Symp., 2018. [Online]. Available:

http://wp.internetsociety.org/ndss/wp-content/uploads/sites/25/2018/02/

ndss2018\_03A-5\_Liu\_paper.pdf

[8] F. Qi et al., “Hidden killer: Invisible textual backdoor attacks with syn-

tactic trigger,” in Proc. 59th Annu. Meeting Assoc. Comput. Linguis-

tics 11th Int. Joint Conf. Natural Lang. Process., 2021, pp. 443–453,

doi: 10.18653/v1/2021.acl-long.37.

[9] F. Qi, Y. Chen, X. Zhang, M. Li, Z. Liu, and M. Sun, “Mind the style of

text! adversarial and backdoor attacks based on text style transfer,” in Proc.

Conf. Empirical Methods Natural Lang. Process., 2021, pp. 4569–4580,

doi: 10.18653/v1/2021.emnlp-main.374.

[7] F. Qi, Y. Chen, M. Li, Y. Yao, Z. Liu, and M. Sun, “ONION: A simple

and effective defense against textual backdoor attacks,” in Proc. 2021

Conf. Empirical Methods Natural Lang. Process., 2021, pp. 9558–9566,

doi: 10.18653/v1/2021.emnlp-main.752.

[14] S. Li et al., “Hidden backdoors in human-centric language mod-

els,” in Proc. ACM SIGSAC Conf. Comput. Commun. Secur., 2021,

pp. 3123–3140, doi: 10.1145/3460120.3484576.

[15] S. Hochreiter and J. Schmidhuber, “Long short-term mem-

ory,” Neural Comput., vol. 9, no. 8, pp. 1735–1780, 1997,

doi: 10.1162/neco.1997.9.8.1735.

[17] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, “Targeted backdoor attacks

on deep learning systems using data poisoning,” 2017, arXiv:1712.05526.

[18] Y. Liu, X. Ma, J. Bailey, and F. Lu, “Reflection backdoor: A natural back-

door attack on deep neural networks,” in Proc. 16th Eur. Conf. Comput.

Vis., 2020, pp. 182–199, doi: 10.1007/978-3-030-58607-2\_11.

[19] A. Turner, D. Tsipras, and A. Madry, “Label-consistent backdoor attacks,”

2019, arXiv:1912.02771.

[20] S. Cheng, Y. Liu, S. Ma, and X. Zhang, “Deep feature space trojan attack

of neural networks by controlled detoxification,” in Proc. 35th AAAI Conf.

Artif. Intell., 33rd Conf. Innov. Appl. Artif. Intell., 11th Symp. Educ. Adv.

Artif. Intell., 2021, pp. 1148–1156. [Online]. Available: https://ojs.aaai.

org/index.php/AAAI/article/view/16201

[21] J. Dumford and W. J. Scheirer, “Backdooring convolutional neural net-

works via targeted weight perturbations,” in Proc. IEEE Int. Joint Conf.

Biometrics, 2020, pp. 1–9, doi: 10.1109/IJCB48548.2020.9304875.

[22] A. S. Rakin, Z. He, and D. Fan, “TBT: Targeted neural network attack

with bit trojan,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.,

Jun. 2020, pp. 13195–13204.

[23] Y. Li, J. Hua, H. Wang, C. Chen, and Y. Liu, “Deeppayload: Black-box

backdoor attack on deep learning models through neural payload injec-

tion,” in Proc. IEEE/ACM 43rd Int. Conf. Softw. Eng., 2021, pp. 263–274,

doi: 10.1109/ICSE43902.2021.00035.

[24] J. Dai, C. Chen, and Y. Li, “A backdoor attack against LSTM-based text

classification systems,” IEEE Access, vol. 7, pp. 138872–138878, 2019,

doi: 10.1109/ACCESS.2019.2941376.

[25] K. Kurita, P. Michel, and G. Neubig, “Weight poisoning attacks on pre-

trained models,” in Proc. 58th Annu. Meeting Assoc. Comput. Linguistics,

2020, pp. 2793–2806, doi: 10.18653/v1/2020.acl-main.249.

[27] X. Pan, M. Zhang, B. Sheng, J. Zhu, and M. Yang, “Hidden trigger back-

door attack on NLP models via linguistic style manipulation,” in Proc. 31st

USENIX Secur. Symp., 2022, pp. 3611–3628. [Online]. Available: https:

//www.usenix.org/conference/usenixsecurity22/presentation/pan-hidden

[28] C. Chen and J. Dai, “Mitigating backdoor attacks in LSTM-based text clas-

sification systems by backdoor keyword identification,” Neurocomputing,

vol. 452, pp. 253–262, 2021, doi: 10.1016/j.neucom.2021.04.105.

[29] C. Fan et al., “Defending against backdoor attacks in natural language

generation,” 2021, arXiv:2106.01810.

[30] K. Shao, J. Yang, Y. Ai, H. Liu, and Y. Zhang, “BDDR: An effective

defense against textual backdoor attacks,” Comput. Secur., vol. 110, 2021,

Art. no. 102433, doi: 10.1016/j.cose.2021.102433.