[6] John Brooke et al. 1996. SUS-A quick and dirty usability scale. *Usability evaluation in industry* 189, 194 (1996), 4–7.

[7] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *EuroS&P*. IEEE, 39–57.

[8] Minhao Cheng, Jinfeng Yi, Pin-Yu Chen, Huan Zhang, and Cho-Jui Hsieh. 2018. Seq2Sick: Evaluating the Robustness of Sequence-to-Sequence Models with Adversarial Examples. (2018). arXiv:cs.LG/1803.01128

[9] Gregory Cohen, Saeed Afshar, Jonathan Tapson, and André van Schaik. 2017. EMNIST: an extension of MNIST to handwritten letters. *arXiv preprint arXiv:1702.05373* (2017).

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[11] Antreas Dionysiou, Michalis Agathocleous, Chris Christodoulou, and Vasilis Promponas. 2018. Convolutional Neural Networks in Combination with Support Vector Machines for Complex Sequential Data Classification. In *ICANN*. Springer, 444–455.

[12] Antreas Dionysiou and Elias Athanasopoulos. 2020. SoK: Machine vs. Machine–A Systematic Classification of Automated Machine Learning-Based CAPTCHA Solvers. *Computers & Security* (2020), 101947.

[13] Antreas Dionysiou, Vassilis Vassiliades, and Elias Athanasopoulos. 2021. *HoneyGen: Generating Honeywords Using Representation Learning*. Association for Computing Machinery, New York, NY, USA, 265–279. https://doi.org/10.1145/3433210.3453092

[14] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. HotFlip: White-Box Adversarial Examples for Text Classification. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (2018), 31–36.

[15] Anthony Y Fu, Xiaotie Deng, Liu Wenyin, and Greg Little. 2006. The methodology and an application to fight against unicode attacks. In *SOUPS*. 91–101.

[16] Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yanjun Qi. 2018. Black-box generation of adversarial text sequences to evade deep learning classifiers. In *SPW*. IEEE, 50–56.

[17] Zhitao Gong, Wenlu Wang, Bo Li, Dawn Song, and Wei-Shinn Ku. 2018. Adversarial texts with gradient methods. *arXiv preprint arXiv:1801.07175* (2018).

[18] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2014. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (2014).

[19] Daniel Holden, Jun Saito, Taku Komura, and Thomas Joyce. 2015. Learning motion manifolds with convolutional autoencoders. In *SIGGRAPH*. ACM, 18.

[20] Hossein Hosseini, Sreeram Kannan, Baosen Zhang, and Radha Poovendran. 2017. Deceiving google's perspective api built for detecting toxic comments. *arXiv preprint arXiv:1702.08138* (2017).

[21] Minqing Hu and Bing Liu. 2004. Mining and summarizing customer reviews. In *SIGKDD*. ACM, 168–177.

[22] Ling Huang, Anthony D Joseph, Blaine Nelson, Benjamin IP Rubinstein, and JD Tygar. 2011. Adversarial machine learning. In *AISec*. ACM, 43–58.

[23] Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. *in EMNLP* (2017), 2021–2031.

[24] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. FastText.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651* (2016).

[25] Yoon Kim. 2014. Convolutional Neural Networks for Sentence Classification. *in EMNLP* (10 2014), 1746–1751.

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *NeurIPS*. 1097–1105.

[27] Brandon Laughlin, Christopher Collins, Karthik Sankaranarayanan, and Khalil El-Khatib. 2019. A Visual Analytics Framework for Adversarial Text Generation. *arXiv preprint arXiv:1909.11202* (2019).

[28] Yann LeCun, Yoshua Bengio, et al. 1995. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks* 3361, 10 (1995), 1995.

[29] Yann LeCun, LD Jackel, Leon Bottou, A Brunot, Corinna Cortes, JS Denker, Harris Drucker, I Guyon, UA Muller, Eduard Sackinger, et al. 1995. Comparison of learning algorithms for handwritten digit recognition. In *ICANN*, Vol. 60. Perth, Australia, 53–60.

[30] Jinfeng Li, Shouling Ji, Tianyu Du, Bo Li, and Ting Wang. 2019. TextBugger: Generating Adversarial Text Against Real-world Applications. *in NDSS* (2019). https://doi.org/10.14722/ndss.2019.23138

[31] Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi. 2018. Deep Text Classification Can Be Fooled. *Proceedings of the 27th International Joint Conference on Artificial Intelligence* (2018), 4208–4215.

[32] Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *49th ACL: Human language technologies*. ACL, 142–150.

[33] Bo Pang and Lillian Lee. 2004. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts. In *42nd ACL*. ACL, 271–es.

[34] Bo Pang and Lillian Lee. 2005. Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales. In *43rd ACL*. ACL, 115–124.

[35] Bo Pang and Lillian Lee. 2008. Opinion Mining and Sentiment Analysis. *Foundations and Trends in Information Retrieval* 2 (2008), 1–135.

[36] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. 2017. Practical Black-Box Attacks Against Machine Learning. In *ASIACCS*. ACM, New York, NY, USA, 506–519.

[37] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. 2016. The limitations of deep learning in adversarial settings. In *EuroS&P*. IEEE, 372–387.

[38] Nicolas Papernot, Patrick McDaniel, Ananthram Swami, and Richard Harang. 2016. Crafting adversarial input sequences for recurrent neural networks. In *MILCOM*. IEEE, 49–54.

[39] Sanglee Park and Jungmin So. 2020. On the Effectiveness of Adversarial Training in Defending against Adversarial Example Attacks for Image Classification. *Applied Sciences* 10, 22 (2020), 8079.

[40] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*. 1532–1543.

[41] Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C Duchi, and Percy Liang. 2019. Adversarial training can hurt generalization. *arXiv preprint arXiv:1906.06032* (2019).

[42] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1985. *Learning internal representations by error propagation*. Technical Report. California Univ San Diego La Jolla Inst for Cognitive Science.

[43] Suranjana Samanta and Sameep Mehta. 2017. Towards crafting text adversarial samples. *arXiv preprint arXiv:1707.02812* (2017).

[44] David Sculley, Gabriel Wachman, and Carla E Brodley. 2006. Spam Filtering Using Inexact String Matching in Explicit Feature Space with On-Line Linear Classifiers.. In *TREC*.

[45] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).

[46] Han Xu, Yao Ma, Haochen Liu, Debayan Deb, Hui Liu, Jiliang Tang, and Anil Jain. 2019. Adversarial attacks and defenses in images, graphs and text: A review. *arXiv preprint arXiv:1909.08072* (2019).

[47] Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *NIPS*. 649–657.

[48] Ye Zhang and Byron Wallace. 2017. A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification. In *IJCNLP*. AFNLP, Taipei, Taiwan, 253–263.

[49] Zhengli Zhao, Dheeru Dua, and Sameer Singh. 2018. Generating natural adversarial examples. *in ICLR* (2018).