

## REFERENCES

- [1] Sahar Abdelnabi and Mario Fritz. 2021. Adversarial Watermarking Transformer: Towards Tracing Text Provenance with Data Hiding. In *42nd IEEE Symposium on Security and Privacy, SP 2021, San Francisco, CA, USA, 24-27 May 2021*. IEEE, 121–140. <https://doi.org/10.1109/SP40001.2021.00083>
- [2] Satanejeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 65–72.
- [3] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeswar, Sherjil Ozair, Yoshua Bengio, R. Devon Hjelm, and Aaron C. Courville. 2018. Mutual Information Neural Estimation. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholm, Sweden, July 10-15, 2018 (Proceedings of Machine Learning Research, Vol. 80)*. PMLR, 530–539.
- [4] J.T. Brassil, Steven Low, N.F. Maxemchuk, and Lawrence O’Gorman. 1995. Electronic Marking and Identification Techniques to Discourage Document Copying. *Selected Areas in Communications, IEEE Journal on* 13 (11 1995), 1495 – 1504. <https://doi.org/10.1109/49.464718>
- [5] Chang Ching-Yun and Clark Stephen. 2010. Practical linguistic steganography using contextual synonym substitution and vertex colour coding. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. 1194–1203.
- [6] Ingemar Cox, Matthew Miller, Jeffrey Bloom, Jessica Fridrich, and Ton Kalker. 2007. *Digital watermarking and steganography*. Morgan kaufmann.
- [7] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 4171–4186. <https://doi.org/10.18653/v1/n19-1423>
- [8] Abbas El Gamal and Young-Han Kim. 2011. *Network Information Theory*. Cambridge university press.
- [9] Josh A. Goldstein, Girish Sastry, Micah Musser, Renee DiResta, Matthew Gentzel, and Katerina Sedova. 2023. Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations. *CoRR abs/2301.04246* (2023). <https://doi.org/10.48550/arXiv.2301.04246>
- [10] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*. 2672–2680.
- [11] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.).
- [12] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A Watermark for Large Language Models. *CoRR abs/2301.10226* (2023). <https://doi.org/10.48550/arXiv.2301.10226>
- [13] Xingyuan Liang and Shijun Xiang. 2020. Robust reversible audio watermarking based on high-order difference statistics. *Signal Process.* 173 (2020), 107584. <https://doi.org/10.1016/j.sigpro.2020.107584>
- [14] Aiwei Liu, Leyi Pan, Xuming Hu, Shuang Li, Lijie Wen, Irwin King, and Philip S. Yu. 2024. An Unforgeable Publicly Verifiable Watermark for Large Language Models. In *The 20th International Conference on Learning Representations (ICLR)*.
- [15] 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 *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 19-24 June, 2011, Portland, Oregon, USA*, Dekang Lin, Yuji Matsumoto, and Rada Mihalcea (Eds.). The Association for Computer Linguistics, 142–150. <https://aclanthology.org/P11-1015/>
- [16] Hasan Mesut Meral, Bülent Sankur, A Sumru Özsoy, Tunga Güngör, and Emre Sevinç. 2009. Natural language watermarking via morphosyntactic alterations. *Computer Speech & Language* 23, 1 (2009), 107–125.
- [17] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer Sentinel Mixture Models. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.
- [18] Mehdi Mirza and Simon Osindero. 2014. Conditional Generative Adversarial Nets. *CoRR abs/1411.1784* (2014). <http://arxiv.org/abs/1411.1784>
- [19] Pierre Moulin and Joseph A. O’Sullivan. 2003. Information-theoretic analysis of information hiding. *IEEE Trans. Inf. Theory* 49, 3 (2003), 563–593. <https://doi.org/10.1109/TIT.2002.808134>
- [20] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 3980–3990. <https://doi.org/10.18653/v1/D19-1410>
- [21] Yunpeng Ren, Ziao Wang, Yiyuan Wang, and Xiaofeng Zhang. 2021. Generating Long Financial Report using Conditional Variational Autoencoders with Knowledge Distillation. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 15879–15880.
- [22] Mercan Topkara, Umut Topkara, and Mikhail J. Atallah. 2006. Words are not enough: sentence level natural language watermarking. In *Proceedings of the 4th ACM international workshop on Contents protection and security*. 37–46.
- [23] Umut Topkara, Mercan Topkara, and Mikhail J. Atallah. 2006. The Hiding Virtues of Ambiguity: Quantifiably Resilient Watermarking of Natural Language Text through Synonym Substitutions. In *Proceedings of the 8th Workshop on Multimedia and Security (Geneva, Switzerland) (MM&Sec ’06)*. Association for Computing Machinery, New York, NY, USA, 164–174. <https://doi.org/10.1145/1161366.1161397>
- [24] Liang Wang, Wei Zhao, Ruoyu Jia, Sujian Li, and Jingming Liu. 2019. Denoising based Sequence-to-Sequence Pre-training for Text Generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (Eds.). Association for Computational Linguistics, 4001–4013. <https://doi.org/10.18653/v1/D19-1412>
- [25] Lingyun Xiang, Yan Li, Wei Hao, Peng Yang, and Xiaobo Shen. 2018. Reversible natural language watermarking using synonym substitution and arithmetic coding. *Computers, Materials & Continua* 55, 3 (2018).
- [26] Xi Yang, Jie Zhang, Kejiang Chen, Weiming Zhang, Zehua Ma, Feng Wang, and Nenghai Yu. 2022. Tracing Text Provenance via Context-Aware Lexical Substitution. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022*. AAAI Press, 11613–11621.
- [27] KiYoon Yoo, Wonhyuk Ahn, Jiho Jang, and Nojun Kwak. 2023. Robust Natural Language Watermarking through Invariant Features. *CoRR abs/2305.01904* (2023). <https://doi.org/10.48550/arXiv.2305.01904>
- [28] Jiren Zhu, Russell Kaplan, Justin Johnson, and Li Fei-Fei. 2018. HiDDeN: Hiding Data With Deep Networks. In *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XV (Lecture Notes in Computer Science, Vol. 11219)*. Springer, 682–697. [https://doi.org/10.1007/978-3-030-01267-0\\_40](https://doi.org/10.1007/978-3-030-01267-0_40)