

DAFTAR PUSTAKA

- “ Ozde G ” Ul, G., & ,ahin, S. (2022). To Augment or Not to Augment? A Comparative Study on Text Augmentation Techniques for Low-Resource NLP under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International (CC BY-NC-ND 4.0) license. *Computational Linguistics*. <https://doi.org/10.1162/COLI-3-Permendikbud-Nomor-6-Tahun-2022>. (t.t.).
- Aryanto, R., Alfian Rosid, M., & Busono, S. (2023). Penerapan Deep Learning untuk Pengenalan Tulisan Tangan Bahasa Aksara Lota Ende dengan Menggunakan Metode Convolutional Neural Networks. *Jurnal Informasi dan Teknologi*, 258–264. <https://doi.org/10.37034/jidt.v5i1.313>
- Baharuddin, F., & Naufal, M. F. (2023). Fine-Tuning IndoBERT for Indonesian Exam Question Classification Based on Bloom’s Taxonomy. *Journal of Information Systems Engineering and Business Intelligence*, 9(2), 253–263. <https://doi.org/10.20473/jisebi.9.2.253-263>
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (t.t.). *Language Models are Few-Shot Learners*. Diambil <https://commoncrawl.org/the-data/>
- Bülbül, M. A., Harirchian, E., Işık, M. F., Aghakouchaki Hosseini, S. E., & Işık, E. (2022). A Hybrid ANN-GA Model for an Automated Rapid Vulnerability Assessment of Existing RC Buildings. *Applied Sciences (Switzerland)*, 12(10). <https://doi.org/10.3390/app12105138>
- Chollet, F. (t.t.). *M A N N I N G*.
- Devlin, J., Chang, M.-W., Lee, K., Google, K. T., & Language, A. I. (t.t.). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Diambil <https://github.com/tensorflow/tensor2tensor>

- Dhole, Y. B., Rampravesh Sharma, D., Patil, Dr. S., & Chavan, Prof. D. (2024). Unveiling Market Sentiments: Finbert-Powered Analysis of Stock News Headlines. *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, 08(10), 1–6. <https://doi.org/10.55041/ijsrem38254>
- Emanuel, R. H. K., Docherty, P. D., Lunt, H., & Möller, K. (2024). The effect of activation functions on accuracy, convergence speed, and misclassification confidence in CNN text classification: a comprehensive exploration. *Journal of Supercomputing*, 80(1), 292–312. <https://doi.org/10.1007/s11227-023-05441-7>
- Fajri, F., Tutuko, B., Sukemi, S., Studi, P., Informasi, S., Sains, F., Teknologi, D., Kunci, K., & Bert, : (t.t.). *Membandingkan Nilai Akurasi BERT dan DistilBERT pada Dataset Twitter*. 8(2), 71–80.
- Fajri, F., Tutuko, B., Sukemi, S., Studi, P., Informasi, S., Sains, F., Teknologi, D., Kunci, K., & Bert, : (2022). Membandingkan Nilai Akurasi BERT dan DistilBERT pada Dataset Twitter. *Jurnal Sistem Informasi*, 8(2), 71–80.
- Francisca Adoma, A., Henry, N., & Chen, W. (2020). COMPARATIVE ANALYSES OF BERT, ROBERTA, DISTILBERT, AND XLNET FOR TEXT-BASED EMOTION RECOGNITION. *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*. [https://doi.org/10.1109/ICCWAMTIP51612.2020.9317379/20/\\$31.00](https://doi.org/10.1109/ICCWAMTIP51612.2020.9317379/20/$31.00)
- Franke, M., & Degen, J. (t.t.). *The softmax function: Properties, motivation, and interpretation* *.
- Gemini Team, Georgiev, P., Lei, V. I., Burnell, R., Bai, L., Gulati, A., Tanzer, G., Vincent, D., Pan, Z., Wang, S., Mariooryad, S., Ding, Y., Geng, X., Alcober, F., Frostig, R., Omernick, M., Walker, L., Paduraru, C., Sorokin, C., ... Vinyals, O. (2024). *Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context*. <http://arxiv.org/abs/2403.05530>

- Guo, Z., & Le Nguyen, M. (t.t.). *Document-level Neural Machine Translation Using BERT as Context Encoder*. Diambil <https://github.com/sameenmaruf/selective-attn>
- Hegghammer, T. (2022). OCR with Tesseract, Amazon Textract, and Google Document AI: a benchmarking experiment. *Journal of Computational Social Science*, 5(1), 861–882. <https://doi.org/10.1007/s42001-021-00149-1>
- Jayoma, J. M., Moyon, E. S., & Morales, E. M. O. (2020, Desember 3). OCR Based Document Archiving and Indexing Using PyTesseract: A Record Management System for DSWD Caraga, Philippines. *2020 IEEE 12th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment, and Management, HNICEM 2020*. <https://doi.org/10.1109/HNICEM51456.2020.9400000>
- Jessica Angelina, S., Bijaksana Putra Negara, A., Muhardi, H., Nawawi, J. H., & Barat, K. (2023). *Analisis Pengaruh Penerapan Stopword Removal Pada Performa Klasifikasi Sentimen Tweet Bahasa Indonesia Analyzing The Impact Of Applying Stopword Removal On Indonesian Tweet Sentiment Classification*. 02(1). <https://doi.org/10.26418/juara.v2i1.69680>
- Jurafsky, D., & Martin, J. H. (t.t.). *Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models Third Edition draft*.
- Khadivi, P., Tandon, R., & Ramakrishnan, N. (2016). *Flow of Information in Feed-Forward Deep Neural Networks*. <http://arxiv.org/abs/1603.06220>
- Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications*, 82(3), 3713–3744. <https://doi.org/10.1007/s11042-022-13428-4>

- Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (t.t.). *IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP*. Online. Diambil <https://huggingface.co/>
- Labd, Z., Bahassine, S., Housni, K., Aadi, F. Z. A. H., & Benabbes, K. (2024). Text classification supervised algorithms with term frequency inverse document frequency and global vectors for word representation: A comparative study. *International Journal of Electrical and Computer Engineering*, 14(1), 589–599. <https://doi.org/10.11591/ijece.v14i1.pp589-599>
- Lee, C., Lin, J., Prokop, A., Gopalakrishnan, V., Hanna, R. N., Papa, E., Freeman, A., Patel, S., Yu, W., Huhn, M., Sheikh, A. S., Tan, K., Sellman, B. R., Cohen, T., Mangion, J., Khan, F. M., Gusev, Y., & Shameer, K. (2022). StarGazer: A Hybrid Intelligence Platform for Drug Target Prioritization and Digital Drug Repositioning Using Streamlit. *Frontiers in Genetics*, 13. <https://doi.org/10.3389/fgene.2022.868015>
- Li, J., Sun, A., Han, J., & Li, C. (2022). A Survey on Deep Learning for Named Entity Recognition. Dalam *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING* (Vol. 34, Nomor 1). <https://www.mpi-inf.mpg.de/departments/>
- Markoulidakis, I., Kopsiaftis, G., Rallis, I., & Georgoulas, I. (2021a). Multi-Class Confusion Matrix Reduction method and its application on Net Promoter Score classification problem. *ACM International Conference Proceeding Series*, 412–419. <https://doi.org/10.1145/3453892.3461323>
- Markoulidakis, I., Kopsiaftis, G., Rallis, I., & Georgoulas, I. (2021b). Multi-Class Confusion Matrix Reduction method and its application on Net Promoter Score classification problem. *ACM International Conference Proceeding Series*, 412–419. <https://doi.org/10.1145/3453892.3461323>
- Mursari, L. R., & Wibowo, A. (2021). The Effectiveness of Image Preprocessing on Digital Handwritten Scripts Recognition with The

- Implementation of OCR Tesseract. *Computer Engineering and Applications*, 10(3).
- Nabiilah, G. Z., Alam, I. N., Purwanto, E. S., & Hidayat, M. F. (2024a). Indonesian multilabel classification using IndoBERT embedding and MBERT classification. *International Journal of Electrical and Computer Engineering*, 14(1), 1071–1078. <https://doi.org/10.11591/ijece.v14i1.pp1071-1078>
- Nabiilah, G. Z., Alam, I. N., Purwanto, E. S., & Hidayat, M. F. (2024b). Indonesian multilabel classification using IndoBERT embedding and MBERT classification. *International Journal of Electrical and Computer Engineering*, 14(1), 1071–1078. <https://doi.org/10.11591/ijece.v14i1.pp1071-1078>
- Nadeau, D., & Sekine, S. (t.t.). *A survey of named entity recognition and classification*. Diambil <http://projects.ldc.upenn.edu/gale/>
- Nair, A. R., Singh, R. P., Gupta, D., & Kumar, P. (2024). Evaluating the Impact of Text Data Augmentation on Text Classification Tasks using DistilBERT. *Procedia Computer Science*, 235, 102–111. <https://doi.org/10.1016/j.procs.2024.04.013>
- Naseem, U., Razzak, I., & Eklund, P. W. (2021). A survey of pre-processing techniques to improve short-text quality: a case study on hate speech detection on twitter. *Multimedia Tools and Applications*, 80(28–29), 35239–35266. <https://doi.org/10.1007/s11042-020-10082-6>
- OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., ... Zoph, B. (2024). *GPT-4 Technical Report*. <http://arxiv.org/abs/2303.08774>
- Putra, T. D., Utami, E., & Kurniawan, M. P. (t.t.). *Klasifikasi penderita kanker Paru Paru Menggunakan Algoritma Artificial Neural Network (ANN)*.
- R Documentation. (t.t.). *Confusion Matrices (Contingency Tables)*. <https://search.r->

- project.org/CRAN/refmans/qwraps2/html/confusion_matrix.html.
 Diambil 24 April 2025, dari https://search.r-project.org/CRAN/refmans/qwraps2/html/confusion_matrix.html
- Ramshaw, L. A., & Marcus, M. P. (t.t.). *Text Chunking using Transformation-Based Learning*.
- Raschka, Sebastian., Liu, Yuxi., Mirjalili, Vahid., & Dzhulgakov, Dmytro. (2022). *Machine learning with Pytorch and Scikit-Learn : develop machine learning and deep learning models with scikit-learn and PyTorch*. Packt Publishing, Limited.
- Sarker, I. (2021). Machine Learning: Algorithms, Real-World Applications and Research Directions. *SN Computer Science*, 2. <https://doi.org/10.1007/s42979-021-00592-x>
- Sharma, S. K., Sharma, N. K., & Potter, P. P. (2020). Fusion approach for document classification using random forest and SVM. *Proceedings of the 2020 9th International Conference on System Modeling and Advancement in Research Trends, SMART 2020*, 231–234. <https://doi.org/10.1109/SMART50582.2020.9337131>
- Silva Barbon, R., & Akabane, A. T. (2022a). Towards Transfer Learning Techniques—BERT, DistilBERT, BERTimbau, and DistilBERTimbau for Automatic Text Classification from Different Languages: A Case Study. *Sensors*, 22(21). <https://doi.org/10.3390/s22218184>
- Silva Barbon, R., & Akabane, A. T. (2022b). Towards Transfer Learning Techniques—BERT, DistilBERT, BERTimbau, and DistilBERTimbau for Automatic Text Classification from Different Languages: A Case Study. *Sensors*, 22(21). <https://doi.org/10.3390/s22218184>
- Sporici, D., Cuşnir, E., & Boiană, C. A. (2020). Improving the accuracy of Tesseract 4.0 OCR engine using convolution-based preprocessing. *Symmetry*, 12(5). <https://doi.org/10.3390/SYM12050715>
- Sukarna, R. H., Safira, F., Marzuliyanti, H. D., & Artikel, R. (t.t.). *International Jurnal of Informatics Studi Komparatif Model OCR Berbasis AI untuk Dokumen Cetak dan Tulisan Tangan*.

- Tran, Q. D. L., & Le, A. C. (2023). Exploring Bi-Directional Context for Improved Chatbot Response Generation Using Deep Reinforcement Learning. *Applied Sciences (Switzerland)*, 13(8). <https://doi.org/10.3390/app13085041>
- Vanderplas, J. T. . (2023). *Python data science handbook : essential tools for working with data*. O'Reilly Media, Incorporated.
- Wang, Y., Qu, W., & Ye, X. (2024). *Selecting Between BERT and GPT for Text Classification in Political Science Research*. <http://arxiv.org/abs/2411.05050>
- Wang, Z. (2023). *Probabilistic Linguistic Knowledge and Token-level Text Augmentation*. <http://arxiv.org/abs/2306.16644>
- Wilie, B., Vincentio, K., Winata, G. I., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., & Purwarianti, A. (2020). *IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding*. <http://arxiv.org/abs/2009.05387>
- Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., ... Rush, A. M. (t.t.). *Transformers: State-of-the-Art Natural Language Processing*. Diambil <https://github.com/huggingface/>
- Wu, W., Joloudari, J. H., Jagatheesaperumal, S. K., Rajesh, K. N. V. P. S., Gaftandzhieva, S., Hussain, S., Rabih, R., Haqjoo, N., Nazar, M., Vahdat-Nejad, H., & Doneva, R. (2024). Deep Transfer Learning Techniques in Intrusion Detection System-Internet of Vehicles: A State-of-the-Art Review. Dalam *Computers, Materials and Continua* (Vol. 80, Nomor 2, hlm. 2785–2813). Tech Science Press. <https://doi.org/10.32604/cmc.2024.053037>
- Zhang, X., Ji, Y., Nguyen, C., & Wang, T. (2018). DeepClean: Data cleaning via question asking. *Proceedings - 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA 2018*, 283–292. <https://doi.org/10.1109/DSAA.2018.00039>

Zhou, P., XIE Zhouchen LIN Shuicheng YAN, X., Xie, X., Lin, Z., & Yan, S. (2024). *Towards understanding convergence and generalization of AdamW*. https://ink.library.smu.edu.sg/sis_research