

DAFTAR PUSTAKA

- Amur, Z. H., Kwang Hooi, Y., Bhanbhro, H., Dahri, K., & Soomro, G. M. (2023). Short-Text Semantic Similarity (STSS): Techniques, Challenges and Future Perspectives. In *Applied Sciences (Switzerland)* (Vol. 13, Number 6). MDPI. <https://doi.org/10.3390/app13063911>
- Atoum, I. (2025). Evolving Information Retrieval: From Traditional Models to Emerging Paradigms. *Journal of Advances in Information Technology*, 16(9), 1277–1294. <https://doi.org/10.12720/jait.16.9.1277-1294>
- Bruch, S., Gai, S., & Ingber, A. (2024). An Analysis of Fusion Functions for Hybrid Retrieval. *ACM Transactions on Information Systems*, 42(1), 20:1-20:35. <https://doi.org/10.1145/3596512>
- Cahyawijaya, S., Indra Winata, G., Wilie, B., Vincentio, K., Li, X., Kuncoro, A., Ruder, S., Lim, Z. Y., Bahar, S., Khodra, M. L., Purwarianti, A., & Fung, P. (2021). IndoNLG: Benchmark and Resources for Evaluating Indonesian Natural Language Generation. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 8875–8898. <https://doi.org/10.18653/v1/2021.emnlp-main.699>
- Chang, S., Ahn, G. J., & Park, S. (2024). Improving Performance of Neural IR Models by Using a Keyword-Extraction-Based Weak-Supervision Method. *IEEE Access*, 12, 46851–46863. <https://doi.org/10.1109/ACCESS.2024.3382190>
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, 4171–4186. <https://github.com/tensorflow/tensor2tensor>
- Diana, D. (2023): Indonesian SBERT for Semantic Textual Similarity Task [Online]. Tersedia: <https://huggingface.co/denaya/indoSBERT-large>
- Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P.-E., Lomeli, M., Hosseini, L., & Jégou, H. (2025). The Faiss library. *ArXiv*. <http://arxiv.org/abs/2401.08281>
- Gao, L., Dai, Z., Chen, T., Fan, Z., Van Durme, B., & Callan, J. (2021). Complementing Lexical Retrieval with Semantic Residual Embeddings. *ArXiv Preprint ArXiv:2004.13969 (Updated 2021)*. <http://arxiv.org/abs/2004.13969>
- Hambarde, K. A., & Proenca, H. (2023). Information Retrieval: Recent Advances and beyond. *IEEE Access*, 11, 76581–76604. <https://doi.org/10.1109/ACCESS.2023.3295776>

- Ibrahim, I. (2021). Indonesian News Title Dataset [Online]. Tersedia: <https://www.kaggle.com/datasets/ibamibrahim/indonesian-news-title>
- Järvelin, K., & Kekäläinen, J. (2002). Cumulated Gain-Based Evaluation of IR Techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4), 422–446. <https://doi.org/10.1145/582415.582418>
- Joseph, M. H., & Ravana, S. D. (2024). Reliable Information Retrieval Systems Performance Evaluation: A Review. *IEEE Access*, 12, 51740–51751. <https://doi.org/10.1109/ACCESS.2024.3377239>
- Kodri, W. A. G., Haris, M., & Fitriadi, R. (2025). Fine-Hybrid: Integration of BM25 And Finetuned SBERT to Enhance Search Relevance. *TEKNIKA*, 14(2), 213–222. <https://doi.org/10.34148/teknika.v14i2.1229>
- Kurniasih, A., & Manik, L. P. (2022). On the Role of Text Preprocessing in BERT Embedding-based DNNs for Classifying Informal Texts. *IJACSA International Journal of Advanced Computer Science and Applications*, 13(6), 927–934. www.ijacsa.thesai.org
- Lasminingrum, D. P., Puspaningrum, E. Y., & Mulyo, B. M. (2026). Perbandingan metode lexical dan semantic retrieval : BM25, IndoSBERT baseline, dan IndoSBERT fine-tuned pada pencarian dokumen berbahasa Indonesia. *Prosiding Seminar Nasional Penelitian Dan Pengabdian Kepada Masyarakat (LPPM Universitas 'Aisyiyah Yogyakarta)*., 4, 245–253. <https://repository.upnjatim.ac.id/48873/>
- Manning, C. D. ., Raghavan, Prabhakar., & Schütze, Hinrich. (2008). *Introduction to information retrieval*. Cambridge University Press.
- Merchant, B., Khazi, A., & Sonawane, S. S. (2026, February 24). Reciprocal Rank Fusion Based Hybrid Dense-Sparse Information Retrieval on Code-Mixed Banglish Social Media Text. *CEUR Workshop Proceedings (FIRE 2025: Forum for Information Retrieval Evaluation)*.
- Mutlu, M. A., Ulku, E. E., & Yildiz, K. (2024). A web scraping app for smart literature search of the keywords. *PeerJ Computer Science*, 10. <https://doi.org/10.7717/PEERJ-CS.2384>
- Nomoto, T. (2023). A Survey of Keyword Extraction Methods. *SN Computer Science*, 4(92), 1–19. <https://doi.org/10.1007/s42979-022-01481-7>
- Papagiannopoulou, E., & Tsoumakas, G. (2019). A Review of Keyphrase Extraction. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*. <http://arxiv.org/abs/1905.05044>
- Pimentel, J. F., Murta, L., Braganholo, V., & Freire, J. (2019). A Large-scale Study about Quality and Reproducibility of Jupyter Notebooks. *2019 IEEE/ACM 16th International Conference on Mining Software Repositories (MSR)*.

- Pratiwi, I. Y. R., & Nugraha, A. F. (2022). Hoax news identification using machine learning model from online media in Bahasa Indonesia. *MATRIX: Jurnal Manajemen Teknologi Dan Informatika*, 12(2), 58–67. <https://doi.org/10.31940/matrix.v12i2.58-67>
- Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 3982–3992. <https://github.com/UKPLab/>
- Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4), 333–389. <https://doi.org/10.1561/1500000019>
- Rossi, J., Wirth, M., & Kanoulas, E. (2019). *Query Generation for Patent Retrieval with Keyword Extraction based on Syntactic Features*. 1–10. <http://arxiv.org/abs/1906.07591>
- Rozaq, A., Fitri, R., Akbar, M. S., Sumari, M. H., Khafiyya, N., Wardiah, I., & Sholihin, F. (2025). Legal Literacy in Indonesia: Leveraging Semantic-Based AI and NLP for Enhanced Civil Law Access. *E3S Web of Conferences 622 (INTERCONNECTS 2024)*, 622. <https://doi.org/10.1051/e3sconf/202562203002>
- Rule, A., Birmingham, A., Zuniga, C., Altintas, I., Huang, S.-C., Knight, R., Moshiri, N., Nguyen, M. H., Rosenthal, S. B., Pérez, F., & Rose, P. W. (2019). Ten Simple Rules for Reproducible Research in Jupyter Notebooks. *PLOS Computational Biology*, 15(6). <https://www.github.com>
- Sharma, S., & Panda, S. P. (2023). Efficient information retrieval model: overcoming challenges in search engines-an overview. *Indonesian Journal of Electrical Engineering and Computer Science*, 32(2), 925–932. <https://doi.org/10.11591/ijeecs.v32.i2.pp925-932>
- Shen, Y., He, X., Gao, J., Deng, L., & Mesnil, G. (2014). A latent semantic model with convolutional-pooling structure for information retrieval. *CIKM 2014 - Proceedings of the 2014 ACM International Conference on Information and Knowledge Management*, 101–110. <https://doi.org/10.1145/2661829.2661935>
- Stabler, M., Turnbull, S., Müller, T., Langdon, C., Marx-Gomez, J., & Köster, F. (2025). The Impact of Chunking Strategies on Domain-Specific Information Retrieval in RAG Systems. *IEEE Conference Proceedings*.
- Stankovic, M. (2025). Cross-Document Topic-Aligned Chunking for Retrieval-Augmented Generation. *ArXiv*. <http://arxiv.org/abs/2601.05265>
- Steck, H., Ekanadham, C., & Kallus, N. (2024). Is Cosine-Similarity of Embeddings Really About Similarity? *WWW 2024 Companion - Companion*

- Proceedings of the ACM Web Conference*, 887–890.
<https://doi.org/10.1145/3589335.3651526>
- Stokes, N. (2006). TREC : experiment and evaluation in information retrieval. *Computational Linguistics*, 32, 563–567.
- Syahrizal, H., & Jailani, M. S. (2023). Jenis-Jenis Penelitian Dalam Penelitian Kuantitatif dan Kualitatif. *QOSIM: Jurnal Pendidikan, Sosial & Humaniora*, 1(1), 13–23.
- Thamrin, H., Oktafiani, D., Rasyid, I. I., & Fauzi, I. M. (2024). Classification of SWOT Statements Employing BERT Pre-Trained Model Embedding. *Jurnal Sistem Informasi Bisnis*, 14(2), 143–152.
<https://doi.org/10.21456/vol14iss2pp143-152>
- Ubaidillah, M., Adikara, P. P., & Perdana, R. S. (2025). Pemerolehan Informasi Terjemahan Al-Quran Berbahasa Indonesia Menggunakan IndoSBERT. *Jurnal Pengembangan Teknologi Informasi Dan Ilmu Komputer (J-PTIHK)*, 9(8).
<http://j-ptiik.ub.ac.id>
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention Is All You Need. *31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA USA*.
- Verma, P. (2025). S2 Chunking: A Hybrid Framework for Document Segmentation Through Integrated Spatial and Semantic Analysis. *ArXiv*.
<http://arxiv.org/abs/2501.05485>
- Wang, S., Zhuang, S., & Zucco, G. (2021). BERT-based Dense Retrievers Require Interpolation with BM25 for Effective Passage Retrieval. *ICTIR 2021 - Proceedings of the 2021 ACM SIGIR International Conference on Theory of Information Retrieval*, 317–324. <https://doi.org/10.1145/3471158.3472233>
- Wang, Z., Gao, C., Xiao, C., Huang, Y., Si, S., Luo, K., Bai, Y., Li, W., Duan, T., Lv, C., Lu, G., Chen, G., Qi, F., & Sun, M. (2025). Document Segmentation Matters for Retrieval-Augmented Generation. *Findings of the Association for Computational Linguistics: ACL 2025*, 8063–8075.
- Wijaya, I. N. S. W., Seputra, K. A., & Parwita, W. G. S. (2021). Comparison of the BM25 and rabinkarp algorithm for plagiarism detection. *Journal of Physics: Conference Series*, 1810(1). <https://doi.org/10.1088/1742-6596/1810/1/012032>
- Wu, C., Wu, F., Huang, Y., & Xie, X. (2023). Personalized News Recommendation: Methods and Challenges. *ACM Transactions on Information Systems*, 41(1).
<https://doi.org/10.1145/3530257>
- Wu, F., Qiao, Y., Chen, J.-H., Wu, C., Qi, T., Lian, J., Liu, D., Xie, X., Gao, J., Wu, W., Zhou, M., & Research, M. (2020). MIND: A Large-scale Dataset for News Recommendation. *Proceedings of the 58th Annual Meeting of the Association*

for *Computational Linguistics (ACL 2020)*, 3597–3606.
<https://www.kaggle.com/gspmoreira/news-portal-user->

Zhou, Y., Wang, S., Koopman, B., & Zuccon, G. (2026). Beyond Chunk-Then-Embed: A Comprehensive Taxonomy and Evaluation of Document Chunking Strategies for Information Retrieval. *ArXiv Preprint*.
<http://arxiv.org/abs/2602.16974>

