

## DAFTAR PUSTAKA

- Arifin, Y., Isa, S. M., Wulandhari, L. A., & Abdurachman, E. (2018). Plagiarism Detection for Indonesian Language using Winnowing with Parallel Processing. *Journal of Physics: Conference Series*, 978(1). <https://doi.org/10.1088/1742-6596/978/1/012082>
- Bilgin, M., & Köktaş, H. (2019). Sentiment Analysis with Term Weighting and Word Vectors Article in International Arab Journal of Information Technology. In *The International Arab Journal of Information Technology* (Vol. 16, Issue 5). <https://www.researchgate.net/publication/335827612>
- Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8), 861–874. <https://doi.org/10.1016/j.patrec.2005.10.010>
- Goswami, M., & Babu, A. (2018). *A Comparative Analysis of Similarity Measures to find Coherent Documents*. <http://www.ijamtes.org/gallery/101.%20nov%20ijmte%20-%20as.pdf>
- Gurusamy, V., Kannan, S., & Professor, A. (2014). *Preprocessing Techniques for Text Mining Agent based distributed framework for Big data View project Multilingual Natural Language Processing View project Preprocessing Techniques for Text Mining*. <https://www.researchgate.net/publication/273127322>
- Han, J., Kamber, M., & Pei, J. (2012). Getting to Know Your Data. In *Data Mining* (pp. 39–82). Elsevier. <https://doi.org/10.1016/B978-0-12-381479-1.00002-2>
- Jurafsky, D., & Martin, J. H. (2008). *Speech And Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*.
- Kalyan, K. S., Rajasekharan, A., & Sangeetha, S. (2021). *AMMUS: A Survey of Transformer-based Pretrained Models in Natural Language Processing*. <http://arxiv.org/abs/2108.05542>
- Kamaruddin, S. S., Yusof, Y., Azzah, N., Bakar, A., Tayie, M. A., Abdulsattar, G., & Alkubaisi, A. J. (2018). Graph-based Representation for Sentence

- Similarity Measure : A Comparative Analysis. In *International Journal of Engineering & Technology* (Vol. 7, Issue 2).
- Kim, S.-E., Baek, G.-H., Ahn, K.-H., & Kim, S.-K. (2018). *A METHOD OF SIMILARITY MEASURE WITH DIFFERENT DOMAINS' DOCUMENTS FOR SEMANTIC CLASSIFICATION USING WORD2VEC AND JACCARD INDEX* (Vol. 23).
- Lau, J. H., & Baldwin, T. (2016). *An Empirical Evaluation of doc2vec with Practical Insights into Document Embedding Generation*. <http://arxiv.org/abs/1607.05368>
- Le, Q., & Mikolov, T. (2014). *Distributed Representations of Sentences and Documents*.
- Leskovec, J., Rajaraman, A., & Ullman, J. D. (2014). Finding Similar Items. In *Mining of Massive Datasets* (pp. 68–122). Cambridge University Press. <https://doi.org/10.1017/CBO9781139924801.004>
- Lovins, J. B. (1968). Development of a Stemming Algorithm\*. In *Mechanical Translation and Computational Linguistics* (Vol. 11, Issue 2).
- Mahendra, R., & Setya, K. N. (2018). *Semi-supervised Textual Entailment on Indonesian Wikipedia Data Indonesian Morphological Analysis View project Consumer Health Question Answering System for Indonesian Language View project*. <https://doi.org/10.13140/RG.2.2.18820.27521>
- Marjai, P., Lehotay-Kéry, P., & Kiss, A. (2021). Document similarity for error prediction. *Journal of Information and Telecommunication*, 5(4), 407–420. <https://doi.org/10.1080/24751839.2021.1893496>
- Maxwell-Smith, Z., Kohler, M., & Suominen, H. (2022). *Scoping natural language processing in Indonesian and Malay for education applications*. <https://doi.org/10.18653/v1/2022.acl-srw.15>
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed Representations of Words and Phrases and Their Compositionality. *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, 3111–3119.

- Rajagopal, S., PK, R., & V.L, L. (2019). *Document and Query Similarity Indexing using VSM for Unstructured Documents and Natural Language Queries*. <http://www.jctjournal.com/gallery/117-dec2019.pdf>
- Rehurek, R., & Sojka, P. (2010). Software framework for topic modelling with large corpora. *Proceedings of LREC 2010 Workshop New Challenges for NLP Frameworks*. Valletta, Malta: University of Malta, 46–50.
- Salton, G., & McGill, M. J. (1983). *Introduction to modern information retrieval*. Inc., New York.
- Samsuri, B. N. (2013). *Pers Berkualitas, Masyarakat Cerdas* (B. N. Samsuri, Ed.; 1st ed.). DEWAN PERS, Perpustakaan Nasional RI: Katalog Dalam Terbitan (KDT).
- Scherrer, Y. (2020). *TaPaCo: A Corpus of Sentential Paraphrases for 73 Languages*. 11–16. <https://doi.org/10.5281/zenodo.3707949>
- Sinaga, D. D., & Hansun, S. (2018). Indonesian text document similarity detection system using rabin-karp and confix-stripping algorithms. *International Journal of Innovative Computing, Information and Control*, 14(5), 1893–1903. <https://doi.org/10.24507/ijicic.14.05.1893>
- Singh, R., & Singh, S. (2021). Text Similarity Measures in News Articles by Vector Space Model Using NLP. *Journal of The Institution of Engineers (India): Series B*, 102(2), 329–338. <https://doi.org/10.1007/s40031-020-00501-5>
- Sitikhu, P., Pahi, K., Thapa, P., & Shakya, S. (2019). *A Comparison of Semantic Similarity Methods for Maximum Human Interpretability*. <http://arxiv.org/abs/1910.09129>
- Snir, I. (2021, March 8). *Get the full news story with Full Coverage in Search*. <https://blog.google/products/news/get-full-news-story-full-coverage-search/>
- Sutikno, H. (2021). IMPLEMENTASI ALGORITMA COSINE SIMILARITY UNTUK MENDETEKSI KEMIRIPAN TOPIK JUDUL. In *JECSIT* (Vol. 1, Issue 1).

- Tikhonov, A., Malkhasov, A., Manoshin, A., Dima, G., Cserháti, R., Hossain Asif, M., & Sárdi, M. (2022). *EENLP: Cross-lingual Eastern European NLP Index*. <https://github.com/altsoph/EENLP/tree/main/eval>
- Upstill, T. (2018, May 8). *The new Google News: AI meets human intelligence*. <https://blog.google/products/news/new-google-news-ai-meets-human-intelligence/>
- Wilie, B., Vincentio, K., Indra Winata, G., Cahyawijaya, S., Li, X., Lim, Z. Y., Soleman, S., Mahendra, R., Fung, P., Bahar, S., Purwarianti, A., & Bandung, I. T. (2020). IndoNLU: Benchmark and Resources for Evaluating Indonesian Natural Language Understanding. *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, 843–857.