DAFTAR PUSTAKA

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2019. [Online]. Available: https://arxiv.org/abs/1810.04805

[2] DataReportal, "Digital 2025: Indonesia," Jan. 2025, online; accessed 14 May 2025. [Online]. Available: https://datareportal.com/reports/digital-2025-indonesia

[3] Microsoft Indonesia, "Studi terbaru dari microsoft menunjukkan peningkatan digital civility (keadaban digital) di seluruh kawasan asia pacific selama masa pandemi," February 2021. [Online]. Available: https://bit.ly/studi-terbaru-dari-microsoft

[4] N. Makassar and S. M. Habibah, "Building digital civility of indonesian net citizens from a digital citizenship perspective," *Digital Theory, Culture & Society*, vol. 2, no. 2, pp. 79–86, 2024. [Online]. Available: https://journal.c-dics.com/index.php/dtcs/article/view/30/47

[5] A. Indonesia, "Laporan pemantauan ujaran kebencian terhadap kelompok rentan pada pemilu 2024," Aug. 2024, online; accessed 14 May 2025. [Online]. Available: https://bit.ly/laporan-pemantauan-ujaran-kebencian

[6] K. Mahendra and A. Purwarianti, "Hate speech detection in the indonesian language on twitter using machine learning approaches," in *Proceedings of the International Conference on Electrical Engineering and Informatics*, 2019.

[7] M. O. Ibrohim and I. Budi, "Multi-label hate speech and abusive language detection in indonesian twitter," in *Proceedings of the Third Workshop on Abusive Language Online*. Association for Computational Linguistics, 2019, pp. 46–57. [Online]. Available: https://aclanthology.org/W19-3506.pdf

[8] S. Jahan and M. Oussalah, "Transformers for hate speech detection: A comprehensive survey," *IEEE Access*, vol. 11, pp. 23 456–23 478, 2023.

[9] Z. Fadhli, K. Kurniawan, F. Ikhwantri, and T. Baldwin, "Indolem and indobert: A benchmark dataset and pre-trained language model for indonesian nlp," in *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*. International Committee on Computational Linguistics, 2020, pp. 757–770. [Online]. Available: https://aclanthology.org/2020.coling-main.66

[10] R. Wibowo, I. Sibaroni, and S. W. Manurung, "Deteksi ujaran kebencian bahasa indonesia menggunakan indobert," *Jurnal RESTI*, vol. 5, no. 6, 2021.

[11] S. Dharmawan, V. C. Mawardi, and N. J. Perdana, "Klasifikasi ujaran kebencian pada media sosial twitter menggunakan algoritma feedforward neural network berbasis indobert," *Jurnal Informatika dan Sistem Informasi (JIKSI)*, vol. 4, no. 2, pp. 94–100, 2022. [Online]. Available: https://journal.untar.ac.id/index.php/jiksi/article/view/24066/14543

[12] F. Koto, E. Baldwin, and T. Cohn, "Indobertweet: A pretrained language model for indonesian twitter with effective domain-specific vocabulary initialization," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021.

[13] J. Kusuma and A. Chowanda, "Indobertweet and bilstm for hate speech and abusive language detection in indonesian," *JOIV: International Journal on Informatics Visualization*, no. 4, pp. 290–295, 2022. [Online]. Available: https://joiv.org/index.php/joiv/article/view/1035

[14] Council of Europe, "Recommendation no. r (97) 20 of the committee of ministers to member states on "hate speech"," 1997, adopted by the Committee of Ministers on 30 October 1997. [Online]. Available: https://rm.coe.int/1680505d5b

[15] C. Wardle and H. Derakhshan, "Information disorder: Toward an interdisciplinary framework for research and policy making," Council of Europe report, 2017. [Online]. Available: https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-researc/168076277c

[16] "ujaran kebencian," Kamus Besar Bahasa Indonesia Daring, 2025, "ujaran yang menyerukan kebencian terhadap orang atau kelompok tertentu". [Online]. Available: https://kbbi.kemdikbud.go.id/entri/ujaran%20kebencian

[17] K. N. R. Indonesia, "Surat edaran kapolri nomor se/6/x/2015 tentang penanganan ujaran kebencian (hate speech)," 2015, diakses dari https://berkas.dpr.go.id/pusaka/files/info_singkat/Info%20Singkat-VII-21-I-P3DI-November-2015-28.pdf.

[18] D. RI, "Undang-undang republik indonesia nomor 40 tahun 2008 tentang penghapusan diskriminasi ras dan etnis," 2008, diakses dari https://peraturan.bpk.go.id/Details/39733/uu-no-40-tahun-2008.

[19] I. Gagliardone, D. Gal, T. Alves, and G. Martinez, *Countering online hate speech*. UNESCO, 2015.

[20] J. T. Nockleby, "Hate speech," in *Encyclopedia of the American Constitution*. Macmillan Reference USA, 2000, pp. 1277–1279.

[21] A. M. Kaplan and M. Haenlein, "Users of the world, unite! the challenges and opportunities of social media," *Business horizons*, vol. 53, no. 1, pp. 59–68, 2010.

[22] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed. Prentice Hall, 2020, draft available at: https://web.stanford.edu/ jurafsky/slp3/.

[23] B. Wilie, K. Vincentio, G. I. Winata, S. Cahyawijaya, X. Li, Z. Y. Lim, S. Soleman, R. Mahendra, P. Fung, S. Bahar, and A. Purwarianti, "IndoNLU: Benchmark and resources for evaluating Indonesian natural language understanding," in *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*, K.-F. Wong, K. Knight, and H. Wu, Eds. Suzhou, China: Association for Computational Linguistics, Dec. 2020, pp. 843–857. [Online]. Available: https://aclanthology.org/2020.aacl-main.85/

[24] R. B. Hadiprakoso, H. Setiawan, R. N. Yasa, and Girinoto, "Text preprocessing for optimal accuracy in indonesian sentiment analysis using a deep learning model with word embedding," in *AIP Conference Proceedings*, vol. 2601. AIP, 2021.

[25] F. Kurniawan, "Natural language processing dalam bahasa indonesia," *Jurnal Linguistik Komputasional*, vol. 8, no. 1, pp. 50–60, 2020.

[26] I. Sakti, "Sastrawi: Indonesian stemmer," https://github.com/sastrawi/sastrawi, 2018.

[27] D. M. W. Powers, "Evaluation: From precision, recall and f-measure to roc, informedness, markedness and correlation," *Journal of Machine Learning Technologies*, vol. 2, no. 1, pp. 37–63, 2011.

[28] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427–437, 2009.

[29] H. A. M. Ahmed and A. Ghosh, "Efficacy of imbalanced data handling methods on deep learning based toxic comment classification," *SN Computer Science*, vol. 1, no. 5, pp. 1–10, 2020. [Online]. Available: https://doi.org/10.1007/s42979-020-00211-1