DAFTAR PUSTAKA

Abdelhamid, M., & Desai, A. (2024). Balancing the Scales: A Comprehensive Study on Tackling Class Imbalance in Binary Classification. *Arxiv*. http://arxiv.org/abs/2409.19751

Aditi Pancholi, R. J. D. J. M. P. (2021). A Review Paper on Applications of Natural Language Processing-Transformation from Data-Driven to Intelligence-Driven 1 Aditi Pancholi. *International Journal of Engineering Research & Technology (IJERT)*. www.ijert.org

Bell, S. J., Sánchez, E., Dale, D., Stenetorp, P., Artetxe, M., & Costa-jussà, M. R. (2025). Translate, then Detect: Leveraging Machine Translation for Cross-Lingual Toxicity Classification. *Proceedings of the Tenth Conference on Machine Translation (WMT 2025) Association for Computational Linguistics*, *1*, 253–268. https://doi.org/10.18653/v1/2025.wmt-1.15

Bogatinovski, J., Todorovski, L., Džeroski, S., & Kocev, D. (2021). Comprehensive Comparative Study of Multi-Label Classification Methods. *IEEE*. http://arxiv.org/abs/2102.07113

Cao, K., Wei, C., Gaidon, A., Arechiga, N., & Ma, T. (2019). *Learning Imbalanced Datasets with Label-Distribution-Aware Margin Loss*. http://arxiv.org/abs/1906.07413

Cheng, S. (2024). Dataset Augmentation for Counteracting Bias in Toxic Comment Classification. Dalam *Highlights in Science, Engineering and Technology CSIC* (Vol. 2023).

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., & Stoyanov, V. (2020). Unsupervised Cross-lingual Representation Learning at Scale. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, 8440–8451. https://github.com/facebookresearch/cc

Devlin, J., Chang, M.-W., Lee, K., Google, K. T., & Language, A. I. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL-HLT 2019*. https://github.com/tensorflow/tensor2tensor

Dinarta, F., & Wicaksana, A. (2025). Enhanced Hate Speech Detection in Indonesian-English Code-Mixed Texts Using XLM-RoBERTa. *Informatica (Slovenia)*, *49*(21), 45–56. https://doi.org/10.31449/inf.v49i21.7713

Dr. Sheshang Degadwala, & Dhairya Vyas. (2024). Theoretical Evaluation of Machine Learning and Deep Learning Applications in Various Domain. *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, *10*(3), 567–575. https://doi.org/10.32628/cseit24103208

El-Alami, F. zahra, Ouatik El Alaoui, S., & En Nahnahi, N. (2022). A multilingual offensive language detection method based on transfer learning from transformer fine-tuning model. *Journal of King Saud University - Computer and Information Sciences*, *34*(8), 6048–6056. https://doi.org/10.1016/j.jksuci.2021.07.013

Esposito, C., Landrum, G. A., Schneider, N., Stiefl, N., & Riniker, S. (2021). GHOST: Adjusting the Decision Threshold to Handle Imbalanced Data in Machine Learning. *Journal of Chemical Information and Modeling*, *61*(6), 2623–2640. https://doi.org/10.1021/acs.jcim.1c00160

Groenendijk, R., Karaoglu, S., Gevers, T., & Mensink, T. (2020). Multi-Loss Weighting with Coefficient of Variations. *Arxiv*. http://arxiv.org/abs/2009.01717

Grotti, L. (2024). Italian Linguistic Features for Toxic Language Detection in Social Media. *Italian Journal of Computational Linguistics*, *10*(1), 65–94. https://doi.org/10.4000/125no

Guo, S., & Xing, D. (2025). Text Information Data Mining Method in Natural Language Processing Tasks. Dalam *IJACSA) International Journal of Advanced Computer Science and Applications* (Vol. 16, Nomor 10). www.ijacsa.thesai.org

Han, J. (2025). Multi-source Text Mining for Risk Signal Detection in Asset-Backed Securities Market: An NLP-driven Data Analytics Approach. *Proceedings of 2025 International Symposium on Machine Learning and Social Computing, MLSC 2025*, 497–506. https://doi.org/10.1145/3778450.3778528

Hasanin, T., Khoshgoftaar, T. M., Leevy, J. L., & Seliya, N. (2019). Examining characteristics of predictive models with imbalanced big data. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0231-2

Hatvani, P., & Yang, Z. G. (2025). Automated detection of toxic comments in Hungarian. *Annales Mathematicae et Informaticae*, *61*, 108–117. https://doi.org/10.33039/ami.2025.10.007

He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, *21*(9), 1263–1284. https://doi.org/10.1109/TKDE.2008.239

Heydari, A. A., Thompson, C. A., & Mehmood, A. (2019). SoftAdapt: Techniques for Adaptive Loss Weighting of Neural Networks with Multi-Part Loss Functions. *Arxiv*. http://arxiv.org/abs/1912.12355

Howard, J., & Ruder, S. (2018). Universal Language Model Fine-tuning for Text Classification. *Arxiv*. http://arxiv.org/abs/1801.06146

Jurafsky, D., & Martin, J. H. (2008). *Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition Second Edition*. Prentice Hall.

Kalyan, K. S., Rajasekharan, A., & Sangeetha, S. (2021). AMMUS : A Survey of Transformer-based Pretrained Models in Natural Language Processing. *Arxiv*. http://arxiv.org/abs/2108.05542

Kaplan, A. M., & Haenlein, M. (2010). Users of the world, unite! The challenges and opportunities of Social Media. *Business Horizons*, *53*(1), 59–68. https://doi.org/10.1016/j.bushor.2009.09.003

Kivlichan, I., Sorensen, J., Elliott, J., Vasserman, L., Görner, M., & Culliton, P. (2020). Jigsaw Multilingual Toxic Comment Classification. Dalam *Kaggle*.

Lashkarashvili, N., & Tsintsadze, M. (2022). Toxicity detection in online Georgian discussions. *International Journal of Information Management Data Insights*, *2*(1). https://doi.org/10.1016/j.jjimei.2022.100062

Lin, T., Wang, Y., Liu, X., & Qiu, X. (2021). A Survey of Transformers. *Arxiv*. http://arxiv.org/abs/2106.04554

Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2018, Februari 7). Focal Loss for Dense Object Detection. *Proceedings of the IEEE International Conference on Computer Vision*. http://arxiv.org/abs/1708.02002

Liu, X., Wang, L., Ma, L., & Wang, C. (2024). DRFL: Dynamic-Recall Focal Loss for Image Classification and Segmentation. *Applied Artificial Intelligence*, *38*(1). https://doi.org/10.1080/08839514.2024.2411845

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019, Juli 26). RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ICLR 2020*. http://arxiv.org/abs/1907.11692

Made, N., Sikiandani, D., Made, I., Dwi Suarjaya, A., & Putra, Y. P. (2025). Browser-Based Detection of Harmful Content with Deep Learning Model. Dalam *Journal of Applied Informatics and Computing (JAIC)* (Vol. 9, Nomor 4). http://jurnal.polibatam.ac.id/index.php/JAIC

Mengxin, C., & Shaolin, Z. (2025). UD-KD: A Structure-Aware Framework for Zero-Shot Cross-Lingual Offensive Language Detection of Large Language Models. *Data Intelligence (Preprint submission)*. https://doi.org/https://doi.org/10.3724/2096-7004.di.2025.0176

Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., & Gao, J. (2024). Large Language Models: A Survey. *Arxiv*. http://arxiv.org/abs/2402.06196

Miyajiwala, A., Ladkat, A., Jagadale, S., & Joshi, R. (2022). On Sensitivity of Deep Learning Based Text Classification Algorithms to Practical Input Perturbations. *Arxiv*. https://doi.org/10.1007/978-3-031-10464-0_42

Modi, M. K., & Patel, V. B. (2025). The Role of Social Media in Shaping Public Opinion: A Comprehensive Literature Review. *INTERNATIONAL JOURNAL OF LATEST TECHNOLOGY IN ENGINEERING,MANAGEMENT & APPLIED SCIENCE (IJLTEMAS)*. https://doi.org/10.51583/IJLTEMAS

Mozafari, M., Farahbakhsh, R., & Crespi, N. (2022). Cross-Lingual Few-Shot Hate Speech and Offensive Language Detection Using Meta Learning. *IEEE Access*, *10*, 14880–14896. https://doi.org/10.1109/ACCESS.2022.3147588

Nabiilah, G. Z., Alam, I. N., Purwanto, E. S., & Hidayat, M. F. (2024). 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

Neog, M., & Baruah, N. (2025). A Multi-Layer Hybrid Deep Learning Model for Toxic Comment Detection in Assamese. *IEEE Access*. https://doi.org/10.1109/ACCESS.2017.DOI

Oliveira, L. (2025). Mining Hidden Behavioral Patterns from Multi Source Social Media Interaction Data. *International Journal of Data Mining and Knowledge Discovery*, *5*(2), 1–5. https://qitpress.com/journals/QITP-IJDMKDFullText:https://qitpress.com/articles/

Peters, M. E., Ruder, S., & Smith, N. A. (2019). To Tune or Not to Tune? Adapting Pretrained Representations to Diverse Tasks. *Arxiv*. http://arxiv.org/abs/1903.05987

Prof. Dr. Zohreh Dehdashti Shahrokh, & Seyed Mojtaba Miri. (2019). *A Short Introduction to Comparative Research*. Allameh Tabataba'i University.

Qian Li, Hao Peng, Jianxin Li, Congying Xia, Renyu Yang, Lichao Sun, Philip S. Yu, & Lifang He. (2021). A Survey on Text Classification: From Traditional to Deep Learning. *ACM International Conference Proceeding Series*, 349–354. https://doi.org/https://doi.org/10.1145/1122445.1122456

Salsabila Azzahra, N., Triantoro Murdiansyah, D., & Lhaksmana, K. M. (2021). Toxic Comment Classification on Social Media Using Support Vector Machine and Chi Square Feature Selection. *Intl. Journal on ICT*, *7*(1), 64–76. https://doi.org/10.34818/ijoict.v7il.552

Schmitt, H. S., Sindermann, C., & Montag, C. (2025). Toxic disinhibition makes a versatile rulebreaker in cyberspace: Mapping online deviance with exploratory graph analysis and structural equation modeling. *Acta Psychologica*, *260*. https://doi.org/10.1016/j.actpsy.2025.105542

Shahid, M., Umair, M., Iqbal, M. A., Rashid, M., Akram, S., & Zubair, M. (2024). Leveraging deep learning for toxic comment detection in cursive languages. *PeerJ Computer Science*, *10*, 1–33. https://doi.org/10.7717/peerj-cs.2486

Sheikholeslami, S., Ghasemirahni, H., Payberah, A. H., Wang, T., Dowling, J., & Vlassov, V. (2025). Utilizing Large Language Models for Ablation Studies in Machine Learning and Deep Learning. *EuroMLSys 2025 - Proceedings of the 2025 5th Workshop on Machine Learning and Systems*, 230–237. https://doi.org/10.1145/3721146.3721957

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing and Management*, *45*(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

Song, G., Huang, D., & Xiao, Z. (2021). A study of multilingual toxic text detection approaches under imbalanced sample distribution. *Information (Switzerland)*, *12*(5). https://doi.org/10.3390/info12050205

Song, G., Huang, D., & Zhang, Y. (2021). A hybrid model for monolingual and multilingual toxic comment detection. *Tehnicki Vjesnik*, *28*(5), 1667–1673. https://doi.org/10.17559/TV-20210325125414

Sulistiyawati, P., Alzami, F., Prabowo, D. P., Pramunendar, R. A., Megantara, R. A., Purinsyira, N., & Irawan, E. (2022). PREDIKSI KATA KASAR BERBAHASA INDONESIA MENGGUNAKAN MACHINE LEARNING

BERBASIS MOBILE INFRASTRUCTURE. *Transmisi*, *24*(2), 55–61. https://doi.org/10.14710/transmisi.24.2.55-61

Sun, C., Qiu, X., Xu, Y., & Huang, X. (2020). How to Fine-Tune BERT for Text Classification? *Arxiv*. http://arxiv.org/abs/1905.05583

Sushma, S., Nayak, S. K., & Krishna, M. V. (2025). Enhanced toxic comment detection model through Deep Learning models using Word embeddings and transformer architectures. *Future Technology*, *4*(3), 76–84. https://doi.org/10.55670/fpll.futech.4.3.8

Tan, K. L., Lee, C. P., & Lim, K. M. (2023). A Survey of Sentiment Analysis: Approaches, Datasets, and Future Research. *Applied Sciences (Switzerland)*, *13*(7). https://doi.org/10.3390/app13074550

Thabtah, F., Hammoud, S., Kamalov, F., & Gonsalves, A. (2020). Data imbalance in classification: Experimental evaluation. *Information Sciences*, *513*, 429–441. https://doi.org/10.1016/j.ins.2019.11.004

Tidake, V. S., & Sane, S. S. (2018). Multi-label Classification: a survey. Dalam *International Journal of Engineering & Technology*. www.sciencepubco.com/index.php/IJET

Valizadehaslani, T., Shi, Y., Wang, J., Ren, P., Zhang, Y., Hu, M., Zhao, L., & Liang, H. (2022). Two-Stage Fine-Tuning: A Novel Strategy for Learning Class-Imbalanced Data. *ArXiv*.

Vaswani, A., Brain, G., 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*.

Waysi Naaman, D., Berivan Tahir Ahmed, & Ibrahim Mahmood Ibrahim. (2025). Architectural Evolution of Transformer Models in NLP: A Comparative Survey of Recent Developments. *The Indonesian Journal of Computer Science*, *14*(5). https://doi.org/10.33022/ijcs.v14i5.4984

Xu, Y., Hu, L., Zhao, J., Qiu, Z., XU, K., Ye, Y., & Gu, H. (2024). A Survey on Multilingual Large Language Models: Corpora, Alignment, and Bias. *IEEE Transactions on Neural Networks and Learning Systems (TNNLS)*. https://doi.org/10.1007/s11704-024-40579-4

Yu, H., Liu, J., Zhang, X., Wu, J., & Cui, P. (2024). A Survey on Evaluation of Out-of-Distribution Generalization. *Arxiv*. http://arxiv.org/abs/2403.01874

Yuan, S., Wu, P., Chen, Y., & Li, Q. (2023). A Survey of Methods for Handling Disk Data Imbalance. *International Journal on Cybernetics & Informatics*, *12*(6), 83–93. https://doi.org/10.5121/ijci.2023.120607

Yuliyanti, S., Septi Asriani, A., Purwayoga, V., & Gusnadi, Z. (2026). HyRoBERTa: Hybrid Robustly Optimized BERT Approach Model for Sentiment and Sarcasm Detection in Post-Flood Social Media Analysis. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, *10*(1), 142–149. https://doi.org/10.29207/resti.v10i1.6963

Zhang, H., & Shafiq, M. O. (2024). Survey of transformers and towards ensemble learning using transformers for natural language processing. *Journal of Big Data*, *11*(1). https://doi.org/10.1186/s40537-023-00842-0