UNIVERSITAS
GADJAH MADA
Medical Concept Normalization pada Data Teks Media Sosial Berbasis SapBERT dan Contextual Information Fusion Similarity
Dyandra Aulia, Yunita Sari, S.Kom., M.Sc., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

DAFTAR PUSTAKA

Arias López, M. D. P., Ong, B. A., Frigola, X. B., Fernández, A. L., Hicklent, R. S., Obeles, A. J. T., Rocimo, A. M., & Celi, L. A. (2023). Digital literacy as a new determinant of health: A scoping review. *PLOS Digital Health, 2*(10). https://doi.org/10.1371/journal.pdig.0000279

Basaldella, M., Liu, F., Shareghi, E., & Collier, N. (2020). COMETA: A Corpus for Medical Entity Linking in the Social Media. In B. Webber, T. Cohn, Y. He, & Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 3122–3137). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.253

Braga, M., Milanese, G. C., & Pasi, G. (2025). Investigating Large Language Models' Linguistic Abilities for Text Preprocessing. *Proceedings of the IEEE/WIC International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT)*. https://doi.org/10.48550/arXiv.2510.11482

Bromley, J., Guyon, I., LeCun, Y., Säckinger, E., & Shah, R. (1993). Signature Verification using a "Siamese" Time Delay Neural Network. In J. Cowan, G. Tesauro, & J. Alspector (Eds.), *Advances in Neural Information Processing Systems* (Vol. 6). Morgan-Kaufmann. https://proceedings.neurips.cc/paper_files/paper/1993/file/288cc0ff022877bd3df94bc9360b9c5d-Paper.pdf

Camacho-Collados, J., & Pilehvar, M. T. (2018). On the Role of Text Preprocessing in Neural Network Architectures: An Evaluation Study on Text Categorization and Sentiment Analysis. In T. Linzen, G. Chrupala, & A. Alishahi (Eds.), *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 40–46). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-5406

Campanozzi, L. L., Gibelli, F., Bailo, P., Nittari, G., Sirignano, A., & Ricci, G. (2023). The role of digital literacy in achieving health equity in the third

UNIVERSITAS
GADJAH MADA
Medical Concept Normalization pada Data Teks Media Sosial Berbasis SapBERT dan Contextual Information Fusion Similarity
Dyandra Aulia, Yunita Sari, S.Kom., M.Sc., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

millennium society: A literature review. *Frontiers in Public Health*, *Volume 11-2023*. https://doi.org/10.3389/fpubh.2023.1109323

Chai, C. P. (2023). Comparison of text preprocessing methods. *Natural Language Engineering*, *29*(3), 509–553. https://doi.org/10.1017/S1351324922000213

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A Simple Framework for Contrastive Learning of Visual Representations. In H. D. III & A. Singh (Eds.), *Proceedings of the 37th International Conference on Machine Learning* (pp. 1597–1607). PMLR. https://proceedings.mlr.press/v119/chen20j.html

Ciampiconi, L., Elwood, A., Leonardi, M., Mohamed, A., & Rozza, A. (2026). A survey and taxonomy of loss functions in machine learning. *AI*, *7*, 128. https://doi.org/10.3390/ai7040128

Dang, T. H., & Thang Nguyen, D. (2024). Enhancing Biomedical Entity Normalization Through Re-ranking Based on Convolutional Neural Networks. *2024 International Conference on Multimedia Analysis and Pattern Recognition (MAPR)*, 1–6. https://doi.org/10.1109/mapr63514.2024.10660948

Dash, A., Darshana, S., Yadav, D. K., & Gupta, V. (2024). A clinical named entity recognition model using pretrained word embedding and deep neural networks. *Decision Analytics Journal*, *10*, 100426. https://doi.org/10.1016/j.dajour.2024.100426

De Silva, T. S., MacDonald, D., Paterson, G., Sikdar, K. C., & Cochrane, B. (2011). Systematized nomenclature of medicine clinical terms (SNOMED CT) to represent computed tomography procedures. *Computer Methods and Programs in Biomedicine*, *101*(3), 324–329. https://doi.org/10.1016/j.cmpb.2011.01.002

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (Vol. 1, pp. 4171–

UNIVERSITAS
GADJAH MADA
Medical Concept Normalization pada Data Teks Media Sosial Berbasis SapBERT dan Contextual Information Fusion Similarity
Dyandra Aulia, Yunita Sari, S.Kom., M.Sc., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

Dirkson, A. R. (2022). *Knowledge discovery from patient forums: gaining novel medical insights from patient experiences* [SIKS Dissertation Series, Leiden Institute of Advanced Computer Science (LIACS)]. https://hdl.handle.net/1887/3492655

Douze, M., Guzhva, A., Deng, C., Johnson, J., Szilvasy, G., Mazaré, P. E., Lomeli, M., Hosseini, L., & Jégou, H. (2026). The Faiss Library. *IEEE Transactions on Big Data*, *12*(2), 346–361. https://doi.org/10.1109/TBDATA.2025.3618474

Flores, L. J., Huang, H., Shi, K., Chheang, S., & Cohan, A. (2023). Medical Text Simplification: Optimizing for Readability with Unlikelihood Training and Reranked Beam Search Decoding. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 4859–4873). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.322

Gnecco, D. P., Serrano, J., Puertas, E., & Carlos Martinez-Santos, J. (2025). Hybrid Re-ranking for Biomedical Entity Linking using SapBERT Embeddings: A High-Performance System for BioNNE-L 2025-1 Notebook for the VerbaNex AI Lab at CLEF 2025. *CEUR Workshop Proceedings*, *4038*, 497–508. https://ceur-ws.org/Vol-4038/paper_35.pdf

Grigonytė, G., Kvist, M., Velupillai, S., & Wirén, M. (2014). Improving Readability of Swedish Electronic Health Records through Lexical Simplification: First Results. In S. Williams, A. Siddharthan, & A. Nenkova (Eds.), *Proceedings of the 3rd Workshop on Predicting and Improving Text Readability for Target Reader Populations (PITR)* (pp. 74–83). Association for Computational Linguistics. https://doi.org/10.3115/v1/W14-1209

Jati, B. S., Widyawan, S. T., & Muhammad Nur Rizal, S. T. (2020). Multilingual Named Entity Recognition Model for Indonesian Health Insurance Question Answering System. *2020 3rd International Conference on Information and*

UNIVERSITAS
GADJAH MADA
Medical Concept Normalization pada Data Teks Media Sosial Berbasis SapBERT dan Contextual Information Fusion Similarity
Dyandra Aulia, Yunita Sari, S.Kom., M.Sc., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

Communications Technology, ICOIACT 2020, 180–184. https://doi.org/10.1109/ICOIACT50329.2020.9332027

Kalyan, K. S., & Sangeetha, S. (2021). BertMCN: Mapping colloquial phrases to standard medical concepts using BERT and highway network. *Artificial Intelligence in Medicine*, *112*, 102008. https://doi.org/10.1016/j.artmed.2021.102008

Kamil, R., & Laksmi, L. (2024). Health Information Behavior of Indonesians During the COVID-19 Pandemic: A Sensemaking Perspective. *JOURNAL OF INFORMATION SCIENCE THEORY AND PRACTICE*, *12*(2). https://doi.org/10.1633/JISTaP.2024.12.2.4

Kobayashi, T. (2025). Temperature in Cosine-based Softmax Loss. *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, 22199–22208. https://doi.org/10.1109/ICCV51701.2025.02061

Kong, L., Winestock, C., & Bhatia, P. (2021). Zero-shot Medical Entity Retrieval without Annotation: Learning From Rich Knowledge Graph Semantics. In C. Zong, F. Xia, W. Li, & R. Navigli (Eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 2401–2405). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-acl.212

Kusumawardani, R. P., & Kusumawati, K. N. (2024). Named entity recognition in the medical domain for Indonesian language health consultation services using bidirectional-lstm-crf algorithm. *Procedia Computer Science*, *245*, 1146–1156. https://doi.org/10.1016/j.procs.2024.10.344

Lee, K., Hasan, S. A., Farri, O., Choudhary, A., & Agrawal, A. (2017). Medical Concept Normalization for Online User-Generated Texts. *2017 IEEE International Conference on Healthcare Informatics (ICHI)*, 462–469. https://doi.org/10.1109/ICHI.2017.59

Limsopatham, N., & Collier, N. (2016). Normalising Medical Concepts in Social Media Texts by Learning Semantic Representation. In K. Erk & N. A. Smith (Eds.), *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 1014–1023).

UNIVERSITAS
GADJAH MADA
Medical Concept Normalization pada Data Teks Media Sosial Berbasis SapBERT dan Contextual Information Fusion Similarity
Dyandra Aulia, Yunita Sari, S.Kom., M.Sc., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

Association for Computational Linguistics. https://doi.org/10.18653/v1/P16-1096

Liu, F., Shareghi, E., Meng, Z., Basaldella, M., & Collier, N. (2021). Self-Alignment Pretraining for Biomedical Entity Representations. In K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tur, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, & Y. Zhou (Eds.), *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4228–4238). Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.naacl-main.334

Luo, G., Shi, N., Wang, G., & Tang, B. (2025). Contextual information contributes to biomedical named entity normalization. *Journal of Biomedical Informatics*, *165*, 104806. https://doi.org/10.1016/J.JBI.2025.104806

Miftahutdinov, Z., Kadurin, A., Kudrin, R., & Tutubalina, E. (2021). Medical concept normalization in clinical trials with drug and disease representation learning. *Bioinformatics*, *37*(21), 3856–3864. https://doi.org/10.1093/BIOINFORMATICS/BTAB474

Mondal, I., Purkayastha, S., Sarkar, S., Goyal, P., Pillai, J. K., Bhattacharyya, A., & Gattu, M. (2019). Medical Entity Linking using Triplet Network. In A. Rumshisky, K. Roberts, S. Bethard, & T. Naumann (Eds.), *Proceedings of the 2nd Clinical Natural Language Processing Workshop* (pp. 95–100). Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-1912

Moreno-Barea, F. J., López-García, G., Mesa, H., Ribelles, N., Alba, E., Jerez, J. M., & Veredas, F. J. (2025). Named entity recognition for de-identifying Spanish electronic health records. *Computers in Biology and Medicine*, *185*, 109576. https://doi.org/10.1016/J.COMPBIOMED.2024.109576

Mou, Y., Chen, H., Lode, G. I., Truhn, D., Sowe, S., & Decker, S. (2024). RadLink: Linking Clinical Entities from Radiology Reports. *2024 2nd International Conference on Foundation and Large Language Models (FLLM)*, 443–449. https://doi.org/10.1109/FLLM63129.2024.10852450

UNIVERSITAS
GADJAH MADA
Medical Concept Normalization pada Data Teks Media Sosial Berbasis SapBERT dan Contextual Information Fusion Similarity
Dyandra Aulia, Yunita Sari, S.Kom., M.Sc., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

Ningtyas, A. M. (2024). *Supporting Laypeople in Learning Formal Medical Terminology* [Dissertation, Doctoral dissertation, Technische Universität Wien]. https://doi.org/10.34726/hss.2024.125702

Obi, J. C. (2023). A comparative study of several classification metrics and their performances on data. *World Journal of Advanced Engineering Technology and Sciences*, *8*(1), 308–314. https://doi.org/10.30574/wjaets.2023.8.1.0054

Oord, A. van den, Li, Y., & Vinyals, O. (2018). Representation Learning with Contrastive Predictive Coding. *ArXiv Preprint ArXiv:1807.03748*. https://doi.org/10.48550/arXiv.1807.03748

Peng, H., Xiong, Y., Xiang, Y., Wang, H., Xu, H., & Tang, B. (2022). Biomedical named entity normalization via interaction-based synonym marginalization. *Journal of Biomedical Informatics*, *136*, 104238. https://doi.org/10.1016/j.jbi.2022.104238

Petrič, G., Atanasova, S., & Kamin, T. (2017). Ill Literates or Illiterates? Investigating the eHealth Literacy of Users of Online Health Communities. *Journal of Medical Internet Research*, *19*(10), e331. https://doi.org/10.2196/jmir.7372

Phatak, A., Savage, D. W., Ohle, R., Smith, J., & Mago, V. (2022). Medical Text Simplification Using Reinforcement Learning (TESLEA): Deep Learning–Based Text Simplification Approach. *JMIR Medical Informatics*, *10*(11), e38095. https://doi.org/10.2196/38095

Siino, M., Tinnirello, I., & La Cascia, M. (2024). Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on Transformers and traditional classifiers. *Information Systems*, *121*, 102342. https://doi.org/10.1016/j.is.2023.102342

Tutubalina, E., Miftahutdinov, Z., Nikolenko, S., & Malykh, V. (2018). Medical concept normalization in social media posts with recurrent neural networks. *Journal of Biomedical Informatics*, *84*, 93–102. https://doi.org/10.1016/J.JBI.2018.06.006

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is All you Need. In I. Guyon, U.

UNIVERSITAS
GADJAH MADA
Medical Concept Normalization pada Data Teks Media Sosial Berbasis SapBERT dan Contextual Information Fusion Similarity
Dyandra Aulia, Yunita Sari, S.Kom., M.Sc., Ph.D.
Universitas Gadjah Mada, 2026 | Diunduh dari http://etd.repository.ugm.ac.id/

Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (pp. 5998–6008). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Voorhees, E. M., Tice, D. M., & others. (1999). The TREC-8 question answering track evaluation. *TREC, 1999*, 82.

Wilhelm, A. G. (2006). *Digital nation: Toward an inclusive information society*. Cambridge: MIT Press.

Xu, D., Gopale, M., Zhang, J., Brown, K., Begoli, E., & Bethard, S. (2020). Unified medical language system resources improve sieve-based generation and bidirectional encoder representations from transformers (BERT)–based ranking for concept normalization. *Journal of the American Medical Informatics Association*, *27*(10), 1510–1519. https://doi.org/10.1093/JAMIA/OCAA080

Xu, D., & Miller, T. (2022). A simple neural vector space model for medical concept normalization using concept embeddings. *Journal of Biomedical Informatics*, *130*, 104080. https://doi.org/10.1016/J.JBI.2022.104080

Yao, Z., Kantu, N. S., Wei, G., Tran, H., Duan, Z., Kwon, S., Yang, Z., & Yu, H. (2024). README: Bridging Medical Jargon and Lay Understanding for Patient Education through Data-Centric NLP. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024* (pp. 12609–12629). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.737

Zhang, Q., Sun, Y., Zhang, L., Jiao, Y., & Tian, Y. (2021). Named entity recognition method in health preserving field based on BERT. *Procedia Computer Science*, *183*, 212–220. https://doi.org/10.1016/J.PROCS.2021.03.010