TOPIC MODELING TWITTER BAHASA INDONESIA DAN JAWA MENGGUNAKAN METODE BERTOPIC DENGAN MODEL EMBEDDINGS NUSABERT

SKRIPSI

Disusun Oleh:

ANDHIKA ACHMAD FAHREZA

09020621024

PROGRAM STUDI SISTEM INFORMASI

FAKULTAS SAINS DAN TEKNOLOGI

UNIVERSITAS ISLAM NEGERI SUNAN AMPEL

SURABAYA

2026

LEMBAR PERNYATAAN KEASLIAN

Saya yang bertanda tangan di bawah ini,

Nama	:	ANDHIKA ACHMAD FAHREZA
NIM	:	09020621024
Prodi	:	SISTEM INFORMASI
Angkatan	:	2021

Menyatakan bahwa saya tidak melakukan plagiat dalam penulisan skripsi saya yang berjudul "TOPIC MODELING TWITTER BAHASA INDONESIA DAN JAWA MENGGUNAKAN METODE BERTOPIC DENGAN MODEL EMBEDDINGS NUSABERT". Apabila suatu saat nanti terbukti saya melakukan tindakan plagiat maka bersedia menerima sanksi yang telah ditetapkan.

Demikian pernyataan keaslian ini saya buat dengan sebenar-benarnya.

Surabaya, 15 Januari 2026

Yang menyatakan,

Andhika Achmad Fahreza
NIM. 09020621024

LEMBAR PERSETUJUAN PEMBIMBING

Skripsi oleh:

Nama	:	ANDHIKA ACHMAD FAHREZA
NIM	:	09020621024
Judul	:	TOPIC MODELING TWITTER BAHASA INDONESIA DAN JAWA MENGGUNAKAN METODE BERTOPIC DENGAN MODEL EMBEDDINGS NUSABERT

Ini telah diperiksa dan disetujui untuk diujikan.

Surabaya, 2 Januari 2026

Dosen Pembimbing 1	Dosen Pembimbing 2
Khalid, M.Kom	Subhan Nooriansyah, M.Kom
NIP. 197906092014031002	NIP. 199012282020121010

LEMBAR PENGESAHAN TIM PENGUJI SKRIPSI

Skripsi Andhika Achmad Fahreza ini telah dipertahankan

Di depan tim penguji Skripsi

Di Surabaya, 7 Januari 2026

Mengesahkan,

Dewan Penguji

Dosen Penguji 1

Dwi Rolliawati, M.T

NIP. 197909212014032001

Dosen Penguji 2

Yunita Ardilla, S. Kom., M.MT

NIP. 199206042020122015

Dosen Penguji 3

Khalid, M.Kom

NIP. 197906092014031002

Dosen Penguji 4

Subhan Nooriansyah, M.Kom

NIP. 199012282020121010

Mengetahui,

Dekan Fakultas Sains dan Teknologi

UIN Sunan Ampel Surabaya

Dr. A. Saepul Hamdani, M.Pd

NIP. 196507312000031002

KEMENTERIAN AGAMA
UNIVERSITAS ISLAM NEGERI SUNAN AMPEL SURABAYA
PERPUSTAKAAN
Jl. Jend. A. Yani 117 Surabaya 60237 Telp. 031-8431972 Fax.031-8413300
E-Mail: perpus@uinsby.ac.id

UIN SUNAN AMPEL
S U R A B A Y A

LEMBAR PERNYATAAN PERSETUJUAN PUBLIKASI KARYA ILMIAH UNTUK KEPENTINGAN AKADEMIS

Sebagai sivitas akademika UIN Sunan Ampel Surabaya, yang bertanda tangan di bawah ini, saya:

Nama : ANDHIKA ACHMAD FAHREZA
NIM : 09020621024
Fakultas/Jurusan : SAINS DAN TEKNOLOGI / SISTEM INFORMASI
E-mail address : ootsutsukimarkoka@gmail.com

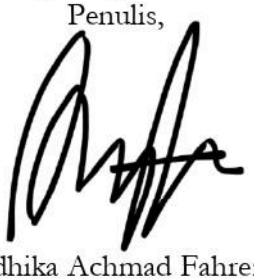
Demi pengembangan ilmu pengetahuan, menyetujui untuk memberikan kepada Perpustakaan UIN Sunan Ampel Surabaya, Hak Bebas Royalti Non-Eksklusif atas karya ilmiah :
■ Sekripsi ☐ Tesis ☐ Desertasi ☐ Lain-lain (.....)
yang berjudul :
TOPIC MODELING TWITTER BAHASA INDONESIA DAN JAWA MENGGUNAKAN METODE BERTOPIC DENGAN MODEL EMBEDDINGS NUSABERT

beserta perangkat yang diperlukan (bila ada). Dengan Hak Bebas Royalti Non-Ekslusif ini Perpustakaan UIN Sunan Ampel Surabaya berhak menyimpan, mengalih-media/format-kan, mengelolanya dalam bentuk pangkalan data (database), mendistribusikannya, dan menampilkan/mempublikasikannya di Internet atau media lain secara ***fulltext*** untuk kepentingan akademis tanpa perlu meminta ijin dari saya selama tetap mencantumkan nama saya sebagai penulis/pencipta dan atau penerbit yang bersangkutan.

Saya bersedia untuk menanggung secara pribadi, tanpa melibatkan pihak Perpustakaan UIN Sunan Ampel Surabaya, segala bentuk tuntutan hukum yang timbul atas pelanggaran Hak Cipta dalam karya ilmiah saya ini.

Demikian pernyataan ini yang saya buat dengan sebenarnya.

Surabaya 15, Januari 2026
Penulis,



Andhika Achmad Fahreza

ABSTRAK

TOPIC MODELING TWITTER BAHASA INDONESIA DAN JAWA MENGGUNAKAN METODE BERTOPIC DENGAN MODEL EMBEDDINGS NUSABERT

Oleh:
Andhika Achmad Fahreza

Pertumbuhan pesat media sosial, khususnya Twitter, telah menghasilkan volume data teks yang sangat besar dalam berbagai bahasa lokal, termasuk Bahasa Indonesia dan Bahasa Jawa. Keragaman linguistik, penggunaan bahasa informal, serta fenomena code-switching menimbulkan tantangan dalam proses analisis teks dan ekstraksi informasi tematik. Penelitian ini bertujuan untuk menerapkan dan mengevaluasi metode BERTopic dengan model embeddings NusaBERT dalam melakukan pemodelan topik pada data Twitter berbahasa Indonesia dan Jawa, serta membandingkan kinerjanya dengan IndoSBERT. Data diperoleh dari platform Hugging Face dan melalui tahapan preprocessing yang meliputi case folding, penghapusan URL, mention, hashtag, escape sequence, tanda baca, serta normalisasi kata tidak baku. Representasi teks dibangun menggunakan embeddings dari NusaBERT dan IndoSBERT, kemudian direduksi dimensinya menggunakan UMAP dan dikelompokkan menggunakan HDBSCAN dalam kerangka BERTopic. Kualitas topik dievaluasi menggunakan metrik Topic Coherence, Topic Diversity, dan analisis visualisasi seperti intertopic distance map dan hierarchical clustering. Hasil penelitian menunjukkan bahwa kombinasi BERTopic–NusaBERT mampu menghasilkan topik yang lebih koheren dan beragam dibandingkan IndoSBERT, khususnya pada data yang mengandung variasi bahasa daerah dan bahasa informal. Hal ini membuktikan bahwa NusaBERT lebih adaptif dalam merepresentasikan karakteristik linguistik Bahasa Indonesia dan Jawa, sehingga lebih efektif untuk pemodelan topik pada data Twitter multibahasa. Penelitian ini diharapkan dapat menjadi referensi dalam pengembangan analisis topik berbasis transformer untuk bahasa lokal di Indonesia.

Kata kunci: *Topic Modeling*, BERTopic, NusaBERT, Twitter, Jawa

ABSTRACT

TOPIC MODELING ON TWITTER IN INDONESIAN AND JAVANESE USING THE BERTOPIC METHOD WITH THE NUSABERT EMBEDDINGS MODEL

By:
Andhika Achmad Fahreza

The rapid growth of social media, particularly Twitter, has generated a massive volume of textual data in various local languages, including Indonesian and Javanese. Linguistic diversity, informal language usage, and code-switching phenomena pose significant challenges for text analysis and thematic extraction. This study aims to implement and evaluate the BERTopic method using NusaBERT embeddings for topic modeling on Indonesian and Javanese Twitter data, as well as to compare its performance with IndoSBERT. The dataset was obtained from the Hugging Face platform and processed through several preprocessing steps, including case folding, removal of URLs, mentions, hashtags, escape sequences, punctuation, and slang normalization. Text representations were generated using NusaBERT and IndoSBERT embeddings, then reduced in dimensionality using UMAP and clustered using HDBSCAN within the BERTopic framework. Topic quality was evaluated using Topic Coherence, Topic Diversity, and visualization analyses such as the intertopic distance map and hierarchical clustering. The results show that the BERTopic–NusaBERT combination produces more coherent and diverse topics than IndoSBERT, particularly for data containing regional language variations and informal expressions. This demonstrates that NusaBERT is more adaptive in capturing the linguistic characteristics of Indonesian and Javanese, making it more effective for topic modeling on multilingual Twitter data. This study is expected to contribute to the development of transformer-based topic modeling for local languages in Indonesia.

Kata kunci: *Topic Modeling*, BERTopic, NusaBERT, Twitter, *Javanese*

DAFTAR ISI

DAFTAR GAMBAR

DAFTAR TABEL

DAFTAR PUSTAKA

Abusive Language and Hate Speech Detection for Javanese and Sundanese Languages in Tweets: Dataset and Preliminary Study. (2021). *Proceedings of 2021 the 11th International Workshop on Computer Science and Engineering*. 2021 the 11th International Workshop on Computer Science and Engineering. https://doi.org/10.18178/wcse.2021.02.011

Adam, R., & Kogler, M. L. (2024). *Bidirectional Topic Matching: Quantifying Thematic Overlap Between Corpora Through Topic Modelling* (No. arXiv:2412.18376). arXiv. https://doi.org/10.48550/arXiv.2412.18376

Aggarwal, C. C., & Zhai, C. (Ed.). (2012). *Mining Text Data*. Springer US. https://doi.org/10.1007/978-1-4614-3223-4

Alfina, I., Yuliawati, A., Tanaya, D., Dinakaramani, A., & Zeman, D. (2022). A Gold Standard Dataset for Javanese Tokenization, POS Tagging, Morphological Feature Tagging, and Dependency Parsing. *Forum for Linguistic Studies*, *6*(5), 131–148. https://doi.org/10.30564/fls.v6i5.6957

Ananta, M., Utari, R., Akhyar, A., & Putri Saptawati, G. A. (2024). Integrating BERTopic and Large Language Models for Thematic Identification of Indonesian Legal Documents. *2024 11th International Conference on Advanced Informatics: Concept, Theory and Application (ICAICTA)*, 1–6. https://doi.org/10.1109/ICAICTA63815.2024.10762984

Angelov, D. (2020). *Top2Vec: Distributed Representations of Topics* (Versi 1). arXiv. https://doi.org/10.48550/ARXIV.2008.09470

Aryani, D., Lucia Kharisma, I., Sujjada, A., & Kamdan, K. (2024). Topic Modeling of the 2024 Election Using the BERTopic Method on Detik.com News Articles. *Inform : Jurnal Ilmiah Bidang Teknologi Informasi Dan Komunikasi*, *9*(2), 171–180. https://doi.org/10.25139/inform.v9i2.8429

Baah, I., & Ahmed, M. M. (2025). *Analysis of Crash Contributing Factors in Worker-Involved Crashes in Ohio Work Zones: A Topic Modeling Approach*. SSRN. https://doi.org/10.2139/ssrn.5449914

Cambria, E., & White, B. (2014). Jumping NLP Curves: A Review of Natural Language Processing Research [Review Article]. *IEEE Computational*

Intelligence Magazine, *9*(2), 48–57. https://doi.org/10.1109/MCI.2014.2307227

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. Dalam J. Pei, V. S. Tseng, L. Cao, H. Motoda, & G. Xu (Ed.), *Advances in Knowledge Discovery and Data Mining* (Vol. 7819, hlm. 160–172). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-37456-2_14

Chakraborty, G., Pagolu, M., & Garla, S. (2014). *Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS*. SAS Institute. https://books.google.co.id/books?id=SUKDAgAAQBAJ

Choirinnisa, D., Alzami, F., Indrayani, H., Rohmani, A., Nugraini, S. H., Zulfiningrumi, R., & Susanti, F. (2025). LDA Topic Modeling: Twitter-Based Public Opinion on Indonesian Ministry of Finance. *Sinkron*, *9*(2), 849–863. https://doi.org/10.33395/sinkron.v9i2.14719

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (No. arXiv:1810.04805). arXiv. https://doi.org/10.48550/arXiv.1810.04805

Dieng, A. B., Ruiz, F. J. R., & Blei, D. M. (2020). Topic Modeling in Embedding Spaces. *Transactions of the Association for Computational Linguistics*, *8*, 439–453. https://doi.org/10.1162/tacl_a_00325

Fathan (Ahmad Fathan Hidayatullah). (2025, Februari 14). https://huggingface.co/fathan

Griffiths, T. L., Jordan, M. I., Tenenbaum, J. B., & Blei, D. M. (t.t.). *Hierarchical Topic Models and the Nested Chinese Restaurant Process*.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(suppl_1), 5228–5235. https://doi.org/10.1073/pnas.0307752101

Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure* (No. arXiv:2203.05794). arXiv. https://doi.org/10.48550/arXiv.2203.05794

Gupta, V., & Lehal, G. S. (2009). A Survey of Text Mining Techniques and Applications. *Journal of Emerging Technologies in Web Intelligence*, *1*(1), 60–76. https://doi.org/10.4304/jetwi.1.1.60-76

Han, B., Cook, P., & Baldwin, T. (2013). Lexical normalization for social media text. *ACM Transactions on Intelligent Systems and Technology*, *4*(1), 1–27. https://doi.org/10.1145/2414425.2414430

Harris, Z. S. (1954). Distributional Structure. WORD, *10*(2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Hidayatullah, A. F. (2024). Code-Mixed Sentiment Analysis on Indonesian-Javanese-English Text Using Transformer Models. *2024 8th International Conference on Information Technology, Information Systems and Electrical Engineering (ICITISEE)*, 340–345. https://doi.org/10.1109/ICITISEE63424.2024.10730138

Hotho, A., Nürnberger, A., & Paaß, G. (2005). A Brief Survey of Text Mining. *Journal for Language Technology and Computational Linguistics*, *20*(1), 19–62. https://doi.org/10.21248/jlcl.20.2005.68

Hoyle, A., Goel, P., Peskov, D., Hian-Cheong, A., Boyd-Graber, J., & Resnik, P. (2021). *Is Automated Topic Model Evaluation Broken?: The Incoherence of Coherence* (Versi 3). arXiv. https://doi.org/10.48550/ARXIV.2107.02173

Indurkhya, N., & Damerau, F. J. (Ed.). (2010). *Handbook of Natural Language Processing* (0 ed.). Chapman and Hall/CRC. https://doi.org/10.1201/9781420085938

Javanese-and-Sundanese-Stopwords/local_languages_stopwords.csv at main · bimarakajati/Javanese-and-Sundanese-Stopwords · GitHub. (t.t.). Diambil 15 Desember 2025, dari https://github.com/bimarakajati/Javanese-and-Sundanese-Stopwords/blob/main/local_languages_stopwords.csv

Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., & Zhao, L. (2019). Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications*, *78*(11), 15169–15211. https://doi.org/10.1007/s11042-018-6894-4

Kiss, T., & Strunk, J. (2006). Unsupervised Multilingual Sentence Boundary Detection. *Computational Linguistics*, *32*(4), 485–525. https://doi.org/10.1162/coli.2006.32.4.485

Kobak, D., & Berens, P. (2019). The art of using t-SNE for single-cell transcriptomics. *Nature Communications*, *10*(1), 5416. https://doi.org/10.1038/s41467-019-13056-x

Kurniawan, F. W., & Maharani, W. (2020). Indonesian Twitter Sentiment Analysis Using Word2Vec. *2020 International Conference on Data Science and Its Applications (ICoDSA)*, 1–6. https://doi.org/10.1109/ICoDSA50139.2020.9212906

Laurens van der Maaten & Geoffrey Hinton. (2008). *Visualizing Data using t-SNE*.

Leland McInnes, John Healy, & Steve Astels. (2016). *Parameter Selection for HDBSCAN*—Hdbscan 0.8.1 documentation*. https://hdbscan.readthedocs.io/en/latest/parameter_selection.html

Listyawan, B. R. N., Setiawan, N. Y., & Saputra, M. C. (2024). Topic Modelling Pada Aktivitas Pengembangan Perangkat Lunak Menggunakan BERTopic. *Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer*, *8*(8). https://j-ptiik.ub.ac.id/index.php/j-ptiik/article/view/14051

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Springer International Publishing. https://doi.org/10.1007/978-3-031-02145-9

Mahfudiyah, N., & Alamsyah, A. (2023). Understanding User Perception of Ride-Hailing Services Sentiment Analysis and Topic Modelling using IndoBERT and BERTopic. *2023 International Conference on Digital Business and Technology Management (ICONDBTM)*, 1–6. https://doi.org/10.1109/ICONDBTM59210.2023.10327320

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to Information Retrieval* (1 ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511809071

McInnes, L., Healy, J., & Melville, J. (2018). *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction* (Versi 3). arXiv. https://doi.org/10.48550/ARXIV.1802.03426

Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. *Ain Shams Engineering Journal*, *5*(4), 1093–1113. https://doi.org/10.1016/j.asej.2014.04.011

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space* (Versi 3). arXiv. https://doi.org/10.48550/ARXIV.1301.3781

Millar, J. R. (t.t.). *Document Clustering and Visualization with Latent Dirichlet Allocation and Self-Organizing Maps*.

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (t.t.). *Optimizing Semantic Coherence in Topic Models*.

Mustafa, M., Zeng, F., Ghulam, H., & Muhammad Arslan, H. (2020). Urdu Documents Clustering with Unsupervised and Semi-Supervised Probabilistic Topic Modeling. *Information*, *11*(11), 518. https://doi.org/10.3390/info11110518

NusaBERT - Lazarus NLP. (t.t.). Diambil 15 Desember 2025, dari https://lazarusnlp.github.io/projects/nusabert/

Park, Y. H., & Choi, Y. S. (2025). Analysis of Tourism Research Trends Using Big Data: Focused on Topic Modeling. *Journal & Article Management System*, *58*(0), 43–61. https://doi.org/10.31927/asec.21.2.3

Prisscilya, V., & Girsang, A. S. (2024). Classification of Indonesia False News Detection Using Bertopic and Indobert. *Jurnal Indonesia Sosial Teknologi*, *5*(8), 3061–3079. https://doi.org/10.59141/jist.v5i8.1310

Reimers, N., & Gurevych, I. (2019a). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks* (Versi 1). arXiv. https://doi.org/10.48550/ARXIV.1908.10084

Reimers, N., & Gurevych, I. (2019b). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks* (No. arXiv:1908.10084). arXiv. https://doi.org/10.48550/arXiv.1908.10084

Riyantoni, N. H. (2024). *Topic modeling skripsi program studi sistem informasi menggunakan kombinasi bertopic dan indosbert* [Undergraduate, UIN Sunan Ampel Surabaya]. http://digilib.uinsa.ac.id/70568/

Röder, M., Both, A., & Hinneburg, A. (2015a). Exploring the Space of Topic Coherence Measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 399–408. https://doi.org/10.1145/2684822.2685324

Röder, M., Both, A., & Hinneburg, A. (2015b). Exploring the Space of Topic Coherence Measures. *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, 399–408. https://doi.org/10.1145/2684822.2685324

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, *24*(5), 513–523. https://doi.org/10.1016/0306-4573(88)90021-0

Saputri, M. S., Mahendra, R., & Adriani, M. (2018). Emotion Classification on Indonesian Twitter Dataset. *2018 International Conference on Asian Language Processing (IALP)*, 90–95. https://doi.org/10.1109/IALP.2018.8629262

Sarkar, D. (2016). *Text Analytics with Python*. Apress. https://doi.org/10.1007/978-1-4842-2388-8

Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, *34*(1), 1–47. https://doi.org/10.1145/505282.505283

Sia, S., Dalmia, A., & Mielke, S. J. (2020). *Tired of Topic Models? Clusters of Pretrained Word Embeddings Make for Fast and Good Topics too!* (Versi 2). arXiv. https://doi.org/10.48550/ARXIV.2004.14914

Sparck Jones, K. (1972). A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL. *Journal of Documentation*, *28*(1), 11–21. https://doi.org/10.1108/eb026526

Stewart, G., & Al-Khassaweneh, M. (2022). An Implementation of the HDBSCAN* Clustering Algorithm. *Applied Sciences*, *12*(5), 2405. https://doi.org/10.3390/app12052405

Strobl, M., Sander, J., Campello, R. J. G. B., & Zaïane, O. (2021). Model-Based Clustering with HDBSCAN*. Dalam F. Hutter, K. Kersting, J. Lijffijt, & I. Valera (Ed.), *Machine Learning and Knowledge Discovery in Databases*

(Vol. 12458, hlm. 364–379). Springer International Publishing. https://doi.org/10.1007/978-3-030-67661-2_22

Syed, S., & Spruit, M. (2017). Full-Text or Abstract? Examining Topic Coherence Scores Using Latent Dirichlet Allocation. *2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, 165–174. https://doi.org/10.1109/DSAA.2017.61

Taufiq Dwi Purnomo & Joko Sutopo. (2024). COMPARISON OF PRE-TRAINED BERT-BASED TRANSFORMER MODELS FOR REGIONAL LANGUAGE TEXT SENTIMENT ANALYSIS IN INDONESIA. *International Journal Science and Technology*, *3*(3), 11–21. https://doi.org/10.56127/ijst.v3i3.1739

Umamaheswaran, S., Dar, V., Sharma, E., & Kurian, J. S. (2023). Mapping Climate Themes From 2008-2021—An Analysis of Business News Using Topic Models. *IEEE Access*, *11*, 26554–26565. https://doi.org/10.1109/ACCESS.2023.3256530

University of Liechtenstein, Debortoli, S., Müller, O., IT University of Copenhagen, Junglas, I., Florida State University, Vom Brocke, J., & University of Liechtenstein. (2016). Text Mining for Information Systems Researchers: An Annotated Topic Modeling Tutorial. *Communications of the Association for Information Systems*, *39*, 110–135. https://doi.org/10.17705/1CAIS.03907

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2023). *Attention Is All You Need* (No. arXiv:1706.03762). arXiv. https://doi.org/10.48550/arXiv.1706.03762

Wongso, W., Setiawan, D. S., Limcorn, S., & Joyoadikusumo, A. (2024). *NusaBERT: Teaching IndoBERT to be Multilingual and Multicultural* (No. arXiv:2403.01817). arXiv. https://doi.org/10.48550/arXiv.2403.01817

Xu, R., & WunschII, D. (2005). Survey of Clustering Algorithms. *IEEE Transactions on Neural Networks*, *16*(3), 645–678. https://doi.org/10.1109/TNN.2005.845141

Yau, C.-K., Porter, A., Newman, N., & Suominen, A. (2014). Clustering scientific documents with topic modeling. *Scientometrics*, *100*(3), 767–786. https://doi.org/10.1007/s11192-014-1321-8

Zhai, C., & Massung, S. (2016). *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. Association for Computing Machinery and Morgan & Claypool. https://doi.org/10.1145/2915031

Zhao, Y., & Karypis, G. (2004). Empirical and Theoretical Comparisons of Selected Criterion Functions for Document Clustering. *Machine Learning*, *55*(3), 311–331. https://doi.org/10.1023/B:MACH.0000027785.44527.d6