PENGELOMPOKKAN ARTIKEL ILMIAH MENGGUNAKAN MULTIBERT DAN K-MEANS

Diajukan Sebagai Syarat Untuk Menyelesaikan
Pendidikan Program Strata-1 Pada
Jurusan Teknik Informatika

Oleh :

Risky Armansyah
NIM : 09021282126055

Jurusan Teknik Informatika

FAKULTAS ILMU KOMPUTER UNIVERSITAS SRIWIJAYA

2025

HALAMAN PENGESAHAN

SKRIPSI

Pengelompokkan Artikel Ilmiah Menggunakan MultiBert dan K-Means

Sebagai salah satu syarat untuk penyelesaian studi di

Program Studi S1 Teknik Informatika

Oleh:

RISKY ARMANSYAH

09021282126055

Pembimbing 1	**:**	**Novi Yusliani, S.Kom., M.T.**
		NIP. 198211082012122001
Pembimbing 2	**:**	**Muhammad Naufal Rachmatullah, S.Kom., M.T.**
		NIP. 199212012022031008

Mengetahui

Ketua Jurusan Teknik Informatika

Hadipurnawan Satria, Ph.D

198004182020121001

TANDA LULUS UJIAN KOMPREHENSIF SKRIPSI

Pada hari Jumat tanggal 14 Maret 2025 telah dilaksanakan ujian komprehensif skripsi oleh Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya

Nama	:	Risky Armansyah
NIM	:	09021282126055
Judul	:	Pengelompokkan Artikel Ilmiah Menggunakan MultiBert dan K-Means

dan dinyatakan **LULUS.**

1. Ketua

   Rizki Kurniati, M.T. .....
   NIP. 199107122019032016

2. Penguji I

   Alvi Syahrini Utami, M.Kom . .....
   NIP. 197812222006042003

3. Pembimbing I

   Novi Yusliani, S.Kom., M.T. .....
   NIP. 198211082012122001

4. Pembimbing II

   M. Naufal Rachmatullah, M.T. .....
   NIP. 199212012022031008

Mengetahui,
Ketua Jurusan Teknik Informatika

Hadipurnawan Satria, Ph.D.
NIP. 198004182020121001

HALAMAN PERNYATAAN

Yang bertanda tangan di bawah ini:

Nama	:	Risky Armansyah
NIM	:	09021282126055
Program Studi	:	Teknik Informatika
Judul Skripsi	:	Pengelompokkan Artikel Ilmiah Menggunakan MultiBert dan K-Means

Hasil pengecekan *Software ~~iThenticate~~/Turnitin:*

Menyatakan bahwa laporan penelitian saya merupakan hasil karya sendiri dan bukan hasil penjiplakan/plagiat Apabila ditemukan unsur penjiplakan/plagiat dalam laporan penelitian ini, maka saya bersedia menerima sanksi akademik Universitas Sriwijaya sesuai dengan ketentuan yang berlaku.

Demikian pernyataan ini saya buat dengan sebenarnya dan tidak ada paksaan oleh siapa pun.

Palembang, 14 Maret 2025

Risky Armansyah

NIM. 09021282126055

MOTTO DAN PERSEMBAHAN

"Akulah yang terhebat. Telah aku katakan bahkan sebelum aku tahu sebelumnya."

— Muhammad Ali

Ku persembakan karya tulis ini kepada:

- Allah SWT
- Nabi Muhammad SAW
- Orang Tua dan Keluarga
- Fakultas Ilmu Komputer
- Universitas Sriwijaya

ABSTRACT

The publication rate of scientific articles has significantly increased over time. This presents a challenge for journal administrators and academics in organizing and sorting these articles to align with the journal's scope. This study aims to address this issue by developing a scientific article clustering system utilizing MultiBERT as the data representation model and K-Means for cluster identification based on the representation results. The model was tested using article data from the Science and Technology Index (SINTA) 1 journals. The evaluation results for each journal yielded a silhouette score of 0.571, indicating well-clustered representations. Furthermore, testing across two journals with diverse topics yielded clusters that accurately corresponded to their respective subject areas.

Keywords : Scientific Article Clustering, MultiBert, K-Means, Silhouette Score

ABSTRAK

Angka penerbitan artikel ilmiah seiring waktu meningkat secara signifikan. Hal ini menjadi tantangan bagi pengelola dan akademisi dalam mengorganisir dan memilah artikel-artikel tersebut agar sesuai dengan cakupan jurnal. Penelitian ini bertujuan untuk membantu menyelesaikan permasalahan tersebut dengan mengembangkan sistem pengelompokan artikel ilmiah yang menggunakan MultiBert sebagai model representasi data dan K-Means untuk mengidentifikasi klaster berdasarkan hasil representasi. Pengujian model dilakukan dengan menggunakan data artikel dari jurnal *Science and Technology Index* (SINTA) 1. Hasil pengujian pada tiap jurnal menunjukkan nilai *silhouette score* sebesar 0.571 yang mengindikasikan hasil representasi terklaster dengan baik. Kemudian pengujian antara dua jurnal dengan topik yang beragam menghasilkan klaster-klaster yang sesuai dengan topiknya masing-masing.

Kata Kunci : Klasterisasi Artikel Ilmiah, MultiBert, K-Means, Skor Silhouette

KATA PENGANTAR

Puji syukur ke hadirat Allah SWT atas limpahan nikmat, rahmat dan karunia-Nya, sehingga penulis dapat menyelesaikan Skripsi yang berjudul "Kemiripan Semantik Dokumen Tugas Akhir Terhadap Ontologi Bidang Ilmu Informatika Menggunakan Metode Wu Palmer" ini dengan baik. Skripsi ini disusun untuk memenuhi salah satu syarat untuk memperoleh gelar Sarjana pada Program Studi Teknik Informatika, Fakultas Ilmu Komputer, Universitas Sriwijaya.

Dalam menyelesaikan skripsi ini, penulis telah menerima bantuan, bimbingan, serta dukungan dari banyak pihak. Oleh karena itu, penulis ingin menyampaikan ucapan terima kasih yang tulus kepada:

1. Kedua orang tua saya, Jalalludin dan Hainani, Kakak saya Edwin Jastin dan Istrinya Hesty Oxcelia.
2. Bapak Prof. Dr. Erwin, S.Si., M.Si. selaku Dekan Fakultas Ilmu Komputer Universitas Sriwijaya.
3. Bapak Hadipurnawan Satria, Ph.D. selaku Ketua Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.
4. Ibu Novi Yusliani, S.Kom., M.T. dan Bapak Muhammad Naufal Rachmatullah, S.Kom., M.T. selaku ke dua dosen pembimbing skripsi saya yang telah memberikan bimbingan selama proses kegiatan perkuliahan.
5. Ibu Yunita, S.Si., M.Cs. selaku dosen pembimbing akademik yang telah memandu saya selama perkuliahan.

6. Seluruh Dosen, Admin, dan Staff Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.
7. Vanya Terra Ardani, teman, sahabat, dan pasangan saya yang telah mewarnai hari-hari saya selama perkuliahan.
8. Teman seperjuangan Project Penting Ahmad Azhari, Dellin Irawan, Hanif Syahri Ramadhani, dan Yolendri Anisyahfitri yang telah menemani, memberikan motivasi, dan dukungan penulis.
9. Teman – teman Teknik Informatika angkatan 2021 yang telah menemani selama perkuliahan.
10. Pihak – pihak lain yang tidak dapat penulis sebutkan satu-persatu.

Penulis menyadari bahwa dalam penyusunan skripsi ini masih banyak sekali kekurangan dikarenakan kurangnya pengalaman dan pengetahuan penulis. Oleh karena itu penulis mengharapkan kritik dan saran yang membangun guna kemajuan penelitian selanjutnya. Semoga skripsi ini dapat bermanfaat. Terima kasih.

Indralaya, 14 Maret 2025
Penulis

Risky Armansyah

DAFTAR ISI

Halaman

DAFTAR TABEL

Halaman

DAFTAR GAMBAR

Halaman

DAFTAR LAMPIRAN

BAB I

PENDAHULUAN

1.1 Pendahuluan

Bab ini akan menjelaskan tentang latar belakang masalah, rumusan masalah, tujuan penelitian, manfaat penelitian, batasan masalah, dan sistematika penulisan yang akan menjadi landasan dasar pada bab selanjutnya. Penelitian ini dilakukan untuk membahas bagaimana sebaran artikel pada suatu jurnal dan klasterisasi untuk mengelompokkan artikel berdasarkan cakupan jurnal menggunakan MultiBERT dan K-Means. Pendekatan ini dapat mengefisiensi proses klasterisasi artikel dan memudahkan penentuan artikel yang berada di luar cakupan jurnal.

1.2 Latar Belakang Masalah

Perkembangan ilmu pengetahuan dari waktu ke waktu membawa peningkatan angka penerbitan artikel ilmiah secara pesat. Hal ini menjadi tantangan bagi pengelola dan akademisi dalam mengorganisir dan memilah artikel-artikel tersebut agar sesuai dengan cakupan jurnal. Pengelompokan artikel ilmiah secara manual berdasarkan cakupan topiknya merupakan langkah yang dapat dilakukan untuk mengorganisir artikel ilmiah lebih baik. Akan tetapi, langkah ini tidak efisien dan membutuhkan alokasi waktu yang tidak sedikit (Subakti et al., 2022). Dengan demikian, dibutuhkan sistem yang dapat melakukan pengelompokan artikel ilmiah secara otomatis. Pengelompokan artikel ilmiah secara otomatis tersebut dapat dilakukan dengan menggunakan metode pengklasteran teks (*Text Clustering*).

Pengklasteran teks merupakan suatu metode yang digunakan untuk melakukan pengorganisasian pada sekumpulan teks yang serupa ke dalam kelompok-kelompok (klaster) berdasarkan kesamaan fitur pada teks (Petukhova et al., 2024). Pada pengklasteran teks, dibutuhkan model *word representation* (representasi kata) yang dapat mengubah teks bahasa menjadi bilangan numerik. Salah satu model *word representation* yang dapat digunakan adalah BERT. BERT merupakan model pemrosesan bahasa alami yang diajukan pada berapa tahun belakang dan memiliki hasil yang memuaskan dalam menyelesaikan berbagai permasalahan pemrosesan bahasa alami seperti klasifikasi teks, pencocokan kalimat, dan ekstraksi informasi (Hu et al., 2021). Berdasarkan pernyataan Devlin (2019) BERT dapat menghasilkan *embedding text* dengan representasi bahasa berkualitas tinggi melalui proses *encoding* pada elemen terkecil dokumen. Namun, terdapat kelemahan ketika menggunakan BERT, yaitu proses inisialisasi bobot yang hanya berasal dari satu *checkpoint*, sehingga representasi kata yang dihasilkan akan berbeda-beda tiap pelatihan.

Hal tersebut dapat teratasi dengan salah satu varian dari BERT yaitu MultiBERT. MultiBERT sendiri merupakan rangkaian 25 model berbasis BERT dengan penggunaan parameter yang mirip, tetapi memiliki variasi dalam inisialisasi bobot acak dan data pelatihan teracak. Model ini bertujuan untuk memperkuat penarikan kesimpulan dengan mengatasi variasi dari proses pelatihan stokastik. Dengan demikian, kesimpulan dan representasi kata yang dihasilkan menjadi lebih universal dan representatif (Sellam et al., 2021).

Selain model *word representation,* pengklasteran teks memerlukan proses klasterisasi, salah satu metode yang dapat digunakan untuk proses klasterisasi adalah K-Means. K-Means merupakan sebuah metode untuk membagi populasi N-dimensi menjadi K set klaster berdasarkan sampel dengan tujuan untuk menghasilkan partisi yang efisien (Xia et al., 2022). Algoritma ini memiliki keunggulan dalam segi kesederhanaan, efisiensi dalam waktu komputasi, skalabilitasnya terhadap data yang besar dan fleksibilitas dalam berbagai pengaplikasian, sehingga telah banyak diaplikasikan pada berbagai bidang ilmu pengetahuan (Ahmed et al., 2020). Akan tetapi, Kinerja dari algoritma K-Means akan menurun ketika berhadapan dengan data berdimensi tinggi, hal ini dinamakan dengan *curse of dimensionality* (Keogh Eamonnand Mueen, 2017). Untuk menghindari hal tersebut, diperlukan suatu metode untuk mereduksi dimensi data. Salah satu metode yang dapat digunakan adalah *Principal Component Analysis.*

Principal Component Analysis (PCA) merupakan metode statistik multivariat yang bertujuan untuk mereduksi dimensi dataset dengan membuat "*principal components"* yang dapat menginterpretasi "informasi" keseluruhan data dengan maksimal (Greenacre et al., 2023). PCA memiliki beberapa fungsi seperti reduksi dimensi data, meningkatkan interpretabilitas data, dan mengefisiensi proses komputasi (Hasan & Abdulazeez, 2021). Dengan fungsi-fungsi tersebut, kinerja K-Means dalam melakukan klasterisasi dapat ditingkatkan. Selain itu, dengan memanfaatkan PCA, hasil klasterisasi tersebut dapat divisualisasi dalam bentuk grafik dua dimensi. Dengan demikian, tiap

klaster dan representasi data dapat dilihat, dianalisis, dan diorganisir sesuai dengan cakupan topik jurnalnya. Oleh karena itu, penelitian ini dibuat dengan tujuan untuk membuat sistem pengelompokan artikel ilmiah menggunakan MultiBERT dan K-Means dengan bantuan PCA sebagai pereduksi dimensi data.

1.3 Rumusan Masalah

Berdasarkan latar belakang yang telah dikemukakan, rumusan masalah dari penelitian ini adalah sebagai berikut:

1. Bagaimana mengimplementasikan MultiBERT, PCA, dan K-Means untuk mengelompokkan teks artikel ilmiah?
2. Bagaimana kinerja dari pengelompokan teks artikel ilmiah MultiBERT dan K-Means berdasarkan *Silhouette Score*?

1.4 Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah sebagai berikut:

1. Menghasilkan sebuah sistem yang dapat mengelompokkan artikel ilmiah menggunakan MultiBERT, PCA, dan K-Means.
2. Mengetahui kinerja dari pengelompokan artikel ilmiah menggunakan MultiBERT dan K-Means berdasarkan *Silhouette Score*.

1.5 Manfaat Penelitian

Berikut adalah manfaat dari penelitian ini:

1. Sistem dapat membantu pengelolaan dan pemilahan artikel ilmiah sesuai dengan cakupan topiknya.

2. Diharapkan penelitian ini dapat menjadi referensi untuk penelitian atau pengembangan selanjutnya.

1.6 Batasan Masalah

Adapun batasan masalah pada penelitian ini adalah data yang digunakan merupakan data artikel ilmiah ter-indeks SINTA berbahasa Indonesia dan Inggris.

1.7 Sistematika Penulisan

Sistematika penulisan tugas akhir disusun berdasarkan standar penulisan tugas akhir yang ditetapkan oleh Fakultas Ilmu Komputer Universitas Sriwijaya, yaitu:

BAB I. PENDAHULUAN

Bab ini akan menguraikan latar belakang, rumusan masalah, tujuan dari penelitian, manfaat penelitian, dan batasan masalah. Pokok-pokok yang dibahas dalam bab ini akan menjadi pijakan utama bagi bab selanjutnya.

BAB II. KAJIAN LITERATUR

Bab ini menyajikan landasan teori yang mendukung penelitian. Di dalamnya terdapat tinjauan literatur dan penelitian terdahulu yang relevan dengan penelitian ini, termasuk penjelasan mengenai model MultiBERT, K-Means, serta penjelasan lain yang terkait.

BAB III. METODOLOGI PENELITIAN

Bab ini akan menguraikan metode dan langkah-langkah yang digunakan dalam penelitian ini mulai dari pengumpulan data, perancangan

dari sistem yang dibuat, dan tahapan dalam melakukan penelitian sesuai dengan perancangan.

BAB IV. PENGEMBANGAN PERANGKAT LUNAK

Bab ini akan menjelaskan tentang metode pengembangan perangkat lunak yang digunakan yaitu metode *Rational Unified Process* (RUP) untuk membuat sistem pengelompokan artikel ilmiah menggunakan MultiBERT dan K-Means.

BAB V. HASIL DAN ANALISIS

Bab ini akan membahas tentang hasil penelitian dari sistem yang telah dibuat dan direncanakan sebelumnya. Hasil analisis penelitian tersebut kemudian akan digunakan sebagai dasar dalam pengambilan kesimpulan penelitian.

BAB VI. KESIMPULAN DAN SARAN

Bab ini memaparkan kesimpulan dari penelitian yang dilakukan berdasarkan uraian pada bab-bab sebelumnya dan memuat saran yang diharapkan dapat membuat sistem lebih baik lagi ke depannya.

1.8 Kesimpulan

Bab ini telah menguraikan terkait penelitian yang akan dilakukan mencakup latar belakang, rumusan masalah, tujuan penelitian, manfaat penelitian, serta batasan masalah dari penulisan yang akan dibuat sebagai dasar pemikiran peneliti.

DAFTAR PUSTAKA

Ahmed, M., Seraj, R., & Islam, S. M. S. (2020). The k-means algorithm: A comprehensive survey and performance evaluation. In *Electronics (Switzerland)* (Vol. 9, Issue 8, pp. 1–12). MDPI AG. https://doi.org/10.3390/electronics9081295

Ankerst, M., Breunig, M. M., Kriegel, H.-P., & Sander, J. (1999). OPTICS: ordering points to identify the clustering structure. *Proceedings of the 1999 ACM SIGMOD International Conference on Management of Data*, 49–60. https://doi.org/10.1145/304182.304187

Apidianaki, M. (2023). *From Word Types to Tokens and Back: A Survey of Approaches to Word Meaning Representation and Interpretation*. https://doi.org/10.1162/coli

Aquije, L., Carranza, L., Pena, G., Cabanillas-Carbonell, M., & Andrade-Arenas, L. (2022). Design of a Mobile Application for the Logistics Process of a Fire Company. *International Journal of Advanced Computer Science and Applications*, *13*. https://doi.org/10.14569/IJACSA.2022.0130982

Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2016). Enriching Word Vectors with Subword Information. *CoRR, abs/1607.04606*. http://arxiv.org/abs/1607.04606

Cardot, H., Cénac, P., & Monnez, J.-M. (2012). A fast and recursive algorithm for clustering large datasets with k-medians. *Computational Statistics & Data Analysis*, *56*(6), 1434–1449. https://doi.org/https://doi.org/10.1016/j.csda.2011.11.019

Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). *What Does BERT Look At? An Analysis of BERT's Attention*. http://arxiv.org/abs/1906.04341

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

Ellickson, P. B., Kar, W., Reeder, J. C., & Zeng, G. (2024). *Using Contextual Embeddings to Predict the Effectiveness of Novel Heterogeneous Treatments*. https://writings.stephenwolfram.com/2023/02/what-is-chatgpt-doing-and-why-does-it-work/

Gao, C. X., Dwyer, D., Zhu, Y., Smith, C. L., Du, L., Filia, K. M., Bayer, J., Menssink, J. M., Wang, T., Bergmeir, C., Wood, S., & Cotton, S. M. (2023). An overview of clustering methods with guidelines for application in mental health research. In *Psychiatry Research* (Vol. 327). Elsevier Ireland Ltd. https://doi.org/10.1016/j.psychres.2023.115265

Garcia-Dias, R., Vieira, S., Lopez Pinaya, W. H., & Mechelli, A. (2019). Clustering analysis. In *Machine Learning: Methods and Applications to Brain Disorders* (pp. 227–247). Elsevier. https://doi.org/10.1016/B978-0-12-815739-8.00013-4

Greenacre, M., Groenen, P. J. F., Hastie, T., Iodice D'enza, A., Markos, A., Tuzhilina, E., & Iodice D'enza, A. (2023). *Economics Working Paper Series Principal component analysis Principal Component Analysis*.

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. In *Source: Journal of the Royal Statistical Society. Series C (Applied Statistics)* (Vol. 28, Issue 1).

Hasan, B. M. S., & Abdulazeez, A. M. (2021). A Review of Principal Component Analysis Algorithm for Dimensionality Reduction. *Journal of Soft Computing and Data Mining*, *2*(1), 20–30. https://doi.org/10.30880/jscdm.2021.02.01.003

Hendrastuty, N. (2024). *Penerapan Data Mining Menggunakan Algoritma K-Means Clustering Dalam Evaluasi Hasil Pembelajaran Siswa*. https://doi.org/10.58602/jima-ilkom.v3i1.26

Hu, W., Xu, D., & Niu, Z. (2021). Improved K-Means Text Clustering Algorithm Based on BERT and Density Peak. *2021 2nd Information Communication Technologies Conference, ICTC 2021*, 260–264. https://doi.org/10.1109/ICTC51749.2021.9441505

Keogh Eamonn and Mueen, A. (2017). Curse of Dimensionality. In G. I. Sammut Claude and Webb (Ed.), *Encyclopedia of Machine Learning and Data Mining* (pp. 314–315). Springer US. https://doi.org/10.1007/978-1-4899-7687-1_192

Kherif, F., & Latypova, A. (2020). Principal component analysis. In *Machine Learning* (pp. 209–225). Elsevier. https://doi.org/10.1016/B978-0-12-815739-8.00012-2

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP. *CoRR*, *abs/2011.00677*. https://arxiv.org/abs/2011.00677

Lenssen, L., & Schubert, E. (2022). Clustering by Direct Optimization of the Medoid Silhouette. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *13590 LNCS*, 190–204. https://doi.org/10.1007/978-3-031-17849-8_15

Liétard, B., Denis, P., & Keller, M. (2024). *To Word Senses and Beyond: Inducing Concepts with Contextualized Language Models*. http://arxiv.org/abs/2406.20054

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly

Optimized BERT Pretraining Approach. *CoRR*, *abs/1907.11692*. http://arxiv.org/abs/1907.11692

Milligan, G. W. (1980). An examination of the effect of six types of error perturbation on fifteen clustering algorithms. *Psychometrika*, *45*(3), 325–342. https://doi.org/10.1007/BF02293907

Omran, M. G. H., Engelbrecht, A. P., & Salman, A. (2007). An overview of clustering methods. In *Intelligent Data Analysis* (Vol. 11, Issue 6, pp. 583–605). IOS Press. https://doi.org/10.3233/ida-2007-11602

Oyewole, G. J., & Thopil, G. A. (2023). Data clustering: application and trends. *Artificial Intelligence Review*, *56*(7), 6439–6475. https://doi.org/10.1007/s10462-022-10325-y

Peng, K., Leung, V. C. M., & Huang, Q. (2018). Clustering Approach Based on Mini Batch Kmeans for Intrusion Detection System over Big Data. *IEEE Access*, *6*, 11897–11906. https://doi.org/10.1109/ACCESS.2018.2810267

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global Vectors for Word Representation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 1532–1543). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). *Deep contextualized word representations*. http://arxiv.org/abs/1802.05365

Petukhova, A., Matos-Carvalho, J. P., & Fachada, N. (2024). *Text Clustering with LLM Embeddings*.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). *A Primer in BERTology: What We Know About How BERT Works*. https://doi.org/10.1162/tacl

Sarıtaş, K., Öz, C. A., & Güngör, T. (2024). A comprehensive analysis of static word embeddings for Turkish. *Expert Systems with Applications*, *252*. https://doi.org/10.1016/j.eswa.2024.124123

Sellam, T., Yadlowsky, S., Wei, J., Saphra, N., D'Amour, A., Linzen, T., Bastings, J., Turc, I., Eisenstein, J., Das, D., Tenney, I., & Pavlick, E. (2021). *The MultiBERTs: BERT Reproductions for Robustness Analysis*. http://arxiv.org/abs/2106.16163

Shahapure, K. R., & Nicholas, C. (2020). Cluster quality analysis using silhouette score. *Proceedings - 2020 IEEE 7th International Conference on Data Science and Advanced Analytics, DSAA 2020*, 747–748. https://doi.org/10.1109/DSAA49011.2020.00096

Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., & Liu, J. (2021). A quantitative discriminant method of elbow point for the optimal number of clusters in clustering algorithm. *Eurasip Journal on*

Wireless Communications and Networking, *2021*(1). https://doi.org/10.1186/s13638-021-01910-w

Siregar, R., & Prayudha, A. S. (2024). Implementation Of The Rational Unified Process Method In The Web-Based Profile Information System At The Ica Aquarium. *Acceleration, Quantum, Information Technology and Algorithm Journal*, *1*(1), 28–36. https://doi.org/10.62123/aqila.v1i1.28

Subakti, A., Murfi, H., & Hariadi, N. (2022). The performance of BERT as data representation of text clustering. *Journal of Big Data*, *9*(1). https://doi.org/10.1186/s40537-022-00564-9

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need*. http://arxiv.org/abs/1706.03762

Wadud, M. A. H., Mridha, M. F., & Rahman, M. M. (2022). Word Embedding Methods for Word Representation in Deep Learning for Natural Language Processing. *Iraqi Journal of Science*, *63*(3), 1349–1361. https://doi.org/10.24996/ijs.2022.63.3.37

Xia, S., Peng, D., Meng, D., Zhang, C., Wang, G., Giem, E., Wei, W., & Chen, Z. (2022). Ball k-Means: Fast Adaptive Clustering With No Bounds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *44*(1), 87–99. https://doi.org/10.1109/TPAMI.2020.3008694

Xu, Q., Gu, H., & Ji, S. W. (2023). Text clustering based on pre-trained models and autoencoders. *Frontiers in Computational Neuroscience*, *17*. https://doi.org/10.3389/fncom.2023.1334436

Yuan, C., & Yang, H. (2019). Research on K-Value Selection Method of K-Means Clustering Algorithm. *J*, *2*(2), 226–235. https://doi.org/10.3390/j2020016

Zheng, X., Lei, Q., Yao, R., Gong, Y., & Yin, Q. (2018). Image segmentation based on adaptive K-means algorithm. *Eurasip Journal on Image and Video Processing*, *2018*(1). https://doi.org/10.1186/s13640-018-0309-3