ANALISIS SEBARAN TOPIK ARTIKEL ILMIAH MENGGUNAKAN INDOBERT DAN K-MEANS

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

Oleh :

SITI MARITZA AQILA
NIM: 09021282126048

Jurusan Teknik Informatika
FAKULTAS ILMU KOMPUTER UNIVERSITAS SRIWIJAYA
2024

LEMBAR PENGESAHAN SKRIPSI

ANALISIS SEBARAN TOPIK ARTIKEL ILMIAH MENGGUNAKAN INDOBERT DAN K-MEANS

Oleh :

Siti Maritza Aqila
NIM: 09021282126048

Indralaya, 31 Desember 2024

Mengetahui,
Ketua Jurusan Teknik Informatika

Hadipurnawan Satria, Ph.D.
NIP. 198004182020121001

Pembimbing,

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

TANDA LULUS UJIAN KOMPREHENSIF

Pada hari Selasa tanggal 31 Desember 2024 telah dilaksanakan ujian komprehensif skripsi oleh Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.

Nama	:	Siti Maritza Aqila
Nim	:	09021282126048
Judul	:	Analisis Sebaran Topik Artikel Ilmiah Menggunakan IndoBERT dan K-Means

Dan dinyatakan LULUS.

1. Ketua Penguji

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

2. Penguji I

   M. Qurhanul Rizqie, M.T., Ph.D.
   NIP. 198712032022031006 .....

3. Pembimbing

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

Mengetahui,
Ketua Jurusan Teknik Informatika

Alvi Syahrini Utami, S.Kom., M.Kom.
NIP. 197503182001122002

HALAMAN PERNYATAAN

Yang bertanda tangan di bawah ini :

Nama	: Siti Maritza Aqila
NIM	: 09021282126048
Program Studi	: Teknik Informatika
Judul Skripsi	: Analisis Sebaran Topik Artikel Ilmiah Menggunakan IndoBERT dan K-Means

Hasil Pengecekan *Software iThenticate/Turnitin*: 5%

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

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

Palembang, 6 Januari 2024

Siti Maritza Aqila

NIM. 09021282126048

MOTTO DAN PERSEMBAHAN

"Setiap halaman yang ditulis adalah bukti perjuangan, doa, dan dukungan tulus dari orang-orang terdekat."

Saya persembahkan karya tulis ini kepada:

- Orangtua
- Keluarga Besar
- Sahabat
- Teman Seperjuangan
- Fakultas Ilmu Komputer
- Universitas Sriwijaya

Abstract

The significant increase in the number of scientific articles published along with advances in science and technology poses challenges in managing and organizing articles to support the efficiency of literature analysis. This research aims to develop a scientific article topic distribution analysis system by utilizing the IndoBERT model and the K-Means algorithm. The dataset used comes from 12 nationally accredited scientific journals on the Science and Technology Index (SINTA). Clustering evaluation was conducted using silhouette score, with the highest value of 0.932656 in journals with EISSN 25799258. This clustering model is then applied to predict new articles based on the relevance of their journal topics, with inscope or outscope categories determined based on the threshold value.

Keywords: *topic distribution analysis of scientific articles, IndoBERT, K-Means, silhouette score, article prediction.*

Abstrak

Peningkatan jumlah artikel ilmiah yang diterbitkan secara signifikan seiring dengan kemajuan ilmu pengetahuan dan teknologi menimbulkan tantangan dalam pengelolaan dan pengorganisasian artikel untuk mendukung efisiensi analisis literatur. Penelitian ini bertujuan untuk mengembangkan sistem analisis sebaran topik artikel ilmiah dengan memanfaatkan model IndoBERT dan algoritma K-Means. Dataset yang digunakan berasal dari 12 jurnal ilmiah terakreditasi nasional pada *Science and Technology Index* (SINTA). Evaluasi klasterisasi dilakukan menggunakan *silhouette score*, dengan nilai tertinggi sebesar 0,932656 pada jurnal dengan EISSN 25799258. Model klasterisasi ini selanjutnya diterapkan untuk memprediksi artikel baru berdasarkan relevansi topik jurnalnya, dengan kategori *inscope* atau *outscope* yang ditentukan berdasarkan nilai ambang batas (*threshold*).

Kata kunci: analisis sebaran topik artikel ilmiah, IndoBERT, K-Means, *silhouette score*, prediksi artikel.

KATA PENGANTAR

Segala puji dan syukur penulis panjatkan kepada Allah SWT atas limpahan rahmat dan karunia Nya sehingga penulis dapat menyelesaikan skripsi ini dengan baik. Penulisan skripsi ini dilakukan sebagai salah satu syarat untuk menyelesaikan pendidikan program Strata-1 di Fakultas Ilmu Komputer Universitas Sriwijaya. Dalam proses penyelesaian skripsi ini, penulis memperoleh berbagai bantuan, bimbingan, serta dukungan dari berbagai pihak, baik secara langsung maupun tidak langsung. Oleh karena itu, penulis ingin menyampaikan rasa terima kasih kepada:

1. Allah SWT atas rahmat dan karunia-Nya penulis dapat menyelesaikan skripsi ini dengan baik.
2. Kedua orang tua dan keluarga besar yang telah mendoakan, memberi semangat, motivasi dan nasihat dalam menyelesaikan skripsi ini.
3. Bapak Hadipurnawan Satria, S.Kom. M.Sc, Ph.D. selaku Ketua Jurusan Teknik Informatika Universitas Sriwijaya.
4. Bapak Rifkie Primartha, S.T., M.T. selaku Dosen Pembimbing Akademik yang telah memberikan bantuan dan arahan kepada penulis selama perkuliahan.
5. Ibu Novi Yusliani, S.Kom., M.T. selaku Dosen Pembimbing yang telah membimbing serta memberikan arahan kepada penulis selama proses pengerjaan skripsi.
6. Seluruh dosen program studi serta admin Jurusan Teknik Informatika Fakultas Ilmu Komputer Universitas Sriwijaya.
7. Seluruh Staf Administrasi dan Pegawai Fakultas Ilmu Komputer yang telah membantu dalam urusan administrasi tugas akhir penulis.
8. Seluruh sahabat dan teman-teman yang telah membantu, memberikan saran, motivasi, dan semangat kepada penulis khususnya Rachma, Valyssa, Yolen, Oca, Ori, dan Meutia.
9. Pihak-pihak lain yang tidak dapat penulis sebutkan satu per satu.

Penulis menyadari bahwa dalam penyusunan skripsi ini masih banyak sekali kekurangan terhadap penyusunan tugas akhir ini. Oleh karena itu, penulis mengharapkan saran dan kritik yang membangun guna kemajuan penelitian kedepannya. Semoga tugas akhir ini dapat bermanfaat. Terima kasih.

Palembang, 24 Desember 2024

Penulis,

Siti Maritza Aqila

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

BAB I

PENDAHULUAN

1.1 Pendahuluan

Bab pendahuluan ini akan menguraikan mengenai latar belakang masalah, rumusan masalah, tujuan penelitian, manfaat penelitian, batasan masalah penelitian, dan sistematika penulisan yang akan menjadi landasan bagi pembahasan di bab berikutnya. Secara keseluruhan, penelitian ini bertujuan untuk menganalisis sebaran topik artikel ilmiah berdasarkan kesamaan konten menggunakan IndoBERT dan K-Means. Penelitian ini diharapkan dapat memberikan wawasan tentang pola distribusi topik artikel ilmiah sehingga mendukung efisiensi pengelolaan literatur ilmiah secara lebih akurat dan terstruktur.

1.2 Latar Belakang

Jumlah artikel ilmiah yang diterbitkan mengalami peningkatan pesat dalam beberapa tahun terakhir, sejalan dengan kemajuan ilmu pengetahuan dan teknologi. Banyaknya artikel ilmiah ini menimbulkan tantangan besar dalam hal pengelolaan dan pengorganisasian artikel-artikel tersebut. Analisis sebaran topik artikel berdasarkan relevansi isi merupakan langkah krusial dalam dunia penelitian untuk meningkatkan efisiensi dalam analisis literatur yang relevan. Akan tetapi, proses pengelompokan atau analisis sebaran secara manual sering kali memakan waktu lama dan berisiko mengalami kesalahan manusia (*human error*). Oleh karena itu, dibutuhkan sistem yang mampu membantu dalam memahami pola distribusi topik artikel ilmiah.

Salah satu teknologi yang banyak digunakan dalam analisis data untuk pengelompokan dan memahami sebaran topik adalah metode K-Means. Algoritma ini berfungsi dengan mengelompokkan data ke dalam beberapa grup berdasarkan kesamaan atribut yang dimiliki. K-Means dapat digunakan untuk mengidentifikasi pola distribusi artikel berdasarkan kesamaan topik atau kata kunci yang muncul dalam teks artikel tersebut (MacQueen, 1967). Algoritma ini dikenal karena kemampuannya dalam memproses data dalam volume besar dengan waktu komputasi yang relatif cepat (Hartigan & Wong, 1979). Selain itu, K-Means juga memiliki implementasi yang sederhana sehingga membuat algoritma ini mudah dipahami dan diaplikasikan, sehingga cocok untuk aplikasi yang membutuhkan analisis pola data yang cepat dan jelas seperti dalam penelitian pasar dan pengelompokan data sosial (Jain, 2010).

Dalam analisis sebaran topik menggunakan K-Means, penggunaan model bahasa adalah hal yang sangat penting. Pada penelitian ini, penulis menggunakan model bahasa IndoBERT yang merupakan salah satu varian dari BERT. BERT telah menjadi salah satu model terkemuka dalam pemrosesan bahasa alami karena kemampuannya untuk menangani tugas-tugas seperti klasifikasi teks, pencocokan kalimat, dan ekstraksi informasi dengan tingkat akurasi yang tinggi. Model ini dilatih menggunakan metode pra-pelatihan dan fine-tuning, yang memungkinkan adaptasi terhadap berbagai jenis tugas bahasa yang berbeda dengan data yang spesifik. BERT dipilih karena mampu menghasilkan embedding teks dengan kualitas tinggi melalui proses encoding yang mendalam pada elemen terkecil dokumen (Devlin et al., 2018). Oleh karena itu, penelitian ini menggunakan model

IndoBERT untuk memahami teks dan K-Means untuk menganalisis pola distribusi topik artikel ilmiah.

1.3 Rumusan Masalah

Berdasarkan latar belakang yang telah dijelaskan, rumusan masalah yang dapat diidentifikasi dalam penelitian ini adalah:

1. Bagaimana mengimplementasikan metode IndoBERT dan K-Means untuk menganalisis sebaran topik artikel ilmiah?
2. Bagaimana kualitas hasil analisis sebaran topik artikel ilmiah menggunakan IndoBERT dan K-Means berdasarkan *Silhouette Score?*

1.4 Tujuan Penelitian

Adapun tujuan dari penelitian ini adalah sebagai berikut:

1. Menghasilkan sebuah sistem yang mampu menganalisis sebaran topik artikel ilmiah menggunakan IndoBERT dan K-Means.
2. Menghasilkan kualitas hasil analisis sebaran topik artikel ilmiah menggunakan IndoBERT dan K-Means berdasarkan *Silhouette Score.*

1.5 Manfaat Penelitian

Manfaat penelitian ini adalah:

1. Sistem dapat membantu menganalisis sebaran topik artikel imiah berdasarkan konten yang sesuai.

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

1.6 Batasan Masalah

Adapun batasan masalah pada penelitian ini adalah:

1. Data yang digunakan berupa data artikel ilmiah berbahasa Indonesia.
2. Data yang digunakan berasal dari 12 jurnal yaitu E-Journal Of Cultural Studies, E-Jurnal Akuntansi, Gadjah Mada International Journal Of Business, International Journal Of Basic And Applied Science, International Jounal Of Elementary Education, Jurnal Akuntansi dan Keuangan, Jurnal Bisnis dan Akuntansi, Jurnal Cendekia : Jurnal Pendidikan Matematika, Jurnal Kesehatan Andalas, Jurnal Kesehatan Masyarakat, Jurnal Obsesi: Jurnal Pendidikan Anak Usia Dini, Jurnal Pendidikan Teknik Mesin Undiksha.

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 IndoBERT, 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 menguraikan proses perancangan perangkat lunak, dimulai dari analisis kebutuhan hingga tahap pengujian untuk mengevaluasi hasil pengembangan perangkat lunak tersebut.

BAB V. HASIL DAN PENELITIAN

Bab ini memaparkan hasil penelitian yang dilaksanakan berdasarkan langkah-langkah dan metode yang telah dirancang. Analisis yang disajikan dalam bab ini akan menjadi dasar dalam merumuskan kesimpulan dari penelitian ini.

BAB VI. KESIMPULAN DAN SARAN

Bab ini memaparkan kesimpulan dan saran dari penelitian yang dilakukan berdasarkan uraian pada bab sebelumnya. Hasil dari penelitian ini diharapkan dapat menjadi bahan referensi atau acuan untuk penelitian selanjutnya dan menghasilkan sistem yang lebih baik lagi.

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

Chandra, C. M., & Rachmadi, M. (2024). Perbandingan Pengembangan Sistem Dengan Pendekatan Konvensional dan Low-Code Pada Sistem Pendukung Keputusan Penilaian Kinerja Pegawai. *JURNAL PENELITIAN SISTEM INFORMASI (JPSI)*, *2*(2), 14–27. https://doi.org/10.54066/jpsi.v2i2.1729

Chang, D.-H., Wang, Y.-H., Hsieh, C.-Y., Chang, C.-W., Chang, K.-C., & Chen, Y.-S. (2021). Incorporating Patient Preferences into a Decision-Making Model of Hand Trauma Reconstruction. *International Journal of Environmental Research and Public Health*, *18*(21), 11081. https://doi.org/10.3390/ijerph182111081

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

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR*, *abs/1810.04805*. http://arxiv.org/abs/1810.04805

Dewi, D. A. I. C., & Pramita, D. A. K. (2019). *Analisis Perbandingan Metode Elbow dan Silhouette pada Algoritma Clustering K-Medoids dalam Pengelompokan Produksi Kerajinan Bali | Matrix: Jurnal Manajemen Teknologi dan Informatika*. https://ojs.pnb.ac.id/index.php/matrix/article/view/1662

Dewi, S., & Pakereng, M. A. I. (2023). IMPLEMENTASI PRINCIPAL COMPONENT ANALYSIS PADA K-MEANS UNTUK KLASTERISASI TINGKAT PENDIDIKAN PENDUDUK KABUPATEN SEMARANG. *JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)*, *8*(4), Article 4. https://doi.org/10.29100/jipi.v8i4.4101

Dharmawan, S., Mawardi, V. C., & Perdana, N. J. (2023). Klasifikasi Ujaran Kebencian Menggunakan Metode FeedForward Neural Network (IndoBERT). *Jurnal Ilmu Komputer Dan Sistem Informasi*, *11*(1), Article 1. https://doi.org/10.24912/jiksi.v11i1.24066

Dutta, A., Pal, A., Bhadra, M., Khan, M. A., & Chakraborty, R. (2021). An Improved K-Means Algorithm for Effective Medical Image Segmentation. In J. K. Mandal, S. Mukhopadhyay, A. Unal, & S. K. Sen (Eds.), *Proceedings of International Conference on Innovations in Software Architecture and Computational Systems* (pp. 169–182). Springer. https://doi.org/10.1007/978-981-16-4301-9_13

Guntara, M., & Lutfi, N. (2023). Optimasi Cacah Klaster pada Klasterisasi dengan Algoritma KMeans Menggunakan Silhouette Coeficient dan Elbow Method. *JuTI "Jurnal Teknologi Informasi,"* *2*(1), Article 1. https://doi.org/10.26798/juti.v2i1.944

Gustientiedina, G., Adiya, M. H., & Desnelita, Y. (2019). Penerapan Algoritma K-Means Untuk Clustering Data Obat-Obatan. *Jurnal Nasional Teknologi dan Sistem Informasi*, *5*(1), Article 1. https://doi.org/10.25077/TEKNOSI.v5i1.2019.17-24

Hananto, A. L., Assiroj, P., Priyatna, B., Nurhayati, Fauzi, A., Rahman, A. Y., & Hilabi, S. S. (2021). Analysis Of Drug Data Mining With Clustering Technique Using K-Means Algorithm. *Journal of Physics: Conference Series*, *1908*(1), 012024. https://doi.org/10.1088/1742-6596/1908/1/012024

Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A K-Means Clustering Algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, *28*(1), 100–108. https://doi.org/10.2307/2346830

Hediyati, D., & Suartana, I. M. (2021). Penerapan Principal Component Analysis (PCA) Untuk Reduksi Dimensi Pada Proses Clustering Data Produksi Pertanian Di Kabupaten Bojonegoro. *JIEET (Journal of Information Engineering and Educational Technology)*, *5*(2), 49–54. https://doi.org/10.26740/jieet.v5n2.p49-54

Hidayat, W. A., & Nastiti, V. R. S. (2024). PERBANDINGAN KINERJA PRE-TRAINED INDOBERT-BASE DAN INDOBERT-LITE PADA KLASIFIKASI SENTIMEN ULASAN TIKTOK TOKOPEDIA SELLER CENTER DENGAN MODEL INDOBERT. *JSiI (Jurnal Sistem Informasi)*, *11*(2), Article 2. https://doi.org/10.30656/jsii.v11i2.9168

Ikotun, A. M., Ezugwu, A. E., Abualigah, L., Abuhaija, B., & Heming, J. (2023). K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences*, *622*, 178–210. https://doi.org/10.1016/j.ins.2022.11.139

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, *31*(8), 651–666. https://doi.org/10.1016/j.patrec.2009.09.011

Koto, F., Rahimi, A., Lau, J. H., & Baldwin, T. (2020). *IndoLEM and IndoBERT: A Benchmark Dataset and Pre-trained Language Model for Indonesian NLP* (No. arXiv:2011.00677). arXiv. https://doi.org/10.48550/arXiv.2011.00677

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics: Vol. 5.1* (pp. 281–298). University of California Press. https://projecteuclid.org/ebooks/berkeley-symposium-on-mathematical-statistics-and-probability/Proceedings-of-the-Fifth-Berkeley-Symposium-on-Mathematical-Statistics-and/chapter/Some-methods-for-classification-and-analysis-of-multivariate-observations/bsmsp/1200512992

Muhr, D., Affenzeller, M., & Küng, J. (2023). A Probabilistic Transformation of Distance-Based Outliers. *Machine Learning and Knowledge Extraction*, *5*(3), 782–802. https://doi.org/10.3390/make5030042

Paembonan, S., & Abduh, H. (2021). Penerapan Metode Silhouette Coefficient untuk Evaluasi Clustering Obat. *PENA TEKNIK: Jurnal Ilmiah Ilmu-Ilmu Teknik*, *6*(2), Article 2. https://doi.org/10.51557/pt_jiit.v6i2.659

Patel, H. (2023). *Customer Segmentation using RFM Analysis and K-Means Clustering to enhance Marketing Strategies; Analysis of Algorithmic Bias in Customer Segmentation Models* [University of Virginia]. https://doi.org/10.18130/ZY3K-SX64

Rianti, R., Andarsyah, R., & Awangga, R. M. (2024). Penerapan PCA dan Algoritma Clustering untuk Analisis Mutu Perguruan Tinggi di LLDIKTI Wilayah IV. *NUANSA INFORMATIKA*, *18*(2), Article 2. https://doi.org/10.25134/ilkom.v18i2.211

Rodriguez, M. Z., Comin, C. H., Casanova, D., Bruno, O. M., Amancio, D. R., Costa, L. da F., & Rodrigues, F. A. (2019). Clustering algorithms: A comparative approach. *PLOS ONE*, *14*(1), e0210236. https://doi.org/10.1371/journal.pone.0210236

Rozeva, A., & Zerkova, S. (2017). Assessing semantic similarity of texts – Methods and algorithms. In *AIP Conference Proceedings* (Vol. 1910). https://doi.org/10.1063/1.5014006

Shutaywi, M., & Kachouie, N. N. (2021). Silhouette Analysis for Performance Evaluation in Machine Learning with Applications to Clustering. *Entropy*, *23*(6), Article 6. https://doi.org/10.3390/e23060759

Situmorang, G. F., & Purba, R. (2024). Deteksi Potensi Depresi dari Unggahan Media Sosial X Menggunakan IndoBERT. *Building of Informatics, Technology and Science (BITS)*, *6*(2), Article 2. https://doi.org/10.47065/bits.v6i2.5496

Suraya, G. R., & Wijayanto, A. W. (2022). Comparison of Hierarchical Clustering, K-Means, K-Medoids, and Fuzzy C-Means Methods in Grouping Provinces in Indonesia according to the Special Index for Handling Stunting: Perbandingan Metode Hierarchical Clustering, K-Means, K-Medoids, dan Fuzzy C-Means dalam Pengelompokan Provinsi di Indonesia Menurut

Indeks Khusus Penanganan Stunting. *Indonesian Journal of Statistics and Its Applications*, *6*(2), Article 2. https://doi.org/10.29244/ijsa.v6i2p180-201

Temnenco, V. (2019). *UML , RUP , and the Zachman Framework: Better together*. https://www.semanticscholar.org/paper/UML-%2C-RUP-%2C-and-the-Zachman-Framework-%3A-Better-Temnenco/f23cb95b0245f99dda5c951765a23c91087685a9

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. ukasz, & Polosukhin, I. (2017). Attention is All you Need. *Advances in Neural Information Processing Systems*, *30*. https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Vysala, A., & Gomes, J. (2020). *Evaluating and Validating Cluster Results* (No. arXiv:2007.08034). arXiv. https://doi.org/10.48550/arXiv.2007.08034

Wang, X., Wu, P., Xu, Q., Zeng, Z., & Xie, Y. (2021). Joint image clustering and feature selection with auto-adjoined learning for high-dimensional data. *Knowledge-Based Systems*, *232*, 107443. https://doi.org/10.1016/j.knosys.2021.107443

Zubair, Md., Iqbal, MD. A., Shil, A., Chowdhury, M. J. M., Moni, M. A., & Sarker, I. H. (2022). An Improved K-means Clustering Algorithm Towards an Efficient Data-Driven Modeling. *Annals of Data Science*. https://doi.org/10.1007/s40745-022-00428-2