Analisis Topic Modeling Journal ArXiv Menggunakan Metode K-Means dengan Algoritma Dimensionality Reduction dan t-SNE Model

TUGAS AKHIR

Diajukan Untuk Memenuhi

Persyaratan Guna Meraih Gelar Sarjana

Informatika Universitas Muhammadiyah Malang

Dema Riandi Akhmad

201710370311116

Data Science

PROGRAM STUDI INFORMATIKA

FAKULTAS TEKNIK

UNIVERSITAS MUHAMMADIYAH MALANG

2024

LEMBAR PERSETUJUAN

Analisis Topic Modelling Journal Arxiv Menggunakan Metode K-Means Dengan Algoritma Dimensionality Reduction Dan t-SNE Model

TUGAS AKHIR

Sebagai Persyaratan Guna Meraih Gelar Sarjana Strata 1
Informatika Universitas Muhammadiyah Malang

Menyetujui,
Malang, *29 April 2024*

Dosen Pembimbing 1	Dosen Pembimbing 2
Ir. Ilyas Nuryasin S.Kom., M.Kom.	**Ir. Wildan Suharso S.Kom., M.Kom**
NIP. 10814100561PNS.	**NIP. 10817030596PNS.**

LEMBAR PENGESAHAN

Analisis Topic Modelling Journal Arxiv Menggunakan Metode K-Means Dengan Algoritma Dimensionality Reduction Dan t-SNE Model

TUGAS AKHIR

Sebagai Persyaratan Guna Meraih Gelar Sarjana Strata 1

InformatikaUniversitas Muhammadiyah Malang

Disusun Oleh :

Dema Riandi Akhmad

201710370311116

Tugas Akhir ini telah diuji dan dinyatakan lulus melalui sidang majelis penguji pada tanggal 29 April 2024

Menyetujui,

Dosen Penguji 1	Dosen Penguji 2
Ir. Galih Wasis Wicaksono S.kom. M.Cs.	**Hardianto Wibowo S.Kom, MT.**
NIP. 10814100541PNS.	**NIP. 10816120592PNS.**

Mengetahui,
Ketua Jurusan Informatika

Ir. Galih Wasis Wicaksono S.kom. M.Cs.
NIP. 10814100541PNS.

LEMBAR PERNYATAAN

Yang bertanda tangan dibawah ini :

NAMA : **Dema Riandi Akhmad**

NIM : **201710370311116**

FAK./JUR. : **Informatika**

Dengan ini saya menyatakan bahwa Tugas Akhir dengan judul **"Analisis Topic Modelling Journal Arxiv Menggunakan Metode K-Means Dengan Algoritma Dimensionality Reduction Dan t-SNE Model"** beserta seluruh isinya adalah karya saya sendiri dan bukan merupakan karya tulis orang lain, baik sebagian maupun seluruhnya, kecuali dalam bentuk kutipan yang telah disebutkan sumbernya.

Demikian surat pernyataan ini saya buat dengan sebenar-benarnya. Apabila kemudian ditemukan adanya pelanggaran terhadap etika keilmuan dalam karya saya ini, atau ada klaim dari pihak lain terhadap keaslian karya saya ini maka saya siap menanggung segala bentuk resiko/sanksi yang berlaku.

Mengetahui,
Dosen Pembimbing

Ir. Ilyas Nuryasin S.Kom., M.Kom.

Malang, 29 April 2024
Yang Membuat Pernyataan

Dema Riandi Akhmad

ABSTRAK

Sinergi antara *Cluster* K-means dan *Principal Component Analysis* (PCA) dalam pemodelan topik menghadirkan pendekatan yang ampuh untuk menyaring pola dari data tekstual. *Cluster* K-means, sebuah teknik pembelajaran tanpa pengawasan, unggul dalam mengelompokkan titik data yang serupa, menjadikannya penting dalam mengelompokkan konten tekstual ke dalam topik yang koheren. Proses ini memungkinkan ekstraksi tema atau subjek mendasar dalam kumpulan data yang luas. PCA berfungsi sebagai metode reduksi dimensi, mengungkap pola-pola penting dengan mengurangi kompleksitas data berdimensi tinggi. Ketika diterapkan pada pemodelan topik, PCA membantu mengidentifikasi fitur atau dimensi yang paling berpengaruh dalam kumpulan data tekstual, meningkatkan kemampuan interpretasi dan memfasilitasi analisis yang mendalam. Integrasi *Cluster* K-means dan PCA menawarkan kerangka kerja yang kuat untuk pemodelan topik yang efisien. Dengan menggunakan K-means untuk mengkategorikan data teks ke dalam kelompok yang mewakili topik berbeda dan selanjutnya memanfaatkan PCA untuk reduksi dimensi, metodologi gabungan ini memberdayakan peneliti untuk mengungkap dan memahami tema laten secara efektif. Kesimpulannya, penggabungan *Cluster* K-means dengan PCA mewakili jalan yang menjanjikan bagi para peneliti untuk mencari wawasan berbeda dari kumpulan data tekstual. Pendekatan terpadu ini memfasilitasi ekstraksi topik yang komprehensif, membantu penemuan pengetahuan dan proses pengambilan keputusan di berbagai domain. Penelitian kali ini menunjukan antara metode Kmeans dengan menggunakan algoritma pereduksi dimensi *PCA* memberikan hasil yang sangat memuaskan dengan mengkombinasi visualisasi antara algortma t-SNE dan UMAP.

Kata Kunci: *Cluster, Topic Modelling, Kmeans, Principal Component Analysis (PCA), t-SNE, UMAP.*

ABSTRACT

The synergy between Cluster K-means and Principal Component Analysis (PCA) in topic modeling presents a powerful approach to filter patterns from textual data. K-means clusters, an unsupervised learning technique, excel at grouping similar data points, making it important in grouping textual content into coherent topics. This process allows the extraction of underlying themes or subjects within a broad data set. PCA functions as a dimensionality reduction method, revealing important patterns by reducing the complexity of high-dimensional data. When applied to topic modeling, PCA helps identify the most influential features or dimensions in a textual dataset, improving interpretability and facilitating in-depth analysis. The integration of Cluster K-means and PCA offers a powerful framework for efficient topic modeling. By using K-means to categorize text data into groups representing different topics and further utilizing PCA for dimensionality reduction, this combined methodology empowers researchers to uncover and understand latent themes effectively. In conclusion, combining Cluster K-means with PCA represents a promising avenue for researchers to extract different insights from textual datasets. This integrated approach facilitates comprehensive topic extraction, aiding knowledge discovery and decision-making processes across multiple domains. This research shows that the Kmeans method using the PCA dimension reduction algorithm provides very satisfactory results by combining visualization between the t-SNE and UMAP algorithms.

Keywords: Cluster, Topic Modeling, Kmeans, Principal Component Analysis (PCA), t-SNE, UMAP.

KATA PENGANTAR

Puji syukur dipanjatkan kepada Allah SWT atas berkah, Rahmat dan karunia-Nya sehingga penulis dapat menyelesaikan Tugas Akhir yang berjudulkan, "Analisis Topic Modelling Journal Arxiv Menggunakan Metode K-Means Dengan Algoritma Dimensionality Reduction Dan T-SNE Model". Banyak terima kasih yang ingin saya ucapkan terhadap semua rekan yang telah ikut andil dalam membantu serta memberikan dukungan maupun inspirasi dalam proses penulisan Tugas Akhir ini. Berikut adalah beberapa rekan, kerabat dan entitas yang ingin saya beri persembahan:

1. Bapak Ilyas Nuryasin, S.kom, M.Kom. selaku Dosen Pembimbing 1 serta Dosen wali kelas dan Bapak Briansyah Ir Wildan Suharso, S.Kom, M.Kom selaku Dosen Pembimbing 2 yang telah bersedia meluangkan waktu untuk membimbing, membantu dan memberikan arahan kepada penulis sehingga dapat menyelesaikan penelitian ini.
2. Bapak/Ibu Dekan Fakultas Teknik Universitas Muhammadiyah Malang.
3. Bapak/Ibu Ketua Jurusan Teknik Informatika Universitas Muhammadiyah Malang.
4. Kepada kedua Ayah dan Ibu saya serta keluarga yang selalu memberikan dukungan moral, finansial hingga motivasi, dan terima kasih juga atas kesabaran yang diberikan sepanjang proses pengerjaan tugas akhir ini dari awal hingga akhir.
5. Terimakasih juga untuk saudara saya Sahedi Ceking dan orang terdekat saya Lita Anggraini yang telah memberikan dukungan berupa bantuan, semangat hingga motivasi selama proses penelitian ini.
6. Terimakasih kepada rekan-rekan Informatika Angkatan 2017 yang sudah memberikan banyak memori serta dukungan selama menempuh Pendidikan di Universitas Muhammadiyah Malang.
7. Tak luput terimakasih juga penulis ucapkan kepada pihak yang belum dapat tersebutkan.

Tanpa para pihak-pihak penting tersebut, penelitian yang dilakukan untuk menyelesaikan tugas akhir ini tidak dapat diselesaikan dengan mudah dan berhasil. Penulis juga mengakui bahwa penulisan ilmiah makalah ini masih jauh dari kata baik dan masih terdapat beberapa kekurangan. Penulis sangat berharap agar penelitian ini dapat memberikan manfaat yang besar bagi pihak-pihak penting lainnya di masa yang akan datang.

DAFTAR ISI

DAFTAR PUSTAKA

[1] J. others MacQueen, "Some methods for classification and analysis of multivariate observations," *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, vol. 1, no. 14, pp. 281–297, 1967.

[2] M. E. Eren, N. Solovyev, E. Raff, C. Nicholas, and B. Johnson, "COVID-19 Kaggle Literature Organization," *Proceedings of the ACM Symposium on Document Engineering, DocEng 2020*, 2020, doi: 10.1145/3395027.3419591.

[3] N. Jangamreddy, "Visualizing and Understanding the Relationship between K-Means Clustering , PCA and Linear Auto encoder," *Researchgate*, no. July, 2019, doi: 10.13140/RG.2.2.17341.82406.

[4] K. Kahloot and P. Ekler, "Improving t-SNE Visualization and Clustering," 2019.

[5] H. Bock and R. Aachen, "A history of the k-means algorithm," *Selected contributions in data analysis and classification. Springer Berlin Heidelberg*, pp. 161–172, 2007.

[6] A. G. Lalayan, "Data Compression-Aware Performance Analysis of Dask and Spark for Earth Observation Data Processing," *Mathematical Problems of Computer Science*, vol. 59, May 2023, doi: 10.51408/1963-0100.

[7] M. Moreno, R. Vilaça, and P. G. Ferreira, "Scalable transcriptomics analysis with Dask: applications in data science and machine learning," *BMC Bioinformatics*, vol. 23, no. 1, Dec. 2022, doi: 10.1186/s12859-022-05065-3.

[8] Y. Li and B. Shen, "Research on Sentiment Analysis of Microblogging Based on LSA and TF-IDF," 2017.

[9] Z. Zhou, J. Qin, X. Xiang, Y. Tan, Q. Liu, and N. N. Xiong, "News text topic clustering optimized method based on TF-iDF algorithm on spark," *Computers, Materials and Continua*, vol. 62, no. 1, pp. 217–231, 2020, doi: 10.32604/cmc.2020.06431.

[10] Y. Yang, "Research and realization of internet public opinion analysis based on improved TF - IDF algorithm," 2017, doi: 10.1109/DCABES.2017.24.

[11] A. Abdulhafedh, "Incorporating K-means, Hierarchical Clustering and PCA in Customer Segmentation," *Journal of City and Development*, vol. 3, no. 1, pp. 12–30, 2021, doi: 10.12691/jcd-3-1-3.

[12] T. D. Tsvetkov and I. G. Iliev, "Channel Activity Analysis of Cognitive Radio with PCA Preprocessing and Different Clustering Methods," in *2021 29th National Conference with International Participation (TELECOM)*, IEEE, Oct. 2021, pp. 20–23. doi: 10.1109/TELECOM53156.2021.9659676.

[13] B. M. Devassy, S. George, and P. Nussbaum, "Unsupervised clustering of hyperspectral paper data using T-SNE," *J Imaging*, vol. 6, no. 5, 2020, doi: 10.3390/JIMAGING6050029.

[14] F. Gong, F. Bu, Y. Zhang, Y. Yan, R. Hu, and M. Dong, "Visual Clustering Analysis of Electricity Data Based on t-SNE," *2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics, ICCCBDA 2020*, pp. 234–240, 2020, doi: 10.1109/ICCCBDA49378.2020.9095574.

[15] L. P. Liu *et al.*, "Application of K-Means ++ Algorithm Based on t-SNE Dimension Reduction in Transformer District Clustering," *2020 Asia Energy and Electrical Engineering Symposium, AEEES 2020*, no. 2, pp. 74–78, 2020, doi: 10.1109/AEEES48850.2020.9121438.

[16] S. Arora, W. Hu, and P. K. Kothari, "An Analysis of the t-SNE Algorithm for Data Visualization," vol. 75, no. 2008, pp. 1–8, 2018, [Online]. Available: http://arxiv.org/abs/1803.01768

[17] D. Kobak, G. Linderman, S. Steinerberger, Y. Kluger, and P. Berens, "Heavy-Tailed Kernels Reveal a Finer Cluster Structure in t-SNE Visualisations," *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11906 LNAI, pp. 124–139, 2020, doi: 10.1007/978-3-030-46150-8_8.

[18] B. Dash, D. Mishra, A. Rath, and M. Acharya, "A hybridized K-means clustering approach for high dimensional dataset," *International Journal of Engineering, Science and Technology*, vol. 2, no. 2, pp. 59–66, 2010, doi: 10.4314/ijest.v2i2.59139.

[19] R. A. Indraputra and R. Fitriana, "K-Means Clustering Data COVID-19," *Jurnal Teknik Industri*, vol. 10, no. 3, pp. 275–282, 2020.

[20] C. Ding and X. He, "K-means clustering via principal component analysis," *Proceedings, Twenty-First International Conference on Machine Learning, ICML 2004*, pp. 225–232, 2004, doi: 10.1145/1015330.1015408.

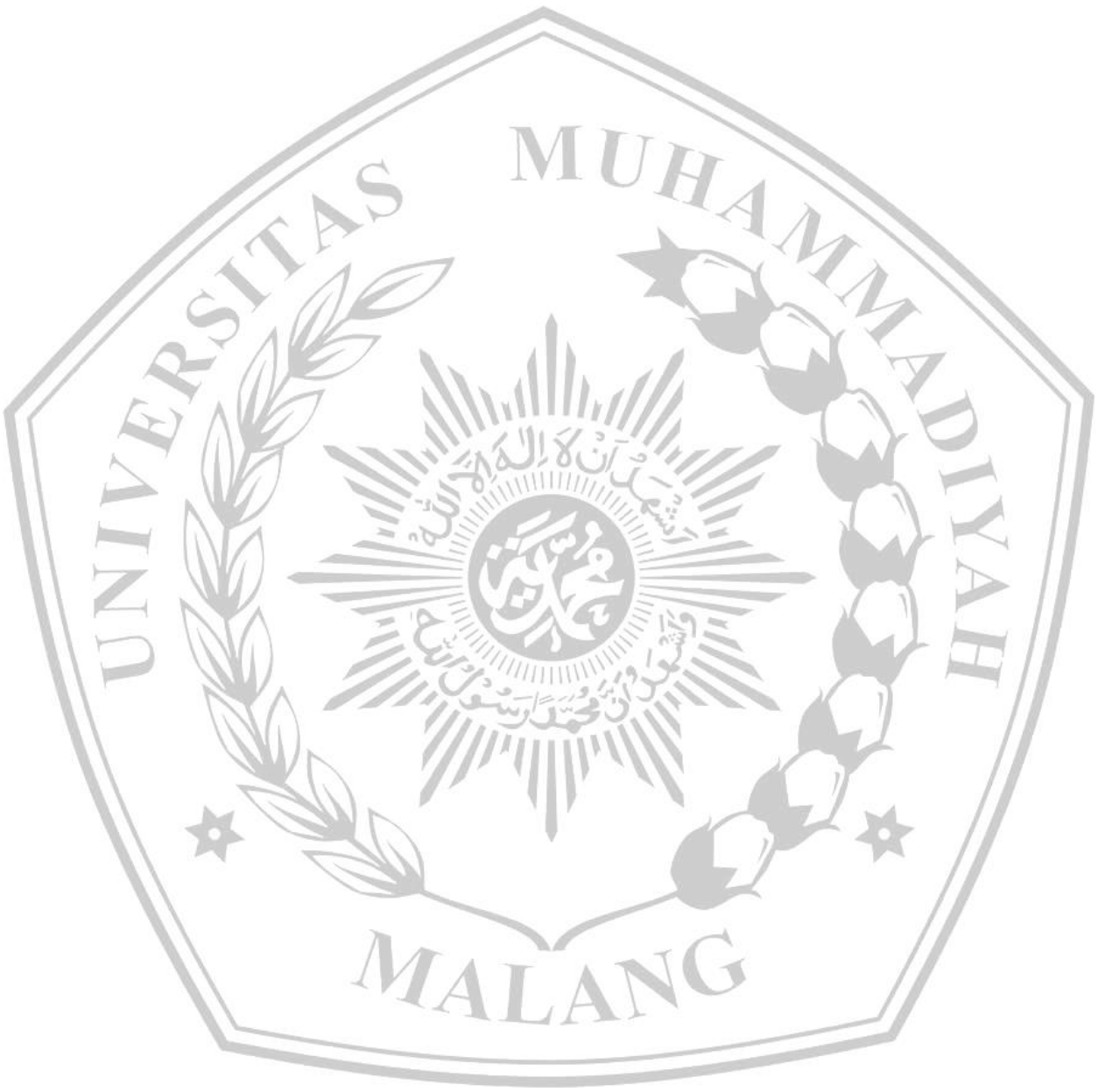
[21] Y. Hozumi, R. Wang, C. Yin, and G. W. Wei, "UMAP-assisted K-means clustering of large-scale SARS-CoV-2 mutation datasets," *Comput Biol Med*, vol. 131, 2021, doi: 10.1016/j.compbiomed.2021.104264.

[22] A. Halgekar, A. Rao, D. Khankhoje, I. Khetan, and K. Bhowmick, "Topic Modelling-Based Approach for Clustering Legal Documents," 2023, pp. 163–173. doi: 10.1007/978-981-19-0095-2_17.

[23] C. Byrne, D. Horak, K. Moilanen, and A. Mabona, "Topic Modeling With Topological Data Analysis," in *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA, USA: Association for Computational Linguistics, 2022, pp. 11514–11533. doi: 10.18653/v1/2022.emnlp-main.792.

[24] D. Mazzei and R. Ramjattan, "Machine Learning for Industry 4.0: A Systematic Review Using Deep Learning-Based Topic Modelling," *Sensors*, vol. 22, no. 22, p. 8641, Nov. 2022, doi: 10.3390/s22228641.

[25] M. Jafarzadegan, F. Safi-Esfahani, and Z. Beheshti, “Combining hierarchical clustering approaches using the PCA method,” *Expert Syst Appl*, vol. 137, pp. 1–10, Dec. 2019, doi: 10.1016/j.eswa.2019.06.064.



UNIVERSITAS MUHAMMADIYAH MALANG

FAKULTAS TEKNIK

INFORMATIKA

informatika.umm.ac.id | informatika@umm.ac.id

FORM CEK PLAGIARISME LAPORAN TUGAS AKHIR

Nama Mahasiswa : **Dema Riandi Akhmad**

NIM : **201710379311116**

Judul TA : **Analisis Topic Modeling Journal ArXiv Menggunakan Metode K-Means dengan Algoritma Dimensionality Reduction dan t-SNE Model**

Hasil Cek Plagiarisme dengan Turnitin

No.	Komponen Pengecekan	Nilai Maksimal Plagiarisme (%)	Hasil Cek Plagiarisme (%) *
1.	Bab 1 – Pendahuluan	10 %	0 %
2.	Bab 2 – Daftar Pustaka	25 %	0 %
3.	Bab 3 – Analisis dan Perancangan	25 %	4 %
4.	Bab 4 – Implementasi dan Pengujian	15 %	0 %
5.	Bab 5 – Kesimpulan dan Saran	5 %	5 %
6.	Makalah Tugas Akhir	20%	0 %

*) *Hasil cek plagiarism diisi oleh pemeriksa (staf TU)*

*) Maksimal 5 kali (4 Kali sebelum ujian, 1 kali sesudah ujian)

Mengetahui,

Pemeriksa (Staff TU)

(.....)


UNIVERSITAS MUHAMMADIYAH
MALANG