KLASIFIKASI DAN DETEKSI KEMIRIPAN ARTIKEL BERBAHASA INDONESIA DI JURNAL BERKALA ILMIAH PADA GARBA RUJUKAN DIGITAL (GARUDA) MENGGUNAKAN METODE NAÏVE BAYES DAN COSINE SIMILARITY

OLEH :

NYIMAS SABILINA CAHYANI

09012682125018

PROGRAM STUDI MAGISTER ILMU KOMPUTER
FAKULTAS ILMU KOMPUTER
UNIVERSITAS SRIWIJAYA
2025

KLASIFIKASI DAN DETEKSI KEMIRIPAN ARTIKEL BERBAHASA INDONESIA DI JURNAL BERKALA ILMIAH PADA GARBA RUJUKAN DIGITAL (GARUDA) MENGGUNAKAN METODE NAÏVE BAYES DAN COSINE SIMILARITY

TESIS

Diajukan Untuk Melengkapi Salah Satu Syarat
Memperoleh Gelar Magister Ilmu Komputer

OLEH :

NYIMAS SABILINA CAHYANI

09012682125018

PROGRAM STUDI MAGISTER ILMU KOMPUTER
FAKULTAS ILMU KOMPUTER
UNIVERSITAS SRIWIJAYA
2025

LEMBAR PENGESAHAN

KLASIFIKASI DAN DETEKSI KEMIRIPAN ARTIKEL BERBAHASA INDONESIA DI JURNAL BERKALA ILMIAH PADA GARBA RUJUKAN DIGITAL (GARUDA) MENGGUNAKAN METODE NAÏVE BAYES DAN COSINE SIMILARITY

TESIS

Diajukan untuk Melengkapi Salah Satu Syarat
Memperoleh Gelar Magister Ilmu Komputer

OLEH :
NYIMAS SABILINA CAHYANI
09012682125018

Palembang, Mei 2025

Pembimbing I	Pembimbing II
Prof. Deris Stiawan, M.T., Ph.D	**Dr. Abdiansah, S.Kom., M.Cs.**
NIP. 197806172006041002	**NIP.** 198410012009121005

Mengetahui,
Koordinator Program Studi Magister Ilmu Komputer

Dr. Firdaus, M.Kom.
NIP. 197801212008121003

HALAMAN PERSETUJUAN

Pada hari Jumat tanggal 23 Mei 2025 telah dilaksanakan ujian sidang komprehensif tesis oleh Magister Ilmu Komputer Fakultas Ilmu Komputer Universitas Sriwijaya.

Nama : Nyimas Sabilina Cahyani
NIM : 09012682125018
Judul : Klasifikasi dan Deteksi Kemiripan Artikel Berbahasa Indonesia di Jurnal Berkala Ilmiah Pada Garba Rujukan Digital (GARUDA) Menggunakan Metode Naive Bayes dan Cosine Similarity

1. Ketua Sidang

   Dr. Rossi Passarella, M.Eng .....
   NIP. 197811062010121004

2. Penguji I

   Dian Palupi Rini, M.Kom., Ph.D .....
   NIP. 197802232006042002

3. Penguji II

   Dr. M. Fachrurrozi, M.T .....
   NIP. 198005222008121002

4. Pembimbing I

   Prof. Deris Stiawan, M.T., Ph.D .....
   NIP. 197806172006041002

5. Pembimbing II

   Dr. Abdiansah, S.Kom., M.Cs .....
   NIP. 198410012009121005

Mengetahui,
Koordinator Program Studi Magister Ilmu Komputer

Dr. Firdaus, M.Kom.
NIP. 197801212008121003

LEMBAR PERNYATAAN

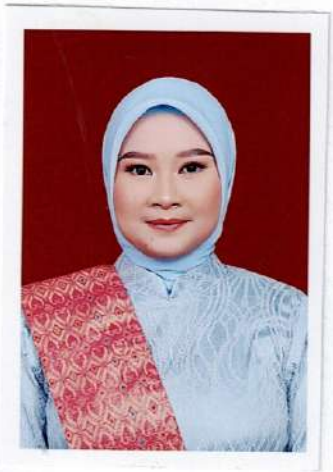
Yang bertanda tangan di bawah ini :

Nama	: Nyimas Sabilina Cahyani
NIM	: 09012682125018
Program Studi	: Magister Ilmu Komputer
Judul Tesis	: Klasifikasi dan Deteksi Kemiripan Artikel Berbahasa Indonesia di Jurnal Berkala Ilmiah Pada Garba Rujukan Digital (GARUDA) Menggunakan Metode Naïve Bayes dan Cosine Similarity

Hasil Pengecekan Software ~~iThenticate~~/Turnitin : **5 %**

Menyatakan bahwa laporan tesis saya merupakan hasil karya sendiri dan bukan hasil penjiplakan/plagiat. Apabila ditemukan unsur penjiplakan/plagiat dalam laporan tesis 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, 01 Juli 2025

Nyimas Sabilina Cahyani

NIM. 09012682125018

KATA PENGANTAR

Puji dan syukur atas kehadirat Allah SWT yang telah memberikan rahmat-Nya sehingga penulis bisa menyelesaikan tesis yang berjudul "**Klasifikasi dan Deteksi Kemiripan Artikel Berbahasa Indonesia di Jurnal Berkala Ilmiah Pada Garba Rujukan Digital (GARUDA) Menggunakan Metode Naive Bayes dan Cosine Similarity**".

Selama proses penulisan tesis ini, penulis banyak mendapat informasi data maupun pengarahan, baik secara langsung maupun tidak langsung dari berbagai pihak. Pada kesempatan kali ini, penulis ingin menyampaikan ucapan terima kasih kepada pihak-pihak yang telah membantu dalam menyelesaikan tesis ini, yaitu:

1. Kedua orang tua, adik-adik, Ari Winarso dan keluarga yang telah memberikan dukungan dan motivasi kepada penulis.
2. Bapak Prof. Dr. Taufiq Marwa, S.E., M.Si. selaku Rektor Universitas Sriwijaya.
3. Bapak Prof. Dr. Erwin, S.Si., M.Si. selaku Dekan Fakultas Ilmu Komputer.
4. Bapak Dr. Firdaus, M.Kom. selaku Koordinator Program Studi Magister Ilmu Komputer.
5. Bapak Prof. Deris Stiawan, M.T., Ph.D selaku Pembimbing I.
6. Bapak Dr. Abdiansah, S.Kom., M.Cs selaku Pembimbing II.
7. Bapak Abdillahi Romadhona, S.E. selaku admin jurusan Magister Ilmu Komputer.
8. Nurul Afifah, Dendi Renaldo Permana, Septiani Kusuma Ningrum (Team CoE) yang telah membantu dan memberikan saran dalam pengerjaan tesis ini.
9. Teman-teman Magister Ilmu Komputer angkatan 2021, terutama Sari Nuzulastri yang telah memberikan bantuan serta dukungan kepada penulis.
10. Owner & Management Kenzo Live Rajawali yang telah memberikan dukungan dan kemudahan waktu kepada penulis dalam perkuliahan dan pengerjaan tesis ini.
11. Semua pihak yang telah membantu dalam penyelesaian tesis ini.

Semoga tesis ini dapat bermanfaat bagi semua pihak yang membutuhkan, serta dapat menambah wawasan dan mengembangkan ilmu pengetahuan khususnya bagi Penulis. Akhirnya semoga Allah merahmati semua pihak yang telah membantu penulis selama ini.

Palembang, Mei 2025

Penulis

Classification and Detection of Similarity of Indonesian Articles In Scientific Periodical Journals On Garba Rujukan Digital (GARUDA) Using Naïve Bayes and Cosine Similarity Methods

Nyimas Sabilina Cahyani

ABSTRACT

The classification of articles into several categories has been done using the Naive Bayes method, with an F1-score result of 98% on balanced data based on titles and abstracts. The results show a high level of classification accuracy, with a processing time of less than 60 minutes. Similarity detection between articles was carried out using the Cosine Similarity method, and a similarity score of 0.071 was obtained, reflecting the low similarity level between articles. In this research, the score range used was 0 to 1, where a score close to 1 indicates the highest level of similarity. The search for similar scientific articles was conducted using the Cosine Similarity method based on titles and abstracts, by sorting articles by highest similarity score. The most similar articles are shown in the top order, and the search process takes 44 to 50 seconds to search time. The results of this research show that the method used can increase the accuracy in the classification process, similarity detection, and article search on the Garba Rujukan Digital (GARUDA) platform accurately and efficiently.

Keywords: Classification, Similarity, Naïve Bayes, Cosine Similarity, GARUDA

Klasifikasi dan Deteksi Kemiripan Artikel Berbahasa Indonesia di Jurnal Berkala Ilmiah Pada Garba Rujukan Digital (GARUDA) Menggunakan Metode Naïve Bayes Dan Cosine Similarity

Nyimas Sabilina Cahyani

ABSTRAK

Klasifikasi artikel ke dalam beberapa kategori telah dilakukan menggunakan metode Naive Bayes, dengan hasil *F1-Score* sebesar 98% pada balanced data berdasarkan judul dan abstrak. Hasil menunjukkan tingkat akurasi klasifikasi yang tinggi, dengan waktu pemrosesan kurang dari 60 menit. Deteksi kemiripan antar artikel telah dilakukan menggunakan metode Cosine Similarity dan mendapatkan hasil skor kemiripan sebesar 0,071, mencerminkan tingkat kemiripan yang rendah antar artikel. Dalam penelitian ini rentang skor yang digunakan 0 hingga 1, dimana skor yang mendekati 1 menunjukkan tingkat kemiripan yang tertinggi. Pencarian artikel ilmiah serupa dilakukan menggunakan metode Cosine Similarity berdasarkan judul dan abstrak, dengan mengurutkan artikel berdasarkan skor kemiripan tertinggi. Artikel yang paling mirip ditampilkan pada urutan atas, dan proses pencarian memerlukan waktu pencarian 44 hingga 50 detik. Hasil penelitian ini menunjukkan bahwa metode yang digunakan mampu meningkatkan akurasi dalam proses klasifikasi, deteksi kemiripan dan pencarian artikel pada platform Garba Rujukan Digital (GARUDA) secara akurat dan efisien.

Kata kunci: Klasifikasi, Similarity, Naïve Bayes, Cosine Similarity, GARUDA

DAFTAR ISI

Halaman

DAFTAR GAMBAR

Halaman

DAFTAR TABEL

Halaman

BAB I. PENDAHULUAN

Pada bagian ini menjelaskan latar belakang penelitian yang berjudul: "Klasifikasi dan Deteksi Kemiripan Artikel Berbahasa Indonesia di Jurnal Berkala Ilmiah Pada Garba Rujukan Digital (GARUDA) Menggunakan Metode Naïve Bayes dan Cosine Similarity". Berdasarkan latar belakang yang ditulis, diperoleh permasalahan yang akan diambil dan diberikan batasan masalah. Permasalahan yang ada akan memiliki solusi atau penyelesaian, melalui penelitian ini. Penelitian yang dilakukan memiliki tujuan dan manfaat, serta menggunakan metodologi atau cara-cara penyelesaian dalam bentuk tahapan-tahapan.

1.1 Latar Belakang

Perkembangan teknologi yang semakin cepat pada layanan aggregator mempermudah pencarian jurnal ilmiah sebagai referensi atau daftar pustaka dalam menentukan topik penulisan artikel. Layanan aggregator adalah platform yang mengumpulkan dan menyusun informasi dari berbagai sumber berbeda untuk menyediakan akses yang lebih mudah dan terorganisir bagi penggunanya. Ada beberapa layanan aggregator yang digunakan untuk melakukan pencarian jurnal ilmiah, salah satunya adalah aggregator nasional Garba Rujukan Digital (GARUDA) Kemendikbud, sebuah platform digital yang dikembangkan oleh Kementerian Pendidikan dan Kebudayaan (Kemendikbud) Republik Indonesia. Platform ini bertujuan untuk memberikan akses kepada berbagai sumber informasi, dokumen, data dan layanan yang berkaitan dengan pendidikan dan kebudayaan di Indonesia. Garuda Kemendikbud dapat digunakan oleh berbagai pihak yang memiliki hubungan dalam dunia pendidikan, termasuk guru, siswa, peneliti, dan masyarakat umum sebagai material atau bahan bacaan. Jenis informasi yang tersedia di Garuda Kemendikbud mencakup jurnal ilmiah, artikel penelitian, dan dokumen akademis lainnya. Selain itu, platform ini juga mendukung perkembangan

ilmu pengetahuan dan pendidikan di Indonesia. Garuda Kemendikbud memiliki 3.105.110 *articles*, 4.163 *publishers*, 22.569 *journals*, 329 *conferences* dan 40 *subject*.

Aggregator Garba Rujukan Digital (GARUDA) Kemendikbud memiliki kepadanan database dan jaringan yang terhubung dengan SINTA, Bima, Arjuna, PDDIKTI, Risbang, Scopus dan Rama. Menurut penelitian yang dilakukan oleh Lukman et al. (2018) penelitian dari semua dosen di Indonesia dikumpulkan. Serta, dimasukkan ke dalam portal *Science and Technology Index* atau Indeks Sains dan Teknologi (SINTA[1]). SINTA sebuah sistem indeks yang digunakan di Indonesia untuk mengukur dan memantau kinerja penelitian ilmiah yang dilakukan oleh para peneliti, perguruan tinggi, dan lembaga penelitian di berbagai bidang ilmu pengetahuan dan teknologi. Indeks ini dikelola oleh Kementrian Riset, Teknologi, dan Pendidikan Tinggi Republik Indonesia. Serta, memiliki tujuan untuk memantau publikasi ilmiah, transparansi dan akuntabilitas, mengukur kualitas publikasi, dan mendukung pengambilan keputusan. SINTA memiliki kepadanan database dan jaringan terhadap aggregator Garba Rujukan Digital (GARUDA[2]).

Menurut Lukman (2017) munculnya jurnal elektronik dan penerbitan akses terbuka bisa meningkatkan akses terhadap data digital untuk setiap artikel, sehingga data tersebut dapat diukur dan diintegrasi dengan berbagai basis data dan indexer seperti Scopus, *Web of Science*, Google Schoolar, dan lainnya. Maka, penelitian ini akan melakukan pengukuran kemiripan sebuah artikel utama dengan artikel lainnya berdasarkan judul dan abstrak. Serta, klasifikasi dengan cara mengategorikan atau melabeli *dataset* dari aggregator nasional Garba Rujukan Digital (GARUDA) Kemendikbud untuk mendapatkan hasil pencarian artikel ilmiah sesuai dengan kategori masing-masing.

Beberapa penelitian terdahulu mengenai klasifikasi jurnal telah dilakukan pada penelitian Nisha (2021) yang menggunakan metode Neighbor Weighted K-Nearest Neighbor dan Rakasiwi et al. (2021) yang menggunakan metode Improved K-Nearest telah mendapatkan tingkat akurasi yang tinggi sebesar 90% dan

[1] https://sinta.kemdikbud.go.id/
[2] https://garuda.kemdikbud.go.id/

membutuhkan waktu yang cukup lama. Penelitian yang dilakukan oleh Lumbanraja et al. (2021) memiliki hasil akurasi tertinggi dengan penggunaan 205 fitur dan algoritma SVM Linear kernel sebesar 58,3%. Penelitian Latif et al. (2018) bertujuan untuk mengklasifikasikan konten abstrak berdasarkan jumlah kata terbanyak dalam abstrak jurnal bahas Inggris dengan menggunakan algoritma Naïve Bayes. Menurut Parlak et al. (2019) klasifikasi berdasarkan abstrak pada jurnal *medical* dengan membandingkan berbagai metode, metode Naïve Bayes Multinomial yang memberikan performa klasifikasi lebih akurat. Maka, algoritma ini sangat cepat dan efisien, terutama memproses data teks dalam jumlah yang sangat besar. Menurut Devita et al. (2018) salah satu metode klasifikasi yaitu metode Naïve Bayes memiliki hasil klasifikasi dengan tingkat akurasi yang maksimal sebesar 70% dan waktu komputasi yang lebih cepat dengan data *training* yang sedikit. Metode klasifikasi Naïve Bayes berguna untuk mengkategorikan dan mengklasifikasikan teks.

Metode ini biasa digunakan dalam pemrosesan bahasa alami (*Natural Language Processing)* untuk memahami dan menganalisis data bahasa manusia. Menurut penelitian yang dilakukan oleh Chang (2021) *Natural Language Processing* (NLP) menyediakan teknik-teknik dan algoritma untuk memproses serta, memahami bahasa manusia secara efektif. NLP memiliki keunggulan dalam mengelola klasifikasi teks atau dokumen dengan jumlah yang besar. Menurut Yasaswi et al. (2022) membahas mengenai klasifikasi berita yang ada pada web dan internet dalam penyebaran berita Covid-19 asli atau palsu. Serta, menemukan cara yang efisien untuk mengklasifikasikan dengan mengembangkan model yang dapat diandalkan dan akurat menggunakan model ML dan NLP.

Menurut Qomariyah et al., 2022 klasifikasi teks menggunakan NLP bisa melakukan deteksi otomatis Covid-19 dari laporan radiologi dalam bahasa Indonesia. Menurut Kumar et al. (2021) klasifikasi teks adalah kategorisasi secara terorganisasi untuk interpretasi informasi sensitif dari teks, sedangkan pemodelan topik adalah menemukan topik abstrak untuk kumpulan teks atau dokumen. NLP digunakan tujuannya untuk memahami, memproses dan menghasilkan bahasa manusia. Pada penelitian ini, menggunakan pemrosesan teks untuk klasifikasi teks

berdasarkan judul dan abstrak artikel kedalam kategori tertentu. Teks diubah menjadi bentuk numerik menggunakan teknik Term Fequency-Inverse Document Frequency (TF-IDF) (Hizqil & Ruldeviani, 2024).

Penelitian Mardatillah et al. (2021) membahas analisis kutipan pada artikel menggunakan Cosine Similarity. Hasil pengujian menunjukkan bahwa 70% kutipan pada artikel ilmiah memiliki tingkat kemiripan yang tinggi dengan sumber rujukan yang di acu, sedangkan 30% kutipan memiliki tingkat kemiripan rendah dan tidak ada hubungan dengan sumber rujukan. Menurut Islam et al. (2023) melakukan deteksi *plagiarism* dalam konten teks Bengali dengan Cosine Similarity berhasil menentukan kesamaan dengan membandingkan vektor yang berupa nilai numerik menggunakan dataset hampir 112.184 kalimat. Penelitian terdahulu mengenai pengukuran kemiripan berdasarkan judul dan abstrak serta klasifikasi pada jurnal ekonomi oleh Putri et al. (2019) menunjukkan bahwa algoritma Cosine Similarity berguna untuk mengukur kemiripan di antara dua vektor dengan menghitung besar sudut kosinus di antara keduanya. Cosine similairty digunakan untuk sistem rekomendasi buku untuk memberikan hasil yang relevan dengan topik kursus dengan presisi 0.7 dan recall 0.73 (Nuipian & Chuaykhun, 2023). Cosine similarity juga dapat mengklasifikasikan artikel berdasarkan judul dan abstrak dengan teknik K-Fold Cross Validation. Menurut Rinjeni et al. (2024) dalam penelitian pencocokan judul artikel ilmiah menggunakan algoritma Cosine Similarity dan Jaccard Similarity, bahwa CS memberikan kinerja yang baik dibandingkan Jaccard Similarity, terutama dalam skenario tertentu.

Penelitian ini dilakukan pengklasifikasian untuk mengukur hasil kinerja dari metode Naïve Bayes menggunakan dataset dari Garuda Kemendikbud. Membangun model klasifikasi menggunakan metode Naïve Bayes dan akan dilakukan klasifikasi *multi-class* untuk memprediksi kelas atau label dari suatu data yang dapat dikelompokkan kedalam lebih dari dua kelas atau kedalam kategori yang berbeda. Serta, deteksi sejauh mana tingkat kemiripan dari artikel utama dengan artikel ilmiah yang lain berdasarkan fitur tertentu yaitu, judul dan abstrak menggunakan metode Cosine Similarity.

1.2 Perumusan Masalah

Dalam penelitian ini membutuhkan kategori, deteksi kemiripan dan fitur pencarian artikel untuk menemukan artikel yang mirip. Sehingga, memiliki rumusan masalah sebagai berikut:

1. Bagaimana pengklasifikasian artikel ilmiah menggunakan metode Naïve Bayes?
2. Bagaimana hasil kinerja dari pengklasifikasian menggunakan metode Naïve Bayes?
3. Bagaimana proses deteksi kemiripan artikel ilmiah utama dengan artikel ilmiah yang lainnya berdasarkan judul dan abstrak menggunakan metode Cosine Similarity?
4. Bagaimana hasil kinerja dari deteksi kemiripan artikel ilmiah menggunakan metode Cosine Similarity?
5. Bagaimana hasil pencarian artikel ilmiah utama dengan artikel ilmiah lainnya menggunakan metode Cosine Similarity?

1.3 Tujuan

Adapun tujuan dari hasil penulisan penelitian ini adalah sebagai berikut:

1. Melakukan pengklasifikasian artikel ilmiah menggunakan metode Naïve Bayes.
2. Mengetahui hasil kinerja pengklasifikasian menggunakan metode Naïve Bayes.
3. Melakukan deteksi kemiripan artikel utama dengan artikel ilmiah lainnya berdasarkan judul dan abstrak menggunakan metode Cosine Similarity.
4. Mengetahui hasil kinerja deteksi kemiripan menggunakan metode Cosine Similarity
5. Mengetahui hasil pencarian artikel menggunakan metode Cosine Similarity.

1.4 Batasan Masalah

Beberapa batasan masalah dalam melakukan penulisan penelitian pengklasifikasian ini sebagai berikut :

1. Penelitian ini menggunakan dataset Aggregator Garba Rujukan Digital (GARUDA) tahun 2022.
2. Pengklasifikasian hanya dibidang komputer Sistem Informasi Manajerial dan berbahasa Indonesia.
3. Metode Cosine Similarity yang digunakan untuk melakukan deteksi kemiripan dan pencarian artikel utama dengan artikel lainnya.
4. Atribut dataset yang digunakan judul dan abstrak.

1.5 Manfaat

Adapun manfaat yang didapatkan dari penelitian ini adalah sebagai berikut:

1. Mendapatkan hasil pengklasifikasian artikel ilmiah menggunakan metode Naïve Bayes.
2. Mendapatkan hasil kinerja dari penelitian ini menggunakan metode Naïve Bayes.
3. Mendapatkan hasil deteksi kemiripan artikel ilmiah utama dengan artikel ilmiah lainnya menggunakan metode Cosine Similarity.
4. Mendapatkan hasil pencarian artikel ilmiah utama dengan artikel ilmiah lainnya menggunakan metode Cosine Similarity.
5. Mendapatkan hasil kinerja dari penelitian ini menggunakan Cosine Similarity.

1.6 Sistematika Penulisan

Sistematika penulisan dirancang untuk mempermudah penyusunan penelitian ini dan menjelaskan isi setiap bab yang ada dalam penelitian. Maka, disusunlah sistematika penulisan sebagai berikut:

1. **BAB I. PENDAHULUAN**

   Bab ini mencakup latar belakang, perumusan masalah, batasan masalah, serta, tujuan dan manfaat penelitian yang memiliki hubungan dengan penelitian ini, serta sistematika penulisan.

2. **BAB II. TINJAUAN PUSTAKA**

   Bab ini mencakup seluruh uraian dasar teori yang berkaitan dengan permasalahan yang diangkat serta, penelitian sebelumnya yang berkaitan dengan penelitian.

3. **BAB III. METODOLOGI PENELITIAN**

   Bab ini mencakup secara rinci dan bertahap mengenai metodologi atau langkah-langkah yang diterapkan dalam membangun kerangka kerja penelitian. Serta, alur penelitian yang digunakan untuk menyelesaikan penelitian. Selain itu, dijelaskan mengenai dataset yang digunakan dalam penelitian ini.

4. **BAB IV. HASIL DAN PEMBAHASAN**

   Bab ini mencakup tentang hasil dan penjelasan dari proses klasifikasi, deteksi kemiripan dan pencarian artikel ilmiah utama dengan artikel ilmiah lainnya yang telah dilakukan dan analisa mengenai F1-Score dan Score CS yang didapatkan.

5. **BAB V. KESIMPULAN**

   Bab ini mencakup kesimpulan dari hasil yang telah dicapai, yang merupakan jawaban atas tujuan penelitian yang dirumuskan pada Bab 1 Pendahuluan. Selain itu, disampaikan saran yang dapat menjadi acuan untuk pengembangan penelitian di masa mendatang.

2. Melakukan penambahan *dataset* yang lebih banyak dalam bidang keilmuan.
3. Penelitian selanjutnya bisa membangun penerapan pada sistem *software* dengan membuat website pencarian artikel ilmiah.

DAFTAR PUSTAKA

Chang, I. C. (2021). Applying text mining, clustering analysis, and latent dirichlet allocation techniques for topic classification of environmental education journals. *Sustainability (Switzerland)*, *13*(19). https://doi.org/10.3390/su131910856

Chen, H., Hu, S., Hua, R., & Zhao, X. (2021). *Improved naive Bayes classification algorithm for traffic risk management*. *6*.

Devita, R. N., Herwanto, H. W., & Wibawa, A. P. (2018). Perbandingan Kinerja Metode Naive Bayes dan K-Nearest Neighbor untuk Klasifikasi Artikel Berbahasa indonesia. *Jurnal Teknologi Informasi Dan Ilmu Komputer*, *5*(4), 427. https://doi.org/10.25126/jtiik.201854773

Hizqil, A., & Ruldeviani, Y. (2024). Sentiment analysis of online licensing service quality in the energy and mineral resources sector of the Republic of Indonesia. *Computer Science and Information Technologies*, *5*(1), 63–71. https://doi.org/10.11591/csit.v5i1.pp63-71

Islam, A., Rahman, E., Chowdhury, A. A., & Mojumder, M. A. N. (2023). A Deep Learning Approach to Detect Plagiarism in Bengali Textual Content using Similarity Algorithms. *Proceedings of IEEE InC4 2023 - 2023 IEEE International Conference on Contemporary Computing and Communications*, *1*, 1–5. https://doi.org/10.1109/InC457730.2023.10262998

Kang, Y., Cai, Z., Tan, C. W., Huang, Q., & Liu, H. (2020). Natural language processing (NLP) in management research: A literature review. *Journal of Management Analytics*, *7*(2), 139–172. https://doi.org/10.1080/23270012.2020.1756939

Kim, S. W., & Gil, J. M. (2019). Research paper classification systems based on

TF-IDF and LDA schemes. *Human-Centric Computing and Information Sciences*, *9*(1). https://doi.org/10.1186/s13673-019-0192-7

Kumar, N., Suman, R. R., & Kumar, S. (2021). Text Classification and Topic Modelling of Web Extracted Data. *2021 2nd Global Conference for Advancement in Technology, GCAT 2021*, 1–8. https://doi.org/10.1109/GCAT52182.2021.9587459

Latif, S., Suwardoyo, U., & Wihelmus Sanadi, E. A. (2018). Content Abstract Classification Using Naive Bayes. *Journal of Physics: Conference Series*, *979*(1). https://doi.org/10.1088/1742-6596/979/1/012036

Liu, Z., Zhu, J., Cheng, X., & Lu, Q. (2023). Optimized Algorithm Design for Text similarity Detection Based on Artificial Intelligence and Natural Language Processing. *Procedia Computer Science*, *228*, 195–202. https://doi.org/10.1016/j.procs.2023.11.023

Lukman, L., Dimyati, M., Rianto, Y., Subroto, I., Sutikno, T., Hidayat, D., Nadhiroh, I., Stiawan, D., Haviana, S., Heryanto, A., & Yuliansyah, H. (2018). *Proposal of the S-score for measuring the performance of researchers, institutions, and journals in Indonesia*.

Lumbanraja, F. R., E, F., Ardiansyah, A, J., & Rizky, P. (2021). *Abstract Classification Using Support Vector Machine Algorithm (Case Study : Abstract in a Computer Science Journal) Abstract Classification Using Support Vector Machine Algorithm (Case Study : Abstract in a Computer Science*. https://doi.org/10.1088/1742-6596/1751/1/012042

Malik, N., Bilal, A., Ilyas, M., Razzaq, S., Maqbool, F., & Abbas, Q. (2021). Plagiarism Detection Using Natural Language Processing Techniques. *Technical Journal, University of Engineering and Technology (UET)*, *26*(1), 2313–7770.

Mardatillah, U., Zulfikar, W. B., Atmadja, A. R., Taufik, I., & Uriawan, W. (2021). Citation Analysis on Scientific Articles Using Cosine Similarity. *Proceeding*

of 2021 7th International Conference on Wireless and Telematics, ICWT 2021, 0–3. https://doi.org/10.1109/ICWT52862.2021.9678402

Nisha, A. C. (2021). Klasifikasi Abstrak Jurnal Repositor di Teknik Informatika UMM Menggunakan Metode Neighbor Weighted K-Nearest Neighbor. *Jurnal Repositor*, *3*(3), 295–304. https://doi.org/10.22219/repositor.v2i3.1225

Nuipian, V., & Chuaykhun, J. (2023). Book Recommendation System based on Course Descriptions using Cosine Similarity. *ACM International Conference Proceeding Series*, 273–277. https://doi.org/10.1145/3639233.3639335

Osman, A. S. (2019). Data mining techniques: Review. *International Journal of Data Science Research*, *2*(1), 1–4.

Parlak, B. (2019). *On classification of abstracts obtained from medical journals*. https://doi.org/10.1177/0165551519860982

Qomariyah, N. N., Araminta, A. S., Reynaldi, R., Senjaya, M., Asri, S. D. A., & Kazakov, D. (2022). NLP Text Classification for COVID-19 Automatic Detection from Radiology Report in Indonesian Language. *2022 5th International Seminar on Research of Information Technology and Intelligent Systems, ISRITI 2022*, 565–569. https://doi.org/10.1109/ISRITI56927.2022.10053077

Rakasiwi, T., Rahayudi, B., & Ridok, A. (2021). *Klasifikasi Artikel Publikasi berdasarkan Judul pada Jurnal Teknologi Informasi dan Ilmu Komputer (JTIIK) Universitas Brawijaya dengan menggunakan Metode Improved K-Nearest Neighbor*. *5*(10), 4510–4516.

Rinjeni, T. P., Indriawan, A., & Rakhmawati, N. A. (2024). Matching Scientific Article Titles using Cosine Similarity and Jaccard Similarity Algorithm. *Procedia Computer Science*, *234*(2023), 553–560. https://doi.org/10.1016/j.procs.2024.03.039

Ristanti, P. Y., Wibawa, A. P., & Pujianto, U. (2019). Cosine Similarity for Title

and Abstract of Economic Journal Classification. *Proceeding - 2019 5th International Conference on Science in Information Technology: Embracing Industry 4.0: Towards Innovation in Cyber Physical System, ICSITech 2019*, *July*, 123–127. https://doi.org/10.1109/ICSITech46713.2019.8987547

Sadikin, M., Rosnelly, R., & Gunawan, T. S. (2020). *Perbandingan Tingkat Akurasi Klasifikasi Penerimaan Dosen Tetap Menggunakan Metode Naive Bayes Classifier dan C4 . 5*. *4*, 1100–1109. https://doi.org/10.30865/mib.v4i4.2434

Saritas, M. M., & Yasar, A. (2019). *Intelligent Systems and Applications in Engineering Performance Analysis of ANN and Naive Bayes Classification Algorithm for Data Classification*. 0–1. https://doi.org/10.1039/b000000x

Setiawan, A., Santoso, L. W., & Adipranata, R. (2020). Klasifikasi Artikel Berita Bahasa Indonesia Dengan Naive Bayes Classifier. *Jurnal Infra*, *8*(1), 146–151.

Singh, R., & Singh, S. (2021). Text Similarity Measures in News Articles by Vector Space Model Using NLP. *Journal of The Institution of Engineers (India): Series B*, *102*(2), 329–338. https://doi.org/10.1007/s40031-020-00501-5

Yasaswi, K., Kambala, V. K., Pavan, P. S., Sreya, M., & Jasmika, V. (2022). News Classification using Natural Language Processing. *Proceedings of 3rd International Conference on Intelligent Engineering and Management, ICIEM 2022*, 63–67. https://doi.org/10.1109/ICIEM54221.2022.9853174

Yuan, H., Tang, Y., Sun, W., & Liu, L. (2020). A detection method for android application security based on TF-IDF and machine learning. *PLoS ONE*, *15*(9 September), 1–19. https://doi.org/10.1371/journal.pone.0238694