Jicom : Journal Information & Computer
Volume 3, Nomor 2, Juli 2025, hlm.118-125
DOI : 10.32493/jicomisc.v2i1.37666

P-ISSN : 3031-2779
E-ISSN : 3026-4197

Pemodelan Topik Jurnal Informatika Menggunakan Bag of Words dan Latent Dirichlet Allocation

Verrino Adityya[*1], Ivander Destian Luis[2], Abdul Rahman[3], Hafiz Irsyad[4]
[1,2,4)]Informatika, Fakultas Ilmu Komputer dan Rekayasa, Universitas Multi Data Palembang
[3)]Teknik Elektro, Fakultas Ilmu Komputer dan Rekayasa, Universitas Multi Data Palembang
Email: [*1]verrinoadityya_2226250053@mhs.mdp.ac.id, [2]ivanderdestianluis_2226250029@mhs.mdp.ac.id, [3]arahman@mdp.ac.id, [4]hafizirsyad@mdp.ac.id,

(Naskah masuk: *4 Juni 2025*, diterima untuk diterbitkan: *31 Juli 2025*)

Abstrak: *Penyebaran jurnal penelitian secara online, khususnya jurnal informatika, seringkali menyajikan topik yang mirip dan berubah sangat cepat, sehingga menyulitkan pembaca memahami konteks jurnal secara utuh. Pemodelan topik menjadi penting untuk mengelompokkan jurnal berdasarkan kemiripan konteks secara semantik, sehingga jurnal menjadi terstruktur dan mudah dipahami sebab-akibatnya. Penelitian ini bertujuan untuk memodelkan topik jurnal dari yang dikumpulkan dari sumber, seperti UMDP dan UIGM. menggunakan Bag of Words (BoW) untuk ekstraksi fitur dan Latent Dirichlet Allocation (LDA) untuk pemodelan topiknya. Data konten jurnal informatika dikumpulkan dari beberapa sumber jurnal informatika dan melalui tahap preprocessing meliputi, penghapusan kalimat dan kata unik, tokenisasi, penghapusan stop words, dan stemming. Setiap token akan dibentuk menjadi unigram dan bigram dan diberi pembobotan dengan BoW. Evaluasi dilakukan dengan mengukur nilai koherensi untuk rentang jumlah topik 2 hingga 10. Hasil penelitian menunjukkan bahwa model LDA mampu mengidentifikasi 4 topik optimal dengan nilai koherensi sebesar 52.1%. Penelitian ini menunjukkan bahwa kombinasi BoW dan LDA efektif untuk menemukan maksud tersembunyi dari setiap topik jurnal informatika secara semantik.*

Kata Kunci – *bag of words; latent dirichlet allocation; pemodelan topik*

Abstract: *The rapid proliferation of journals, particularly informatic journals, often presents similar and quickly changing topics, making it challenging for readers to fully grasp the overall context. Topic modeling is crucial for grouping news based on semantic contextual similarity, thereby structuring the news and facilitating an understanding of its causal relationships. This research aims to model news topics from the news portal using Bag of Words (BoW) for feature extraction and Latent Dirichlet Allocation (LDA) for topic modeling. Informatics journal content data is collected from several informatics journal sources and subjected to preprocessing, which included the removal of unique sentences and words, tokenization, stop word removal, and stemming. These tokens were then formed into unigrams and bigrams, weighted using the BoW model. Evaluation was performed by measuring coherence scores for a range of 2 to 9 topics. The results showed that the LDA model successfully identified 4 optimal topics, achieving a coherence score of 52.1%. This research highlights the effectiveness of combining BoW and LDA for semantically uncovering latent themes within educational news topics.*

Keywords – *bag of words; latent dirichlet allocation; topic modelling*

1. PENDAHULUAN

Di era teknologi yang maju saat ini, bidang pendidikan mengalami perkembangan yang pesat, ditandai dengan peningkatan volume publikasi jurnal ilmiah yang signifikan. Berawal dari hadirnya sebuah temuan atau isu-isu permasalahan secara global yang berdampak pada lingkungan, hal ini memberikan sebuah peluang bagi para peneliti dari berbagai penjuru dunia untuk menghadirkan sebuah temuan baru [1]. Setiap perguruan tinggi memiliki ketentuan tersendiri dalam penyelesaian tugas akhir, yang dapat berupa skripsi, tesis, jurnal ilmiah, artikel, prototipe, dan berbagai bentuk lainnya [2]. Seringkali dalam mencari jurnal yang relevan, para peneliti harus melakukan pencarian di dalam mesin pencarian dalam kurun waktu yang relatif lama dikarenakan jurnal - jurnal yang direkomendasikan kurang sesuai dengan topik yang dicari [3]. Pesatnya

pertumbuhan pada publikasi sebuah jurnal menuntut metode yang efektif dan cepat dalam memahami konteks pada dokumen besar. Salah satu pendekatan yang dapat digunakan yaitu pemodelan topik.

Pemodelan topik adalah salah satu metode *unsupervised machine learning* untuk menemukan konteks tersembunyi pada kumpulan dokumen besar dan digunakan untuk mengelompokkan dokumen-dokumen menjadi satu konteks [4]. Dalam melakukan pemodelan topik, metode pengelompokkan berdasarkan kedekatan data dapat menggunakan model *Latent Dirichlet Allocation* (LDA) [5]. LDA adalah model probabilistik yang menghasilkan output berupa topik yang terdiri dari kata-kata dan probabilitasnya [6]. Model LDA mengasumsikan bahwa dokumen terdiri dari distribusi topik dan topik tersusun dari distribusi kata-kata yang secara semantik [7]. Penelitian yang berhubungan dengan LDA, seperti untuk pemodelan topik berita daring [8], pemodelan topik untuk *headline* berita *online* [9] pemodelan topik untuk klasifikasi komentar kuliah [10], dan peringkasan teks berita [11].

Dalam proses pemodelan topik, data jurnal dengan topik informatika pada tahun 2024 akan dikumpulkan dari melalui beberapa sumber jurnal, seperti UMDP dan UIGM dengan menggunakan teknik *Text Mining*. *Text Mining*, disebut juga penambangan data teks atau penemuan pengetahuan dari *database*, merujuk pada proses menemukan pola menarik dan berharga dari dokumen teks [12]. Konteks sebuah jurnal biasanya membutuhkan satu kata atau lebih supaya konteks jurnal dapat diterima dengan lebih jelas. Metode ekstraksi fitur dengan *N-Gram* dapat memecah kata menjadi satu kata (*unigram*), dua kata (*bigram*), tiga kata (trigram), atau lebih. Teknik berbasis *N-gram* sangat cocok untuk klasifikasi teks, terutama untuk kategorisasi bahasa [13]. *N-Gram* akan menjadi sangat efektif untuk membentuk topik-topik pada jurnal. Kata-kata yang dipecah dengan *N-Gram* akan diberi pembobotan dengan *Bag of Words* (BoW). BoW adalah sebuah metode untuk melakukan ekstraksi fitur untuk mengklasifikasi dokumen dengan mengevaluasi frekuensi dari kata unik yang muncul pada dokumen [14].

Pada penelitian ini, kami menggunakan sebuah model LDA yang dapat memodelkan topik jurnal berdasarkan judul dari jurnal. Kami melakukan *text mining* pada portal jurnal digital dengan menyaring jurnal dengan topik informatika. Setiap dokumen akan berisi jurnal, dan kata-kata akan dibentuk dengan teknik *N-Gram*. Bentuk kata *N-Gram* akan dihitung frekuensi menghasilkan BoW yang akan menjadi data latih untuk model LDA.

2. METODE PENELITIAN

Metode Penelitian meliputi pengumpulan data, *preprocessing*, ekstraksi fitur, model *training*, dan evaluasi model. Tahapan penelitian yang dilakukan dapat dilihat pada Gambar 1.

Gambar 1. Tahap penelitian

2.1 Pengumpulan Data

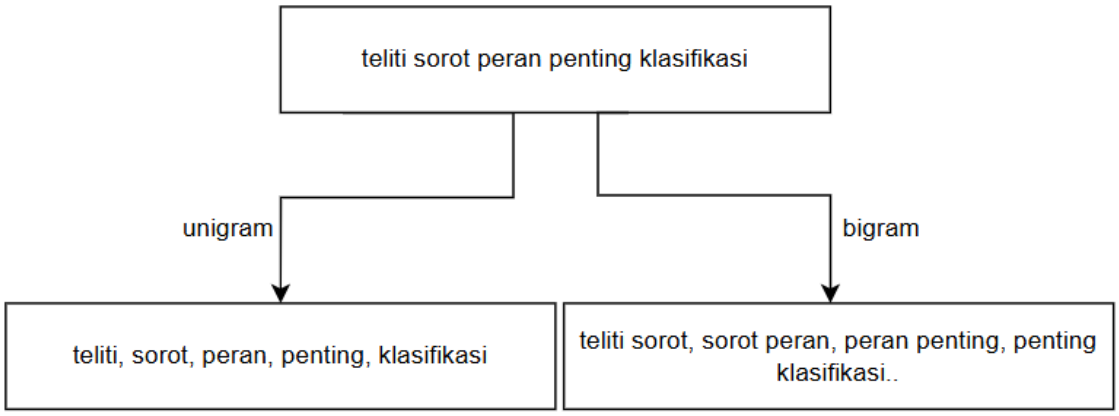
Data pada jurnal akan dikumpulkan melalui beberapa sumber jurnal, seperti UMDP (Algoritme dan JATISI) dan UIGM (Jurnal Ilmiah Informatika Global) dengan mengambil bagian judul jurnal dan abstrak jurnal. Isi jurnal tidak diambil karena isi jurnal biasanya berisi latar belakang permasalah atau penjelasan dari metode yang dilakukan, yang di mana akan menyebabkan redudansi yang dapat mengurangi makna semantik dari jurnal.

2.2 Preprocessing Data

Setelah proses pengumpulan data dilakukan data yang masih kotor akan dilakukan proses *preprocessing*. *Preprocessing* dilakukan dikarenakan data pada jurnal masih buruk untuk digunakan sebagai pelatihan. Tahapan ini meliputi proses menghapus kata-kata tertentu, menghapus *stop words*, melakukan *stemming* untuk setiap kata pada teks, dan menghapus kata yang memiliki panjang kurang dari 2.

2.3 Ekstraksi Fitur

Data yang telah dilakukan *preprocessing* akan melewati proses ekstraksi fitur, yaitu *N-Gram*. *N-Gram* adalah metode ekstraksi fitur dengan menggunakan satu atau beberapa kata. Bentuk kata *N-Gram* akan disusun secara berurutan dari urutan kata pada dokumen. Penggunaan *N-Gram* dapat mempertahankan semantik dari frasa yang memerlukan beberapa kata, seperti contoh bigram "sistem informasi", "neural network", "naive bayes" dan sebagainya. Contoh proses *N-Gram* dapat dilihat pada Gambar 2.



Gambar 2. Contoh Proses N-Gram

Data yang telah dilakukan proses *N-Gram* kata yang telah dipisah akan dihitung jumlah kata berdasarkan corpus yang dibuat, proses ini disebut sebagai *Bag of Words*. BoW adalah sebuah metode untuk melakukan ekstraksi fitur untuk mengklasifikasi dokumen dengan mengevaluasi frekuensi dari kata unik yang muncul pada dokumen [14].

2.4 Model Training

Untuk menghasilkan topik yang mewakili jurnal informatika, digunakan metode pemodelan topik dengan pendekatan LDA. Dengan LDA, setiap dokumen terdiri dari beberapa topik dan setiap topik direpresentasikan sebagai distribusi probabilitas. Dalam LDA, jumlah topik yang dihasilkan dapat mempengaruhi evaluasi, yaitu nilai koherens. Nilai koherensi berfungsi untuk mengukur seberapa "masuk akal" atau konsisten secara semantik kumpulan kata yang membentuk sebuah topik. Mengutip dari [15] persamaan dari model LDA dapat dituliskan pada persamaan (1).

$$p(D|\alpha,\beta) = \prod_{d=1}^{M} \int p(\theta_d|\, a\,) \left(\prod_{n=1}^{N_d} \sum_{z_{dn}} p(z_{dn}\,|\theta_d) p(w_{dn}|z_{dn},\beta)\right) d\theta_d \qquad (1)$$

2.5 Evaluasi Model

Evaluasi model LDA dilakukan dengan menggunakan nilai koherensi yang dihasilkan. Model akan dilatih dengan jumlah topik yang berbeda, yaitu pada rentang 2-10. Pemilihan jumlah topik akan menghasilkan nilai koherensi yang berbeda. Semakin besar nilai koherensi, maka kata kunci yang mengelompokkan sebuah jurnal agar menjadi semakin "masuk akal" dan konsisten secara semantik.

3. HASIL DAN PEMBAHASAN

3.1 Pengumpulan Data

Data ini dikumpulkan dari beberapa sumber jurnal, seperti UMDP (Algoritme dan JATISI) dan UIGM (Jurnal Ilmiah Informatika Global) dengan mengambil bagian judul jurnal dan abstrak jurnal. Data ini dikumpulkan pada tanggal 3 Juni 2025. Dataset berisi jurnal informatika pada tahun 2024. Hasil dari pengumpulan data dapat dilihat pada Tabel 1.

Tabel 1: hasil dataset yang dikumpulkan

Judul Jurnal	Abstrak Jurnal
Analisis Metode Klasifikasi Pada Data Sewa Sepeda Di Seoul	Penelitian ini menyoroti peran penting klasifikasi berbasis empat musim dalam konteks manajemen mobilitas perkotaan, dengan fokus untuk mengatasi tantangan kemacetan lalu lintas dan mendukung pilihan transportasi yang berkelanjutan…
Identifikasi Tingkat Kesegaran Daging Ayam Kampung Menggunakan Metode KNN Berdasarkan Warna Daging	Ayam kampung merupakan jenis unggas yang masih bersifat alami yang berarti belum mendapatkan perlakuan perbaikan genetik. Penentuan tingkat kesegaran daging ayam kampung merupakan salah satu faktor penting untuk menentukan kualitas daging yang akan dikonsumsi. Untuk itu dilakukan penelitian untuk menentukan tingkat kesegaran daging ayam kampung dengan menggunakan metode K-Nearest Neighbor…
Penerapan Teknik SMOTE Pada Analisis Sentimen Bea Cukai Menggunakan Algoritma Naïve Bayes	Media sosial seperti Youtube sering digunakan masyarakat dan dapat membuat isu menjadi viral dengan cepat. Baru-baru ini, bea cukai menjadi sorotan karena kasus bea masuk yang dianggap terlalu tinggi…

Setelah data telah berhasil dikumpulkan, data akan diteruskan ke dalam tahap *preprocessing* untuk membersihkan dan normalisasi data. Fitur yang diambil adalah judul dari jurnal, karena judul dari sebuah jurnal memiliki konteks yang lebih jelas dibandingkan dengan bagian lainnya dari jurnal.

3.2 Preprocessing Data

Setelah data dikumpulkan, dataset akan memasuki tahap *preprocessing* untuk dapat diterima sebagai input oleh model saat melakukan pelatihan dan pengujian. Dataset jurnal informatika yang dikumpulkan masih buruk untuk melatih model, maka dari itu diperlukan *preprocessing* data untuk membersihkan dan normalisasi data. Untuk fitur yang diambil adalah judul jurnal.

Tahap pertama yang dilakukan, yaitu menghapus kata unik, seperti "\n". Selanjutnya menghapus semua karakter selain huruf, menghapus *stop words*, melakukan *stemming* untuk setiap kata pada teks, dan menghapus kata yang memiliki panjang kurang dari 3. Beberapa kata seperti algoritma, metode, sistem, dan sebagainya dimasukkan ke dalam *stop words* karena kata-kata tersebut biasa ditemukan pada jurnal informatika. Hasil pembersihan data dapat dilihat pada Tabel 2.

Tabel 2: hasil dataset yang telah melewati *preprocessing*

Judul Jurnal Awal	Judul Jurnal Bersih
Analisis Metode Klasifikasi Pada Data Sewa Sepeda Di Seoul	analisis klasifikasi data sewa sepeda seoul
Identifikasi Tingkat Kesegaran Daging Ayam Kampung Menggunakan Metode KNN Berdasarkan Warna Daging	identifikasi tingkat segar daging ayam kampung knn dasar warna daging
Penerapan Teknik SMOTE Pada Analisis Sentimen Bea Cukai Menggunakan Algoritma Naïve Bayes	terap teknik smote analisis sentimen bea cukai bayes

3.3 Ekstraksi Fitur

Setelah data dilakukan *preprocessing*, data teks akan ditokenisasi sehingga membentuk *N-Gram*, yaitu *unigram*. Dari setiap kata *unigram* akan digabungkan menjadi 2 kata sehingga membentuk *bigram*. Data *unigram* dan *bigram* akan digabungkan dan dihapus kata yang duplikat membentuk sebuah kamus kata (*corpus*). Hasil pembentukan *N-Gram* dapat dilihat pada Tabel 3.

Tabel 3: hasil pembentukan *N-Gram* untuk *unigram* dan *bigram*

Judul Jurnal Bersih	***Unigram***	***Bigram***
analisis klasifikasi data sewa sepeda seoul	analisis, klasifikasi, data, sewa, sepeda, seoul	analisis klasifikasi, klasifikasi data, data sewa, sewa sepeda, sepeda seoul
identifikasi tingkat segar daging ayam kampung knn dasar warna daging	identifikasi, tingkat, segar, daging, ayam, kampung, knn, dasar, warna, daging	identifikasi tingkat, tingkat segar, segar daging, daging ayam, ayam kampung, kampung knn, knn dasar, dasar warna, warna daging
terap teknik smote analisis sentimen bea cukai bayes	terap, teknik, smote, analisis, sentimen, bea, cukai, bayes	terap teknik, teknik smote, smote analisis, analisis sentimen, sentimen bea, bea cukai, cukai bayes

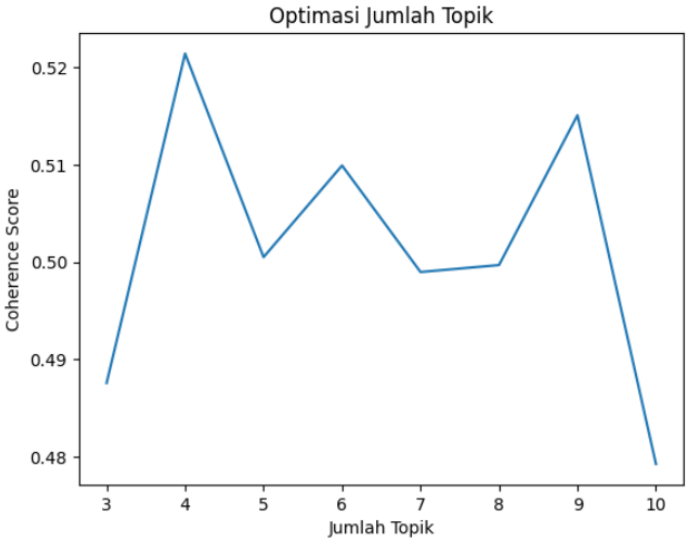
Setiap data pada jurnal akan dihitung jumlah kata berdasarkan *corpus* yang dibuat. Proses ini disebut sebagai *Bag of Words* (BoW). Setiap kata pada corpus akan dihitung frekuensi kemunculan pada setiap data jurnal informatika. Hasil *Bag of Words* akan dijadikan fitur untuk melatih model.

3.4 Model Training

Sebelum melatih data, data corpus kata akan dikonversi menjadi *format gensim* untuk menjadi input pada model *Latent Dirichlet Allocation* (LDA). Dari *format gensim* akan dibentuk *corpus* baru yang berisi *id* dari setiap *token* dan frekuensinya. Pada proses pelatihan, jumlah topik yang dihasilkan akan diuji dari 2-10 dan dilihat nilai koherensi tertinggi yang dihasilkan.

3.5 Evaluasi Model

Evaluasi model LDA dilakukan dengan menggunakan nilai koherensi yang dihasilkan. Model akan dilatih dengan jumlah topik yang berbeda, yaitu rentang 2-10. Grafik nilai koherensi untuk setiap jumlah topik yang dihasilkan dapat dilihat pada Gambar 3.

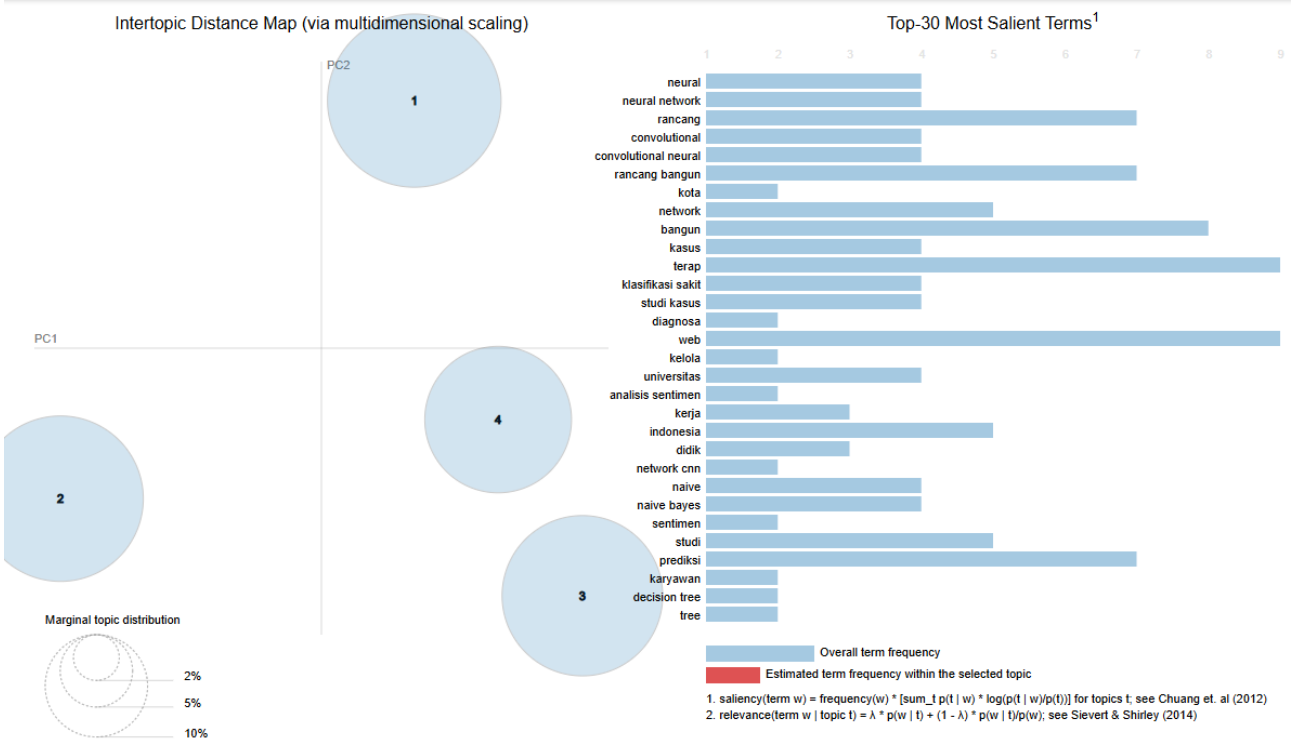


Gambar 3 : Grafik nilai koherens untuk setiap jumlah topik pada model LDA

3.6 Analisis

Dari grafik nilai koherensi pada Gambar 3, hasil pemilihan jumlah topik yang paling optimal adalah sebanyak 4 dengan nilai koherensi sebesar 52.1%. Untuk pemodelan topik pada jurnal informatika satu tahun, jumlah topik yang terlalu sedikit akan menyebabkan jurnal digeneralisir menjadi gabungan topik besar. Jumlah topik yang terlalu banyak juga akan menyebabkan jurnal menjadi terlalu spesifik seperti dipisah untuk setiap algoritma atau metode.

Visualisasi daerah topik yang dibentuk LDA model dengan jumlah topik sebanyak 4 topik dapat dilihat pada Gambar 4. Dapat dilihat juga bahwa terdapat 30 kata yang paling relevan yang menggabungkan keseluruhan topik jurnal yang dihasilkan. Daerah topik (lingkaran) terlihat saling menjauhi antar topik. Ini menunjukkan bahwa tidak ada topik yang memiliki konteks atau semantik yang sama dengan topik yang lain. Ini membuktikan bahwa model LDA berhasil membuat pengelompokkan topik jurnal sesuai dengan konteks atau semantik yang dibahas pada setiap jurnal. 10 kata kunci paling relevan untuk setiap daerah topik dapat dilihat pada Tabel 4. Pada daerah topik 1 terdapat 28.7% jurnal informatika yang secara semantik berkaitan dengan kata kunci yang dikelompokkan, diikuti daerah topik lainnya, yaitu 26.2%, 24.6%, dan 20.5%. Untuk setiap daerah topik akan dibentuk nama topik yang mewakili 10 kata kunci yang paling relevan. Tafsiran nama topik yang dihasilkan dapat dilihat pada Tabel 4.



Gambar 4: Hasil model LDA

Tabel 4 : hasil kata kunci paling relevan, banyak, dan topik yang dapat dihasilkan

Daerah Topik	10 Kata Kunci Paling Relevan	Banyak (%)	Topik
1	sistem informasi, web, klasifikasi, data, bangun, aplikasi, sakit, citra, rancang bangun, rancang	28.7%	Pembangunan Aplikasi Sistem Informasi Berbasis Web
2	rancang, neural, neural network, network, bangun, rancang bangun, convolutional, convolutional neural, klasifikasi, tingkat	26.2%	Perancangan Klasifikasi Citra menggunakan Convolutional Neural Network
3	terap, analisis, website, tingkat, sistem informasi, klasifikasi, ajar, rekomendasi, digital, sentimen	24.6%	Penerapan Sistem Informasi Berbasis Website untuk Analisis Sentimen dan Klasifikasi
4	aplikasi, prediksi, bayes, sakit, analisis, terap, sistem informasi, kota, naive, naive bayes	20.5%	Perancangan Aplikasi Dianogsa Penyakit menggunakan naive bayes

4. KESIMPULAN

Berdasarkan hasil penelitian, didapat kesimpulan sebagai berikut :

1. Pemodelan topik dapat membantu untuk mengelompokkan atau menggabungkan topik-topik jurnal yang mirip, sehingga jurnal menjadi lebih terstruktur dan ada alur proses pengembangan metodenya.
2. Model LDA dengan ekstraksi fitur BoW terbukti mampu untuk memodelkan topik-topik jurnal yang mirip dengan nilai koherensi mencapai 52.1% dengan jumlah topik sebanyak 4 topik.
3. Kekurangan dari penelitian adalah metode ekstraksi fitur yang menggunakan BoW merupakan metode yang kuno. Beberapa metode seperti TF-IDF dapat menghitung seberapa penting kata dalam dokumen, dan *Word2Vec* yang mampu menangkap hubungan semantik antara kata.

DAFTAR PUSTAKA

[1] S. Aryana, A. Y. Wijayanti, and N. Haryati, "P2M STKIP Siliwangi Analisis Trend Topik Penelitian Pendidikan dan Pengajaran pada Jurnal Internasional Bereputasi Q1 Periode 2020-2021," 2022. [Online]. Available: https://www.scimagojr.com

[2] M. Erreza, A. Mustika Rizki, U. Pembangunan Nasional, and J. Timur, "Pencarian Topik Penelitian Pada Studi Kasus Jurnal JIFTI Menggunakan Teknik Hiearchical Dirichlet Processes," vol. 16, pp. 170–182, 2024, [Online]. Available: https://jifti.upnjatim.ac.id/index.php/jifti/issue/archive.

[3] Anisatuzzumara, "FINAL PROJECT LATENT DIRICHLET ALLOCATION (LDA) AND K-NEAREST," Semarang, 2024. Accessed: Jun. 03, 2025. [Online]. Available: http://repository.unissula.ac.id/id/eprint/34043

[4] E. Puspita, D. F. Shiddieq, and F. F. Roji, "Pemodelan Topik pada Media Berita Online Menggunakan Latent Dirichlet Allocation (Studi Kasus Merek Somethinc)," *MALCOM: Indonesian Journal of Machine Learning and Computer Science*, vol. 4, no. 2, pp. 481–489, Feb. 2024, doi: 10.57152/malcom.v4i2.1204.

[5] Y. Matira and I. Setiawan, "Pemodelan Topik pada Judul Berita Online Detikcom Menggunakan Latent Dirichlet Allocation," *Estimasi: Journal of Statistics and Its Application*, vol. 4, no. 1, pp. 2721–379, 2023, doi: 10.20956/ejsa.vi.24843.

[6] W. Wiranto and Mila Rosyida Uswatunnisa, "Topic Modeling for Support Ticket using Latent Dirichlet Allocation," *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, vol. 6, no. 6, pp. 998–1005, Dec. 2022, doi: 10.29207/resti.v6i6.4542.

[7] S. Zhou, P. Kan, Q. Huang, and J. Silbernagel, "A guided latent Dirichlet allocation approach to investigate real-time latent topics of Twitter data during Hurricane Laura," *Journal of Information Science*, vol. 49, no. 2, pp. 465–479, Apr. 2023, doi: 10.1177/01655515211007724.

[8] S. Khoirunnisa, F. Nurdin, F. Sains, and D. Teknologi, "Analisa Pemodelan Topik Berita Daring Menggunakan Semi-supervised Dan Fully Unsupervised Latent Dirichlet Allocation Program Studi Matematika," 2023.

[9] C. Naury, D. H. Fudholi, and A. F. Hidayatullah, "Topic Modelling pada Sentimen Terhadap Headline Berita Online Berbahasa Indonesia Menggunakan LDA dan LSTM," *JURNAL MEDIA INFORMATIKA BUDIDARMA*, vol. 5, no. 1, p. 24, Jan. 2021, doi: 10.30865/mib.v5i1.2556.

[10] B. Subeno, "Topic Modelling Latent Dirichlet Allocation untuk Klasifikasi Komentar Kuliah Pada Twitter X," 2024. [Online]. Available: https://x.com

[11] B. Hamdani,"Sistem Peringkasan Teks Berita Berbahasa Indonesia Menggunakan Latent Dirichlet Allocation Dan Maximum Marginal Relevance Skripsi", 2024 Program Studi Teknik Informatika Fakultas Sains Dan Teknologi Universitas Islam Negeri Maulana Malik Ibrahim Malang .

[12] B. Paula, M. Fawzan, and H. Irsyad, "Analisis Sentiment Masyarakat Terhadap penyebaran Starlink di Indonesia Menggunakan Algoritma Naive Bayes," *Journal Information & Computer JICOM*, vol. 02, no. 2, 2024.

[13] N. Nasser, L. Karim, A. el Ouadrhiri, A. Ali, and N. Khan, "n-Gram based language processing using Twitter dataset to identify COVID-19 patients," *Sustainable Cities and Society*, vol. 72, Sep. 2021, doi: 10.1016/j.scs.2021.103048.

[14] G. Muppala and T. Devi, "Accurate Recasting of Giant Text into Charts Using Rapid Automatic Keyword Extraction Algorithm in Comparison with Bag of Words Algorithm," in Proceedings of International Conference on Contemporary Computing and Informatics, IC3I 2023, Institute of Electrical and Electronics Engineers Inc., 2023, pp. 2548–2552. doi: 10.1109/IC3I59117.2023.10397804.

[15] A. Farkhod, A. Abdusalomov, F. Makhmudov, and Y. I. Cho, "Lda-based topic modeling sentiment analysis using topic/document/sentence (Tds) model," *Applied Sciences (Switzerland)*, vol. 11, no. 23, Dec. 2021, doi: 10.3390/app112311091.