Identifikasi Topik Hangat di Media Berita Menggunakan Latent Dirichlet Allocation

Aulia Cisatra[1], Daffa Daris Mahendra Ansori[2], Dara Nasywa Fathya Afiqah Akmal[3], Slamet Ramadhani[4], Nur Aini Rakhmawati[5]

[1,2,3,4,5] Program studi S-1 Sistem Informasi, Institut Teknologi Sepuluh Nopember
[1]cisatra.205026@mhs.its.ac.id
[2]daffa.205026@mhs.its.ac.id
[3]dara.2050261@mhs.its.ac.id
[4]5998231017@student.its.ac.id
[5]nur.aini@is.its.ac.id

***Abstrak*— Akses informasi di era digital saat ini sebagian besar bersumber dari media daring. Namun, identifikasi topik aktual yang tengah hangat dibicarakan publik kian menjadi tantangan untuk menentukan topik berita yang relevan dan menarik perhatian pembaca. Pada studi ini, dikembangkan sistem cerdas berbasis pembelajaran mesin guna membantu portal berita *online* dalam memahami minat publik terhadap beragam topik. Memanfaatkan pendekatan *natural language processing*, model *latent dirichlet allocation* akan digunakan untuk analisis dan klasifikasi berita berdasarkan topiknya. Data teks digital yang digunakan berasal dari beragam portal berita nasional melalui akses *application programming interface* (API). Melalui *text mining* dan *machine learning*, ekstraksi topik utama secara otomatis dari data teks yang besar dan tidak terstruktur dapat dikenali secara cepat juga akurat. Metode ini memungkinkan pengenalan atas topik-topik yang mendominasi berita daring, sehingga memungkinkan pembaca, peneliti, hingga praktisi media untuk tetap terhubung dengan isu terkini. Hasil studi menunjukkan tiga topik terpopuler berdasarkan dominasi token secara berurutan adalah KPK (36.6%), cawapres (32.8%), dan jakarta (30.6%). Penelitian ini diharapkan dapat meningkatkan daya saing portal berita dalam menyajikan konten aktual sesuai preferensi pembaca. Selain itu, hasil studi dapat menjadi acuan bagi peneliti lain dalam bidang serupa di masa mendatang.**

***Kata Kunci*— Latent Dirichlet Allocation, Media Berita, Natural Language Processing, Topic Modelling, Topik Hangat.**

I. PENDAHULUAN

Merupakan hak setiap manusia atas tersedianya akses ke berita terbaru dari seluruh penjuru dunia. Terlebih di era digital, siaran berita melalui media dalam jaringan pun tidak menjadi pengecualian melainkan salah satu sumber utama dalam persebaran informasi [1]. Fasilitas portal berita daring memungkinkan masyarakat untuk terus terhubung dan mendapatkan pembaruan informasi dalam berbagai bidang termasuk sektor politik, ekonomi, dan budaya secara mudah. Melalui media berita daring, pembaca dapat turut berpartisipasi aktif dalam forum diskusi dan mengutarakan gagasan terkait topik berita melalui kolom komentar dan media sosial. Meskipun diskusi ini dapat saling mencerdaskan antar individu akibat pertukaran perspektif, tantangan dalam mengidentifikasi topih yang mendominasi perbincangan publik menjadi tantangan tersendiri akibat banyaknya informasi yang tersedia. Oleh karena itu, dibutuhkan sistem yang dapat secara cerdas menganalisis data teks digital guna mengetahui topik perbincangan utama.

Berbagai pendekatan telah dikembangkan sebagai upaya dalam mengatasi kendala indetifikasi topik hangat ini. Salah satunya adalah melalui analisis sentimen dan *topic modelling* berbasis pemrosesan bahasa alami atau *natural language processing* [2]. Meskipun demikian, perkembangan dalam siklus pemberitaan serta dinamika minat masyarakat turut menjadi tantangan dalam penelitian sehingga media berita pun memerlukan pendekatan yang lebih cermat dan adaptif terhadap perkembangan terkini. Guna mengatasi tantangan tersebut, penelitian ini mengembangkan sistem idnetifikasi dan klasifikasi topik berita demi mendorong hadirnya sajian informasi intelektual dan berkualitas sesuai minat pembaca.

Pada studi ini dikembangkan sistem cerdas berbasis *natural language processing* (NLP) menggunakan metode *topic modelling* dan pemanfaatan algoritma *latent dirichlet allocation* (LDA) sebagai solusi. Bertujuan untuk merepresentasikan preferensi publik secara nyata, data teks digital bersumber dari kumpulan informasi berita terkini oleh portal berita daring di Indonesia. Dengan memanfaatkan *application programming interface* (API) sebagai sumber data, sistem ini dirancang untuk dapat bekerja secara otomatis dan adaptif dalam analisis sejumlah besar data teks guna ekstraksi daftar topik paling banyak diperbincangkan publik [3].

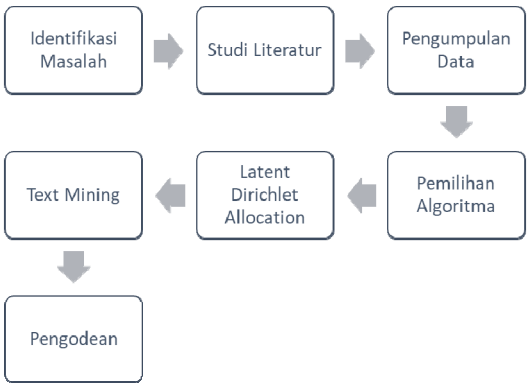
Penelitian serupa terkait identifikasi topik pada portal media daring pernah dikembangkan pada 2022 lalu [4]. Menggunakan *topic modelling* dengan algoritma LDA, penelitian tersebut mengumpulkan data berita dari salah satu akun Twitter milik media berita Detik.com dalam kurun waktu satu tahun. Adapun studi tersebut memiliki potensi pengembangan dengan penggunaan data yang hanya bersumber dari satu portal berita dan rentang waktu yang terbatas. Sehingga pada penelitian ini diupayakan perluasan sumber data yang digunakan dengan tetap terbatas pada situs portal berita dalam negeri oleh situs antaranews.com, cnbcindonesia.com, cnnindonesia.com, jpnn.com, merdeka.com, dan beberapa situs lainnya. Selain itu, penelitian ini pun akan menggunakan rentang serta kategori topik pembicaraan yang lebih luas demi relevansi atas kondisi masyarakat saat ini.

Algoritma LDA ini pun pernah digunakan dan diteliti pada 2022 lalu untuk kasus analisis topik pembahasan dominan melalui pemantauan tagar covidindonesia [5]. Penelitian tersebut dilakukan atas analisis hasil *text mining* 84 takarir (*caption*) dari berbagai pengguna sosial media Instagram. Dalam menentukan jumlah topik yang optimal, penelitian tersebut menggunakan penilaian berdasarkan *perplexity* dan *topic coherence* yang menunjukkan lima topik teratas suatu unggahan video. Identifikasi topik atas konten tersebut antara lain meliputi covidindonesia, covid_19, pandemi di Indonesia, dan pembahasan mutasi virus covid-19.

Selain itu, penelitian lainnya terkait identifikasi konten oleh *virtual YouTuber* (VTuber) terkenal turut menerapkan algoritma LDA dengan mempertimbangkan nilai *perplexity* dan *topic coherence* dalam validasi penentuan topik [6]. Analisis terhadap 4312 video dari 10 kanal VTuber teratas di Indonesia tersebut menghasilkan lima topik tayangan teratas para VTuber Indonesia berupa gim Minecraft dan *reading donation*, gim Apex Legend disertai kolaborasi dengan VTuber lain, siaran langsung vidio gim, *cover* lagu, serta tayangan gim *multiplayer* seperti Raft atau Phasmophobia.

Pemanfaatan *topic modelling* pada jejaring sosial seperti Twitter ataupun portal berita tentu tidak terhitung jumlahnya. Adanya kumpulan penelitian terdahulu menggunakan berbagai pendekatan terhadap area penelitian yang bervariasi pun mendukung gagasan bahwa kombinasi penggunaan *topic modelling* dan algoritma LDA dapat menghasilkan akurasi yang sama baiknya.

Pada penelitian ini, pemanfaatan *topic modelling* dalam identifikasi topik hangat pada berbagai portal berita daring serta penggunaan algoritma LDA yang mampu mengelompokkan, menghubungkan, dan memproses data berita dalam jumlah besar akan digunakan. Selain itu, dipastikan pula bahwa seluruh data berada dalam kategori terbaru demi hasil analisis topik yang relevan dengan masanya. Adapun studi ini diharapkan dapat meningkatkan daya saing media *online* serta memberikan kontribusi dalam pemantauan tren juga isu penting yang tengah berlangsung di Indonesia. Selain itu, diharapkan juga penelitian ini dapat menjadi rujukan dan inspirasi bagi para praktisi akademik untuk penelitian yang berkaitan.



Gbr 1. Metodologi Penelitian

II. METODE PENELITIAN

Pada penelitian ini digunakan pembelajaran mesin, tepatnya pendekatan *natural language processing* (NLP). Metodologi studi ini antara lain meliputi tahapan identifikasi masalah, pembelajaran literatur, pengumpulan data, pemilihan algoritma, penggunaan *latent dirichlet allocation* (LDA), hingga *text mining* dan pengodeaan sebagaimana [Gambar 1].

A. Identifikasi Masalah

Tahapan identifikasi masalah dilakukan guna untuk mengetahui kendala yang melatarbelakangi penelitian, yaitu seiring dengan derasnya peredaran informasi di portal berita, seringkali terjadi kekeliruan dalam penafsiran suatu informasi dan intisari berita dari media sosial. Sehingga pada penelitian ini diajukan pemgembangan sistem identifikasi topik hangat menggunakan *natural language processing.*

B. Studi Literatur

Dalam bagian ini akan diuraikan kajian pustaka yang relevan dengan penelitian identifikasi topik hangat di media berita *online*. Kajian ini membahas beberapa konsep kunci yang mendukung landasan teori dan metodologi penelitian.

1) Data preprocessing

Pra-pemrosesan data merupakan tahapan persiapan untuk menghasilkan kualitas pengelompokan (*clustering*) yang baik [7]. Rangkaian tahapan yang dilakukan pada penelitian ini adalah (1) mengubah kalimat menjadi kata; (2) menghilangkan beberapa kata dalam bahasa Indonesia dan bahasa Inggris yang tidak memiliki arti, seperti kata '*using*', '*of*', '*the*', '*in*', '*on*', '*as*', '*and*', dan '*based*' pada bahasa Inggris dan kata 'berbasis', 'kasus', 'dan', juga 'pada' dalam bahasa Indonesia [12]; (3) mengubah susunan kalimat dalam bentuk bigram, misalnya konversi kalimat "Inovasi Media Pembelajaran Sain Teknologi di SMP Berbasis Mikrokontroler" menjadi [inovasi, media], [pembelajaran, sain], [teknologi, di], [SMP, berbasis], dan juga [mikrokontroller] [8].

2) Natural language processing (NLP)

Pemrosesan bahasa alami atau NLP adalah cabang keilmuan dari teknologi kecerdasan buatan yang berfokus pada pemahaman dan pengolahan bahasa manusia secara alami oleh komputer [2]. Keilmuan ini mencakup sejumlah teknik untuk analisis dan pemrosesan data berbasis teks, seperti tokenisasi, entitas nama, analisa sentimen, hingga pemodelan topik. Dalam konteks identifikasi topik berita, NLP dapat digunakan untuk mengelompokkan artikel berita berdasarkan topik bahasannya.

3) Topic modelling

Pemodelan topik atau *topic modelling* memungkinkan identifikasi kategori dalam koleksi dokumen teks dengan pengelompokan (*clustering*) berdasarkan kemunculan kata kunci untuk setiap topik. Salah satu metode paling populer dalam *topic*

modelling adalah *latent dirichlet allocation* (LDA) [9]. Sebagai model probabilistik, LDA mendasarkan asumsinya bahwa setiap dokumen dalam koleksi adalah campuran dari beberapa topik dan bahwa setiap kata dalam dokumen berasal dari salah satu topik tersebut [10]. Dengan bantuan LDA, penelitian dapat mengungkap topik yang tersembunyi dalam koleksi dokumen, menemukan analisa konten, melakukan pengelompokan dokumen, serta mengidentifikasi tren ataupun isu terkini dalam berita atau teks lainnya

C. *Pengumpulan Data*

Proses koleksi data dilakukan dengan memanfaatkan *application programming interface* (API) oleh API Berita Indonesia yang menampilkan berita oleh portal nasional dari Antara, CNBC, CNN Indonesia JPNN, Kumparan, Merdeka, Okezone, Republika, Sindonews, Suara, Tempo, dan Tribun. Data ini secara langsung didapatkan dari *rich site summary* (RSS) masing-masing situs yang kemudian diubah ke tipe JSON Rest API. Pada penelitian ini, dikumpulkan 574 data berita yang terdiri dari tautan berita, judul, deskripsi, dan tanggal publikasi per tanggal 22 Desember 2023. Alasan pemanfaatan API ini adalah sebab data yang didapatkan bersumber dari berbagai portal berita di Indonesia dan merupakan informasi terbaru pada hari API diakses.

D. *Latent Dirichlet Allocation*

Pendekatan *latent dirichlet allocation* atau LDA merupakan sebuah metode statistika yang digunakan sebagai model untuk analisis suatu dokumen [10]. Dengan upaya melihat dokumen secara mundur untuk menemukan kumpulan topik yang mungkin telah dikoleksi, LDA merepresentasikan dokumen dengan berbagai topik yang dibuat berdasarkan probabilitas tertentu. Di mana probabilitas topik merepresentasikan kejelasan dari suatu dokumen.

E. *Text Mining*

Text Mining adalah metode yang terdiri dari banyak langkah yang memungkinkan Anda menyimpulkan informasi dari data teks yang tidak terstruktur. Proses membersihkan dan mengonversi data teks menjadi format yang dapat digunakan disebut pemrosesan awal teks, dan ini harus dilakukan sebelum Anda dapat menggunakan salah satu dari banyak teknik penambangan teks. Pemrosesan bahasa alami (NLP) adalah komponen kunci dari proses ini, dan untuk menyiapkan data dengan benar untuk analisis

F. *Pengodean*

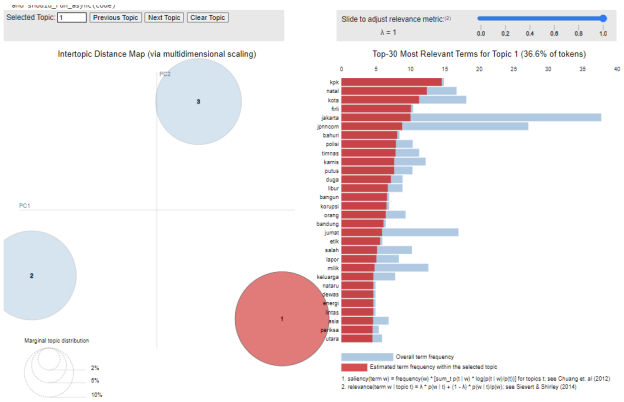
Langkah pengodean melibatkan eksekusi kode program dalam bahasa pemrograman Python menggunakan lingkungan Google Colab. Tujuan tahapan ini adalah untuk menjalankan fungsi dan metode yang relevan dalam penelitian untuk memperoleh hasil analisa yang diharapkan.

III. HASIL DAN PEMBAHASAN

Pengolahan data pada penelitian ini dilakukan dengan *preprocessing* menggunakan metode *latent dirichlet allocation* (LDA). Tujuan dari metode ini adalah untuk menentukan kata kunci yang mewakili tiap topik melalui implementasi *library* Python. Adapun luaran dari pengolahan ini adalah berupa diagram topik populer dari beberapa portal media daring sebagai bentuk representasi konten berita. Hasil pemodelan ini divisualisasikan menggunakan *library* pydavis dan dikaji melalui analisa mendalam. Pada penelitian ini, visualisasi hasil dibagi menjadi tiga topik dengan masing-masing topik terdiri dari setidaknya 30 pembahasan.

A. *Hasil topik 1*

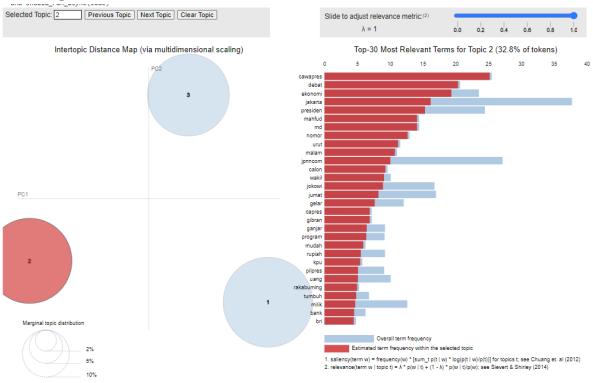
Topik pertama pada penelitian ini didominasi oleh bahasan mengenai berita di Jakarta. Adapun beberapa pembicaraan populer dari topik ini adalah mengenai “kpk”, acara tahunan “natal”, “kota”, “firli”, “jakrta”, dan lain – lain sebagaimana terlampir pada Gambar 2



Gbr 2. Hasil analisis topik berita di Jakarta

B. *Hasil topik 2*

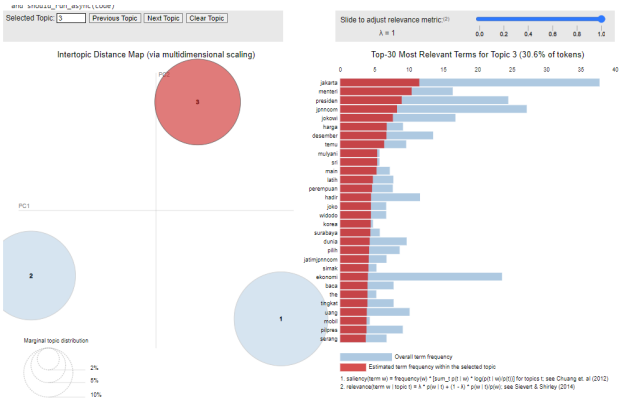
Topik kedua pada penelitian ini didominasi oleh bahasan mengenai pemilu. Adapun beberapa pembicaraan populer dari topik ini adalah “cawapres”, “debat”, “ekonomi”, “jakarta”, “presiden”, dan lain - lain sebagaimana terlampir pada Gambar 3



Gbr 3. Hasil analisis topik mengenai pemilu

C. *Hasil topik 3*

Topik ketiga pada penelitian ini didominasi oleh bahasan mengenai pemerintahan. Adapun beberapa pembicaraan populer dari topik ini adalah "jakarta", "menteri", "presiden", jpnncom", "jokowi", dan lain – lain sebagimana terlampir pada Gambar 4



Gbr 4. Hasil analisis topik mengenai pemerintahan

D. *Word Cloud*

Dikenal sebagai cloud tag, Word Colud adalah representasi visual dari teks tertulis yang diatur berdasarkan referensi. Teknik ini sering digunakan untuk mepresentasikan data tekstual secara grafis, menyerderhanakan proses analisis data tekstual secara grafis, menyederhanakan proses analisis data tekstual seperti esai, tanggapan tertulis dan sejenisnya. Word Cloud juga berfungsi sebagai titik awal untuk analisis mendalam terhadap materi tertulis atau teks.

Gbr 5. Visualisasi Word Cloud topik berita di indonesia

IV. SIMPULAN

Pada penelitian ini metode Latent Dirichlet Allocation (LDA) berhasil diimplementasikan untuk memodelkan topik berita beberapa portal yang di dalamnya memuat berita berbeda dan kategori berita yang berbeda pula. Penelitian ini menghasilkan tiga topik yang hasil pemodelan tersebut diplot untuk mengetahui trend berita atau kategori berita mana yang sering menjadi pembahasan di akun detik.com. Hasil dari ketiga topik adalah topik 1 yang utamanya berkaitan dengan berita di jakarta, diikuti dengan topik 2 yang didominasi bahasan mengenai pemilu 2024, serta topik 3 yang membahas tentang pemerintahan.

REFERENSI

[1] B. Garrison, "Journalists' Perceptions of Online Information-Gathering Problems," Journal. Mass Commun. Q., vol. 77, no. 3, pp. 500–514, Sep. 2000, doi: 10.1177/107769900007700303.

[2] M. Birjali, M. Kasri, and A. Beni-Hssane, "A comprehensive survey on sentiment analysis: Approaches, challenges and trends," Knowl.-Based Syst., vol. 226, p. 107134, Aug. 2021, doi: 10.1016/j.knosys.2021.107134.

[3] L. Chen, H. Zhang, J. M. Jose, H. Yu, Y. Moshfeghi, and P. Triantafillou, "Topic detection and tracking on heterogeneous information," J. Intell. Inf. Syst., vol. 51, no. 1, pp. 115–137, Aug. 2018, doi: 10.1007/s10844-017-0487-y.

[4] S. Karmila and V. I. Ardianti, "METODE LATENT DIRICHLET ALLOCATION UNTUK MENENTUKAN TOPIK TEKS SUATU BERITA," J. Inform. Dan Komputasi Media Bahasan Anal. Dan Apl., vol. 16, no. 01, pp. 36–44, Nov. 2022, doi: 10.56956/jiki.v16i01.100.

[5] K. R. A. P. Santoso, A. Husna, N. W. Putri, and N. A. Rakhmawati, "Analisis Topik Tagar Covidindonesia pada Instagram Menggunakan Latent Dirichlet Allocation," JISKA J. Inform. Sunan Kalijaga, vol. 7, no. 1, pp. 1–9, Jan. 2022, doi: 10.14421/jiska.2022.7.1.1-9.

[6] Akdeas Oktanae Widodo, F. Septiadi, and Nur Aini Rakhmawati, "ANALISIS TREN KONTEN PADA VTUBER INDONESIA MENGGUNAKAN LATENT DIRICHLET ALLOCATION," J. Inform. Dan Rekayasa Elektron., vol. 6, no. 1, pp. 56–63, Apr. 2023, doi: 10.36595/jire.v6i1.718.

[7] D. Wu, R. Yang, and C. Shen, "Sentiment word co-occurrence and knowledge pair feature extraction based LDA short text clustering algorithm," J. Intell. Inf. Syst., vol. 56, no. 1, pp. 1–23, Feb. 2021, doi: 10.1007/s10844-020-00597-7.

[8] K. Watanabe and A. Baturo, "Seeded Sequential LDA: A Semi-Supervised Algorithm for Topic-Specific Analysis of Sentences," Soc. Sci. Comput. Rev., p. 08944393231178605, May 2023, doi: 10.1177/08944393231178605.

[9] K. Bastani, H. Namavari, and J. Shaffer, "Latent Dirichlet allocation (LDA) for topic Modelling of the CFPB consumer complaints," Expert Syst. Appl., vol. 127, pp. 256–271, Aug. 2019, doi: 10.1016/j.eswa.2019.03.001.

[10] D. Blei, A. Ng, and M. Jordan, "Latent Dirichlet Allocation," in Advances in Neural Information Processing Systems, T. Dietterich, S. Becker, and Z. Ghahramani, Eds., MIT Press, 2001. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2001/file/296472c9542ad4d4788d543508116cbc-Paper.pdfi1

[11] Akmal, D. N. F. A., Ansori, D. D. M., Cisatra, A., Ramadhani, S., & Rakhmawati, N. A. (2023). Dataset Indonesian News. https://doi.org/10.5281/ZENODO.8399304

[12] [12] R. Rani and D. K. Lobiyal, "An extractive text summarization approach using tagged-LDA based topic modeling," Multimed. Tools Appl., vol. 80, no. 3, pp. 3275–3305, Jan. 2021, doi: 10.1007/s11042-020-09549-3.