

BAB I

PENDAHULUAN

1.1 Latar Belakang

Perkembangan teknologi informasi yang pesat telah mendorong meningkatnya jumlah berita online berbahasa Indonesia secara eksponensial. Berita online kini menjadi sumber informasi utama masyarakat karena dapat diakses dengan cepat dan mudah melalui berbagai platform digital. Setiap harinya, ribuan artikel berita dipublikasikan oleh berbagai portal berita nasional, menghasilkan volume data yang sangat besar dan bersifat tidak terstruktur. Kondisi ini menyulitkan proses analisis secara manual karena tidak efisien dan membutuhkan waktu yang sangat lama. Oleh karena itu, diperlukan metode pengolahan teks otomatis yang mampu mengekstraksi pola dan struktur tematik dari kumpulan dokumen berita secara efisien (Hou & Huang, 2025; Li et al., 2022).

Salah satu pendekatan yang banyak digunakan untuk menganalisis data teks tidak terstruktur adalah pemodelan topik (*topic modeling*), yaitu teknik yang bertujuan menemukan topik-topik tersembunyi dalam kumpulan dokumen secara otomatis tanpa memerlukan data berlabel (Vayansky & Kumar, 2020; Abdelrazek et al., 2023; Churchill & Singh, 2022). Metode ini bekerja dengan menganalisis pola distribusi kata dalam dokumen untuk mengungkap tema-tema utama yang merepresentasikan isi keseluruhan korpus (Qiang et al., 2020; Zuo et al., 2021). Karena termasuk dalam kategori *unsupervised learning*, pemodelan topik sangat sesuai untuk diterapkan pada data berita online yang umumnya tidak memiliki label kategori yang eksplisit (Murshed et al., 2023; Curiskis et al., 2020).

Pemodelan topik telah banyak diterapkan dalam berbagai domain penelitian, mulai dari analisis media sosial, forum diskusi, dokumen ilmiah, hingga berita online (Muthusami et al., 2024; Ghasiya & Okamura, 2021; Jung et al., 2024; Silva et al., 2021). Dalam konteks berita, metode ini digunakan untuk memahami opini publik, mengidentifikasi isu sosial yang sedang berkembang, serta menganalisis tren informasi dalam periode waktu tertentu (Kumi et al., 2024; Sik et al., 2023; Li &

Chang, 2023). Hal ini membuktikan bahwa pemodelan topik merupakan pendekatan yang relevan dan bernilai tinggi dalam analisis data teks berskala besar.

Meskipun pemodelan topik telah terbukti bermanfaat, terdapat permasalahan mendasar yang belum terselesaikan, yaitu perbedaan kualitas topik yang dihasilkan oleh setiap metode. Setiap metode memiliki karakteristik dan pendekatan yang berbeda sehingga dapat menghasilkan topik dengan tingkat koherensi, kejelasan, dan interpretabilitas yang berbeda pula (Amaro & Bacao, 2024; Rüdiger et al., 2022). Permasalahan ini menjadi semakin kritis pada data berita online berbahasa Indonesia yang memiliki kompleksitas linguistik tinggi, variasi kosakata yang beragam, serta penggunaan kata-kata serapan dan istilah daerah yang tidak ditemukan dalam bahasa lain. Hingga saat ini, belum terdapat acuan yang jelas mengenai metode pemodelan topik yang paling optimal untuk data berita online berbahasa Indonesia, khususnya dalam membandingkan *Latent Dirichlet Allocation* (LDA) dan *Non-negative Matrix Factorization* (NMF) secara langsung (Amaro & Bacao, 2024).

Sejumlah penelitian sebelumnya telah menerapkan LDA maupun NMF secara terpisah dalam pemodelan topik, namun menunjukkan hasil yang bervariasi bergantung pada karakteristik data yang digunakan. Sebagai contoh, beberapa studi menemukan bahwa LDA menghasilkan topik yang lebih representatif pada data yang berdistribusi luas, sementara studi lain menunjukkan bahwa NMF lebih unggul dalam menghasilkan topik yang lebih tajam dan mudah diinterpretasikan (Egger & Yu, 2022; Latif et al., 2021; M'sik & Casablanca, 2020). Perbedaan temuan ini mengindikasikan bahwa performa kedua metode sangat bergantung pada domain dan karakteristik data, sehingga perbandingan langsung pada konteks berita berbahasa Indonesia menjadi penting untuk dilakukan.

Di antara berbagai pendekatan pemodelan topik, LDA menonjol sebagai metode probabilistik yang paling luas diadopsi dalam penelitian text mining. LDA memodelkan dokumen sebagai campuran distribusi topik dan setiap topik sebagai distribusi kata, sehingga mampu menangkap hubungan laten antar kata dan dokumen secara fleksibel (Chauhan & Shah, 2021; Hua et al., 2020). Keunggulan utama LDA terletak pada kemampuannya memodelkan ketidakpastian distribusi

topik secara global sehingga cocok digunakan pada korpus berukuran besar dengan variasi tema yang luas (M'sik & Casablanca, 2020). Namun demikian, LDA memiliki keterbatasan dalam hal interpretabilitas, karena topik yang dihasilkan terkadang mengandung kata-kata yang secara semantik kurang kohesif sehingga sulit dipahami secara langsung.

Sebagai perbandingan, *Non-negative Matrix Factorization* (NMF) menawarkan pendekatan yang berbeda secara fundamental, yaitu berbasis faktorisasi matriks. NMF memecah matriks dokumen-kata menjadi dua matriks non-negatif yang merepresentasikan hubungan antara dokumen dan topik serta antara topik dan kata (Vangara et al., 2021; Ebrahimi & Peltonen, 2025). Karena setiap nilai dalam matriks bersifat non-negatif, representasi yang dihasilkan lebih bersifat parsimonious dan mudah diinterpretasikan. Keunggulan NMF terletak pada kemampuannya menghasilkan topik yang lebih spesifik, terpisah dengan jelas antar satu topik dengan topik lainnya, dan lebih mudah dipahami oleh pengguna akhir (M'sik & Casablanca, 2020). Namun, NMF tidak secara eksplisit memodelkan distribusi probabilistik antar topik seperti pada LDA sehingga representasinya lebih deterministik.

Perbedaan mendasar antara pendekatan probabilistik pada Latent Dirichlet Allocation (LDA) dan pendekatan faktorisasi matriks pada Non-Negative Matrix Factorization (NMF) menjadi dasar dilakukannya perbandingan antara kedua metode tersebut. LDA diketahui lebih unggul dalam memodelkan distribusi topik secara global serta dalam menangani ambiguitas kata, sedangkan NMF dinilai lebih mampu menghasilkan topik yang memiliki batas yang lebih jelas dan mudah diinterpretasikan.

Berdasarkan perbedaan tersebut, diperlukan analisis untuk menentukan metode yang lebih efektif ketika diterapkan pada data berita online berbahasa Indonesia. Penentuan metode yang lebih optimal tidak dapat dilakukan secara teoritis semata, melainkan harus didasarkan pada hasil evaluasi empiris yang dilakukan secara langsung terhadap dataset yang relevan (Egger & Yu, 2022; Latif et al., 2021; M'sik & Casablanca, 2020).

Konteks bahasa Indonesia menambah dimensi tersendiri dalam penelitian ini. Bahasa Indonesia memiliki karakteristik linguistik yang unik, di antaranya sistem afiksasi yang kompleks (penggunaan awalan, akhiran, dan imbuhan sisipan), penggunaan kata-kata serapan dari bahasa daerah maupun bahasa asing, serta gaya penulisan berita yang variatif antarportal. Karakteristik ini berpotensi memengaruhi kualitas representasi teks dan performa masing-masing metode pemodelan topik secara berbeda. Oleh karena itu, hasil penelitian dari konteks bahasa lain tidak dapat langsung digeneralisasikan ke data berbahasa Indonesia, dan studi komparatif yang berfokus pada konteks ini menjadi sangat dibutuhkan.

Evaluasi terhadap hasil pemodelan topik dalam penelitian ini dilakukan menggunakan metrik *topic coherence*, yaitu metrik yang mengukur tingkat keterkaitan semantik antar kata-kata teratas dalam suatu topik (Meaney et al., 2023; Hosseiny Marani & Baumer, 2023; Korenčić et al., 2021). Nilai koherensi yang lebih tinggi mengindikasikan bahwa topik yang dihasilkan lebih konsisten secara semantik dan lebih mudah dipahami. Selain itu, evaluasi juga dilengkapi dengan penilaian interpretabilitas topik secara kualitatif guna memperoleh gambaran yang lebih komprehensif mengenai kualitas hasil pemodelan.

Berdasarkan uraian permasalahan di atas, penelitian ini bertujuan untuk menganalisis dan membandingkan performa metode LDA dan NMF dalam pemodelan topik pada berita online berbahasa Indonesia. Dataset yang digunakan diperoleh melalui proses *web scraping* dari beberapa portal berita nasional dengan total 4.000 berita yang terbagi ke dalam empat kategori, yaitu politik, ekonomi, teknologi, dan olahraga, dengan periode publikasi tahun 2020–2025. Penelitian ini diharapkan mampu memberikan acuan empiris yang jelas mengenai metode pemodelan topik yang paling efektif pada konteks berita berbahasa Indonesia berdasarkan nilai *topic coherence* dan interpretabilitas, serta menjadi referensi bagi penelitian selanjutnya di bidang pengolahan teks berbasis *unsupervised learning*.

1.2 Rumusan Masalah

1.2.1 Permasalahan

Berdasarkan uraian pada latar belakang, permasalahan utama dalam penelitian ini adalah belum diketahuinya metode pemodelan topik yang paling

efektif dalam menghasilkan topik berkualitas pada data berita online berbahasa Indonesia. Meskipun pemodelan topik telah terbukti mampu mengekstraksi struktur tematik dari kumpulan dokumen secara otomatis, setiap metode menghasilkan topik dengan tingkat koherensi dan interpretabilitas yang berbeda-beda. Kondisi ini menyebabkan belum adanya acuan yang jelas mengenai metode yang paling optimal untuk digunakan pada konteks data berita berbahasa Indonesia, khususnya dalam hal perbandingan antara metode *Latent Dirichlet Allocation* (LDA) dan *Non-Negative Matrix Factorization* (NMF). LDA dan NMF merupakan dua metode pemodelan topik yang paling banyak digunakan, namun memiliki pendekatan yang secara fundamental berbeda. LDA berbasis probabilistik dan unggul dalam memodelkan distribusi topik secara global, sedangkan NMF berbasis faktorisasi matriks dan cenderung menghasilkan topik yang lebih spesifik serta mudah diinterpretasikan. Perbedaan pendekatan ini mengakibatkan kedua metode berpotensi menghasilkan kualitas topik yang berbeda ketika diterapkan pada dataset yang sama, khususnya pada berita berbahasa Indonesia yang memiliki kompleksitas linguistik tersendiri. Oleh karena itu, diperlukan analisis perbandingan yang sistematis untuk mengevaluasi performa kedua metode berdasarkan nilai *topic coherence* dan interpretabilitas topik yang dihasilkan (Egger & Yu, 2022; Amaro & Bacao, 2024; M'sik & Casablanca, 2020).

1.2.2 Metode Penelitian

Permasalahan tersebut dapat diatasi dengan menerapkan metode pemodelan topik menggunakan pendekatan unsupervised learning melalui beberapa tahapan yang saling berkaitan. Tahap pertama adalah pengumpulan dataset berita online berbahasa Indonesia yang diperoleh melalui proses web scraping dari beberapa portal berita nasional. Data yang digunakan berupa judul berita dan isi berita yang kemudian dilakukan tahap preprocessing teks yang meliputi pembersihan teks, tokenisasi, stopwords removal, dan stemming agar data siap diproses pada tahap berikutnya.

Tahap kedua adalah pembentukan representasi numerik teks menggunakan metode Term Frequency–Inverse Document Frequency (TF-IDF) sehingga dokumen teks dapat diolah dalam bentuk matriks dokumen-kata. Tahap ketiga adalah penerapan metode *Latent Dirichlet Allocation* (LDA) dan *Non-Negative*

Matrix Factorization (NMF) untuk menghasilkan sekumpulan topik beserta kata-kata dominan yang merepresentasikan isi berita. Kedua metode tersebut diterapkan pada dataset yang sama sehingga hasil yang diperoleh dapat dibandingkan secara objektif. Tahap terakhir adalah evaluasi dan perbandingan hasil pemodelan topik yang dihasilkan oleh metode LDA dan NMF. Evaluasi dilakukan berdasarkan nilai koherensi topik serta interpretabilitas topik yang dihasilkan untuk mengetahui metode yang lebih efektif dalam memodelkan topik berita online berbahasa Indonesia.

1.2.3 Pertanyaan Penelitian

Berapa optimal perbandingan kualitas topik yang dihasilkan oleh metode LDA dan NMF dalam pemodelan topik berita online berbahasa Indonesia berdasarkan nilai topic coherence ?

1.3 Tujuan dan Manfaat

1.3.1 Tujuan

Tujuan penelitian ini adalah untuk menganalisis dan membandingkan kualitas topik yang dihasilkan oleh metode Latent Dirichlet Allocation (LDA) dan Non-Negative Matrix Factorization (NMF) dalam pemodelan topik berita online berbahasa Indonesia berdasarkan nilai topic coherence, interpretabilitas, serta kemampuan representasi tema dokumen.

1.3.2 Manfaat

Penelitian ini memberikan manfaat dalam mengetahui metode pemodelan topik yang lebih efektif antara LDA dan NMF pada data berita online berbahasa Indonesia. Melalui perbandingan yang dilakukan, penelitian ini mampu menunjukkan metode yang menghasilkan topik dengan kualitas lebih baik berdasarkan tingkat koherensi dan kemudahan interpretasi. Selain itu, hasil penelitian ini juga memberikan gambaran pola topik dalam berita online serta dapat menjadi referensi dalam pemilihan metode pemodelan topik yang sesuai pada penelitian serupa.

1.4 Batasan Masalah

1. Penelitian ini menggunakan data berupa berita online berbahasa Indonesia yang diperoleh melalui proses web scraping dari portal berita nasional seperti Kompas, Detik, CNN Indonesia, dan Antaranews dengan periode data tahun 2020–2025.
2. Jumlah data yang digunakan dalam penelitian ini 4000 berita, di mana 1 record merepresentasikan 1 berita yang terdiri dari judul berita dan isi berita.
3. Penelitian ini hanya menganalisis berita berbahasa Indonesia, sehingga berita berbahasa selain Bahasa Indonesia tidak disertakan dalam proses analisis.
4. Topik berita yang dianalisis dalam penelitian ini dibatasi pada kategori politik, ekonomi, teknologi, dan olahraga untuk merepresentasikan keberagaman isu dalam berita online.
5. Penelitian ini berfokus pada perbandingan kinerja metode Latent Dirichlet Allocation (LDA) dan Non-Negative Matrix Factorization (NMF) dalam pemodelan topik, tanpa melibatkan metode pemodelan topik lainnya.
6. Representasi teks yang digunakan dalam penelitian ini adalah TF-IDF (*Term Frequency–Inverse Document Frequency*) untuk membentuk matriks dokumen-kata sebelum diterapkan metode LDA dan NMF.
7. Tahapan preprocessing teks meliputi cleaning, case folding, tokenizing, stemming, stopword removal.
8. Evaluasi hasil pemodelan topik dilakukan menggunakan nilai topic coherence dan interpretabilitas topik.
9. Output penelitian berupa kumpulan topik beserta kata dominan pada setiap topik, dan bukan berupa klasifikasi berita.

1.5 Sistematika Penulisan Skripsi

Penulisan pada laporan ini akan ditulis ke dalam 5 bab, yaitu sebagai berikut :

BAB 1 PENDAHULUAN

Bab ini berisi latar belakang penelitian, perumusan masalah, usulan solusi, pertanyaan penelitian, tujuan penelitian, manfaat penelitian, batasan masalah, serta

sistematika penulisan laporan. Bab ini menjelaskan pentingnya analisis topik pada data berita online berbahasa Indonesia serta alasan pemilihan metode *Latent Dirichlet Allocation* (LDA) dan *Non-Negative Matrix Factorization* (NMF) sebagai metode pemodelan topik yang dibandingkan.

BAB II TINJAUAN PUSTAKA

Bab ini memuat teori-teori yang mendukung penelitian serta penelitian terdahulu yang relevan. Pembahasan dalam bab ini meliputi konsep *Text Mining*, *Natural Language Processing* (NLP), tahapan text preprocessing, pembobotan kata menggunakan TF-IDF, konsep pemodelan topik, metode *Latent Dirichlet Allocation* (LDA), metode *Non-Negative Matrix Factorization* (NMF), serta metode evaluasi model menggunakan topic coherence. Selain itu, bab ini juga membahas penelitian terdahulu yang berkaitan dengan penerapan pemodelan topik pada data teks.

BAB III TINJAUAN PUSTAKA

Bab ini menjelaskan metode penelitian yang digunakan dalam penelitian ini. Pembahasan meliputi proses pengumpulan dataset berita online berbahasa Indonesia, tahap preprocessing data yang terdiri dari case folding, tokenizing, stopword removal, dan stemming, proses pembobotan kata menggunakan TF-IDF, penerapan metode LDA dan NMF untuk ekstraksi topik, serta proses evaluasi hasil pemodelan topik menggunakan nilai topic coherence untuk membandingkan performa kedua metode.

BAB IV HASIL DAN PEMBAHASAN

Bab ini menyajikan seluruh hasil eksperimen pemodelan topik yang dilakukan terhadap dataset berita online berbahasa Indonesia. Pembahasan meliputi hasil preprocessing data, representasi TF-IDF, pencarian jumlah topik optimal (K), hasil pemodelan topik menggunakan metode LDA dan NMF, perbandingan nilai Topic Coherence dan waktu komputasi kedua metode, serta analisis pengaruh stemming terhadap kualitas topik yang dihasilkan.

BAB V PENUTUP

Bab ini berisi kesimpulan dari hasil penelitian yang menjawab pertanyaan penelitian mengenai perbandingan kualitas topik antara metode LDA dan NMF berdasarkan nilai topic coherence, interpretabilitas, dan efisiensi komputasi. Selain itu, bab ini juga memuat saran untuk pengembangan penelitian selanjutnya di bidang pemodelan topik pada data teks berbahasa Indonesia.