

BAB I

PENDAHULUAN

1.1 Latar Belakang

Pemodelan topik merupakan pendekatan yang digunakan untuk menemukan kelompok kata atau topik dari sekumpulan dokumen yang dapat merepresentasikan informasi di dalam dokumen tersebut (Fahlevvi, 2022). Salah satu metode yang banyak digunakan dalam pemodelan topik adalah *Latent Dirichlet Allocation* (LDA)(Abdurrazzaq, 2023). LDA merupakan model *probabilistik generatif* yang memandang dokumen sebagai campuran dari beberapa topik *laten*, sedangkan setiap topik direpresentasikan melalui distribusi kata. Dengan pendekatan tersebut, LDA dapat digunakan untuk menemukan topik tersembunyi dari kumpulan dokumen teks secara otomatis (Blei et al., 2003).

Sebelum dilakukan pemodelan topik, data teks perlu diubah ke dalam bentuk numerik agar dapat diproses oleh model. Hal ini diperlukan karena data teks mentah tidak dapat langsung digunakan sebagai masukan model komputasi. Representasi teks menjadi tahap penting dalam pemrosesan bahasa alami karena mengubah data teks mentah menjadi bentuk numerik yang dapat diproses oleh model. Oleh karena itu, pemilihan representasi teks menjadi bagian penting dalam proses pemodelan topik (Nuradilla et al., 2025).

Bag of Words (BoW) dan *Term Frequency-Inverse Document Frequency* (TF-IDF) merupakan dua pendekatan representasi teks yang dapat digunakan dalam pemodelan topik. BoW merepresentasikan dokumen berdasarkan kemunculan kata tanpa mempertimbangkan urutan kata, sedangkan TF-IDF memberikan pembobotan terhadap kata dengan mempertimbangkan frekuensi kemunculan kata dalam dokumen dan tingkat kemunculannya pada keseluruhan korpus. Perbedaan cara pembobotan tersebut membuat BoW dan TF-IDF dapat menghasilkan representasi dokumen yang berbeda, sehingga keduanya layak dibandingkan untuk melihat pengaruhnya terhadap kualitas topik yang dihasilkan oleh LDA (Nisa et al., 2025).

Beberapa penelitian terdahulu menunjukkan bahwa penggunaan representasi BoW dan TF-IDF pada metode LDA dapat menghasilkan performa yang berbeda. Pada dataset publikasi dosen dan abstrak tugas akhir, TF-IDF menghasilkan nilai coherence lebih tinggi dibandingkan BoW (Nisa et al., 2025)

(Prihatini & Sudiarta, 2024). Sebaliknya, pada dataset ulasan TripAdvisor terkait pariwisata Raja Ampat, BoW menghasilkan nilai *coherence* lebih tinggi dibandingkan TF-IDF (Sugiarto et al., n.d.). Hasil serupa juga ditemukan pada data Twitter terkait kebocoran data, di mana BoW menghasilkan topik yang lebih mudah diinterpretasikan dibandingkan TF-IDF (Ariansyah & Indahyanti, 2024). Berdasarkan hasil tersebut, dapat diketahui bahwa TF-IDF cenderung menghasilkan *coherence* lebih tinggi pada dataset yang memiliki pembahasan lebih beragam. Sebaliknya, BoW cenderung menghasilkan *coherence* lebih tinggi atau topik yang lebih mudah diinterpretasikan pada dataset dengan pembahasan yang lebih sedikit.

Topic coherence digunakan sebagai metrik evaluasi untuk menilai kualitas topik yang dihasilkan oleh LDA. Metrik ini mengukur nilai sebuah topik dengan menghitung derajat kesamaan semantik antara kata-kata yang memiliki nilai tinggi dalam topik. Dengan demikian, semakin tinggi nilai *topic coherence*, semakin baik keterkaitan antar kata dalam topik yang dihasilkan (Abdurrazzaq, 2023).

Media sosial berkembang sebagai sarana interaksi publik yang memungkinkan masyarakat menyampaikan opini terhadap isu sosial dan kebijakan publik. Salah satu platform yang banyak digunakan adalah TikTok, yang tidak hanya berfungsi sebagai media hiburan berbasis video pendek, tetapi juga sebagai ruang diskusi publik untuk menyampaikan pandangan, dukungan, dan kritik terhadap suatu kebijakan. Tingginya interaksi pengguna pada kolom komentar TikTok menjadikan platform ini sebagai sumber data yang bernilai untuk memahami opini publik. Dalam penelitian ini, komentar TikTok terkait RUU Perampasan Aset digunakan sebagai objek penelitian karena memuat tanggapan masyarakat terhadap isu publik. Selain itu, komentar TikTok cenderung bersifat informal, pendek, dan bervariasi, sehingga data tersebut perlu melalui proses pengolahan teks sebelum digunakan dalam pemodelan topik menggunakan LDA (Nugroho & Mujiyono, 2026).

Berdasarkan uraian sebelumnya, perbandingan representasi BoW dan TF-IDF pada metode LDA masih perlu diuji pada karakteristik data yang berbeda. Pada penelitian ini, data yang digunakan berupa komentar TikTok berbahasa Indonesia terkait RUU Perampasan Aset. Penelitian yang secara khusus membandingkan

representasi BoW dan TF-IDF pada data tersebut masih terbatas. Oleh karena itu, penelitian ini membandingkan kedua representasi teks tersebut pada pemodelan topik menggunakan LDA dengan optimasi parameter jumlah topik, *alpha*, dan *beta* melalui *Grid Search*. Model terbaik dari setiap representasi selanjutnya dibandingkan berdasarkan nilai *topic coherence* untuk mengetahui representasi teks yang menghasilkan kualitas topik terbaik pada data yang dianalisis.

1.2 Rumusan Masalah

Berdasarkan latar belakang penelitian yang telah diuraikan sebelumnya, terdapat beberapa permasalahan yang perlu dikaji dalam penelitian ini.

1.2.1 Permasalahan

Permasalahan dalam penelitian ini adalah belum diketahui representasi teks antara BoW dan TF-IDF yang menghasilkan nilai *topic coherence* lebih tinggi pada pemodelan topik menggunakan metode LDA terhadap komentar TikTok berbahasa Indonesia terkait RUU Perampasan Aset.

1.2.2 Metode Usulan

Metode yang digunakan dalam penelitian ini adalah LDA sebagai metode pemodelan topik. Representasi teks yang digunakan yaitu BoW dan TF-IDF. Kemudian dilakukan juga optimasi parameter menggunakan *Grid Search* dengan menguji variasi jumlah topik, *alpha*, dan *beta*. Model terbaik dari setiap representasi ditentukan berdasarkan nilai *topic coherence* tertinggi.

1.2.3 Pertanyaan Penelitian

Representasi teks manakah antara BoW dan TF-IDF yang menghasilkan nilai *topic coherence* lebih tinggi pada pemodelan topik menggunakan metode LDA terhadap dataset komentar TikTok terkait RUU Perampasan Aset dengan karakteristik data yang berfokus pada satu isu utama?

1.3 Tujuan dan Manfaat Penelitian

Berdasarkan rumusan masalah dan pertanyaan penelitian yang telah diuraikan sebelumnya.

1.3.1 Manfaat Penelitian

Penelitian ini dapat menambah referensi dalam bidang text mining dan *Topic Modeling*, khususnya mengenai pengaruh representasi teks BoW dan TF-IDF terhadap kualitas topik yang dihasilkan oleh metode LDA. Selain itu penelitian ini

dapat menjadi referensi bagi peneliti atau pengembang sistem analisis teks dalam memilih representasi teks yang sesuai untuk pemodelan topik pada data media sosial berbahasa Indonesia.

1.4 Batasan Masalah

Untuk menjaga fokus penelitian dan menghindari pembahasan yang terlalu luas, maka batasan masalah dalam penelitian ini adalah sebagai berikut:

1. Data yang digunakan berupa komentar Tiktok berbahasa Indonesia terkait RUU Perampasan Aset yang berjumlah 2.954 komentar.
2. Penelitian hanya menggunakan metode *Latent Dirichlet Allocation* (LDA).
3. Evaluasi model dilakukan menggunakan metrik *topic coherence*.
4. Penelitian tidak membahas analisis sentimen atau aspek hukum dari isi komentar.

1.4.1 Tujuan

Tujuan penelitian ini adalah membandingkan nilai *topic coherence* yang dihasilkan oleh representasi teks BoW dan TF-IDF pada pemodelan topik menggunakan metode LDA terhadap komentar TikTok berbahasa Indonesia terkait RUU Perampasan Aset.

1.5 Sistematika Penulisan

Sistematika penulisan dalam penelitian ini disusun untuk memberikan gambaran mengenai struktur dan alur pembahasan penelitian secara sistematis.

BAB I PENDAHULUAN

Bab ini menjelaskan secara umum mengenai gambaran awal penelitian, yang meliputi latar belakang penelitian, perumusan masalah, tujuan dan manfaat penelitian, batasan masalah, serta sistematika penulisan yang menjadi kerangka dasar dalam pelaksanaan penelitian.

BAB II TINJAUAN PUSTAKA

Bab ini membahas teori-teori pendukung dan penelitian terkait yang relevan dengan topik penelitian, meliputi media sosial Tiktok,

Natural Language Processing, pemodelan topik, LDA, representasi teks BoW, serta metrik evaluasi *topic coherence*.

BAB III METODE PENELITIAN

Bab ini menjelaskan metode penelitian yang digunakan, meliputi sumber dan pengumpulan data, tahapan *preprocessing* data teks, penerapan metode LDA berbasis BoW dan TF-IDF, serta metode evaluasi model menggunakan *topic coherence*.

BAB IV HASIL DAN PEMBAHASAN

Bab ini menjelaskan hasil penelitian yang diperoleh dari setiap tahapan yang telah dilakukan. Pembahasan meliputi hasil *preprocessing* data, pembentukan *N-Gram*, representasi teks menggunakan BoW dan TF-IDF, hasil pemodelan topik menggunakan LDA, hasil evaluasi *topic coherence*, dan perbandingan antara model LDA dengan representasi BoW dan TF-IDF.

BAB V Bab ini berisi kesimpulan dan saran dari penelitian yang telah dilakukan. Kesimpulan disusun berdasarkan hasil pengujian dan analisis yang telah dibahas pada bab sebelumnya.