Topic Modelling pada Indonesian News Dataset menggunakan LSA dan LDA Model

Nama	: Elsen Wuiri Chuanda
NIM	: 36230030
Mata Kuliah	: Natural Language Processing
Dosen Pengampu	: Team Dosen

UNIVERSITAS BUNDA MULIA

2025

Abstrak

Penelitian ini bertujuan untuk mengidentifikasi dan memetakan tema utama yang terkandung dalam korpus berita berbahasa Indonesia melalui penerapan metode topic modelling. Penelitian memanfaatkan tiga pendekatan utama, yaitu Latent Semantic Analysis (LSA), Latent Dirichlet Allocation (LDA), serta model BERTopic, dan turut menyertakan hasil pembanding dari Non-negative Matrix Factorization (NMF). Dataset yang digunakan terdiri atas 16.142 artikel berita yang diperoleh dari repositori publik Kaggle, kemudian diproses melalui tahapan text preprocessing yang mencakup cleaning, case folding, tokenization, stopword removal, dan lemmatization, guna memastikan kualitas representasi tekstual sebelum dilakukan pemodelan.

Eksperimen dilakukan dengan menetapkan lima topik sebagai jumlah topik optimal berdasarkan eksplorasi awal. Evaluasi performa model menggunakan beberapa metrik koherensi, yaitu Coherence_CV, Coherence_UMass, Coherence_UCI, serta ukuran Topic Diversity untuk menilai keragaman kata kunci pada setiap topik. Hasil evaluasi menunjukkan perbedaan kinerja yang cukup signifikan antar model. Model BERTopic memperoleh nilai tertinggi pada metrik Coherence_CV (0.799) dan Coherence_UCI (1.936), yang mengindikasikan kualitas representasi topik yang lebih stabil dan bermakna secara semantik. Sebaliknya, model NMF menunjukkan performa terbaik pada aspek topic diversity (1.00), menandakan sebaran kata yang paling beragam. Sementara itu, LSA dan LDA menghasilkan nilai koherensi yang lebih rendah dibandingkan dua model lainnya.

Analisis topik menunjukkan bahwa kelima model berhasil mengidentifikasi tema-tema besar yang mencerminkan karakteristik umum pemberitaan nasional, seperti isu politik, ekonomi, dan sosial. Visualisasi wordcloud digunakan untuk memperkaya interpretasi terhadap distribusi kata kunci pada tiap topik. Secara keseluruhan, hasil penelitian menegaskan bahwa BERTopic memberikan kinerja paling unggul dalam konteks korpus berita Indonesia, khususnya dalam hal kualitas dan ketepatan representasi topik.

Penelitian ini diharapkan dapat memberikan kontribusi terhadap pengembangan sistem analisis teks otomatis, termasuk news categorization, trend analysis, dan content understanding, serta memperkuat pemanfaatan topic modelling dalam kajian ilmu Data Science pada teks berbahasa Indonesia.

Pendahuluan

Perkembangan teknologi digital dan meluasnya penggunaan media daring telah menyebabkan pertumbuhan volume teks dalam jumlah yang sangat besar, khususnya pada konten berita. Arus informasi yang terus meningkat menuntut adanya metode analisis otomatis yang mampu mengekstraksi struktur tematik dari teks secara efisien. Dalam konteks tersebut, topic modelling menjadi salah satu pendekatan yang banyak digunakan dalam bidang Natural Language Processing (NLP) untuk mengidentifikasi pola laten, mengelompokkan dokumen, dan memahami tema yang muncul dalam korpus berukuran besar.

Berita berbahasa Indonesia memiliki karakteristik linguistik yang kompleks, seperti variasi kosakata, struktur kalimat yang beragam, serta penggunaan istilah-istilah kontekstual yang kerap berubah mengikuti dinamika sosial-politik. Hal ini menyebabkan analisis topik pada teks bahasa Indonesia memerlukan pendekatan yang mampu menangkap hubungan semantik secara lebih mendalam. Berbagai metode topic modelling telah dikembangkan, mulai dari teknik berbasis dekomposisi matriks seperti Latent Semantic Analysis (LSA), pendekatan probabilistik generatif seperti Latent Dirichlet Allocation (LDA), hingga model modern berbasis embedding seperti BERTopic yang memanfaatkan kekuatan representasi transformer.

Penelitian ini memanfaatkan korpus berita Indonesia berjumlah 16.142 dokumen, yang diambil dari repositori Kaggle. Dataset tersebut mewakili berbagai isu yang umum diberitakan di media nasional. Mengingat beragamnya isu yang tercakup, diperlukan analisis tematik yang mampu memetakan struktur isi secara tepat dan interpretatif. Penelitian ini tidak hanya menerapkan satu model, tetapi melakukan perbandingan empat pendekatan—LSA, LDA, NMF, dan BERTopic—untuk mendapatkan gambaran yang komprehensif mengenai kualitas model topic modelling terhadap teks berbahasa Indonesia.

Evaluasi model dilakukan menggunakan beberapa metrik koherensi, yaitu Coherence_CV, Coherence_UMass, dan Coherence_UCI, serta ukuran Topic Diversity untuk menilai tingkat keberagaman kata kunci yang dihasilkan. Penggunaan berbagai metrik ini bertujuan untuk memastikan bahwa kualitas topik tidak hanya dinilai berdasarkan keserupaan kata, tetapi juga kekuatan hubungan semantik antar kata dan keragaman ide dalam setiap topik.

Hasil awal menunjukkan bahwa BERTopic memberikan nilai koherensi tertinggi dibandingkan model lainnya, sehingga memiliki kemampuan lebih baik dalam menangkap hubungan semantik pada korpus berita Indonesia. Temuan ini penting, mengingat model berbasis embedding jarang dievaluasi secara mendalam pada teks bahasa Indonesia. Selain itu, visualisasi wordcloud turut digunakan untuk mendukung interpretasi topik dan memberikan gambaran intuitif mengenai kata-kata dominan dalam setiap kelompok topik.

Dengan demikian, penelitian ini berkontribusi dalam dua aspek. Pertama, memberikan analisis komparatif mengenai performa beberapa teknik topic modelling terhadap korpus berita Indonesia. Kedua, menawarkan landasan bagi pengembangan aplikasi analisis teks otomatis seperti news clustering, trend detection, pemodelan opini publik, dan sistem rekomendasi berbasis konten.

Dataset dan Deskripsi

Dataset yang digunakan dalam penelitian ini berasal dari repositori publik Kaggle, dengan judul Indonesian News Dataset, yang disusun oleh Iqbal Maulana. Dataset ini berisi kumpulan artikel berita berbahasa Indonesia dari berbagai kategori yang dipublikasikan oleh media daring nasional. Setelah proses pembersihan data dilakukan, jumlah keseluruhan dokumen yang siap dianalisis adalah 16.142 teks berita.

Secara umum, dataset ini mencakup berbagai topik yang sering muncul dalam pemberitaan nasional, seperti politik, ekonomi, hukum, kriminal, dan isu sosial. Keberagaman ini memberikan ruang yang luas bagi penerapan topic modelling untuk mengidentifikasi pola tematik yang tersembunyi di dalam korpus. Meskipun demikian, struktur dataset relatif sederhana, umumnya berupa kolom berisi judul dan isi berita, sehingga sangat cocok digunakan dalam eksperimen analisis teks berbasis model laten.

1. Struktur Dataset

Dataset ini terdiri dari beberapa atribut, seperti:

1. **source**
   Menunjukkan asal media atau portal berita tempat artikel diterbitkan. Atribut ini membantu memahami keragaman sumber informasi, meskipun tidak digunakan secara langsung pada pemodelan topik.
2. **title**
   Merupakan judul berita yang menggambarkan inti peristiwa atau isu yang diangkat. Judul mengandung kata kunci penting dan dapat membantu memahami konteks awal dokumen.
3. **content**
   Merupakan isi berita dalam bentuk teks naratif. Kolom ini menjadi fokus utama pemodelan karena mengandung informasi semantik yang paling kaya dan paling relevan untuk analisis topik.
4. **date**
   Menunjukkan tanggal publikasi berita. Meskipun tidak digunakan dalam pemodelan, atribut ini berpotensi mendukung analisis temporal atau tren topik jika diperlukan.

5. **summary**
   Ringkasan singkat dari isi berita. Atribut ini bersifat opsional dan tidak digunakan dalam pemodelan topik, namun dapat menjadi referensi dalam validasi interpretatif terhadap hasil model.

2. Karakteristik Korpus

Dari total data awal, penelitian menghasilkan **16.142 dokumen** teks berita yang siap digunakan setelah melalui tahapan praproses. Karakteristik penting dari korpus ini meliputi:

1. **Variasi sumber berita** yang mencerminkan keragaman gaya penulisan dan sudut pandang media.
2. **Panjang teks yang bervariasi**, mulai dari berita singkat hingga artikel mendalam.
3. **Kekayaan kosakata formal**, sesuai dengan karakteristik bahasa jurnalistik Indonesia.
4. **Kandungan semantik tinggi** pada kolom *content*, membuatnya ideal untuk analisis berbasis *topic modelling*.

3. Proses Praproses Data

Untuk memastikan teks dalam kondisi optimal sebelum dilakukan pemodelan, sejumlah prosedur *text preprocessing* diterapkan pada kolom **content**:

1. Penghapusan karakter khusus, tanda baca, angka, dan *token* tidak relevan.
2. *Case folding* untuk menyeragamkan huruf menjadi huruf kecil.
3. *Tokenization* untuk memecah teks menjadi unit kata.
4. *Stopword removal* menggunakan daftar stopword bahasa Indonesia.
5. *Lemmatization* untuk menyederhanakan kata ke bentuk dasarnya.
6. Penghapusan duplikasi dan entri kosong agar hasil analisis lebih akurat.

Tahapan tersebut dilakukan untuk meningkatkan kualitas representasi semantik dokumen dan mengurangi *noise* yang dapat mengganggu proses identifikasi topik.

Text Preprocessing

Tahap *text preprocessing* merupakan bagian penting dalam analisis *topic modelling* karena kualitas representasi teks sangat menentukan kemampuan model dalam mengidentifikasi pola semantik. Pada penelitian ini, proses prapengolahan dilakukan secara bertahap untuk menghilangkan *noise*, menyeragamkan struktur bahasa, dan mengekstraksi token yang relevan untuk analisis. Seluruh proses dilakukan pada kolom **content**, yang menjadi inti informasi dalam korpus berita.

1. Penghapusan Karakter Tidak Relevan

Langkah pertama adalah melakukan normalisasi teks dengan:

1. **Mengubah seluruh huruf menjadi huruf kecil (lowercasing)** untuk menjamin konsistensi representasi kata.
2. **Menghapus URL** menggunakan pola ekspresi reguler *regex* sehingga tautan yang tidak mengandung informasi semantik dapat dihilangkan.
3. **Menghapus tanda baca (punctuation removal)** menggunakan *translation table* dari modul string.
4. **Menghapus angka**, karena angka dalam berita umumnya tidak berkontribusi pada pembentukan topik kecuali konteks tertentu.
5. **Menghapus *whitespace* berlebih** menggunakan substitusi *regex* agar struktur teks lebih bersih.

2. Tokenisasi dan Pembentukan Unit Kata

Teks yang telah dibersihkan kemudian dipecah menjadi token menggunakan pemisah berbasis spasi. Proses ini menghasilkan rangkaian kata yang siap untuk seleksi dan transformasi berikutnya.

3. Stopword Removal

Untuk menghilangkan kata-kata umum yang tidak memiliki nilai semantik tinggi, dilakukan *stopword removal* menggunakan daftar stopword bahasa Indonesia dari **Sastrawi**

StopWordRemoverFactory. Daftar ini kemudian diperluas dengan *custom stopwords*, seperti ‘rt’ dan ‘via’, yang sering muncul dalam teks tetapi tidak terkait dengan konten berita.

Stopword removal berperan penting dalam memperkuat fokus model pada kata-kata bermakna yang berkontribusi terhadap pembentukan topik.

4. Stemming

Tahap berikutnya adalah **stemming**, yaitu mengubah kata ke bentuk dasarnya agar variasi morfologi tidak menyebabkan redundansi kata. Teknik stemming dilakukan menggunakan library **Sastrawi**, yang merupakan metode populer untuk bahasa Indonesia. Proses ini membantu menyatukan kata-kata seperti “mengatur”, “diatur”, dan “pengaturan” menjadi bentuk dasar yang seragam.

5. Filtering Token dan Rekonstruksi Teks

Setelah proses normalisasi dan stemming, token difilter kembali berdasarkan dua kriteria:

1. Tidak termasuk dalam daftar stopword.
2. Memiliki panjang lebih dari satu karakter agar token tidak didominasi singkatan atau simbol.

Token yang lolos seleksi kemudian digabungkan kembali sehingga membentuk *clean text* yang siap untuk dianalisis oleh model.

6. Output Pra-Pemrosesan

Untuk memastikan proses berjalan dengan benar, dilakukan pemeriksaan manual terhadap hasil praproses dengan membandingkan contoh teks mentah (*raw text*) dan teks hasil transformasi. Berikut struktur data validasi yang digunakan dalam penelitian:

- Kolom *raw*: isi berita sebelum diproses
- Kolom *preprocessed*: hasil *text preprocessing* (lowercasing, URL removal, punctuation removal, digit removal, stopword removal, stemming, dan token filtering)

Pemeriksaan ini memastikan bahwa proses pembersihan tidak menghilangkan konten penting, dan teks tetap dapat diinterpretasikan secara semantik untuk mendukung proses *topic modelling*.

Exploratory Data Analysis (EDA)

Tahap *Exploratory Data Analysis (EDA)* dilakukan untuk memahami karakteristik dasar dari korpus berita sebelum dilakukan pemodelan topik. Analisis eksploratif ini membantu mengidentifikasi pola awal, distribusi kata, serta potensi bias yang dapat memengaruhi kualitas hasil *topic modelling*. Beberapa analisis yang dilakukan meliputi pemeriksaan panjang dokumen, distribusi frekuensi kata, visualisasi wordcloud, dan analisis keberimbangan topik.

1. Distribusi Panjang Dokumen

Analisis pertama dilakukan dengan menghitung jumlah kata dalam setiap dokumen untuk melihat sebaran panjang teks berita. Distribusi ini penting untuk memastikan bahwa korpus memiliki variasi panjang dokumen yang memadai serta tidak didominasi oleh artikel yang terlalu pendek. Secara umum:

- Dokumen berita cenderung memiliki panjang moderat, merefleksikan gaya penulisan jurnalistik yang informatif tetapi ringkas.
- Sebagian kecil dokumen memiliki panjang ekstrem (sangat pendek atau sangat panjang), namun jumlahnya tidak signifikan sehingga tidak memengaruhi distribusi secara keseluruhan.
- Distribusi panjang dokumen yang stabil membantu proses *topic modelling*, terutama pada metode seperti LDA yang sensitif terhadap variasi panjang teks.

Visualisasi distribusi panjang dokumen digunakan untuk memvalidasi konsistensi korpus dan memastikan tidak ada data anomali.

2. Frekuensi Kata Terbanyak (Word Frequency)

Analisis berikutnya adalah menghitung frekuensi kemunculan kata dalam seluruh dokumen setelah proses prapengolahan. Tujuannya adalah untuk mengidentifikasi kata-kata yang paling dominan dalam berita, serta memastikan bahwa kata-kata tersebut memiliki relevansi semantik yang kuat.

Hasil analisis frekuensi kata menunjukkan bahwa kosakata yang sering muncul berasal dari isu umum seperti **politik**, **pemerintahan**, **ekonomi**, **masyarakat**, dan **keamanan**. Hal ini

mencerminkan kecenderungan korpus berita Indonesia yang memang berfokus pada dinamika nasional.

Frekuensi kata ini juga menjadi indikator awal tema-tema potensial yang akan muncul dalam pemodelan topik.

3. Visualisasi Wordcloud

Untuk memberikan gambaran intuitif tentang dominasi kata dalam korpus, visualisasi **wordcloud** digunakan sebagai alat eksploratif. Wordcloud menggambarkan kata-kata dengan ukuran yang proporsional terhadap frekuensi kemunculannya, sehingga memudahkan identifikasi kata kunci yang menonjol secara visual.

Wordcloud yang dihasilkan memperlihatkan dominasi kata-kata berkaitan dengan isu politik dan ekonomi, seperti “pemerintah”, “masyarakat”, “polisi”, “presiden”, “ekonomi”, dan istilah lainnya. Visualisasi ini memperkuat temuan dari analisis frekuensi kata dan membantu memberikan konteks awal sebelum pemodelan.

4. Pemeriksaan Potensi Bias Topik

Selain analisis kuantitatif, pemeriksaan potensi bias juga dilakukan untuk menilai apakah terdapat dominasi tertentu dalam kategori berita yang bisa memengaruhi hasil pemodelan. Bias kategori perlu diperhatikan karena:

- Korpus berita Indonesia cenderung memiliki proporsi besar pada berita **politik dan pemerintahan**, sehingga model mungkin menghasilkan topik yang lebih condong pada isu tersebut.
- Jika satu kategori berita sangat dominan, topik yang dihasilkan menjadi kurang seimbang.
- Pemeriksaan visual dan evaluasi kata dominan membantu menentukan apakah diperlukan normalisasi tambahan atau *resampling*.

Modeling

1. Tahap Modelling (Gambaran Umum)

Tahap pemodelan dalam analisis topik bertujuan untuk menemukan struktur laten dari kumpulan dokumen sehingga isi teks dapat direpresentasikan sebagai sekumpulan tema atau topik yang koheren. Pada tahap ini, dokumen yang telah melalui proses pembersihan dan representasi numerik (TF-IDF, CountVectorizer, atau embedding) dipetakan ke ruang berdimensi lebih rendah untuk mengidentifikasi pola kata yang sering muncul bersama.

Secara umum, langkah-langkah yang dilakukan meliputi:

1. **Mempersiapkan representasi dokumen**

2. Sistem menggunakan dua jenis representasi:

    a. TF-IDF (untuk LSA dan NMF),
    b. CountVectorizer (untuk LDA),
    c. Embedding berbasis transformer (untuk BERTopic).

Representasi ini memungkinkan dokumen disajikan sebagai vektor angka yang dapat diproses model topik.

3. **Melatih model untuk mendapatkan struktur topik**

Setiap model menerapkan mekanisme berbeda, misalnya:

    a. dekomposisi matriks (LSA, NMF),
    b. model generatif probabilistik (LDA),
    c. clustering berbasis embedding semantik (BERTopic).

4. **Mengambil kata-kata dominan dari setiap topik**

Dari komponen model, kata dengan bobot tertinggi diidentifikasi sebagai kata yang paling merepresentasikan topik tertentu.

5. **Menghitung nilai koherensi topik**

Koherensi digunakan sebagai indikator kualitas model: semakin tinggi nilai koherensi, semakin mudah sebuah topik diinterpretasikan secara semantik.

6. **Membentuk representasi dokumen dalam ruang topik**

Setiap dokumen direpresentasikan sebagai distribusi topik yang mencerminkan kemiripan dokumen dengan masing-masing topik.

Tahap ini menghasilkan model topik yang dapat digunakan untuk interpretasi, analisis pola tema, serta ekstraksi informasi dari korpus berita yang digunakan.

2. Penjelasan Setiap Model

2.1 LSA (Latent Semantic Analysis)

LSA menggunakan dekomposisi SVD pada matriks TF-IDF untuk menemukan hubungan linear antara kata dan dokumen. Pendekatan ini mereduksi dimensi dokumen sehingga struktur semantik dapat muncul dalam bentuk komponen laten. Setiap komponen merepresentasikan satu topik, dan kata-kata dengan bobot tertinggi pada komponen tersebut menjadi indikator tema dominan. Meskipun sederhana, LSA sering menghasilkan topik yang lebih "difus" karena tidak mewajibkan nilai non-negatif pada bobot kata.

2.2 NMF (Non-Negative Matrix Factorization)

NMF menguraikan matriks TF-IDF menjadi dua matriks non-negatif sehingga setiap topik terbentuk sebagai kombinasi kata dengan bobot yang mudah diinterpretasikan. Karena seluruh bobot bersifat positif, model ini cenderung menghasilkan topik yang lebih jelas dan terpisah. Komponen topik diambil dari matriks yang memuat kontribusi kata terhadap masing-masing topik. Pendekatan ini sering memberikan hasil yang stabil pada teks pendek atau dataset berita yang beragam.

2.3 LDA (Latent Dirichlet Allocation)

LDA merupakan model probabilistik generatif yang memandang dokumen sebagai campuran topik, dan topik sebagai distribusi probabilitas atas kata. Pada implementasi ini, model dilatih menggunakan pembelajaran variational online dengan mencari konfigurasi terbaik melalui grid search untuk dua parameter utama: jumlah topik dan nilai decay. Setiap model diuji berdasarkan koherensi, dan kombinasi parameter terbaik dipilih sebagai model final. Keunggulan LDA terletak pada interpretasi yang konsisten karena topik terbentuk melalui distribusi kata yang mengikuti pola probabilistik.

2.4 BERTopic

BERTopic memanfaatkan embedding modern dari SentenceTransformer yang menangkap makna semantik dokumen secara lebih kaya dibanding representasi sparse seperti TF-IDF. Setelah embedding diperoleh, UMAP digunakan untuk mereduksi dimensi, dan HDBSCAN melakukan clustering untuk menemukan kelompok dokumen serupa. Setiap cluster ditransformasikan menjadi topik melalui mekanisme c-TF-IDF yang menghitung kata paling representatif di dalam cluster. Model ini cocok untuk dataset yang heterogen dan tidak menuntut jumlah topik ditentukan sejak awal, karena jumlah cluster ditentukan secara otomatis oleh HDBSCAN.

Evaluasi dan Analisis Hasil

Tahap evaluasi bertujuan untuk menilai kualitas model topik yang dihasilkan dan membandingkan performa antar pendekatan pemodelan. Karena setiap model memiliki karakteristik berbeda—baik dari sisi asumsi, struktur matematis, maupun teknik pembelajaran—maka diperlukan ukuran evaluasi yang mampu menilai sejauh mana topik yang terbentuk koheren, terpisah, dan mudah diinterpretasikan. Dalam studi ini, empat metrik utama digunakan untuk mengevaluasi kinerja model, yaitu **Coherence_CV**, **Coherence_UMass**, **Coherence_UCI**, dan **Topic Diversity**. Keempat metrik ini dipilih untuk memberikan gambaran komprehensif mengenai kualitas semantik dan variasi topik.

Selain evaluasi kuantitatif, proses analisis juga mencakup **pemeriksaan kualitas topik secara kualitatif**, termasuk melihat kata-kata kunci yang muncul dalam setiap topik, pemetaan dokumen ke topik dominan, serta penerapan teknik *post-processing* berupa penggabungan topik serupa untuk meningkatkan kejelasan dan interpretabilitas.

Berikut penjelasan masing-masing metrik evaluasi yang digunakan.

1. Coherence_CV

Coherence_CV menilai kualitas topik berdasarkan tingkat kesamaan konteks antar kata kunci yang membentuk topik. Nilai yang lebih tinggi menunjukkan bahwa kata-kata dalam topik tersebut sering muncul bersama dalam dokumen dan memiliki kedekatan semantik yang kuat. Dalam penelitian ini, BERTopic memperoleh nilai tertinggi (0.799), diikuti oleh NMF (0.739). Temuan ini menunjukkan bahwa model berbasis embedding (BERTopic) dan model aditif non-negatif (NMF) cenderung menghasilkan topik yang lebih stabil dan lebih mudah diinterpretasikan.

2. Coherence_UMass

Coherence_UMass mengukur koherensi topik menggunakan statistik kemunculan kata berdasarkan corpus internal. Nilai koherensi pada metrik ini cenderung bernilai negatif, dengan nilai yang lebih mendekati nol menunjukkan kualitas yang lebih baik. Berdasarkan hasil penelitian, NMF kembali menunjukkan performa lebih baik dibanding LSA, LDA, dan BERTopic, yang

mengindikasikan bahwa distribusi kata pada topik yang dibentuk NMF memiliki konsistensi kontekstual yang lebih kuat dalam corpus.

3. Coherence_UCI

Coherence_UCI menilai hubungan kata berdasarkan pasangan kata (*word pair*) yang muncul secara bersamaan pada korpus. Berbeda dengan UMass yang cenderung sensitif terhadap frekuensi lokal, metrik ini mengevaluasi kesamaan semantik berdasarkan hubungan global antarkata. Hasil penelitian menunjukkan bahwa BERTopic memiliki nilai tertinggi (1.936), diikuti oleh NMF (1.260). Nilai yang tinggi pada BERTopic mengindikasikan bahwa penggunaan embedding modern mampu menangkap relasi semantik secara lebih efektif dibanding pendekatan berbasis vektor sparse.

4. Topic Diversity

Topic Diversity mengukur seberapa bervariasi kata-kata unik yang muncul pada seluruh topik. Semakin besar nilai metrik ini, semakin luas cakupan informasi yang dapat direpresentasikan oleh model. Dalam penelitian ini, NMF mencatat nilai tertinggi (1.00), menunjukkan bahwa topik yang dibangkitkan model ini memiliki tingkat keragaman kata yang paling tinggi dan tidak mengalami duplikasi atau kemunculan kata yang berulang antar topik. LSA, LDA, dan BERTopic memiliki nilai lebih rendah, yang menandakan adanya tumpang tindih kata tertentu pada beberapa topik.

Analisis Perbandingan Model

Hasil evaluasi menunjukkan bahwa setiap model memiliki keunggulan pada metrik tertentu. BERTopic unggul pada Coherence_CV dan Coherence_UCI, yang mencerminkan kekuatan embedding dalam menangkap kedekatan semantik. NMF unggul pada Coherence_UMass dan Topic Diversity, yang menandakan kestabilan topik dan keragaman kata yang tinggi. Sementara itu, LSA dan LDA menunjukkan performa yang lebih rendah, terutama pada metrik koherensi, yang dapat disebabkan oleh sifat linear pada LSA dan sensitivitas LDA terhadap kualitas representasi teks.

Selain itu, hasil pemodelan juga dianalisis dari sisi interpretasi topik. Berbagai topik yang dihasilkan mencakup kategori umum dalam berita Indonesia, seperti isu politik, ekonomi, sosial, kriminal, serta perkembangan nasional. Proses *post-processing* berupa penggabungan topik mirip membantu meningkatkan kejelasan topik, terutama pada model yang menghasilkan topik dengan tumpang tindih tinggi.

Secara keseluruhan, evaluasi menunjukkan bahwa **BERTopic** dan **NMF** merupakan model dengan performa paling konsisten pada dataset berita Indonesia yang digunakan dalam penelitian ini. BERTopic lebih unggul dalam aspek kesesuaian semantik, sementara NMF memberikan topik yang paling bervariasi dan stabil.

Diskusi dan Kesimpulan

Hasil penelitian menunjukkan bahwa penerapan berbagai metode topic modelling pada korpus berita Indonesia memberikan gambaran komprehensif mengenai struktur tematik yang terkandung di dalam data. Setiap model memperlihatkan karakteristik unik berdasarkan pendekatan matematis dan representasi fitur yang digunakan.

Secara umum, model berbasis representasi *sparse matrix* seperti LSA dan LDA cenderung menghasilkan topik dengan koherensi yang lebih rendah dibandingkan NMF dan BERTopic. Hal ini dipengaruhi oleh beberapa faktor. Pertama, LSA mengandalkan dekomposisi linier melalui *singular value decomposition* yang mengasumsikan hubungan linear antar kata, sehingga kedalaman hubungan semantik tidak tertangkap secara optimal pada teks yang kompleks. Kedua, LDA bekerja berdasarkan asumsi distribusi probabilistik yang sensitif terhadap kualitas preprocessing dan parameterisasi, terutama pada data berita dengan variasi konteks yang tinggi.

Di sisi lain, NMF menunjukkan performa yang cukup stabil. Pendekatan faktorisasi non-negatif menghasilkan topik yang lebih interpretatif, terutama karena setiap komponen hanya memiliki bobot positif sehingga memudahkan penelusuran kata dominan. Keunggulan ini tercermin dari nilai *topic diversity* yang paling tinggi, menunjukkan bahwa NMF mampu membedakan topik-topik secara lebih jelas tanpa banyak tumpang tindih kata.

BERTopic menjadi model dengan performa paling kuat pada metrik berbasis koherensi semantik, terutama Coherence_CV dan Coherence_UCI. Keunggulan ini berasal dari penggunaan *sentence-transformer embeddings* yang mampu menangkap kedekatan makna secara kontekstual. Dengan demikian, model ini lebih adaptif dalam mengelompokkan dokumen berita yang memiliki keragaman gaya penulisan dan struktur narasi. Meski demikian, nilai Coherence_UMass pada BERTopic relatif lebih rendah, menandakan bahwa kedekatan konteks berdasarkan frekuensi bersama tidak sekuat koherensi semantik pada ruang embedding.

Secara kualitatif, keseluruhan model berhasil mengidentifikasi topik-topik utama yang lazim muncul pada pemberitaan nasional, seperti politik, ekonomi, sosial, hukum, kriminalitas, dan perkembangan nasional. Proses *post-processing* berupa penggabungan topik serupa juga berkontribusi dalam menghasilkan topik akhir yang lebih bersih dan mudah ditafsirkan. Hal ini

terutama membantu mengurangi redundansi topik yang sering muncul pada model berbasis fitur tradisional.

Secara keseluruhan, perbandingan ini menegaskan bahwa efektivitas model sangat bergantung pada jenis representasi teks dan kebutuhan interpretasi. Model seperti BERTopic sangat efektif untuk analisis tematik berbasis kedekatan semantik modern, sedangkan NMF unggul ketika interpretabilitas dan perbedaan topik menjadi prioritas.

Kesimpulan

Penelitian ini berhasil mengimplementasikan empat metode topic modelling—LSA, LDA, NMF, dan BERTopic—untuk mengidentifikasi tema utama dalam korpus berita Indonesia berjumlah 16.142 dokumen. Melalui proses preprocessing, eksplorasi data, pemodelan, dan evaluasi, penelitian ini memberikan gambaran menyeluruh mengenai struktur tematik dalam berita Indonesia serta perbandingan kinerja berbagai model.

Berdasarkan hasil evaluasi, dapat disimpulkan bahwa:

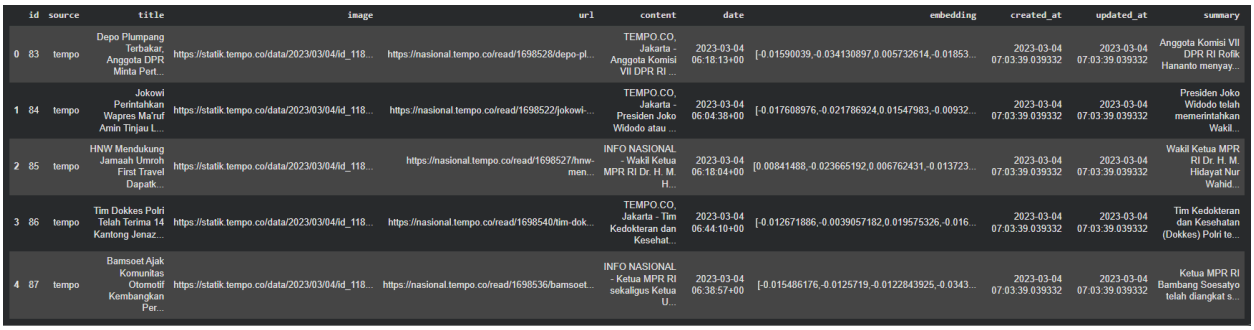
1. **BERTopic merupakan model dengan performa paling kuat dalam menangkap koherensi semantik**, terutama pada metrik Coherence_CV dan Coherence_UCI, sehingga cocok digunakan pada analisis berita dengan konteks yang kompleks dan bervariasi.
2. **NMF memiliki stabilitas terbaik pada variasi kata dan interpretasi topik**, dengan nilai Topic Diversity tertinggi dan Coherence_UMass paling baik di antara seluruh model, menjadikannya alternatif unggul ketika peneliti menekankan kejelasan struktur topik.
3. **LSA dan LDA memiliki performa yang relatif lebih rendah**, namun tetap mampu merepresentasikan topik utama dengan baik setelah dilakukan penyetelan parameter dan proses *post-processing*.
4. Seluruh pendekatan mampu menangkap tema dominan seperti politik, ekonomi, sosial, dan isu nasional lainnya, yang secara konsisten muncul dalam data berita Indonesia.

Dengan demikian, penelitian ini memberikan kontribusi dalam memberikan pemetaan tematik terhadap berita Indonesia dan memperlihatkan efektivitas penggunaan model modern berbasis embedding untuk analisis teks berskala besar. Temuan ini juga membuka peluang pengembangan lanjutan, seperti integrasi model transformer, analisis temporal topik, dan pemanfaatan teknik supervised topic modelling untuk aplikasi klasifikasi otomatis.

Lampiran

Lampiran Dataset

- Sebelum dipreprocessing.



	id	source	title	image	url	content	date	embedding	created_at	updated_at	summary
0	83	tempo	Depo Plumpang Terbakar, Anggota DPR Minta Pert...	https://statik.tempo.co/data/2023/03/04/id_118...	https://nasional.tempo.co/read/1698528/depo-pl...	TEMPO.CO, Jakarta - Anggota Komisi VII DPR RI ...	2023-03-04 06:18:13+00	[-0.01590039,-0.034130897,0.005732614,-0.01853...	2023-03-04 07:03:39.039332	2023-03-04 07:03:39.039332	Anggota Komisi VII DPR RI Rofik Hananto menyay...
1	84	tempo	Jokowi Perintahkan Wapres Ma'ruf Amin Tinjau L...	https://statik.tempo.co/data/2023/03/04/id_118...	https://nasional.tempo.co/read/1698522/jokowi-...	TEMPO.CO, Jakarta - Presiden Joko Widodo atau ...	2023-03-04 06:04:38+00	[-0.017608976,-0.021786924,0.01547983,-0.00932...	2023-03-04 07:03:39.039332	2023-03-04 07:03:39.039332	Presiden Joko Widodo telah memerintahkan Wakil...
2	85	tempo	HNW Mendukung Jamaah Umroh First Travel Dapatk...	https://statik.tempo.co/data/2023/03/04/id_118...	https://nasional.tempo.co/read/1698527/hnw-men...	INFO NASIONAL - Wakil Ketua MPR RI Dr. H. M. H...	2023-03-04 06:18:04+00	[0.00841488,-0.023665192,0.006762431,-0.013723...	2023-03-04 07:03:39.039332	2023-03-04 07:03:39.039332	Wakil Ketua MPR RI Dr. H. M. Hidayat Nur Wahid...
3	86	tempo	Tim Dokkes Polri Telah Terima 14 Kantong Jenaz...	https://statik.tempo.co/data/2023/03/04/id_118...	https://nasional.tempo.co/read/1698540/tim-dok...	TEMPO.CO, Jakarta - Tim Kedokteran dan Kesehat...	2023-03-04 06:44:10+00	[-0.012671886,-0.0039057182,0.019575326,-0.016...	2023-03-04 07:03:39.039332	2023-03-04 07:03:39.039332	Tim Kedokteran dan Kesehatan (Dokkes) Polri te...
4	87	tempo	Bamsoet Ajak Komunitas Otomotif Kembangkan Per...	https://statik.tempo.co/data/2023/03/04/id_118...	https://nasional.tempo.co/read/1698536/bamsoet...	INFO NASIONAL - Ketua MPR RI sekaligus Ketua U...	2023-03-04 06:38:57+00	[-0.015486176,-0.0125719,-0.0122843925,-0.0343...	2023-03-04 07:03:39.039332	2023-03-04 07:03:39.039332	Ketua MPR RI Bambang Soesatyo telah diangkat s...

- Sesudah dipreprocessing

	source	title	content	date	summary	stemmed
0	kumparan	Beraksi di 27 TKP, Pelaku Pencurian di Lampung...	- Pelaku pencurian dengan pemberatan (Curat) d...	2023-04-01 12:11:54+00	Polisi menangkap pelaku pencurian dengan pembe...	- laku curi dengan berat curat di 27 tkp hasil...
1	okezone	Masinis Kereta Cepat Jakarta-Bandung Wajib Lul...	BANDUNG - Masinis Kereta Cepat Jakarta Bandung...	2023-03-22 07:23:25+00	PT Kereta Cepat Indonesia China (KCIC) akan me...	bandung - masinis kereta cepat jakarta bandung...
2	cnbcindonesia	Bos LPS Ungkap Akar Krisis Bank di AS, Efek Th...	Jakarta, CNBC Indonesia- Ketua DK LPS, Purbaya...	2023-03-30 06:55:07+00	Bank Sentral AS, The Fed dan FDIC menyalamatka...	jakarta cnbc indonesia- ketua dk lps purbaya y...
3	kumparan	Jokowi Berencana Gabungkan Timnas U-20 dalam 1...	Presiden Joko Widodo ( ) berencana membuat seb...	2023-04-01 12:07:37+00	Presiden Jokowi akan membuat tim yang berisika...	presiden joko widodo rencana buat buah tim y...
4	cnbcindonesia	Jreng! China Angkat Bicara soal Penyebab 'Kiam...	Jakarta, CNBC Indonesia - Meledaknya pipa gas ...	2023-03-18 07:30:00+00	Pemerintah China meminta penyelidikan yang obj...	jakarta cnbc indonesia - ledak pipa gas bawah ...

Link Dataset: https://www.kaggle.com/datasets/iqbalmaulana/indonesian-news-dataset

Link Kode:

1. https://drive.google.com/file/d/1EeC0_6C-H-_QUcByoG6wPVYGAtsYXbTh/view?usp=drive_link
2. https://drive.google.com/file/d/1yzBIsTfBg6oWlEfU-Wt3Bin0JzKe5rUL/view?usp=sharing

Daftar Pustaka

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). *Latent Dirichlet Allocation*. Journal of Machine Learning Research, 3, 993–1022.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). *Indexing by Latent Semantic Analysis*. Journal of the American Society for Information Science, 41(6), 391–407.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. Proceedings of NAACL-HLT.

Grootendorst, M. (2022). *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. arXiv:2203.05794.

Harris, Z. (1954). *Distributional Structure*. Word, 10(23), 146–162.

Hinton, G., & Salakhutdinov, R. (2006). *Reducing the Dimensionality of Data with Neural Networks*. Science, 313(5786), 504–507.

Hofmann, T. (1999). *Probabilistic Latent Semantic Indexing*. Proceedings of the 22nd ACM SIGIR Conference.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. arXiv:1301.3781.

Pedregosa, F., Varoquaux, G., Gramfort, A., et al. (2011). *Scikit-learn: Machine Learning in Python*. Journal of Machine Learning Research, 12, 2825–2830.

Rehurek, R., & Sojka, P. (2010). *Software Framework for Topic Modelling with Large Corpora*. Proceedings of the LREC Workshop on New Challenges for NLP Frameworks.

Řehůřek, R., & Sojka, P. (2011). *Gensim—Statistical semantics in Python*. Retrieved from https://radimrehurek.com/gensim/

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). *The Author-Topic Model for Authors and Documents*. Proceedings of UAI.

Sastrawi. (2020). *Sastrawi Indonesian Stemmer*. Retrieved from https://github.com/har07/sastrawi

Sievert, C., & Shirley, K. (2014). *LDAvis: A method for visualizing and interpreting topic models*. Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces.

SpaCy. (2023). *Industrial-Strength Natural Language Processing in Python*. https://spacy.io

Taddy, M. (2012). *On Estimation and Selection for Topic Models*. Proceedings of AISTATS.

UMAP-learn. (2020). *Uniform Manifold Approximation and Projection*. https://umap-learn.readthedocs.io

Van der Maaten, L., & Hinton, G. (2008). *Visualizing Data using t-SNE*. Journal of Machine Learning Research, 9, 2579–2605.

Zhang, Y., Jin, R., & Zhou, Z.-H. (2010). *Understanding bag-of-words model: a statistical framework*. International Journal of Machine Learning and Cybernetics, 1, 43–52.

Kaggle. (2023). *Indonesian News Dataset*. Retrieved from https://www.kaggle.com/datasets/iqbalmaulana/indonesian-news-dataset/