Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1367−1379
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.8298

Narasi Presiden Indonesia: Analisis Wacana Politik Menggunakan BERTopic dalam Mengungkap Pola Tematik Pidato Presiden

Uliyatunisa Uliyatunisa [1,*], Tukiyat Tukiyat[2], Arya Adhyaksa Waskita[2], Murni Handayani[2], Rafi Mahmud Zain[2]

[1] Fakultas Ilmu Komputer, Teknik Informatika, Universitas Pamulang, Tangerang Selatan, Indonesia
[2] Program Pascasarjana, Program Studi Teknik Informatika S-2, Universitas Pamulang, Tangerang Selatan, Indonesia
Email: [1,*]uliyatunisa212@gmail.com,[2]dimastuky@gmail.com,[3]aawaskita@unpam.ac.id, [4]murnie_h@yahoo.com, [5]rafizain777@gmail.com
Email Penulis Korespondensi: uliyatunisa212@gmail.com
Submitted: **29/08/2025**; Accepted: **20/09/2025**; Published**: 21/09/2025**

Abstrak−Pidato Presiden Indonesia memiliki peran penting sebagai sarana komunikasi politik, penyampaian kebijakan, serta pembentukan citra kepemimpinan di hadapan publik. Namun, volume pidato yang semakin besar menghadirkan tantangan baru dalam proses analisis secara manual, karena membutuhkan waktu dan rentan subjektivitas peneliti. Penelitian ini menawarkan solusi dengan menggunakan BERTopic, sebuah metode pemodelan topik berbasis transformer yang memanfaatkan representasi semantik dari model embedding modern. Data penelitian berupa transkrip pidato resmi Presiden Joko Widodo yang diperoleh dari portal Sekretariat Kabinet. Untuk meningkatkan kualitas representasi semantik, penelitian ini membandingkan beberapa model embedding berbahasa Indonesia, yaitu DistilBERT, NusaBERT, IndoE5, dan SBERT. Proses analisis dilakukan melalui tahapan preprocessing data, pembentukan embedding, dimensi reduksi, clustering, serta evaluasi model dengan metrik koherensi topik. Tujuan penelitian adalah mengungkap tema yang terkandung dalam pidato Presiden, sekaligus mengevaluasi efektivitas model embedding dalam menghasilkan topik yang lebih koheren. Hasil penelitian menunjukkan dua puluh topik utama yang konsisten muncul, meliputi pembangunan infrastruktur, kebijakan ekonomi, kesehatan dan pandemi, transformasi digital, diplomasi internasional, olahraga, isu nasionalisme, serta pembangunan daerah. Dari sisi kinerja, SBERT memberikan hasil terbaik dengan koherensi senilai UMass = -2.036 dan NPMI = 0.082, yang mengindikasikan hubungan semantik positif. Nilai UMass yang mendekati nol menunjukkan lebih banyak koherensi kata dalam satu topik, sedangkan nilai NPMI di atas nol menunjukkan bahwa keterhubungan antar kata lebih mudah dipahami oleh manusia. Penelitian ini berkontribusi pada pengembangan kajian wacana politik berbasis NLP di Indonesia, memberikan gambaran empiris mengenai pemilihan model embedding yang tepat dalam pemodelan topik dan membuka peluang integrasi metode serupa dalam analisis kebijakan publik.

Kata Kunci: Analisis Wacana; BERTopic; Komunikasi Politik; Pemodelan Topik; Pidato Presiden

Abstract−The speeches of the President of Indonesia play an important role as a means of political communication, policy delivery, and leadership image building in front of the public. However, the increasing volume of speeches presents new challenges in the manual analysis process, as it is time-consuming and prone to researcher subjectivity. This study offers a solution by using BERTopic, a transformer-based topic modelling method that utilises semantic representations from modern embedding models. The research data consists of transcripts of President Joko Widodo's official speeches obtained from the Cabinet Secretariat portal. To improve the quality of semantic representations, this study compares several Indonesian language embedding models, namely DistilBERT, NusaBERT, IndoE5, and SBERT. The analysis process was carried out through the stages of data preprocessing, embedding formation, dimension reduction, clustering, and model evaluation using topic coherence metrics. The objectives of this study were to reveal the themes contained in the President's speeches and to evaluate the effectiveness of embedding models in producing more coherent topics. The results show twenty main themes that consistently appear, including infrastructure development, economic policy, health and the pandemic, digital transformation, international diplomacy, sports, nationalism issues, and regional development. In terms of performance, SBERT provides the best results with a coherence value of UMass = -2.036 and NPMI = 0.082, indicating a positive semantic relationship. A UMass value close to zero indicates greater coherence of words within a topic, while an NPMI value above zero indicates that the connections between words are more easily understood by humans. This research contributes to the development of NLP-based political discourse studies in Indonesia, providing an empirical overview of the selection of appropriate embedding models in topic modelling and opening up opportunities for the integration of similar methods in public policy analysis.

Keywords: Discourse Analysis; BERTopic; Political Communication; Topic Modelling; Presidential Speeches

1. PENDAHULUAN

Akses digital dan kanal komunikasi publik yang lebih luas telah menyebabkan volume teks pidato Presiden Joko Widodo meningkat pesat. Portal berita, media sosial, dan televisi adalah semua platform di mana pernyataan resmi dapat ditemukan. Peningkatan ini menghasilkan korpus teks yang besar, yang membuat analisis manual menjadi sulit[1]. Ini terutama karena pidato mencakup berbagai masalah sosial, politik, dan ekonomi[2]. Pendekatan manual terbatas pada jumlah data kecil dan tidak kuantitatif, sehingga membutuhkan metode komputasional untuk pemetaan topik secara sistematis[2], [3]. Namun, beberapa penelitian sebelumnya berfokus pada strategi retorika, kohesi gramatikal, ideologi tersembunyi, dan simbolisme politik dalam pidato [4][5][6].

Topic modeling adalah teknik yang populer untuk mengekstraksi tema utama dari dokumen tanpa melakukan label manual. Pendekatan metode konvensional seperti *Latent Dirichlet Allocation* (LDA) dan *Non-negative Matrix Factorization* (NMF) memiliki keterbatasan dalam memodelkan teks panjang dan kompleks, karena hanya mengandalkan distribusi kata tanpa mempertimbangkan konteks semantik antar kalimat yang berakibat hasil topik yang tidak konsisten secara semantik[7][8]. Untuk menghasilkan topik yang lebih koheren, BERTopic menggunakan

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1367−1379
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.8298

embedding berbasis *Transformer*, pengurangan dimensi, dan algoritma *clustering*. Dalam analisis pidato politik, hal ini menyebabkan topik yang dihasilkan sering tumpang tindih dan kurang koheren. Studi terbaru menunjukkan bahwa pendekatan berbasis transformer, seperti BERTopic, mampu menghasilkan topik yang lebih akurat karena memanfaatkan representasi semantik dari model *embedding* modern[3].

Keunggulan BERTopic juga terletak pada fleksibilitasnya untuk mengintegrasikan berbagai embedding, metode reduksi dimensi, dan algoritma clustering yang dapat disesuaikan dengan karakteristik data. Dalam konteks ini, pemilihan model embedding menjadi faktor krusial karena menentukan kemampuan representasi semantik kalimat dan kualitas topik yang dihasilkan. Embedding yang baik memungkinkan setiap kalimat pidato ditangkap nuansa maknanya, termasuk implikasi politik, intensi komunikatif, serta penekanan ide tertentu yang sering kali tidak tampak dari sekadar kata kunci. Dengan kata lain, topik yang terbentuk tidak hanya berupa kumpulan kata, tetapi mencerminkan narasi yang lebih utuh dan dapat dipahami secara konseptual[9][10]. Dalam konteks Indonesia, penelitian terkait analisis politik berbasis NLP masih jarang memanfaatkan kekuatan transformer. Sebagian besar penelitian cenderung menggunakan pendekatan klasik seperti LDA atau varian turunannya, yang terbatas dalam menangkap makna semantik pada teks panjang. Padahal, pidato presiden sering kali bersifat multimodal: selain teks, ia disampaikan melalui intonasi, penekanan, serta konteks situasional. Aspek-aspek ini secara tidak langsung juga terekam dalam konstruksi teks pidato. Tanpa pendekatan yang mampu memahami hubungan semantik, analisis bisa kehilangan detail penting yang berhubungan dengan framing politik[11]. Oleh karena itu, penelitian ini memanfaatkan BERTopic untuk memetakan pola tematik dalam pidato Presiden Joko Widodo.

Alasan pemilihan BERTopic dibandingkan pendekatan lain terletak pada kemampuannya menangani volume data besar dengan tetap mempertahankan koherensi semantik topik. Koherensi ini penting karena pidato Presiden bukan hanya berisi kata kunci, melainkan narasi kompleks yang menyampaikan kebijakan, legitimasi politik, hingga citra kepemimpinan. Topik seperti pembangunan infrastruktur, transformasi digital, hingga diplomasi internasional seringkali terjalin dalam satu rangkaian narasi, sehingga diperlukan metode yang mampu membedakan, mengelompokkan, sekaligus mempertahankan makna.Oleh karena itu, penelitian ini memanfaatkan BERTopic untuk memetakan pola tematik dalam pidato Presiden Joko Widodo. Alasan pemilihan BERTopic dibandingkan pendekatan lain terletak pada kemampuannya menangani volume data besar dengan tetap mempertahankan koherensi semantik topik. Hal ini penting karena pidato Presiden bukan hanya berisi kata kunci, melainkan narasi kompleks yang menyampaikan kebijakan, legitimasi politik, hingga citra kepemimpinan. Selain itu, kelebihan BERTopic adalah kemampuannya untuk menghasilkan representasi topik yang dapat divisualisasikan. Visualisasi ini memungkinkan peneliti melihat peta tematik secara menyeluruh, misalnya bagaimana isu ekonomi bersinggungan dengan isu sosial, atau bagaimana tema kesehatan muncul kembali pada periode krisis seperti pandemi COVID-19. Dengan demikian, hasil analisis tidak berhenti pada tingkat deskriptif, tetapi juga bisa memberikan wawasan interpretatif yang mendalam[12]. Dalam penelitian ini, empat model embedding populer untuk Bahasa Indonesia dan multilingual digunakan sebagai pembanding. Pemilihan empat model tersebut didasarkan pada perbedaan kapasitas representasi semantik, ukuran model, serta ketersediaannya dalam domain publik. Variasi embedding ini memungkinkan evaluasi menyeluruh mengenai bagaimana kualitas representasi semantik mempengaruhi koherensi, keberagaman, dan interpretabilitas topik. Jika suatu embedding mampu menangkap makna idiomatik atau konstruksi lokal Bahasa Indonesia, maka hasil topik yang terbentuk cenderung lebih natural dibandingkan embedding yang dilatih hanya pada korpus internasional.

Dengan demikian, tujuan penelitian ini adalah untuk mengidentifikasi tema penting dalam pidato Presiden Joko Widodo dengan menggunakan BERTopic. Selain itu, penelitian ini membandingkan empat model embedding populer untuk Bahasa Indonesia dan multilingual guna mengevaluasi pengaruh variasi embedding terhadap koherensi, keberagaman, dan interpretabilitas topik pidato Presiden. Dengan pendekatan ini, penelitian diharapkan dapat memberikan peta topik yang lebih presisi, mengungkap tema dominan, serta memperluas penggunaan metode topic modeling dalam analisis wacana politik di Indonesia. Hasil penelitian diharapkan menjadi rujukan untuk studi komunikasi politik berbasis data dan pertimbangan kebijakan publik. Selain itu, penelitian ini juga membuka peluang pengembangan aplikasi NLP lebih lanjut, misalnya dalam pemantauan opini publik, analisis sentimen terhadap kebijakan presiden, hingga penyusunan arsip digital yang dapat diakses oleh masyarakat luas. Dengan cara ini, analisis akademik tidak hanya memberi kontribusi pada pengembangan ilmu pengetahuan, tetapi juga berfungsi sebagai sarana partisipasi publik dalam kehidupan demokratis.

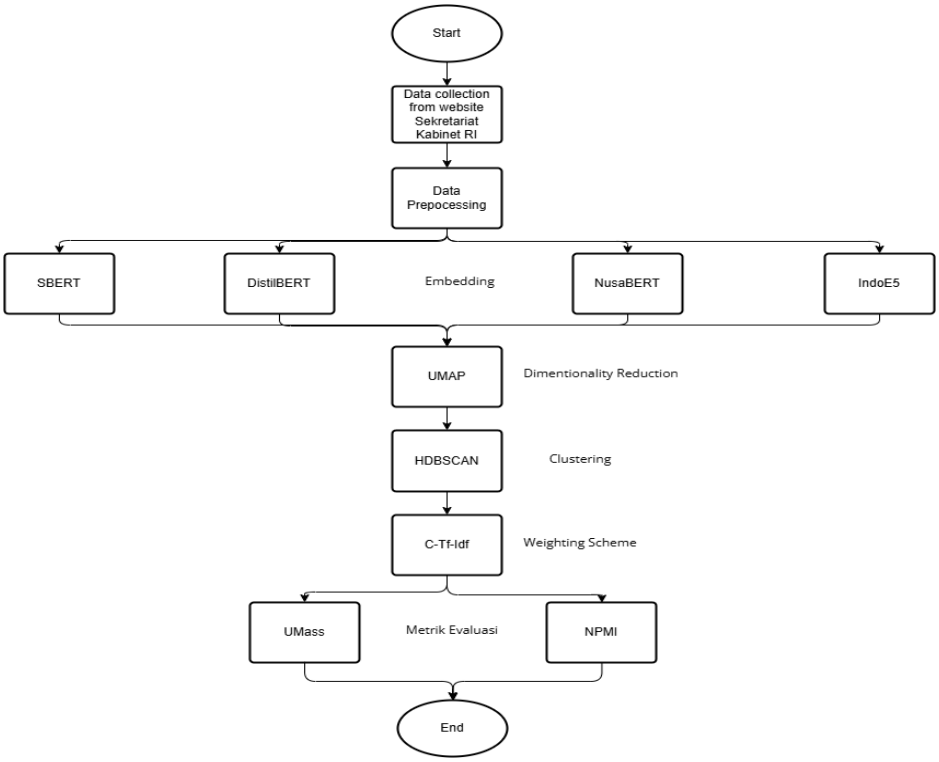
2. METODOLOGI PENELITIAN

2.1 Tahapan Penelitian

Penelitian ini dilaksanakan melalui beberapa tahapan sebagaimana ditunjukkan pada Gambar 1. Data dikumpulkan dari *website* Sekretariat Kabinet Republik Indonesia berupa dokumen teks pidato resmi. Data yang diperoleh kemudian diproses melalui tahap pra-pemrosesan. Selanjutnya, teks direpresentasikan menjadi vektor dengan menggunakan empat model *embedding*, yaitu SBERT[13], DistilBERT[14], NusaBERT, dan IndoE5[15]. Hasil *embedding* kemudian direduksi dimensinya dengan *Uniform Manifold Approximation and Projection* (UMAP) untuk mempermudah proses pengelompokan. Tahap berikutnya adalah *clustering* menggunakan algoritma *Hierarchical Density-Based Spatial Clustering of Applications with Noise* (HDBSCAN) yang mampu mendeteksi distribusi data

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1367−1379
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.8298

tidak beraturan sekaligus mengidentifikasi *noise*. Hasil cluster kemudian diperkaya dengan skema pembobotan c-TF-IDF untuk mengekstraksi kata representatif dari setiap topik. Evaluasi model dilakukan dengan dua metrik utama, yaitu UMass dan NPMI, untuk menilai koherensi dan keterkaitan semantik antar kata dalam topik. Melalui tahapan ini diperoleh gambaran kinerja masing-masing model *embedding* dalam menghasilkan topik yang valid dan bermakna.



Gambar 1. *Flowchart* penelitian yang dilakukan

2.2 Teknik Pengumpulan data

Data untuk penelitian ini diperoleh melalui penerapan metode *web scraping*[16]. Metode ini mengekstrak teks lengkap dari situs *web* yang ditentukan dengan memodifikasi struktur HTML dan menyimpannya secara sistematis dalam basis data MySQL. Proses ekstraksi dilakukan menggunakan Python bersama perpustakaan BeautifulSoup4, yang memungkinkan pengambilan teks secara otomatis dari situs *web* sesuai dengan kerangka HTML-nya. Data yang dikumpulkan disimpan dalam format SQL untuk analisis dan pemrosesan selanjutnya.

2.3 *Pre-Processing*

Sebelum melakukan pemodelan topik, data terlebih dahulu melalui beberapa tahap prapemrosesan. Tujuan prapemrosesan ini adalah untuk mengorganisir atau menormalisasikan data dan memastikan hanya kata yang relevan yang dianalisis[17]. Pra-pemrosesan memainkan peran krusial dalam mengubah teks dari bahasa manusia menjadi format yang dapat dipahami mesin dalam metodologi penambangan teks. Tahap pra-pemrosesan disini termasuk tokenisasi yaitu memecah teks menjadi unit kata, *lowercasing* yaitu menyamakan huruf menjadi huruf kecil, pembersihan karakter non-alfabet, pembersihan *stopword* yaitu menghapus kata umum yang tidak bermakna, dan stemming yaitu mengembalikan kata ke bentuk dasar. Proses ini menghasilkan teks yang lebih ringkas, konsisten, dan menonjolkan istilah penting yang berfungsi sebagai dasar analisis topik.

2.4 *Topic Modeling*

Topic modeling adalah metode pembelajaran mesin tanpa pengawasan yang memfasilitasi identifikasi otomatis topik atau tema abstrak dalam kumpulan dokumen yang luas[18]. Metode ini secara luas digunakan dalam pemrosesan bahasa alami dan penambangan teks untuk mengidentifikasi struktur semantik tersembunyi yang terdapat dalam kumpulan dokumen, tanpa memerlukan data yang telah diberi label sebelumnya. *Topic modeling* beroperasi berdasarkan asumsi bahwa dokumen terdiri dari kombinasi topik laten, masing-masing ditandai oleh distribusi kata[19].

2.5 BERTopic

BERTopic adalah teknik pemodelan topok berbasis *Transformers* yang secara komprehensif memahami konteks teks dalam kedua arah, menghasilkan representasi teks yang kaya akan makna semantik. Penggunaan model *embedding Transformers* dalam BERTopic digunakan untuk membangun representasi topik melalui serangkaian langkah sistematis. Kerangka kerja ini mengintegrasikan *word embedding*, pengurangan dimensi, dan algoritma *clustering*

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1367−1379
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.8298

untuk menciptakan kluster yang secara semantik serupa. Metode ini memanfaatkan kemampuan BERT untuk menghasilkan representasi dokumen yang lebih baik, memudahkan identifikasi topik yang lebih akurat dan relevan[20][21].

2.6 *Transformers Embedding*

Transformers membentuk representasi teks kontemporer, termasuk penelitian ini yang berfokus pada pidato Presiden Republik Indonesia. Menghasilkan representasi kontekstual, di mana makna sebuah kata ditentukan oleh konteks kalimat di sekitarnya, adalah perbedaan utamanya dari metode klasik seperti *Bag-of-Words* atau TF-IDF. Oleh karena itu, *embedding* yang dibuat tidak sekadar menyimpan frekuensi kata, tetapi juga mengidentifikasi hubungan semantik yang lebih halus antar token, kalimat, dan dokumen. Dalam praktiknya, arsitektur model yang digunakan meliputi vektor representasi dan jumlah parameter yang akan mempengaruhi kemampuan model untuk memahami bahasa[22].

Tahap awal dalam proses ini yang krusial adalah mengubah corpus menjadi representasi vektor numerik melalui *embedding model*. *Embedding* berfungsi sebagai proyeksi teks ke dalam ruang vektor berdimensi tinggi sehingga relasi semantik antar kalimat dapat ditangkap secara matematis[23]. Secara umum, *embedding* dapat dipandang sebagai suatu fungsi pemetaan(1):

$$f: D \rightarrow Rd \tag{1}$$

Dengan D adalah himpunan teks, dan d menyatakan dimensi vektor *embedding* yang bergantung pada model yang digunakan. Setiap kalimat atau dokumen $x_i \in D$ diproyeksikan menjadi vektor *embedding* $x_i \in D$ dalam persamaan(2) sebagai berikut:

$$e_i = f(x_i) \in \mathbb{R}D \tag{2}$$

Sebuah *term* atau kata dalam pidato Presiden dinyatakan sebagai urutan token seperti persamaan(3):

$$S = [w_1, w_2, w_3, \dots, w_n] \tag{3}$$

Kemudian dari setiap token kata w_i dipetakan ke vektor *embedding* awal melalui lookup matrix E dalam persamaan(4):

$$w_i = E(w_i) \in \mathbb{R}D \tag{4}$$

Dengan d adalah *dimensi embedding* yang digunakan pada masing – masing model. Selanjutnya *transformer* membangun representasi kontekstual tiap token melalui mekanisme *multi-head self-attention*. Untuk setiap token *embedding* x_i vektor query, key, dan value dihitung dengan proyeksi linear dalam persamaan(5):

$$Q = x_i W_q, K = x_i W_K, V = x_i W_V \tag{5}$$

Di mana matriks $W_q, W_K, W_V \in \mathbb{R}^{d \times d_k}$ merupakan komponen yang dilatih dalam mekanisme perhatian. Pada proses ini, Q $(Query)$ merepresentasikan pencarian dari sebuah token, sedangkan K (Key) merepresentasikan fitur yang dimiliki oleh token tersebut. Sementara itu, V $(Value)$ merepresentasikan informasi aktual antara query dan key. Interaksi antar token dihitung dengan perkalian dot-product antara Query dan Key, kemudian diskalakan dengan $\sqrt{d_k}$: seperti dalam persamaan(6):

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{6}$$

Di mana QK^T menghitung similarity antara query dan semua key. Sementara $\sqrt{d_k}$ menormalisasi untuk menghindari nilai dot-product terlalu besar dan $softmax$: fungsi untuk mengubah skor menjadi distribusi probabilitas.

Untuk mendapatkan representasi dokumen atau pidato secara keseluruhan, digunakan *pooling*. Dalam konteks untuk topic model, digunakan metode *mean pooling*, dengan mengambil nilai rata - rata seluruh *hidden states* dari setiap token seperti dalam persamaan(7):

$$s = \frac{1}{n}\sum_{i=1}^{n} h_i \tag{7}$$

Di mana s adalah representasi kalimat dalam vektor, dan n : jumlah total token dalam kalimat / dokumen, serta h_i : vektor *hidden state* dari token ke-iii pada model transformer. Dengan desain dan kompleksitas model yang berbeda, perbandingan ini sangat penting untuk memahami bagaimana embedding mempengaruhi performa topic modeling dengan BERTopic. Ringkasan arsitektur, dimensi *embedding*, dan jumlah parameter dari keempat model yang digunakan disajikan dalam Tabel 1 berikut:

Tabel 1. Embedding yang digunakan dan Perbandingan Arsitektur Model

Model	Arsitektur	Dimensi Embedded	Parameter
SBERT (m-bert-base-nli-stsb-mean-tokens)	BERT-base + Siamese	768	110M

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1367−1379
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.8298

Model	Arsitektur	Dimensi Embedded	Parameter
DistilBERT (distilbert-multilingual)	Distilled BERT	512	66M
NusaBERT (all-nusabert-base-v4)	BERT-base	768	110M
IndoE5 (all-indo-e5-small-v2)	E5-small (encoder-only)	384	33M

2.7 *Dimentionality Reduction*

Dokumen atau data teks yang telah diubah menjadi vektor akan menjalani proses pengelompokan. Langkah ini melibatkan teknik *dimentionality reduction* atau reduksi dimensi. Teknik ini bertujuan untuk meminimalkan volume data yang tersisa setelah pembentukan *embedding*. Proses ini akan menghasilkan data yang serupa dengan vektor, meskipun dengan dimensi yang lebih rendah. Dalam penelitian ini menggunakan teknik reduksi dimensi yang dikenal dengan nama UMAP yang berbasis teori *manifold* dan *graf* terhubung. Prosesnya mencakup dua tahap utama yaitu kontruksi graf dan reduksi dimensi dengan optimasi graf.

UMAP membangun *graf* berbobot berdasarkan jarak antar data menggunakan kernel berbasis jarak ρ dan σ untuk menentukan hubungan lokal. Probabilitas hubungan dalam graf dimuat berdasarkan probabilitas lokal dalam dimensi tinggi dan dalam dimensi rendah, rumus probabilitas hubungan lokal *p* antar data *i* dan *j* dalam dimensi tinggi didefinisikan dalam persamaan(8):

$$P\ i,j = \exp\left(-\frac{\max(0,d\ (x_i,x_j)-\rho_i)}{\sigma_i}\right) \tag{8}$$

Di mana dx_i, x_j adalah jarak antara data x_i dan x_j, kemudian ρ_i Adalah jarak terdekat yang memastikan setidaknya ada satu tetangga, dan σ_i adalah parameter skala untuk menentukan hubungan lokal.

Sedangkan probabilitas *q* dalam hubungan antar data **i** dan **j** adalah probabilitas ruang dimensi rendah, setelah data direduksi oleh UMAP. $q_{i,j}$ digunakan untuk menggambarkan kesamaan hubungan antara data dalam ruang rendah, dengan mempertimbangkan bahwa data di ruang rendah mungkin terpisah lebih jauh dibandingkan dengan di ruang tinggi sebagaimana didefinisikan dalam persamaan(9):

$$q_{i,j} = \frac{1}{1+\alpha\ \|\ y_i-y_j\|^{2b}} \tag{9}$$

Di mana y_i, y_j merupakan representasi titik data *i* dan *j* dalam ruang dimensi rendah setelah proses *embedding*. Sedangkan *a, b* Adalah parameter yang mengontrol distribusi jarak di ruang dimensi rendah.

UMAP memproyeksikan data ke dimensi rendah Y dengan meminimalkan jarak antara graf pada dimensi tinggi (X) dan dimensi rendah (Y) sebagaimana didefinisikan pada persamaan(10):

$$C(Y) = \sum_{i<j}\left[p_{i,j}\log\frac{p_{i,j}}{q_{i,j}} + \left(1-\ p_{i,j}\right)\log\frac{1-p_{i,j}}{1-q_{i,j}}\right] \tag{10}$$

Di mana pij adalah probabilitas pada *graf* dimensi tinggi, sedangkan qij adalah probabilitas pada *graf* dimensi rendah.

2.7 *Clustering*

Untuk membentuk kluster setelah vektor melalui tahap dimensi reduksi, maka langkah selanjutnya adalah menggunakan teknik *clustering* yang menjadi *pipeline* untuk BERTopic. HDBSCAN merupakan teknik *clustering* yang menghitung densitas lokal untuk setiap titik data[24]. Densitas lokal dihitung dengan menggunakan jarak antar titik dan jumlah titik terdekat di sekitarnya, secara umum untuk menghitung densitas lokal untuk titik *i* didefinisikan dalam persamaan(11):

$$\rho_i = \frac{1}{Volume(B_i)} \tag{11}$$

Di mana ρ_i merupakan densitas lokal titik i dan *Bi* adalah bola atau lingkungan sekitar titik i yang dihitung berdasarkan jarak yang sudah ditentukan sebelumnya. Dalam HDBSCAN[25], mengukur keterhubungan antara dua titik i dan j berdasarkan kedekatannya. Titik yang tidak memenuhi keduanya akan dianggap sebagai *noise*. Pada dasarnya, sebuah titik i akan dianggap *noise* jika densitas lokalnya lebih rendah dari suatu ambang batas yang dihitung menggunakan parameter tersebut. Umumnya, metrik yang digunakan adalah jarak *Euclidean* atau *cosine distance* definisikan dalam persamaan (12):

$$d(i,j) = \|\ X_i - X_j\ \| \tag{12}$$

Di mana d(i, j) adalah jarak antara titik i dan titik j, sedangkan xi dan xj merupakan vektor posisi titik i dan titik j dalam ruang fitur. Dalam HDBSCAN diperkenalkan pula memanfaatkan teknik *hierarki clustering* berbasis densitas untuk menentukan kelompok titik yang membentuk *cluster*. Secara matematis, HDBSCAN membangun pohon hierarki yang disebut *mutual reachability distance* antara dua titik i dan j. Jarak ini dihitung menggunakan persamaan (13):

$$d_{reach}(i,j) = max(core\_distance(i), core\_distance(j), d(i,j)) \tag{13}$$

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1367−1379
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.8298

Di mana d_{reach} (i,j) : jarak *mutual reachability* antara titik i dan j. Sementara $core\_distance(i)$ adalah jarak minimum yang diperlukan untuk i agar memiliki *min_samples* tetangga terdekat, sedangkan $d(i,j)$ merupakan jarak *Euclidean* atau jarak lain antara i dan j.

2.8 *Model Evaluation*

Dalam menilai kualitas topik yang dihasilkan oleh berbagai pendekatan pemodelan topik, penelitian ini menggunakan dua metrik skor koherensi: UMASS dan NPMI. Kedua metrik tersebut umumnya digunakan untuk mengevaluasi koherensi semantik di antara kata dalam suatu topik[8][26].

2.8.1 UMass

Metrik UMass menghitung *coherence score* berdasarkan kemunculan kata-kata dalam topik yang sama di dokumen. UMass mengevaluasi probabilitas kemunculan pasangan kata secara bersamaan dalam suatu dokumen dan membandingkannya dengan probabilitas individu dari salah satu kata dalam pasangan tersebut[22], [27]. Secara matematis, UMass didefinisikan dengan persamaan(14):

$$C_{UMass}(T) = \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{i-1} \log \frac{p(w_i,w_j)+\frac{1}{D}}{p(w_j)} \quad (14)$$

Di mana T={w1,w2,…,wN} adalah daftar kata-kata dalam topik. Pada persamaan p(wi,wj) merepresentasikan probabilitas dua kata wi dan wj yang muncul bersama dalam dokumen (konteks), sementara p(wj) adalah probabilitas kata wj muncul dalam dokumen. Nilai N mengacu pada jumlah kata dalam topik, sedangkan D merupakan jumlah total dokumen dalam korpus.

2.8.2 NPMI

Sebagai pelengkap UMass, *Normalized Pointwise Mutual Information* (NPMI) digunakan untuk mengevaluasi seberapa erat hubungan antar kata dalam suatu topik dibandingkan dengan kemunculan independen kata tersebut. NPMI menghitung *Pointwise Mutual Information* (PMI) yang dinormalisasi agar hasilnya berkisar antara -1 hingga 1[28]. Secara matematis, NPMI didefinisikan dengan persamaan(15):

$$NPMI(w_i, w_j) = \left(\frac{PMI}{-\log(p(w_i,w_j)+\epsilon)} \right) \quad (15)$$

di mana PMI, tingkat di mana kata-kata muncul bersama dalam korpus tertentu dirumuskan dalam persamaan(16):

$$PMI(w_i, w_j) = \log \frac{(p(w_i,w_j)+\epsilon)}{(p(w_i).\ p(w_j)} \quad (16)$$

Di mana p(wi,wj) merupakan persamaan probabilitas dua kata wi dan wj muncul bersama dalam korpus. Dalam persamaan p(wi) merepresentasikan probabilitas kata wi muncul di korpus, sementara p(wj) merepresentasikan probabilitas kata wj muncul di korpus. Sedangkan ϵ : nilai kecil untuk mencegah logaritma dari nol.

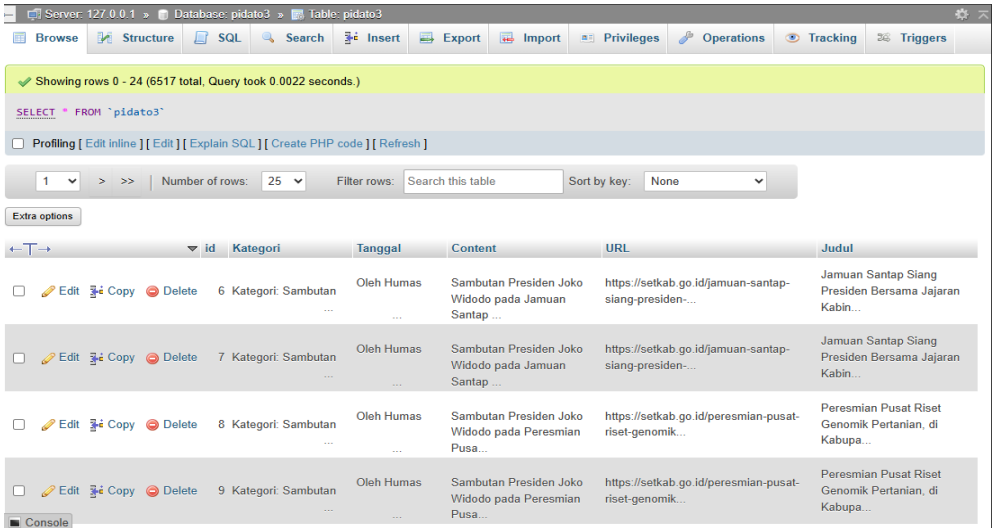
3. HASIL DAN PEMBAHASAN

Dalam bagian ini, hasil dari setiap tahapan penelitian ditunjukkan sesuai dengan penjelasan metodologi. Pra-pemrosesan menghasilkan teks bersih yang siap untuk dianalisis. Tahap penerapan dan pengurangan dimensi mengubah teks menjadi representasi vektor yang terstruktur, dan tahap pembubaran menghasilkan kelompok topik utama dari pidato presiden. Selanjutnya, setiap topik dievaluasi menggunakan metrik koherensi untuk mengevaluasi kualitas hasil.

3.1 Data Penelitian

Data yang digunakan dalam penelitian ini berupa transkrip pidato Presiden Republik Indonesia yang diperoleh dari Sekretariat Kabinet Republik Indonesia dengan alamat website https://setkab.go.id/category/transkrip-pidato/ yang secara konsisten menyediakan dokumentasi pidato Presiden dalam bentuk teks. Proses pengumpulan dilakukan secara sistematis dengan mengunduh, memeriksa, dan menyusun setiap pidato ke dalam format teks yang seragam. Proses pengambilan transkrip pidato dimulai dari periode pertama jabatan presiden RI ke-7 Ir. H. Joko Widodo sejak September 2014 hingga masa berakhirnya periode kedua jabatan presiden bulan September 2024, data dibatasi selama 10 tahun. Transkrip pidato yang dikumpulkan sebanyan 6516 transkrip pidato dan disimpan dalam format penyimpanan database MySQL. Gambar 2 menunjukkan total baris dan kolom yang digunakan.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1367−1379
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.8298



Gambar 2. Data Penelitian dalam database

3.2 *Data Prepocessing Result*

Setelah data disimpan, langkah selanjutnya korpus yang telah disimpan atau *raw data* akan dinormalisasi atau diubah menjadi bentuk yang lebih sederhana dan terstruktur setelah tahapan *preprocessing* dilakukan menggunakan metode yang dijelaskan dalam metode penelitian. Hasil menunjukkan bahwa kata yang memiliki makna utama telah dipertahankan, tetapi kata umum dan tanda baca yang tidak relevan telah dihilangkan. Oleh karena itu, teks yang dihasilkan dari proses *preprocessing* menjadi lebih ringkas, konsisten, dan siap untuk digunakan pada tahap analisis berikutnya. Seperti terlihat pada Tabel 1, pra-pemrosesan berhasil menyaring kata-kata yang kurang bermakna dan hanya menyisakan token penting. Tahap ini krusial karena menentukan kualitas embedding pada langkah selanjutnya

Tabel 1. Hasil *Pre-processing*

RAW Data	*Preprocessed Data*
Sambutan Presiden Joko Widodo pada Jamuan Santap Siang Bersama Jajaran Kabinet Indonesia Maju Jelang Purnatugas, 18 Oktober 2024. Pada kesempatan yang baik ini, saya ingin menyampaikan ucapan terima kasih yang sebesar-besarnya atas dukungan, atas support, atas kerja keras untuk negara ini. Dan saya berharap apa yang sudah kita lakukan dalam sepuluh tahun ini bermanfaat bagi rakyat, bermanfaat bagi negara, bermanfaat bagi bangsa kita yang kita cintai.	sambut presiden joko widodo jamu santap siang jajaran kabinet indonesia maju jelang purnatuga oktober kesempatan baik ingin sampai ucap terima kasih besar dukung support kerja keras negara harap laku sepuluh tahun manfaat rakyat manfaat negara manfaat bangsa cinta

3.3 *Topic Result*

Selama proses pemodelan topik, empat model *embedding* yang berbeda digunakan bersama kerangka kerja BERTopic untuk mengidentifikasi tema tersembunyi dalam korpus. Setiap model dikonfigurasi untuk menghasilkan 20 topik, memungkinkan perbandingan yang adil antar pendekatan yang berbeda.

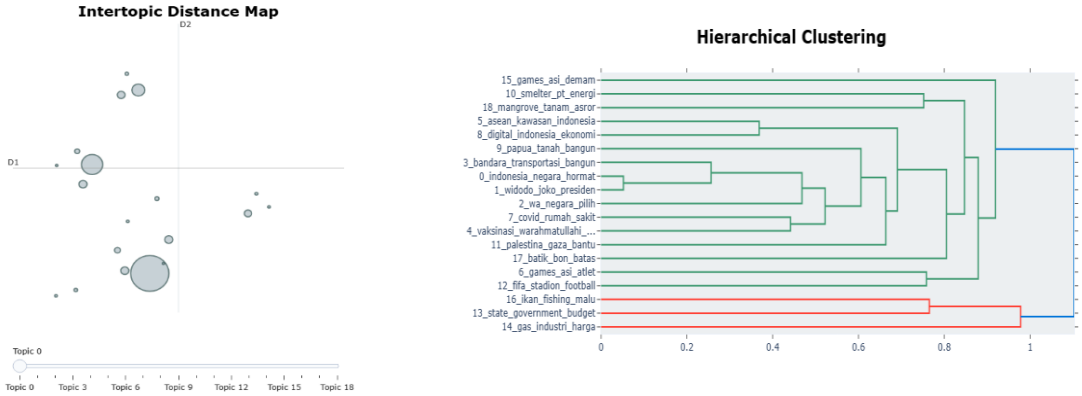
3.3.1 SBERT *Embedding*

Tabel 2 menunjukkan hasil *embedding* dengan SBERT menghasilkan representasi semantik yang kemudian dikelompokkan menjadi sejumlah topik. Menurut analisis kualitatif, isu seperti "*bandara_transportasi_bangun_kota_terminal*" menggambarkan agenda pembangunan infrastruktur transportasi, yang merupakan fokus kebijakan utama Presiden Joko Widodo. Tema "*asean_kawasan_indonesia_ktt_mulia*" berkaitan dengan partisipasi Indonesia dalam forum regional ASEAN dan pertemuan internasional. Sementara itu, hashtag "*palestina_gaza_bantu_oki_manusia*" menunjukkan bahwa Indonesia selalu mendukung masalah kemanusiaan dalam politik luar negerinya. Namun, muncul pula topik yang tampak tidak wajar, misalnya "vaksinasi_warahmatullahi_wabarakatuh_alaikum". Topik ini dapat dipahami sebagai gabungan kata salam pembuka pidato dengan isu vaksinasi, terutama pada periode pandemi COVID-19. Demikian juga topik "widodo_joko_presiden_republik_wartawan" menunjukkan keterkaitan dengan konferensi pers, karena banyak pidato Presiden disampaikan di hadapan wartawan. Kedua contoh ini menunjukkan adanya *noise* akibat *preprocessing* yang belum sepenuhnya menyaring kata sapaan atau kata yang terlalu sering muncul. Meskipun demikian, keberulangan tema besar seperti ekonomi, infrastruktur, kesehatan, dan diplomasi menunjukkan bahwa SBERT mampu menangkap pola wacana utama secara konsisten.

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1367−1379
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.8298

Tabel 2. Hasil SBERT

Topic	Count	Name
-1	791	-1_indonesia_presiden_ya_negara_menteri
0	3305	0_indonesia_negara_hormat_presiden_ekonomi
1	1024	1_widodo_joko_presiden_republik_wartawan
2	375	2_wa_negara_pilih_indonesia_presiden
3	151	3_bandara_transportasi_bangun_kota_terminal
4	149	4_vaksinasi_warahmatullahi_wabarakatuh_alaikum
5	148	5_asean_kawasan_indonesia_ktt_mulia
6	137	6_games_asi_atlet_wartawan_latih
7	121	7_covid_rumah_sakit_sehat_masyarakat
8	78	8_digital_indonesia_ekonomi_umkm_uang
9	57	9_papua_tanah_bangun_pon_muda
10	38	10_smelter_pt_energi_tembaga_listrik
11	28	11_palestina_gaza_bantu_oki_manusia
12	26	12_fifa_stadion_football_sepak_bola
13	19	13_state_government_budget_economic_pandemic
14	16	14_gas_industri_harga_kaca_turun
15	15	15_games_asi_demam_xviii_helat
16	15	16_ikan_fishing_malu_ambon_kapal
17	12	17_batik_bon_batas_sambas_dr



Gambar 3. Visualisasi Topik embedding SBERT

3.3.2 DistilBERT *Embedding*

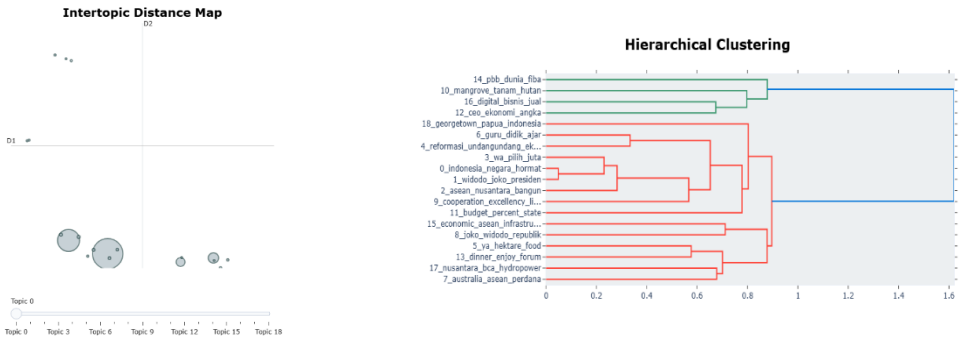
Tabel 3 menunjukkan hasil *embedding* dengan DistilBERT menghasilkan representasi semantik yang kemudian dikelompokkan menjadi sejumlah topik. Secara kualitatif, topik seperti "*reformasi_undang-undang_ekonomi_kawasan_hukum*" terkait dengan agenda reformasi hukum dan regulasi yang sering disampaikan oleh presiden. Transformasi digital dan ekosistem startup ditampilkan dalam topik "*digital_bisnis_jual_indonesia_startup*". Selain itu, "*mangrove_tanam_hutan_covid_rumah*" menunjukkan masalah lingkungan dan kesehatan yang muncul setelah pandemi. Namun, DistilBERT juga menggunakan token yang tidak informatif seperti "*dinner_enjoy_forum_thank_evening*". Kata ini berasal dari pidato di forum internasional yang menggunakan bahasa Inggris. Karena sering digunakan, mereka dianggap sebagai topik tersendiri. Ini menunjukkan bahwa model DistilBERT, meskipun lebih sederhana, lebih mudah menemukan kata asing tanpa konteks.

Tabel 3. Hasil DistilBERT

Topic	Count	Name
-1	752	-1_indonesia_presiden_negara_hormat_republik
0	3173	0_indonesia_negara_hormat_menteri_ekonomi
1	1672	1_widodo_joko_presiden_republik_indonesia
2	368	2_asean_nusantara_bangun_resmi_hormat
3	273	3_wa_pilih_juta_ya_presiden
4	32	4_reformasi_undangundang_ekonomi_kawasan_hukum
5	30	5_ya_hektare_food_rp_pangan
6	26	6_guru_didik_ajar_anakanak_ubah
7	26	7_australia_asean_perdana_albanese_indonesia
8	22	8_joko_widodo_republik_presiden_indonesia

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1367−1379
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.8298

Topic	Count	Name
9	21	9_cooperation_excellency_liberia_ghana_tol
10	18	10_mangrove_tanam_hutan_covid_rumah
11	15	11_budget_percent_state_economic_trillion
12	15	12_ceo_ekonomi_angka_kerja_peluang
13	14	13_dinner_enjoy_forum_thank_evening
14	14	14_pbb_dunia_fiba_arena_konser
15	12	15_economic_asean_infrastructure_investment
16	12	16_digital_bisnis_jual_indonesia_startup
17	12	17_nusantara_bca_hydropower_tarang_investasi



Gambar 4. Visualisasi Topik embedding DistilBERT

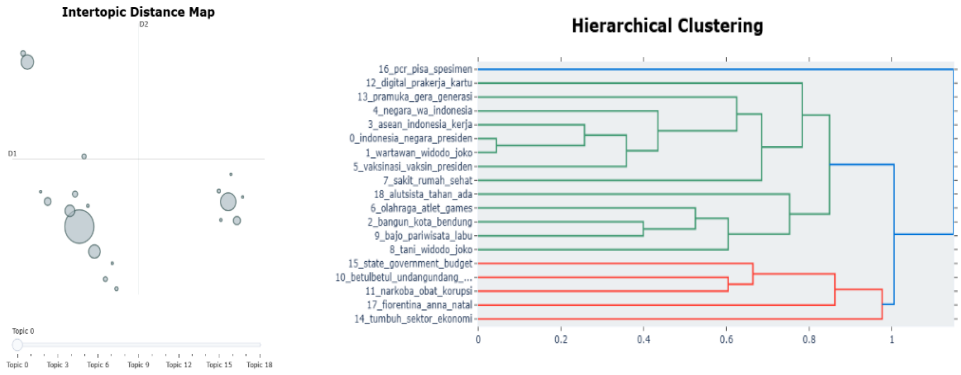
3.3.3 NusaBERT *Embedding*

Tabel 4 menunjukkan hasil embedding dengan *NusaBERT* menghasilkan representasi semantik yang kemudian dikelompokkan menjadi sejumlah topik. Menurut analisis kualitatif, topik "*bangun_kota_bendung_nusantara_selesai*" mencerminkan rencana dan proyek pembangunan infrastruktur Ibu Kota Nusantara. Tema "*vaksinasi_vaksinasi_presiden_terima_widodo*" jelas terkait dengan pidato Presiden selama pandemi COVID-19, ketika vaksinasi menjadi masalah utama di seluruh negeri. Jika ada hubungan antara topik "*olahraga_atlet_games_medali_fifa*" dan acara olahraga internasional seperti Asian Games atau masalah FIFA, maka topik tersebut akan ditunjukkan. NusaBERT juga menangkap topik yang lebih sektoral, seperti "*tani_widodo_joko_hektare_politeknik*", yang berkaitan dengan pertanian dan pendidikan vokasi, dan "*narkoba_obat_korupsi_terorisme_lawan*", yang merujuk pada pidato tentang masalah hukum dan keamanan. Meskipun demikian, beberapa topik tampak tidak koheren, seperti "*fiorentina_anna_natal_sitaniapessy_indone*", karena nama asing digunakan dalam olahraga. Ini menunjukkan bahwa meskipun NusaBERT dapat menangkap semantik bahasa Indonesia dengan baik, model masih terganggu ketika menggunakan istilah asing.

Tabel 4. Hasil NusaBERT

Topic	Count	Name
-1	897	-1_indonesia_presiden_republik_ya_widodo
0	2797	0_indonesia_negara_presiden_ya_republik
1	784	1_wartawan_widodo_joko_presiden_republik
2	515	2_bangun_kota_bendung_nusantara_selesai
3	442	3_asean_indonesia_kerja_dunia_negara
4	320	4_negara_wa_indonesia_hormat_ulama
5	171	5_vaksinasi_vaksin_presiden_terima_widodo
6	151	6_olahraga_atlet_games_medali_fifa
7	86	7_sakit_rumah_sehat_layan_joko
8	72	8_tani_widodo_joko_hektare_politeknik
9	60	9_bajo_pariwisata_labu_kawasan_wisatawan
10	55	10_betulbetul_undangundang_anggar_didik
11	36	11_narkoba_obat_korupsi_terorisme_lawan
12	34	12_digital_prakerja_kartu_indonesia_ekonomi
13	22	13_pramuka_gera_generasi_muda_disiplin
14	20	14_tumbuh_sektor_ekonomi_provinsi_sulawesi
15	15	15_state_government_budget_economic_pandemic
16	14	16_pcr_pisa_spesimen_thousand_lapor
17	14	17_fiorentina_anna_natal_sitaniapessy_indone
18	12	18_alutsista_tahan_ada_industri_ikan

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1367−1379
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.8298



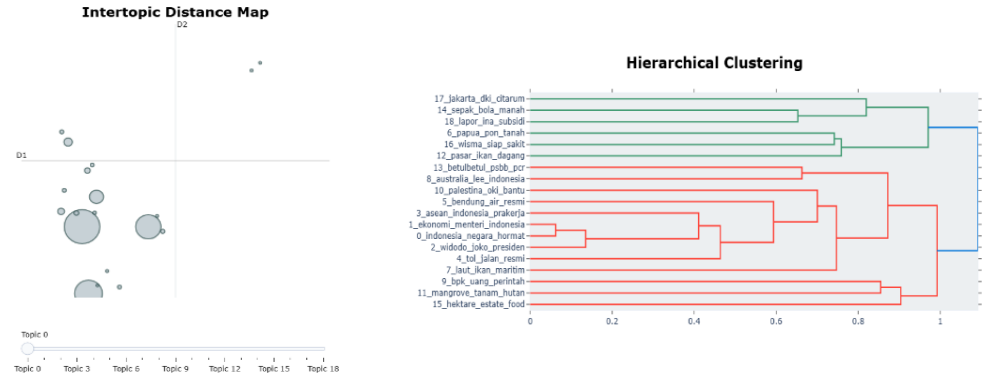
Gambar 5. Visualisasi Topik embedding NusaBERT

3.3.4 IndoE5 *Embedding*

Tabel 5 menunjukkan hasil *embedding* dengan IndoE5 menghasilkan representasi semantik yang kemudian dikelompokkan menjadi sejumlah topik. Menurut analisis kualitatif, topik "*tol_jalan_resmi_jakarta_ruas*" mengacu pada pidato peresmian infrastruktur transportasi dan jalan tol, yang merupakan ciri khas pembangunan era Jokowi. Tema "*laut_ikan_maritim_kapal_nelayan*" menunjukkan bahwa Indonesia adalah pusat maritim global. Selain itu, sikap diplomasi Indonesia yang pro-Palestina ditunjukkan oleh frase "*palestina_oki_bantu_israel_manusia*". Namun, IndoE5 menghasilkan beberapa topik yang bercampur, seperti "*betul_psbb_pcr_banjir_antar*", yang tampaknya menggabungkan masalah pandemi dengan banjir, mungkin karena keduanya dibicarakan dalam konteks yang berbeda tetapi menggunakan istilah yang sering diulang.

Tabel 5. Hasil IndoE5

Topic	Count	Name
-1	868	-1_indonesia_presiden_republik_negara_ya
0	2263	0_indonesia_negara_hormat_presiden_republik
1	1365	1_ekonomi_menteri_indonesia_kerja_negara
2	1143	2_widodo_joko_presiden_wartawan_republik
3	348	3_asean_indonesia_prakerja_kerja_terima
4	129	4_tol_jalan_resmi_jakarta_ruas
5	82	5_bendung_air_resmi_kabupaten_provinsi
6	58	6_papua_pon_tanah_bangun_olahraga
7	40	7_laut_ikan_maritim_kapal_nelayan
8	32	8_australia_lee_indonesia_singapura_singapo
9	28	9_bpk_uang_perintah_bpkp_wtp
10	28	10_palestina_oki_bantu_israel_manusia
11	26	11_mangrove_tanam_hutan_pohon_sampah
12	22	12_pasar_ikan_dagang_rahmat_jaya
13	19	13_betulbetul_psbb_pcr_banjir_antar
14	14	14_sepak_bola_manah_latih_sepakbola
15	14	15_hektare_estate_food_sumatra_eks
16	14	16_wisma_siap_sakit_asi_games
17	12	17_jakarta_dki_citarum_sungai_jabodetabek



Gambar 6. Visualisasi Topik embedding IndoE5

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1367−1379
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.8298

3.4 Evaluasi Model

3.4.1 UMass

UMass merupakan metrik koherensi topik yang mengukur seberapa konsisten kata-kata dalam satu topik muncul bersama dalam dokumen. Nilai yang lebih tinggi mendekati 0 menunjukkan topik yang lebih koheren secara semantik. Berdasarkan Tabel 6, NusaBERT menunjukkan nilai terbaik (-2.016), diikuti SBERT (-2.036), IndoE5 (-2.459), dan DistilBERT (-2.797). Artinya, NusaBERT dan SBERT menghasilkan topik yang lebih koheren dibandingkan model lainnya menurut metrik UMass.

Tabel 6. Evaluasi UMass

Model	Score
SBERT	-2.036
DistilBERT	-2.797
NusaBERT	-2.016
IndoE5	-2.459

3.4.2 NPMI

NPMI digunakan untuk mengukur koherensi topik berdasarkan seberapa sering kata-kata dalam satu topik muncul bersamaan secara relatif terhadap peluang kemunculannya. Nilai NPMI yang lebih tinggi mendekati 1 menunjukkan topik yang lebih koheren. Dari Tabel 7, SBERT memiliki skor tertinggi (0.082), diikuti NusaBERT (0.076), IndoE5 (0.065), dan DistilBERT (0.060), menunjukkan SBERT menghasilkan topik yang paling konsisten secara semantik menurut metrik NPMI.

Tabel 7. Evaluasi NPMI

Model	Score
SBERT	0.082
DistilBERT	0.060
NusaBERT	0.076
IndoE5	0.065

3.5 Pembahasan

Hasil evaluasi kuantitatif menunjukkan bahwa SBERT dan NusaBERT adalah dua model embedding yang paling efektif dalam memodelkan topik pidato Presiden Joko Widodo. SBERT menempati posisi terbaik berdasarkan metrik NPMI yang mencapai nilai 0.082. Angka ini menunjukkan bahwa kata-kata dalam satu topik memiliki distribusi yang konsisten, sehingga sesuai dengan cara manusia memahami hubungan makna antar kata. Sementara itu, NusaBERT berada di posisi terbaik pada metrik UMass dengan skor -2.016, menunjukkan bahwa topiknya lebih kohesif dalam konteks dokumen tertentu. Dalam literatur, NPMI sering dianggap sebagai metrik yang lebih akurat karena lebih dekat dengan penilaian manusia dibandingkan UMass. Oleh karena itu, SBERT dapat disimpulkan sebagai model terbaik secara keseluruhan dalam penelitian ini.

Pada SBERT, topik "transformasi_digital_startup_teknologi" muncul secara konsisten dalam beberapa pidato yang berbeda tetapi membahas ekonomi digital. Ini menunjukkan kemampuan SBERT dalam menghubungkan dokumen yang berbeda dengan tema serupa. Di sisi lain, pada NusaBERT, topik "pertanian_pangan_petani" menunjukkan kohesi internal yang kuat dalam pidato yang fokus pada sektor pertanian, dengan kata-kata yang lebih seragam dan lebih berfokus. Hal ini memperkuat bahwa SBERT lebih unggul dalam menemukan konsistensi lintas dokumen, sementara NusaBERT lebih baik dalam menjaga koherensi internal dalam satu dokumen tertentu. Temuan ini menunjukkan bahwa spesialisasi bahasa dan ukuran model embedding sangat memengaruhi kualitas pemodelan topik.

Model dengan embedding yang besar, seperti SBERT, mampu menunjukkan detail makna yang lebih kaya, meskipun membutuhkan komputer yang lebih kuat. Sementara itu, NusaBERT, yang terlatih khusus pada bahasa Indonesia, mampu menjaga nuansa makna yang lebih spesifik dan lokal. Dengan demikian, pemilihan model embedding harus disesuaikan dengan tujuan analisis, jika tujuan utama adalah memahami makna secara unggul dan konsisten lintas dokumen, maka SBERT lebih tepat digunakan, sedangkan untuk analisis yang fokus pada satu dokumen, NusaBERT tetap menjadi pilihan yang baik.

4. KESIMPULAN

Studi ini menunjukkan bahwa pemodelan topik berbasis BERTopic dapat secara kuantitatif mengungkap tema-tema utama dalam pidato Presiden Joko Widodo. Perbandingan empat model *embedding* SBERT, DistilBERT, NusaBERT, dan IndoE5 menunjukkan bahwa kualitas koherensi topik dan kemampuan model untuk menangkap relasi semantik antar kalimat dipengaruhi secara langsung oleh pilihan embedding. Hasil evaluasi yang dilakukan menggunakan metrik UMass dan NPMI menunjukkan bahwa SBERT dan NusaBERT adalah *embedding* terbaik; SBERT

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1367−1379
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.8298

menghasilkan topik dengan konsistensi semantik terbaik menurut NPMI, dan NusaBERT menunjukkan koherensi internal dokumen UMass tertinggi. Keunggulan ini dipengaruhi oleh dimensi *embedding* yang besar dan arsitektur model yang mampu mendeteksi dengan tepat variasi Bahasa Indonesia.Selain itu, penelitian ini menyoroti adanya *trade-off* antara ukuran *embedding* dan efisiensi komputasi. Model dengan *embedding* besar cenderung memberikan detail semantik yang lebih kaya, tetapi membutuhkan sumber daya komputasi lebih tinggi. Sebaliknya, model lebih kecil lebih efisien, namun berisiko kehilangan beberapa informasi kontekstual yang penting. Temuan ini menekankan pentingnya menyesuaikan pemilihan *embedding* dengan tujuan analisis, apakah fokus pada kualitas koherensi topik atau efisiensi pemrosesan. Namun, penelitian ini memiliki keterbatasan. Analisis ini hanya membahas pidato Presiden Joko Widodo, sehingga tidak dapat digeneralisasi ke pidato pejabat lain atau periode pemerintahan sebelumnya. Selain itu, jumlah topik dan konfigurasi parameter BERTopic dapat memengaruhi hasil akhir, karena itu memasukkan domain tertentu atau perubahan parameter dapat menjadi ide untuk penelitian berikutnya.

REFERENCES

[1] I. L. Alamsyah, N. Aulya, and S. H. Satriya, "Transformasi media dan dinamika komunikasi dalam era digital: Tantangan dan peluang ilmu komunikasi," *J. Ilm. Res. Student*, vol. 1, no. 3, pp. 168–181, 2024, doi: https://doi.org/10.61722/jirs.v1i3.554.

[2] A. K. N. Oktavianaa, N. A. S. ERa, I. B. M. Mahendraa, I. G. S. Astawaa, I. G. A. Wibawaa, and I. K. A. Mogia, "Pemodelan Topik Artikel Berita Menggunakan Structural Topic Model dan Latent Dirichlet Allocation," *J. Elektron. Ilmu Komput. Udayana p-ISSN*, vol. 2301, p. 5373, 2022, doi: 10.24843/JLK.2023.v11.i03.p02.

[3] D. Grootendorst, "BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure," *arXiv Prepr. arXiv2203.05794*, 2022, doi: https://doi.org/10.48550/arXiv.2203.05794.

[4] D. D. Harmoko and P. W. Purwaningrum, "Simbolisme Politik dan Narasi Keberlanjutan: Studi Analisis Wacana Kritis Pidato Presiden RI Ke 7 Joko Widodo," *JURNALISTRENDi J. Linguist. SASTRA, DAN Pendidik.*, vol. 10, no. 1, pp. 54–66, 2025, doi: https://doi.org/10.51673/jurnalistrendi.v10i1.2415.

[5] D. Y. M. Putri and A. S. A. Sabardila, "Kohesi Gramatikal dan Leksikal Pidato Presiden Jokowi Tentang Penyematan Pangkat Jendral Kehormatan Kepada Prabowo Subianto," *J. Pendidik. Rokania*, vol. 9, no. 3, pp. 350–361, 2024, doi: https://doi.org/10.37728/jpr.v9i3.1098 putri2024.

[6] A. M. Alam, "Analisis Wacana Kritis pada Pidato Presiden Tahun 2022: Model Norman Fairclough," *J. Onoma Pendidik.*, vol. 10, no. 1, 2024, doi: https://doi.org/10.30605/onoma.v10i1.3163.

[7] S. Anggai, R. M. Zain, T. Tukiyat, and A. A. Waskita, "Enhancing BERTopic with Neural Network Clustering for Thematic Analysis of US Presidential Speeches," *J. Tek. Inform.*, vol. 6, no. 4, pp. 1957–1970, 2025, doi: https://doi.org/10.52436/1.jutif.2025.6.4.5090.

[8] R. M. Zain, S. Anggai, Tukiyat, A. Musyafa, and A. A. Waskita, "Revealing a Country ' s Government Discourse Through BERT-based Topic Modeling in the US Presidential Speeches," *2024 Int. Conf. Comput. Control. Informatics its Appl.*, vol. 11, pp. 191–196, 2024, doi: 10.1109/IC3INA64086.2024.10732578.

[9] N. C. Hellwig, J. Fehle, M. Bink, T. Schmidt, and C. Wolff, "Exploring Twitter discourse with BERTopic: topic modeling of tweets related to the major German parties during the 2021 German federal election," *Int. J. Speech Technol.*, vol. 27, no. 4, pp. 901–921, 2024, doi: https://doi.org/10.1007/s10772-024-10142-4.

[10] K. Sakiyama, L. de Souza Rodrigues, B. M. Nogueira, E. T. Matsubara, and R. A. F. Romero, "A Framework for Controversial Political Topics Identification Using Twitter Data," in *Brazilian Conference on Intelligent Systems*, 2023, pp. 283–298. doi: https://doi.org/10.1007/978-3-031-45392-2_19.

[11] J. G. Gutiérrez, M. Fernandez-de-Retana, and A. Bilbao-Jayo, "A Transformer-Based Approach to Analyzing Public Opinion and Political Trends," in *2025 10th International Conference on Smart and Sustainable Technologies (SpliTech)*, 2025, pp. 1–6. doi: 10.23919/SpliTech65624.2025.11091652.

[12] E. Księżniak and M. Sawiński, "Political Narratives and Misinformation During the COVID-19 Pandemic: A Comparative Analysis of Polish Political Parties on Twitter," in *International Conference on Business Information Systems*, 2025, pp. 125–133. doi: https://doi.org/10.1007/978-3-031-94193-1_10.

[13] P. S. Suryadjaja and R. Mandala, "Improving the performance of the extractive text summarization by a novel topic modeling and sentence embedding technique using sbert," in *2021 8th International Conference on Advanced Informatics: Concepts, Theory and Applications (ICAICTA)*, 2021, pp. 1–6. doi: https://doi.org/10.1109/ICAICTA53211.2021.9640295.

[14] H. U. Khan, A. Naz, F. K. Alarfaj, and N. Almusallam, "Identifying artificial intelligence-generated content using the DistilBERT transformer and NLP techniques," *Sci. Rep.*, vol. 15, no. 1, p. 20366, 2025, doi: https://doi.org/10.1038/s41598-025-08208-7.

[15] T. D. Purnomo and J. Sutopo, "Comparison of Pre-Trained Bert-Based Transformer Models for Regional Language Text Sentiment Analysis in Indonesia," *Int. J. Sci. Technol.*, vol. 3, no. 3, pp. 11–21, 2024, doi: https://doi.org/10.56127/ijst.v3i3.1739.

[16] A. S. Kazmali and A. Sayar, "Web Scraping: Legal and Ethical Considerations in General and Local Context - A Review," *Procedia Comput. Sci.*, vol. 259, no. C, pp. 1563–1572, Jun. 2025, doi: 10.1016/j.procs.2025.04.111.

[17] M. A. Abdurrazzaq, "Analisis Ulasan Aplikasi MyPertamina Menggunakan Topic Modeling dengan Latent Dirichlet Allocation," *J. Sains dan Teknol.*, vol. 10, no. 1, doi: https://doi.org/10.53008/kalbiscientia.v10i1.694.

[18] D. Maulidiya, "Topic Modelling using Latent Dirichlet Allocation (LDA) to Investigate the Latent Topics of Mathematical Creative Thinking Research in Indonesia," *J. Intell. Comput. Heal. Informatics*, vol. 3, pp. 35–46, 2022, doi: 10.26714/jichi.v3i2.11428.

[19] B. A. Tondang, Muhammad Rizqan Fadhil, Muhammad Nugraha Perdana, Akhmad Fauzi, and Ugra Syahda Janitra, "Analisis pemodelan topik ulasan aplikasi BNI, BCA, dan BRI menggunakan latent dirichlet allocation," *INFOTECH J. Inform. Teknol.*, vol. 4, no. 1, pp. 114–127, Jun. 2023, doi: 10.37373/infotech.v4i1.601.

[20] Z. Kastrati, A. L. I. S. Imran, S. M. Daudpota, M. A. Memon, and M. Kastrati, "Soaring Energy Prices : Understanding Public

Building of Informatics, Technology and Science (BITS)
Volume 7, No 2, September 2025 Page: 1367−1379
ISSN 2684-8910 (media cetak)
ISSN 2685-3310 (media online)
DOI 10.47065/bits.v7i2.8298

Engagement on Twitter Using Sentiment Analysis and Topic Modeling With Transformers," *IEEE Access*, vol. 11, no. February, pp. 26541–26553, 2023, doi: 10.1109/ACCESS.2023.3257283.

[21] H. Son and Y. E. Park, "Agenda-setting effects for covid-19 vaccination: Insights from 10 million textual data from social media and news articles using BERTopic," *Int. J. Inf. Manage.*, vol. 83, no. February, p. 102907, 2025, doi: 10.1016/j.ijinfomgt.2025.102907.

[22] H. S. Jung, H. Lee, and J. H. Kim, "Unveiling Cryptocurrency Conversations: Insights From Data Mining and Unsupervised Learning Across Multiple Platforms," *IEEE Access*, vol. 11, no. November, pp. 130573–130583, 2023, doi: 10.1109/ACCESS.2023.3334617.

[23] H. Lim, Q. Li, S. Yang, and J. Kim, "A BERT-Based Multi-Embedding Fusion Method Using Review Text for Recommendation," *Expert Syst.*, vol. 42, no. 5, p. e70041, 2025, doi: https://doi.org/10.1111/exsy.70041.

[24] Y. An, D. Kim, J. Lee, H. Oh, J. S. Lee, and D. Jeong, "Topic Modeling-Based Framework for Extracting Marketing Information From E-Commerce Reviews," *IEEE Access*, vol. 11, no. December, pp. 135049–135060, 2023, doi: 10.1109/ACCESS.2023.3337808.

[25] A. Mulia and A. R. Dzikrillah, "Analisis Perbedaan Pendapat Netizen Indonesia tentang Presiden Jokowi sebelum dan sesudah Kenaikan Harga BBM Analysis of Indonesian Netizens' Dissent on President Jokowi before and after Fuel Price Increase," *J. Comput. Eng. Syst. Sci.*, vol. 8, no. 2, pp. 318–328, 2023, doi: https://doi.org/10.24114/cess.v8i2.45319.

[26] D. Aryani, I. L. Kharisma, and A. Sujjada, "Topic Modeling of the 2024 Election Using the BERTopic Method on Detik . com News Articles," *Inf. J. Ilm. Bid. Teknol. Inf. dan Komun.*, vol. 9, no. 2, pp. 171–180, 2024, doi: https://doi.org/10.25139/inform.v9i2.8429.

[27] S. Umamaheswaran, V. Dar, E. Sharma, and J. S. Kurian, "Mapping Climate Themes from 2008-2021 - An Analysis of Business News Using Topic Models," *IEEE Access*, vol. 11, no. February, pp. 26554–26565, 2023, doi: 10.1109/ACCESS.2023.3256530.

[28] J. Song, Y. Yuan, K. Chang, B. Xu, J. Xuan, and W. Pang, "Navigating Public Sentiment in the Circular Economy through Topic Modelling and Hyperparameter Optimisation," *Energy AI*, vol. 18, no. May, p. 100433, 2024, doi: 10.1016/j.egyai.2024.100433.