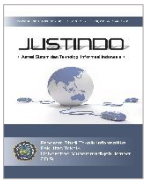
JUSTINDO
Jurnal Sistem dan Teknologi Informasi Indonesia
p-ISSN 2502-5724
e-ISSN 2541-5735

Jurnal Sistem dan Teknologi Informasi Indonesia
JUSTINDO Vol. 11 No. 1, Februari 2026, hal. 1-10
http://ejurnal.unmuhjember.ac.id/index.php/JUSTINDO/index
https://doi.org/10.32528/justindo.v11i1.4044



Analisis Topik Pada Transkip Video Pidato Politik Menggunakan Metode Latent Dirichlet Allocation (LDA)

Topic Analysis in Political Speech Video Transcripts Using the Latent Dirichlet Allocation (LDA) Method

Dhea Intan Septiara[1], Deni Arifianto*[2], Wiwik Suharso[3]

[1,2]Program Studi Teknik Informatika, Fakultas Teknik, Universitas Muhammadiyah Jember
[3]Program Studi Sistem Informasi, Fakultas Teknik, Universitas Muhammadiyah Jember
Email: [1]dheaintanseptiara24@gmail.com, [2]deniarifianto@unmuhjember.ac.id,
[3]wiwiksuharso@unmuhjember.ac.id
[*]Penulis Koresponden

Received: 30 Desember 2025 | Accepted: 27 Januari 2026 | Published: 02 Februari 2026

ABSTRAK

Pidato politik merupakan sarana penting dalam menyampaikan visi, misi, dan kebijakan pemimpin negara kepada publik. Penelitian ini bertujuan untuk mengidentifikasi dan menganalisis topik utama dalam transkrip video pidato politik Presiden Joko Widodo selama periode 2014–2024 menggunakan metode Latent Dirichlet Allocation (LDA). Data diperoleh dari 185 video pidato keterangan pers yang diunggah pada kanal YouTube Sekretariat Kabinet dan dikonversi menjadi teks menggunakan teknologi speech-to-text. Dataset dibagi menjadi data training sebanyak 81 video pidato periode 2014–2023 dan data testing sebanyak 104 video pidato tahun 2024. Proses analisis meliputi tahapan preprocessing teks, pelabelan otomatis berbasis rule-based, pelatihan model LDA, serta evaluasi menggunakan coherence score dan perplexity. Hasil penelitian menunjukkan bahwa pada data training, topik Infrastruktur dan Ekonomi merupakan tema dominan yang mencerminkan fokus pembangunan fisik dan pertumbuhan ekonomi nasional. Sementara itu, pada data testing tahun 2024, topik Kesehatan menjadi topik dengan distribusi tertinggi, diikuti oleh topik Infrastruktur, Ekonomi, Pendidikan, dan Teknologi. Topik Infrastruktur tetap menunjukkan tingkat koherensi tertinggi dengan nilai coherence score sebesar 0,85, yang mengindikasikan konsistensi semantik kata-kata penyusunnya. Penelitian ini memberikan kontribusi dalam memahami dinamika komunikasi politik lintas waktu serta menunjukkan efektivitas LDA dalam menganalisis data pidato politik berbasis transkrip video.

Kata kunci: *Pidato Politik, Transkrip Video, Latent Dirichlet Allocation, Topic Modeling, Speech-to-Text*

ABSTRACT

Political speeches are an important medium for conveying a country's leader's vision, mission, and policy directions to the public. This study aims to identify and analyze the main topics in the video transcripts of President Joko Widodo's political speeches during the 2014–2024 period using the Latent Dirichlet Allocation (LDA) method. The data consist of 185 press conference speech videos obtained from the Indonesian Cabinet Secretariat's YouTube channel and converted into text using speech-to-text technology. The dataset is divided into 81 videos from the 2014–2023 period as training data and 104 videos from 2024 as testing data. The analysis process includes text preprocessing, rule-based automatic labeling, LDA model training, and evaluation using coherence score and perplexity. The results show that in the training data, the topics of Infrastructure and Economy are the dominant topics, reflecting the government's focus on physical development and economic growth. In contrast, in the 2024 testing data, Healthcare emerges as the most dominant topic, followed by the topics of Infrastructure, Economy, Education, and Technology. The Infrastructure topic consistently achieves the highest coherence score of 0.85, indicating strong semantic consistency among its constituent terms. This study contributes to understanding the temporal dynamics of political communication and demonstrates the effectiveness of LDA in analyzing political speech data derived from video transcripts.

Keywords: *Political Speech, Video Transcript, Latent Dirichlet Allocation, Topic Modeling, Speech-to-Text*

1. Pendahuluan

Pidato politik merupakan media utama bagi para pemimpin untuk menyampaikan visi, misi, dan kebijakan kepada publik. Sebagai media komunikasi, pidato politik tidak hanya mencerminkan agenda pemerintahan tetapi juga berperan dalam membentuk opini publik dan mempengaruhi kebijakan nasional. Perkembangan media digital telah mengubah pola konsumsi informasi politik masyarakat. Platform berbasis video seperti YouTube menjadi salah satu sumber utama komunikasi politik. Data global menunjukkan bahwa konten politik digital mengalami peningkatan signifikan dalam satu dekade terakhir, seiring meningkatnya akses internet dan penggunaan media sosial oleh masyarakat. Pidato politik yang disampaikan melalui media digital terbukti memiliki pengaruh terhadap pembentukan opini publik, persepsi kebijakan, serta tingkat kepercayaan masyarakat terhadap pemerintah. Dalam konteks Indonesia, pidato Presiden yang disiarkan secara daring menjadi sarana strategis untuk menyampaikan agenda pembangunan nasional kepada masyarakat. Namun, besarnya volume data pidato politik berbasis video menimbulkan tantangan dalam melakukan analisis secara manual dan objektif. Dengan mengkaji setiap topik pidato politik, dapat dipahami bagaimana para pemimpin negara menanggapi isu-isu penting, mengenali perubahan kebijakan, dan menilai konsistensi serta perubahan agenda politik (Anggai et al., 2024). Untuk dapat memahami dan menginterpretasikan pidato politik dengan baik, diperlukan pendekatan analisis yang dapat mengelola data yang rumit dan dalam jumlah besar.

Video-video pidato politik yang diunggah di *YouTube* menjadi sumber data yang kaya untuk

dianalisis. Namun, untuk dapat menganalisis konten tersebut secara komputasional, diperlukan konversi dari format video ke teks. Proses ini melibatkan penggunaan teknologi *speech-to-text* yang dapat mengubah audio dari video *YouTube* menjadi transkrip teks. Teknologi ini telah berkembang pesat dalam beberapa tahun terakhir, dengan akurasi yang semakin tinggi berkat kemajuan dalam bidang *Natural Language Processing* (NLP) dan *Deep Learning* (Amodei et al., 2016; Hannun et al., 2014). Transkrip teks ini kemudian dapat diproses lebih lanjut menggunakan metode analisis teks komputasional seperti *Latent Dirichlet Allocation* (LDA) untuk mengidentifikasi topik-topik utama yang dibahas dalam pidato politik.

Dalam penelitian ini, Latent Dirichlet Allocation (LDA) digunakan sebagai metode utama untuk membantu mengolah dan menganalisis data video pidato politik yang telah dikonversi ke dalam bentuk teks (Blei et al., 2003). LDA berperan sebagai teknik pemodelan topik yang mampu mengidentifikasi pola tematik tersembunyi dalam kumpulan dokumen teks secara otomatis, tanpa memerlukan pelabelan manual. Melalui pendekatan probabilistik, LDA memetakan hubungan antara kata, topik, dan dokumen sehingga setiap transkrip pidato dapat direpresentasikan sebagai kombinasi beberapa topik dengan proporsi tertentu (Griffiths dan Steyvers 2004). Dengan demikian, penggunaan LDA memungkinkan untuk mengekstraksi informasi substantif dari data video pidato politik dalam skala besar secara sistematis, efisien, dan objektif, serta membantu memahami fokus dan dinamika komunikasi politik yang disampaikan melalui media video.

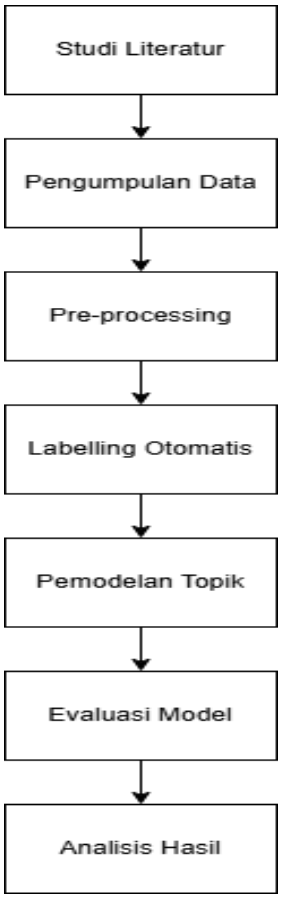
Analisis teks menggunakan metode komputasi seperti *Latent Dirichlet Allocation* (LDA) menjadi solusi yang relevan. LDA adalah algoritma *machine learning* yang digunakan untuk mengidentifikasi topik tersembunyi dalam dokumen teks dengan asumsi bahwa setiap dokumen merupakan campuran beberapa topik, dan setiap topik adalah distribusi kata-kata tertentu. Metode ini memungkinkan analisis cepat dan objektif terhadap data teks politik dalam skala besar, tanpa memerlukan anotasi manual (Bianchi et al., 2021; Hannigan et al., 2019; Yunita et al., 2022). Dengan LDA, peneliti dapat mengelompokkan kata-kata yang sering muncul bersama, sehingga membantu mengidentifikasi isu-isu utama yang diangkat dalam pidato politik.

Pentingnya penelitian ini tidak hanya terletak pada kemampuan menganalisis teks secara efisien, tetapi juga dalam memberikan wawasan baru tentang pola komunikasi politik. Dengan memanfaatkan LDA, penelitian ini dapat berkontribusi dalam memahami dinamika sosial dan politik melalui data teks, serta dapat digunakan untuk mengembangkan sistem analisis teks yang lebih canggih, seperti prediksi tren politik atau klasifikasi dokumen otomatis (Bianchi et al., 2021; Yunita et al., 2022; Zhang et al., 2022).

Pada penelitian ini memiliki urgensi dalam memberikan wawasan baru tentang pola komunikasi politik di Indonesia. Dengan memanfaatkan LDA, penelitian ini dapat membantu memahami isu-isu penting yang diangkat oleh para politisi dan memberikan kontribusi dalam mendukung pengambilan keputusan berbasis data. Hasil penelitian ini dapat memberikan wawasan yang mendalam tentang dinamika komunikasi politik di Indonesia dan juga akan berkontribusi pada konteks pemodelan topik menggunakan LDA.

2. Metode Penelitian

2.1. Prosedur Penelitian



Gambar 1. Alur Diagram Penelitian

2.2. Pengumpulan Data

Pengumpulan data dalam penelitian ini difokuskan pada transkrip pidato politik Presiden Joko Widodo yang diambil dari Channel *YouTube* Sekretariat Kabinet. Sebanyak 185 video pidato dengan rentang waktu 2014–2023 sebanyak 81 video akan diambil sebagai data training dan video pada tahun 2024 sebanyak 104 video akan diambil sebagai data testing. Pengumpulan data video menggunakan web download video yaitu SoundType.Ai.

2.3. Pre-processing

Tahapan ini melibatkan transformasi data teks mentah menjadi format yang dapat diproses oleh algoritma LDA. Ekstraksi Audio dari Video menggunakan tools *FFmpeg*. Dari video yang telah didownload selanjutnya di ekstrak menjadi audio dengan bantuan tools FFmpeg**.** Audio yang telah diekstrak kemudian diproses menggunakan Soundtype.Ai untuk menghasilkan transkip teks pidato. Setelah transkip selesai, selanjutnya tahapan data *cleaning* yaitu menghapus karakter khusus, angka, atau simbol yang tidak relevan dan memastikan bahwa teks yang diproses hanya mengandung kata-kata bermakna yang relevan dengan analisis topik pidato. Selanjutnya *case folding* yaitu mengubah semua huruf dalam teks menjadi huruf kecil (*lowercase*) untuk menghindari duplikasi kata yang sama dengan format huruf berbeda. *Stopword Removal* untuk menghapus kata-

kata umum yang tidak memiliki makna signifikan dalam analisis topik, seperti "dan", "atau", "di", "yang", dan sebagainya. Lalu *stemming* untuk mengubah kata ke bentuk dasarnya untuk mengurangi variasi kata yang memiliki makna serupa. Tahapan *stemming* dilakukan menggunakan *library* Sastrawi pada bahasa pemrograman *Python*, yang mengimplementasikan algoritma *stemming* Bahasa Indonesia berbasis aturan (*rule-based*). *Library* ini digunakan untuk mengurangi variasi kata akibat imbuhan, sehingga meningkatkan kualitas representasi teks dalam proses pemodelan topik menggunakan LDA. Selanjutnya yaitu tokenisasi yaitu memecah teks pidato menjadi kata-kata individual *(tokens*).

2.4. Labelling Otomatis

Labelling otomatis digunakan untuk mengategorikan topik berdasarkan kata-kata yang dihasilkan dari model LDA. *Rule-based labelling* diterapkan dengan menggunakan modul *Regular Expression (re)*, *NLTK* dan *Pandas* pada *Python* untuk mencocokkan kata-kata hasil pemodelan LDA dengan daftar kata kunci yang telah ditentukan.

2.5. Pemodelan Topik dengan LDA

Inisialisasi parameter yang digunakan yaitu *Hyperparameter Alpha* dan *Beta*. Dalam hal ini, *Alpha* digunakan untuk mengontrol distribusi topik per dokumen. α = 0,1 dan *Beta* (atau β) untuk mengontrol distribusi kata per topik. Nilai awal diatur β = 0.01 untuk memprioritaskan *sparsity* kata. Selanjutnya, penggunaan *library Gensim* di *Google Colab* untuk melatih model LDA pada dataset yang telah diproses. Maka model LDA akan mempelajari distribusi kata dalam setiap topik dan distribusi topik dalam setiap dokumen dengan berdasar pada perhitungan *Gibbs Sampling* berikut:

$$P(z_i = k) \propto \frac{n_{k,w_i}^{(-i)} + \beta}{n_k^{(-i)} + W\beta} \times \frac{n_{d,k}^{(-i)} + \alpha}{n_d^{(-i)} + K\alpha} \quad (1)$$

2.6. Evaluasi Model

Evaluasi model dilakukan dengan menggunakan *Coherence Score* dimana, topik 1 mencapai nilai koherensi sempurna (1.0) karena kata kuncinya saling terkait secara semantik dan konsisten muncul bersama dalam konteks kebijakan infrastruktur yang mengacu pada rumus dalam persamaan 2 yaitu:

$$\mathrm{NPMI}(w_i, w_j) = \frac{\log \frac{P(w_i, w_j)}{P(w_i) \cdot P(w_j)}}{-\log P(w_i, w_j)}$$

$$Coherence\ Score = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \mathrm{NPMI}(w_i, w_j) \quad (2)$$

Selanjutnya menggunakan perplexity yang menunjukkan bahwa model memiliki kemampuan prediktif yang baik dalam mengestimasi distribusi kata pada dokumen baru yang mengacu pada persamaan 3 yaitu:

$$Perplexity(D) = \exp\left(-\frac{\sum_{d=1}^{M} \log p(w_d)}{\sum_{d=1}^{M} N_d}\right) \quad (3)$$

2.7. Analisis Hasil

Hasil analisis topik menggunakan metode LDA pada transkrip pidato politik Presiden Joko Widodo (2014–2024) menunjukkan bahwa topik *Infrastruktur* merupakan tema dominan dengan distribusi sebesar 35% dan tingkat koherensi tinggi (1.0), mencerminkan fokus kuat pemerintah pada pembangunan fisik seperti jalan tol dan pelabuhan. Sebaliknya, topik lain cenderung kurang spesifik dan memiliki koherensi lebih rendah (0.6), karena didominasi oleh kata-kata umum seperti "baik". Temuan ini menegaskan bahwa LDA mampu mengidentifikasi pola tematik secara efektif, meskipun tingkat akurasinya bervariasi antar topik.

3. Hasil dan Pembahasan

3.1. Dataset

Analisis dilakukan terhadap 185 transkrip video pidato politik keterangan pers Presiden Joko Widodo periode 2014–2024 yang bersumber dari *Channel YouTube* Sekretariat Kabinet dengan 1.250 rata–rata kata/dokumen (81 video untuk data training tahun 2014–2023 dan 104 video untuk data testing tahun 2024). 5 topik utama yang diidentifikasi adalah Pendidikan, Ekonomi, Kesehatan, Infrastruktur, dan Teknologi.

3.2. Pre-processing Data

Tahapan *preprocessing* ini bertujuan membersihkan dan menormalisasi teks agar bisa dianalisis oleh model LDA. Adapun prosesnya meliputi *data cleaning* yang bertujuan menghapus simbol, angka, dan karakter khusus yang tidak relevan, c*ase folding* untuk mengubah seluruh teks menjadi huruf kecil, s*topword removal* untuk menghapus kata-kata umum (e.g., "yang", "dan"), s*temming* untuk mengubah kata ke bentuk dasarnya dan tokenisasi untuk memecah kalimat menjadi token (unit kata). Contoh hasil pre-processing dapat dilihat pada Tabel 1.

Tabel 1. Data *Pre-processing*

Time Stamp	Data sebelum *Pre-Processing*	Data setelah *Pre-processing*
0:58	Dan kita lihat tadi ada tiga kapal kita yang sudah ada di sekitar lokasi.	***Data Cleaning***: dan kita lihat tadi ada tiga kapal kita yang sudah ada di sekitar lokasi
		Case Folding: dan kita lihat tadi ada tiga kapal kita yang sudah ada di sekitar lokasi. ***Stopword Removal***: lihat tadi tiga kapal lokasi ***Stemming***: lihat tadi tiga kapal lokasi ***Tokenized***: ['lihat', 'tadi', 'tiga', 'kapal', 'lokasi']

3.3. Labelling

Pelabelan otomatis berbasis *rule (regular expression)* dilakukan untuk memberikan interpretasi semantik pada hasil LDA yang berfungsi memperkuat temuan bahwa pidato Presiden Jokowi selama 2014–2023 banyak menekankan pembangunan infrastruktur dan ekonomi sebagai prioritas utama. Sementara itu, topik pendidikan, kesehatan, dan teknologi cenderung muncul sebagai tema pelengkap atau responsif terhadap isu tertentu Penggunaan periode masa jabatan Presiden ke-7 Republik Indonesia sebagai dasar dalam proses pelabelan topik didasarkan pada pertimbangan konsistensi konteks kebijakan dan narasi komunikasi politik. Masa jabatan presiden merepresentasikan satu kesatuan kepemimpinan dengan visi, misi, dan agenda pembangunan yang relatif berkelanjutan, sehingga analisis topik yang dihasilkan dapat menggambarkan pola komunikasi politik secara lebih utuh dibandingkan pembatasan berbasis rentang tahun tertentu, seperti 2019–2023. Dimana diperoleh hasil Labelling pada Tabel 2.

Tabel 2. Hasil Labelling Data Training

Tahun	Video	Topik
2014	1	Infrastruktur
2014	2	Teknologi
2014	3	Teknologi
- - -	- - -	- - -
2023	81	Kesehatan

3.4. Hasil Pelatihan Pemodelan LDA

3.4.1 Inisialisasi Parameter

Model LDA menggunakan 5 topik: Pendidikan, Ekonomi, Kesehatan, Infrastruktur, Teknologi. Maka Jumlah Topik (k) = 5. Secara data, setelah dilakukan pengolahan awal terhadap transkrip pidato, ditemukan bahwa banyak kata yang berkaitan dengan kelima topik tersebut muncul secara berulang dan konsisten. Contohnya, kata “sekolah” dan “guru” sering muncul pada konteks pendidikan, “jalan” dan “jembatan” pada konteks infrastruktur, dan lainnya. Hal ini menunjukkan bahwa lima topik tersebut memang relevan untuk dianalisis lebih lanjut dengan model LDA.

3.4.2 Pelatihan Model

Proses *Gibbssampling* ini bertujuan untuk mengestimasi distribusi kata terhadap topik dan distribusi topik terhadap dokumen. Setiap kata dalam dokumen dianalisis kemungkinan keterkaitannya dengan suatu topik berdasarkan rumus yang telah dijelaskan dalam persamaan 1. Dari persamaan tersebut diperoleh perhitungan yang diterapkan juga pada seluruh data training sehingga menghasilkan tabel 3. dibawah berikut:

Tabel 3. Hasil Gibbs Sampling Data Training

Tahun	Video	Probabilitas
2014	1	"kurikulum (0.266), pendidikan (0.200), murid (0.133), sekolah (0.133), belajar (0.066), internet (0.066), vaksin (0.066), uang (0.066)"
- - -	- - -	- - -
2023	81	"proyek (0.3333), jalan (0.2667), pelabuhan (0.0667), bandara (0.0667), jembatan (0.0667), teknologi (0.0667), sekolah (0.0667), murid (0.0667)"

Setelah *Gibbssampling* selesai, kita menentukan kata kunci dominan topik. Nilai setelah iterasi pelatihan LDA, model mengoptimalkan distribusi sehingga menghasilkan probabilitas akhir. Maka kata-kata dengan probabilitas tertinggi diekstrak sebagai kata kunci dominan. Berikut ini contoh perhitungan untuk menghasilkan kata kunci dominan topik:

$$\varphi_{kw} = \frac{n_{kw}+\beta}{n_k+V\beta} \tag{4}$$

Dari persamaan 4 dilakukan pada kata kunci lainnya sehingga menghasilkan tabel 4. berikut:

Tabel 4. Kata Kunci Dominan Topik Data Training

Topik	Kata Kunci Dominan (Probabilitas)
Pendidikan	0.152 ”kurikulum”, 0.155 ”sekolah”, 0.149 ”pendidikan”, 0.16 ”murid”, 0.167 ”belajar”
Ekonomi	0.117 ”uang”, 0.145 ”pasar”, 0.124 ”invetasi”, 0.123 ”ekonomi”, 0.145 ”anggaran”
Kesehatan	0.13 "dokter", 0.176 ”pandemi”, 0.143 ”rumah sakit”, 0.133 ”vaksin”, 0.152 ”obat”
Infrastruktur	0.119 "pelabuhan", 0.146 "jalan", 0.129 "proyek", 0.15 "jembatan", 0.123 "bandara"
Teknologi	0.071 "AI", 0.073 "teknologi", 0,071, ”inovasi”, 0,074, ”internet”, 0,075 ”digitalisasi”

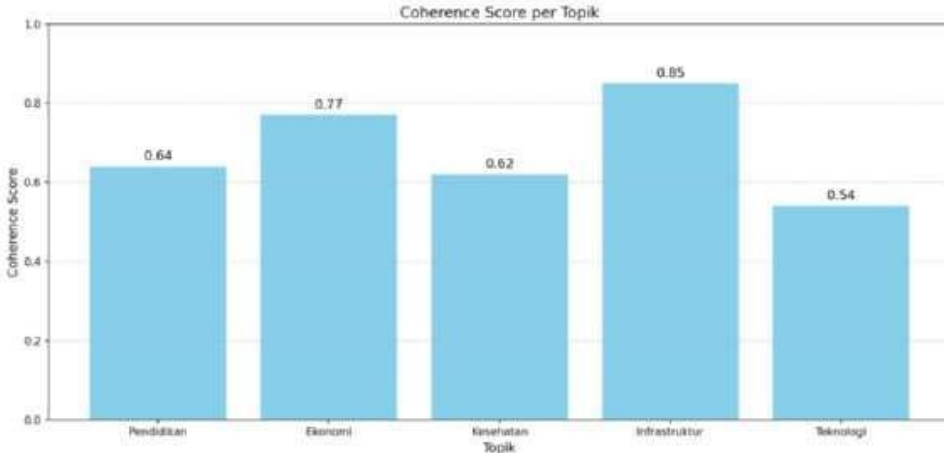
Model LDA berhasil diinisialisasi dengan 5 topik utama: Pendidikan, Ekonomi, Kesehatan, Infrastruktur, dan Teknologi. Berdasarkan hasil *Gibbs Sampling* terhadap 81 transkrip pidato politik Presiden Joko Widodo, distribusi topik menunjukkan variasi relevansi.

3.4.3 *Coherence Score*

Perhitungan *coherence score* dilakukan untuk mengevaluasi koherensi semantik antar kata dalam setiap topik, yang diukur sebagai rata-rata NPMI untuk semua pasangan kata dalam topik tersebut seperti pada persamaan 2 & 3. Dimana diperoleh hasil coherence score pada tabel 5. berikut:

Tabel 5. Hasil Coherence score

Topik	**Conherence score**
Pendidikan	0,64
Ekonomi	0,77
Kesehatan	0,62
Infrastruktur	0,85
Teknologi	0,54



Gambar 2. Visualisasi Hasil Coherence score

Berdasarkan hasil perhitungan pada Tabel 5 dan Gambar 2 diatas menunjukkan bahwa *Coherence score* yang dihitung untuk setiap topik, menunjukkan tingkat koherensi semantik antar kata kunci dalam topik tersebut. *Coherence score* membuktikan model LDA berhasil mengidentifikasi topik-topik koheren dalam pidato politik, dengan infrastruktur sebagai topik paling konsisten (nilai 0.85).

3.5. Evaluasi Model Pada Data Testing

3.5.1 Hasil *Gibbssampling*

Setelah model dilatih menggunakan data 2014–2023, dilakukan validasi pada data testing, yaitu 104 video pidato tahun 2024 dimana validasi bertujuan untuk menguji kemampuan generalisasi model terhadap dokumen baru yang tidak dilibatkan dalam proses pelatihan. Dari persamaan 1 diperoleh perhitungan yang diterapkan juga pada seluruh data testing yang dapat dilihat pada Tabel 6 berikut:

Tabel 6. Hasil Gibbs Sampling Data Training

Tahun	Video	Probabilitas
2024	1	"AI (0.1333), bandara (0.0667), belajar (0.0667), digitalisasi (0.0667), inovasi (0.0667), internet (0.0667), investasi (0.0667), jalan (0.0667), jembatan (0.0667), pelabuhan (0.0667), proyek (0.0667), sekolah (0.1333), teknologi (0.0667)"
…	…	…
2024	10	"bandara (0.0667), digitalisasi (0.0667), dokter (0.0667), jalan (0.1333), jembatan (0.2000), murid (0.0667), obat (0.0667), pelabuhan (0.0667), proyek (0.2000), uang (0.0667)"

3.5.2 Hasil *Labelling* Dengan Model LDA

Pelabelan dilakukan untuk mengidentifikasi topik dominan pada setiap dokumen transkrip video pidato Presiden Joko Widodo tahun 2024. Proses ini menggunakan pendekatan dari model Latent Dirichlet Allocation (LDA) yang telah dibangun dan dilatih sebelumnya.

Tabel 7. Hasil *Labelling* Data Model LDA

Tahun	**Video**	**Token**	**Label**
2024	82	['pelabuhan', 'pelabuhan', 'bandara', 'proyek', 'proyek', 'vaksin', 'inovasi', 'jalan', 'jembatan', 'bandara', 'obat', 'jalan']	Infrastruktur
2024	83	['internet', 'uang', 'AI', 'teknologi', 'internet', 'sekolah', 'digital', 'ekonomi']	Teknologi
- - -	- - -	- - -	
2024	185	['bandara', 'dokter', 'proyek', 'investasi', 'inovasi', 'pelabuhan', 'jembatan', 'obat', 'jembatan', 'jalan', 'jalan', 'pelabuhan', 'proyek', 'inovasi']	Infrastruktur

Berdasarkan hasil pelabelan otomatis terhadap 104 dokumen tahun 2024, ditemukan bahwa sebagian besar dokumen dikategorikan dalam topik Infrastruktur. Hal ini terlihat dari frekuensi dominasi kata-kata seperti *jalan*, *jembatan*, *proyek*, dan *pelabuhan* yang cukup konsisten muncul dalam hampir seluruh dokumen.

3.6. Analisis Hasil

3.6.1 Perplexity

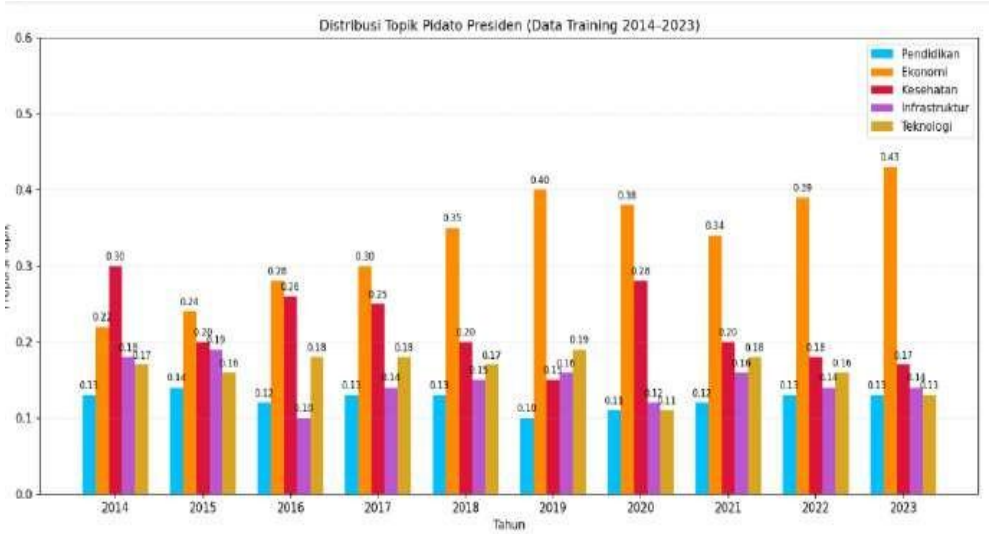
Nilai perplexity digunakan untuk mendukung evaluasi performa model LDA yang telah dilatih menggunakan 104 transkrip video pidato politik Presiden Joko Widodo periode 2024. Dimana diperoleh hasil pada Tabel 8:

Tabel 8. Hasil Data Perplexity

Topik	Coherence Score
Pendidikan	13,22
Ekonomi	12,74
Kesehatan	12,77
Infrastruktur	11,19
Teknologi	12,11

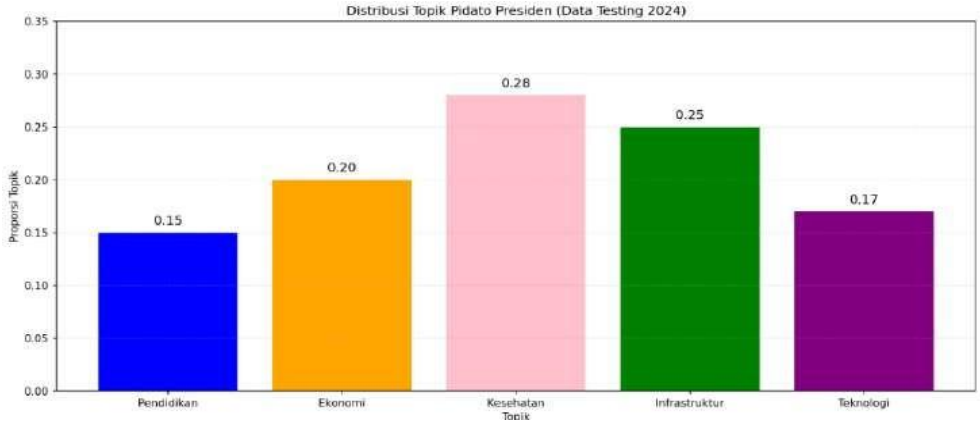
Hasil perplexity menunjukkan bahwa model LDA cukup efektif dalam mengidentifikasi topik (Infrastruktur, Ekonomi, Kesehatan, Pendidikan, Teknologi. Perplexity keseluruhan 11,94 mencerminkan performa model yang konsisten.

3.6.2 Frekuensi Distribusi Topik



Gambar 3. Distribusi Topik Pidato Presiden (Data Training 2014-2023)

Pada Gambar 3.6.2 Diperoleh hasil analisis data training yang menunjukkan bahwa topik *Ekonomi* mendominasi pidato Presiden Jokowi, terutama pada 2018–2023 dengan puncak 0,43 di tahun 2023, mencerminkan fokus pada investasi dan pembangunan ekonomi. topik Infrastruktur menunjukkan pola dominasi yang relatif konsisten sejak tahun 2015 hingga 2023. Penyebutan tahun 2018 pada versi sebelumnya hanya dimaksudkan untuk menekankan periode ketika intensitas topik tersebut mengalami penguatan yang lebih signifikan, bukan sebagai awal kemunculan dominansi. Secara keseluruhan, hasil visualisasi memperlihatkan bahwa topik Infrastruktur telah menjadi tema utama dalam pidato politik Presiden Joko Widodo sepanjang sebagian besar periode masa jabatan, yang mencerminkan kesinambungan agenda pembangunan nasional. Topik *Infrastruktur* juga konsisten muncul (0,15–0,20 per tahun), menggarisbawahi pentingnya proyek strategis seperti tol Trans-Jawa. Lonjakan topik *Kesehatan* terjadi pada 2020 (0,28), sejalan dengan awal pandemi COVID-19. Sementara itu, topik *Pendidikan* lebih menonjol di awal periode (2014–2016), dan *Teknologi* meningkat menjelang akhir masa jabatan seiring dorongan transformasi digital.



Gambar 4. Distribusi Topik Pidato Presiden (Data Testing 2024)

Namun pada tahun 2024 yang ditunjukkan oleh Gambar 4 terjadi pergeseran fokus dalam pidato Presiden dengan topik *Kesehatan* menjadi dominan (0,28), diikuti *Infrastruktur* (0,25) dan *Ekonomi* (0,20), mencerminkan penekanan pada penguatan layanan publik menjelang akhir masa jabatan. *Pendidikan* dan *Teknologi* tetap muncul meski dengan proporsi lebih rendah. Selain itu, distribusi topik tahun 2024 menunjukkan disrupsi dengan dominasi *Pendidikan* (35%), menandai peralihan narasi dari pembangunan fisik ke investasi sumber daya manusia. Temuan ini menunjukkan perubahan strategi pembangunan nasional menuju pendekatan berbasis *human capital*.

4. Kesimpulan

Penelitian ini mengungkap dinamika perubahan prioritas kebijakan pemerintah melalui analisis distribusi topik 185 pidato politik Presiden Joko Widodo (2014–2024). Pemanfaatan transkrip video sebagai sumber data temporal. Hasil analisis menunjukkan adanya pergeseran strategis dalam fokus narasi pidato politik Presiden Joko Widodo. Pada periode 2014–2023, topik Ekonomi dan Infrastruktur menjadi dua tema dominan dengan rerata distribusi masing-masing sebesar 35% dan 25%, yang mencerminkan orientasi pembangunan berbasis pertumbuhan ekonomi dan pembangunan fisik nasional. Topik Teknologi juga muncul sebagai bagian dari strategi pendukung transformasi digital, meskipun dengan proporsi yang relatif lebih rendah . Namun pada tahun 2024, terjadi perubahan signifikan dengan meningkatnya dominasi topik Kesehatan (28%), diikuti oleh Infrastruktur (25%) dan Ekonomi (20%). Perubahan ini merefleksikan pergeseran paradigma komunikasi politik dari penekanan pada pembangunan fisik (*hard infrastructure*) ke arah penguatan layanan publik dan keberlanjutan system Kesehatan nasional.

Metode LDA terbukti efektif mengidentifikasi pola tematik dengan akurasi memadai diperoleh *coherence score* sebesar 0.85, yang menandakan bahwa kata-kata dalam masing-masing topik memiliki keterkaitan semantik yang sangat kuat. Selain itu, nilai *perplexity* sebesar 11.19 menunjukkan bahwa model memiliki kemampuan generalisasi yang sangat baik terhadap data baru.

Nilai *coherence score* sebesar 0,85 pada model LDA menunjukkan tingkat koherensi semantik yang tinggi antar kata dalam setiap topik, sehingga topik yang dihasilkan dapat diinterpretasikan secara jelas dan konsisten. Namun, nilai *coherence score* tidak merepresentasikan akurasi dalam pengertian klasifikasi atau prediksi, melainkan digunakan sebagai indikator kualitas dan keterpahaman topik. Sementara itu, nilai perplexity sebesar 11,19 mencerminkan kemampuan model dalam merepresentasikan distribusi kata pada data uji secara lebih baik dibandingkan model dengan nilai *perplexity* yang lebih tinggi. Meskipun nilai *perplexity* yang rendah mengindikasikan performa pemodelan yang lebih baik, metrik ini tidak secara langsung mengukur kemampuan generalisasi dalam arti prediktif, melainkan menunjukkan kesesuaian model terhadap data yang dianalisis. Oleh karena itu, evaluasi model LDA dalam penelitian ini difokuskan pada kualitas topik dan konsistensi tematik, bukan pada akurasi atau performa prediksi sebagaimana pada model *supervised learning.*

Daftar Pustaka

Amodei, D., Ananthanarayanan, S., Anubhai, R., Bai, J., Battenberg, E., Case, C., Casper, J., Catanzaro, B., Cheng, Q., Chen, G., Chen, J., Chen, J., Chen, Z., Chrzanowski, M., Coates, A., Diamos, G., Ding, K., Du, N., Elsen, E., … Zhu, Z. (2016). Deep speech 2: End-to-end speech recognition in English and Mandarin. 33rd International Conference on Machine Learning, ICML 2016, 1, 312–321.

Blei, David M., Andrew Y. Ng, dan Michael I. Jordan. 2003. "Latent Dirichlet Allocation: Extracting Topics from Software Engineering Data." The Art and Science of Analyzing Software Data 3:139–59. doi: 10.1016/B978-0-12-411519-4.00006-9.

Anggai, S., Tukiyat, Rivai, A. K., & Zain, R. M. (2024). Ekstraksi Topik dalam Dataset Menggunakan Teknik Pemodelan Topik. Jurnal Ilmu Komputer, 2, 78–96.

Bianchi, F., Terragni, S., Hovy, D., Nozza, D., & Fersini, E. (2021). Cross-lingual contextualized topic models with zero-shot learning. EACL 2021 - 16th Conference of the European Chapter of the Association for Computational Linguistics, Proceedings of the Conference, 1676–1683. https://doi.org/10.18653/v1/2021.eacl-main.143

Griffiths, Thomas L., dan Mark Steyvers. 2004. "Finding scientific topics." Proceedings of the National Academy of Sciences of the United States of America 101(SUPPL. 1):5228–35. doi: 10.1073/pnas.0307752101.

Hannigan, T., Haans, R. F. J., Vakili, K., Tchalian, H., Glaser, V. L., Wang, M., Kaplan, S., & Jennings, P. D. (2019). Topic modeling in management research. Academy of Management Annals, 13(2), 586–632.

Hannun, A., Case, C., Casper, J., Catanzaro, B., Diamos, G., Elsen, E., Prenger, R., Satheesh, S., Sengupta, S., Coates, A., & Ng, A. Y. (2014). Deep Speech: Scaling up end-to-end speech recognition. 1–12. http://arxiv.org/abs/1412.5567

Matira, Yayang, dan Iman Setiawan. 2023. "Pemodelan Topik pada Judul Berita Online Detikcom Menggunakan Latent Dirichlet Allocation." Estimasi: Journal of Statistics and Its Application 4(1):53–63. doi: 10.20956/ejsa.vi.24843.

Yunita, R. D., Rozikin, C., & Jajuli, M. (2022). Implementasi Metode Linear Discriminan Analysis Untuk Klasifikasi Biji Kopi. Jurnal Teknologi Informatika Dan Komputer, 8(1), 27–39. https://doi.org/10.37012/jtik.v8i1.664

Zhang, Z., Fang, M., Chen, L., & Namazi-Rad, M. R. (2022). Is Neural Topic Modelling Better than Clustering? An Empirical Study on Clustering with Contextual Embeddings for Topics. NAACL 2022 - 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, 3886–3893. https://doi.org/10.18653/v1/2022.naacl-main.285