JURNAL ILMU KOMPUTER
Volume 2 Nomor 1, Juli, 2024
ISSN: 3031-125X (ONLINE)
Penerbit: Program Studi Teknik Informatika S-2 Universitas Pamulang

# Ekstraksi Topik dalam Dataset Menggunakan Teknik Pemodelan Topik

* Sajarwo Anggai [1], Tukiyat [2], Abu Khalid Rivai [3], Rafi Mahmud Zain [4]

[1,2,3,4)] Teknik Informatika, Program Pascasarjana, Universitas Pamulang, Tangerang Selatan, Banten

Email: [1] sajarwo@gmail.com, [2] dosen02711@unpam.ac.id, [3] dosen01591@unpam.ac.id, [4] rafizain777@gmail.com

***ABSTRACT***

*The issue in this research is the lack of understanding regarding the main topics and their changes in speeches and media publications related to President Joko Widodo. This study aims to identify, analyze, and predict changes in key topics within speeches, statements, and media publications related to President Joko Widodo using Latent Dirichlet Allocation (LDA) topic modeling techniques. The research employs a quantitative approach to analyze President Joko Widodo's speech texts using the Latent Dirichlet Allocation (LDA) method. The process began with scraping documents from the official website of the Republic of Indonesia's Secretariat, resulting in 5,988 speech transcripts from October 20, 2014, to March 2, 2024. Text preprocessing involved tokenization, stopword removal, and stemming/ lemmatization, followed by dictionary-term formation. The findings indicate that the model with k=16 has the highest coherence (0.554) and the best perplexity at k=21 (-13.130). The main topics identified include Nationalism and National Values, Regional Government, and Education and Children. Topic visualization with PyLDAvis aids in the exploration and identification of topics, providing insights for decision-making and policy development. To enhance understanding of topic changes, it is recommended to conduct trend analysis on key topics over time. This will help identify how President Joko Widodo's priorities shift and respond to new issues. By monitoring these trends, the research can provide deeper insights into the evolution of policies and the President's focus.*

*Keywords: Dataset, Evaluation, Latent Dirichlet Allocation, Topic Modeling.*

**ABSTRAK**

Permasalahan dalam penelitian ini adalah kurangnya pemahaman tentang topik-topik utama dan perubahannya dalam pidato serta publikasi media terkait Presiden Joko Widodo. Penelitian ini bertujuan untuk mengidentifikasi, menganalisis, dan memprediksi perubahan topik utama dalam pidato, pernyataan, dan publikasi media terkait Presiden Joko Widodo menggunakan teknik pemodelan topik LDA. Penelitian ini menggunakan pendekatan kuantitatif untuk menganalisis teks pidato Presiden Joko Widodo dengan metode Latent Dirichlet Allocation (LDA). Proses dimulai dengan scraping dokumen dari situs Sekretariat Negara Republik Indonesia, menghasilkan 5988 naskah pidato dari 20 Oktober 2014 hingga 2 Maret 2024. Preprocessing teks melibatkan tokenisasi, penghapusan stopwords, dan stemming/lemmatization, diikuti oleh pembentukan dictionary-term. Temuan penelitian menunjukkan bahwa model dengan k=16 memiliki koherensi tertinggi (0.554) dan perplexity terbaik pada k=21 (-13.130). Topik utama mencakup Nasionalisme dan Nilai-nilai Kebangsaan, Pemerintahan Daerah, serta Pendidikan dan Anak. Visualisasi topik dengan PyLDAvis membantu dalam eksplorasi dan identifikasi topik, memberikan wawasan untuk pengambilan keputusan dan pengembangan kebijakan. Untuk meningkatkan pemahaman mengenai perubahan topik, disarankan untuk melakukan analisis tren terhadap topik-topik utama dari waktu ke waktu. Ini akan membantu mengidentifikasi bagaimana prioritas Presiden Joko Widodo berubah dan merespons isu-isu baru. Dengan memantau tren ini, penelitian dapat memberikan wawasan yang lebih mendalam mengenai evolusi kebijakan dan fokus Presiden.

Kata Kunci: Dataset, Evaluasi, *Latent Dirichlet Allocation*, Topik Model.

## 1. PENDAHULUAN

Konsep pidato presiden menjadi isu penting karena pidato tersebut tidak hanya mencerminkan kebijakan dan prioritas pemerintah, tetapi juga mempengaruhi persepsi publik dan arah kebijakan nasional. Pidato presiden sering kali menjadi sarana untuk menyampaikan visi dan strategi pemerintah, serta menjawab tantangan dan perubahan sosial-politik yang sedang berlangsung. Dengan menganalisis pidato presiden, kita dapat memahami bagaimana pemimpin negara merespons isu-isu penting, mengidentifikasi tren perubahan dalam kebijakan, dan menilai konsistensi serta perubahan dalam agenda politiknya. Selain itu, pidato presiden juga berfungsi sebagai alat komunikasi strategis yang dapat mempengaruhi opini publik dan membentuk narasi nasional. Oleh karena itu, pemahaman mendalam tentang konsep dan konten pidato presiden sangat penting untuk menilai efektivitas komunikasi politik dan dampaknya terhadap masyarakat serta kebijakan negara.

Permasalahan yang diangkat dalam penelitian ini adalah kurangnya pemahaman tentang topik-topik utama dan perubahannya dalam pidato serta publikasi media terkait Presiden Joko Widodo. Meskipun terdapat ketersediaan publik yang luas terhadap pidato dan pernyataan beliau, masih terdapat kesenjangan dalam menganalisis dan melacak bagaimana fokus komunikasi ini berubah seiring waktu. Penelitian ini bertujuan untuk mengisi kesenjangan tersebut dengan menggunakan teknik pemodelan topik Latent Dirichlet Allocation (LDA) untuk mengidentifikasi, menganalisis, dan memprediksi perubahan topik utama dalam kumpulan pidato, pernyataan, dan publikasi media Presiden Widodo. Tujuannya adalah untuk memberikan pandangan menyeluruh tentang pergeseran tematik dan prioritas yang tercermin dalam komunikasi publiknya dari 20 Oktober 2014 hingga 2 Maret 2024, sehingga memberikan wawasan berharga mengenai evolusi fokus kebijakan dan tanggapan beliau terhadap isu-isu yang muncul.

Fenomena transformasi informasi melalui berita elektronik, artikel website, dan jurnal digital telah memudahkan akses informasi namun sekaligus memperbesar risiko pengguna menghadapi informasi yang tidak relevan dan kesulitan dalam memilah informasi yang bermanfaat. Hal ini juga berlaku pada dokumen teks yang ditemukan di jejaring sosial, seperti komentar, ulasan, dan pesan. Dengan pertumbuhan pesat platform jejaring sosial, volume konten tekstual yang dihasilkan pengguna setiap hari sangat besar.

Meski topik-topik ini tersebar luas, mengekstrak topik dari teks tersebut tetap merupakan tantangan karena adanya masalah ketersebaran data, di mana kata-kata yang jarang muncul dalam dokumen individual dapat menyebabkan kesulitan dalam penerapan model tradisional. Dalam konteks pidato Presiden Joko Widodo, tantangan ini mencerminkan perlunya teknik analisis yang dapat menangani kompleksitas dan volume data untuk memahami dan menafsirkan pesan serta perubahan topik dalam pidato secara efektif.

Pemodelan topik adalah salah satu tugas pembelajaran tanpa pengawasan yang bertujuan untuk mengungkap struktur semantik tersembunyi dalam dokumen. Proses ini mengasumsikan bahwa setiap dokumen terdiri dari campuran topik, dan setiap topik diwakili oleh sekumpulan kata. Ekstraksi topik dari dataset sangat penting untuk memahami konten dan mengelompokkan data berdasarkan kesamaan topiknya. Hal ini meningkatkan analisis data, pengambilan keputusan, dan penemuan informasi berharga dengan memberikan pemahaman yang lebih mendalam tentang elemen tematik dalam dataset [1], [2], [3].

Salah satu metode yang digunakan dalam ekstraksi topik adalah Latent Dirichlet Allocation (LDA), yang merupakan teknik standar dalam pemrosesan bahasa alami (NLP) [4]. LDA memungkinkan identifikasi pola tersembunyi dalam data teks, memfasilitasi pengelompokan kata-kata ke dalam topik tertentu. Metode ini tidak hanya menyederhanakan analisis konten tetapi juga meningkatkan kualitas fitur dalam model pembelajaran mesin. Dengan menerapkan LDA, peneliti dapat mengungkap topik laten dalam dataset, yang mendukung pengembangan sistem rekomendasi dan meningkatkan hasil pencarian topik [5], [6]. Oleh karena itu, pemodelan topik adalah langkah awal yang penting dalam analisis data, menawarkan wawasan signifikan untuk memahami dan memodelkan data [7].

## 2. TUJUAN PENELITIAN

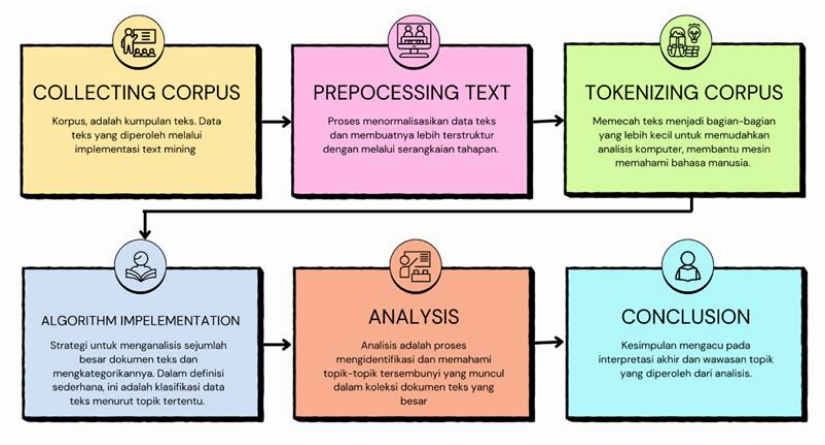
Tujuan penelitian ini adalah untuk mengidentifikasi, menganalisis, dan memprediksi perubahan topik utama dalam pidato, pernyataan, dan publikasi media terkait Presiden Joko Widodo menggunakan teknik pemodelan topik Latent Dirichlet Allocation (LDA). Dengan menerapkan pendekatan kuantitatif, penelitian ini bertujuan untuk memahami bagaimana prioritas dan fokus komunikasi Presiden Joko Widodo berubah seiring waktu. Selain itu, penelitian ini bertujuan untuk memberikan wawasan

mendalam mengenai pergeseran tematik dalam pidato dan publikasi media, serta untuk mengungkap pola-pola penting yang dapat membantu dalam memahami evolusi kebijakan dan tanggapan terhadap isu-isu baru yang muncul.

## 3. METODOLOGI

### 3.1. Metode Penelitian

Penelitian ini menggunakan metode kuantitatif dengan algoritma Latent Dirichlet Allocation (LDA) untuk menganalisis teks secara statistik, dengan tujuan mengidentifikasi dan mengukur distribusi topik dalam dokumen. LDA memodelkan teks sebagai kombinasi topik yang masing-masing terdiri dari sejumlah kata, memungkinkan peneliti untuk mengungkap tema-tema dominan seperti ekonomi, kebijakan luar negeri, kesehatan, dan pendidikan dalam pidato presiden. Metode ini efektif untuk analisis data besar, memberikan hasil yang objektif dan dapat diulang, serta memberikan wawasan mendalam tentang prioritas dan perubahan agenda politik presiden serta respons terhadap isu-isu tertentu [4]. Secara konseptual metodologi proses dan prosedur penelitian dapat ditunjukkan dengan Gambar 1.




Gambar 1. Tahapan melakukan *Topic Modeling*

Latent Dirichlet Allocation (LDA) adalah algoritma generatif yang menemukan topik tersembunyi dalam kumpulan dokumen teks. Prosesnya melibatkan pengumpulan korpus, pemrosesan teks dengan menghilangkan kata umum, tanda baca, dan melakukan stemming atau lemmatization, serta tokenisasi. LDA memodelkan dokumen sebagai campuran topik dan topik sebagai campuran kata, memberikan distribusi probabilistik dari topik dalam dokumen dan kata dalam topik. Metode ini memungkinkan identifikasi dan pengukuran tema-tema dominan dalam teks, menyediakan alat yang kuat untuk mengeksplorasi dan memahami struktur topik dalam dokumen besar secara statistik.

### *3.2. Web Scrapping*

Secara teoritis, web scraping adalah proses mengumpulkan data dari situs web menggunakan program otomatis yang berinteraksi dengan server, meminta data (biasanya dalam bentuk HTML), dan menguraikan informasi tersebut [8]. Dalam penelitian ini, teknik scraping digunakan untuk mengambil data dari situs web Sekretariat Kabinet Presiden Republik Indonesia.

### *3.3. Preprocessing Text*

*Prepocessing text* adalah adalah serangkaian langkah yang dilakukan untuk mengubah teks mentah menjadi format yang lebih bersih dan seragam sehingga dapat digunakan secara lebih efektif dalam analisis dan pemodelan data [9]. Proses ini dimulai dengan pembersihan teks, di mana karakter atau simbol yang tidak diperlukan, seperti tanda baca atau angka, serta spasi ekstra, dihapus untuk memastikan teks lebih mudah diproses. Selanjutnya, dilakukan normalisasi yang mengonversi semua teks menjadi bentuk yang konsisten, seperti mengubah semua huruf menjadi huruf kecil atau melakukan *stemming* dan *lemmatization* untuk mengurangi kata ke bentuk dasarnya.

Langkah penting berikutnya adalah penghapusan *stopwords*, yaitu kata-kata umum yang tidak memberikan makna signifikan dalam analisis, seperti "dan", "atau", "yang" dalam bahasa Indonesia. Setelah itu, teks dipecah menjadi unit-unit yang lebih kecil, seperti kata atau frasa, melalui proses tokenisasi. Terakhir, teks yang telah diproses ini diubah menjadi representasi numerik, seperti *bag-of-words*, TF-IDF, atau embedding kata, sehingga dapat digunakan dalam algoritma pembelajaran mesin [10]. Keseluruhan proses ini penting untuk memastikan bahwa teks yang dianalisis berada dalam bentuk yang optimal, memungkinkan alat dan algoritma untuk bekerja lebih efisien dan akurat.

### *3.4. Topic Modeling*

Pemodelan topik adalah teknik analisis data yang mengelompokkan dokumen ke dalam berbagai tema. Pendekatan ini menemukan topik-topik yang muncul dan menentukan distribusi topik di seluruh materi. Pendekatan ini efektif untuk mengatur dan mengekstrak informasi dari dokumen yang tidak terstruktur, membantu menemukan pola, dan mengekstrak atribut penting yang dapat diklasifikasikan atau dikelompokkan [6]. Topic modeling adalah metode statistik yang diterapkan untuk menemukan tema dan tren laten dalam kumpulan dokumen teks. Data dari *web* forum internet, juga dapat dicirikan

sebagai dokumen teks yang mengandung tema laten [10]. Model topik membuat "topik" yang terdiri dari peringkat kata-kata berdasarkan seberapa relevan topik dengan dokumen. Misalnya, topik-topik tertentu terkait dengan keuangan dan perdagangan, sedangkan topik-topik lain terkait dengan pipa gas dan kontrak. Dengan menggunakan model topik, setiap topik dapat dikaitkan dengan dokumen tertentu sehingga lebih mudah memahami konteksnya. Misalnya, dokumen yang membahas reaksi utilitas California terhadap pasar listrik jangka pendek dapat dikaitkan dengan topik tertentu. Pada dasarnya, model topik terdiri dari berbagai algoritma dan formulasi matematis, tetapi fokus utamanya adalah pada pendekatan probabilistik seperti LDA [10].

### *3.5. Latent Dirichlet Allocation*

LDA adalah model probabilistik generatif dari kumpulan teks yang disebut *corpus*. Metode ini memiliki ide dasar bahwa setiap dokumen mewakili campuran topik yang bersifat acak dan tersembunyi dan karakter setiap topik ditentukan oleh distribusi kata per topik. LDA dapat meringkas, mengelompokkan, menghubungkan, dan menangani volume data yang sangat besar dengan menghasilkan daftar topik berbobot untuk setiap dokumen. Distribusi Dirichlet digunakan untuk menentukan distribusi subjek dokumen. Proses pembangkitan yang dihasilkan oleh *Dirichlet* kemudian digunakan untuk menetapkan kata-kata dalam dokumen ke subjek yang berbeda. Dokumen dapat dilihat pada LDA, sedangkan struktur laten meliputi distribusi kata pada subjek, distribusi topik pada dokumen, dan kategorisasi setiap kata pada tema tertentu [4]. LDA akan mengikuti proses generatif berikut untuk setiap dokumen w dalam sebuah *corpus*.

1. Untuk setiap topik $k = \{1,...,K\}$
   Membuat distribusi atas kosakata $V$, $\beta_k \sim Dir_{(\eta)}$
2. Untuk setiap dokumen *d*
   a. Membuat distribusi atas berbagai topik, $\theta_d \sim Dir_{(\alpha)}$
   b. Untuk setiap kata *w* didalam dokumen *d*
      i. Membuat distribusi topik $Z_{d\,n} \sim \text{Mult}(\theta_d)$, dengan ketentuan $Z_{d\,n} \in \{1\ldots,K\}$
      ii. Membuat distribusi kata $w_{d\,n} \sim \text{Mult}\ (\beta z_{d,n})$, dengan ketentuan $w_{d\,n} \in \{1\ldots,V)$

Model dasar ini memiliki beberapa asumsi, beberapa di antaranya akan dibahas di bagian selanjutnya. Pertama, dimensi distribusi *Dirichlet* k (dan dengan demikian dimensi variabel topik z) dianggap diketahui dan tetap. Kedua, probabilitas kata diparameterkan oleh matriks k×V β, di mana βij = p(wj = 1| zi = 1), yang diperlakukan sebagai kuantitas tetap yang harus diestimasi. Akhirnya, asumsi *Poisson* tidak diperlukan untuk semua hal berikutnya, dan distribusi panjang dokumen yang lebih realistis dapat digunakan sesuai keinginan. Penting untuk diperhatikan bahwa N tidak dipengaruhi oleh faktor-faktor lain yang menghasilkan data (θ dan z). Oleh karena itu, N merupakan variabel tambahan, dan biasanya akan mengabaikan ketidakpastiannya dalam evolusi selanjutnya yang dituliskan dalam persamaan (1)

$$p(\theta|\alpha) = \frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta_1^{\alpha_i - 1} \dots \theta_k^{\alpha_k - 1} \quad (1)$$

Dimana :

- Distribusi *Dirichlet* $p(\theta|\alpha)$ : Ini adalah distribusi probabilitas untuk vektor θ yang terdiri dari k variabel. α adalah parameter bentuk (*shape parameter*) dari distribusi *Dirichlet*.
- Fungsi *Gamma* Γ (·) : Fungsi *Gamma* adalah generalisasi dari faktorial untuk bilangan real. Untuk bilangan bulat n, Γ(n) = (n−1)!
- $\frac{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}{\prod_{i=1}^{k} \Gamma(\alpha_i)}$: Ini adalah konstanta normalisasi yang memastikan bahwa total probabilitas dari distribusi *Dirichlet* adalah 1. Untuk $\sum_{i=1}^{k} \alpha_i$ adalah jumlah dari semua elemen dalam vektor α. Sedangkan $\prod_{i=1}^{k} \Gamma(\alpha_i)$ adalah hasil kali dari fungsi *Gamma* untuk setiap elemen dalam α
- $\theta_1^{\alpha_i - 1} \dots \theta_k^{\alpha_k - 1}$ : Ini adalah fungsi distribusi yang sebenarnya untuk θ. Setiap θi dipangkatkan dengan αi−1. θ adalah vektor dengan k elemen yang merupakan probabilitas, sehingga θi ≥0 dan $\sum_{i=1}^{k} \theta_i = 1$

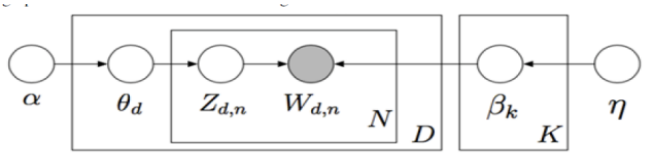
Diberikan parameter α dan β, distribusi gabungan dari campuran topik θ, sekumpulan N topik z, dan sekumpulan N kata w diberikan oleh persamaan (2)

$$p(\theta, \mathbf{Z}, \mathbf{W}|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p(Z_n|\theta) p(w_k\ |Z_n, \beta); \quad (2)$$

Dimana :

- Parameter α dan β : α (alpha) adalah parameter Dirichlet untuk distribusi topik dalam dokumen. sedangkan β (beta) adalah parameter Dirichlet untuk distribusi kata dalam topik.
- Distribusi gabungan $p(\theta, \mathbf{Z}, \mathbf{W}|\alpha, \beta)$ : Distribusi gabungan ini mencakup semua parameter variabel tersembunyi dan teramati dalam model LDA.
- $p(\theta|\alpha)$ : Distribusi *Dirichlet* dari campuran topik θ yang diberikan parameter α. Ini menentukan distribusi probabilitas topik dalam dokumen.
- $\prod_{n=1}^{N} p(Z_n|\theta)$ : Produk dari probabilitas topik Zn yang diberikan distribusi topik θ untuk setiap kata nnn dalam dokumen. Ini menunjukkan probabilitas pemilihan topik Zn dari distribusi topik θ.
- $p(w_n\ |Z_n, \beta)$; : Probabilitas kata wn yang diberikan topik Zn dan parameter β. Ini menunjukkan probabilitas pemilihan kata wn dari topik Zn.

Model LDA direpresentasikan sebagai model grafis probabilistik pada Gambar 1. Seperti yang terlihat jelas pada Gambar 2, terdapat tiga level dalam representasi LDA. Parameter α dan β adalah parameter tingkat *corpus*, yang diasumsikan disampel satu kali proses untuk menghasilkan *corpus*. Kotak-kotak tersebut adalah " *plate* " yang mewakili replikasi. *Plate* luar mewakili dokumen, sedangkan *plate* dalam mewakili pilihan topik dan kata yang diulang-ulang dalam sebuah dokumen [3].




Gambar 2. Representasi model grafis dari LDA

Setelah data dimasukkan dan di proses dengan metode LDA. Corpora berfungsi untuk melakukan pemetaan pada setiap kata dengan menggunakan id kemudian diubah menjadi *term dictionary* [6].

### *3.6. Topic Coherence dan Perplexity*

Koherensi topik menentukan jumlah model dalam pemodelan topik. Nilai koherensi topik yang lebih besar menyiratkan bahwa model akan menghasilkan hasil yang lebih baik [5]. *Topic coherence* mengukur nilai setiap topik dengan cara mengukur tingkat kesamaan semantik antarkata dengan nilai tinggi dalam topik. Pengukuran ini

dapat digunakan untuk membedakan antara topik hasil temuan inferensi statistik dengan topik yang dapat diinterpretasi secara *semantic* [11]. Koherensi topik menilai nilai sebuah topik dengan membandingkan kesamaan semantik dari istilah-istilah bernilai tinggi di dalamnya. Pengukuran ini membedakan antara isu-isu yang didasarkan pada inferensi statistik dan isu-isu yang dapat dipahami secara semantik. Model yang sangat baik menghasilkan tema-tema yang kohesif dengan nilai koherensi yang tinggi. Memberikan label sederhana untuk mendefinisikan isu dapat mengindikasikan relevansi dan kegunaannya. Model yang baik akan menghasilkan topik yang koheren, yaitu topik dengan skor koherensi yang tinggi. Salah satu indikator bahwa topik yang dihasilkan baik atau bermakna adalah kemudahan sesorang dalam memberikan label pendek untuk menggambarkan topik tersebut. Topik yang tidak koheren mungkin mengandung istilah yang tidak terkait secara semantik. Nilai koherensi topik dapat dihitung dengan menggunakan persamaan perhitungan CV, Umass, UCI, NPMI. Untuk perhitungan *coherence score* CV, dapat dilihat pada persamaan (3).

$$CV = \sum i,j \; score(vi, vj, \epsilon) \tag{3}$$

Dimana :

- ∑i,j: Simbol sum (penjumlahan) untuk semua pasangan kata vi dan vj dalam topik. Ini berarti kita menghitung *score* untuk setiap pasangan kata dalam topik dan menjumlahkan hasilnya.
- *score*(vi,vj,ϵ): Fungsi skor yang mengukur koherensi antara pasangan kata vi dan vj dengan mempertimbangkan parameter ϵ. vi dan vj adalah representasi vektor dari kata-kata dalam topik. Representasi ini bisa berupa frekuensi kemunculan kata dalam dokumen atau representasi lain seperti *embedding* kata. Skor dihitung berdasarkan seberapa sering pasangan kata vi dan vj muncul bersama dalam dokumen yang sama. ϵ adalah parameter yang digunakan untuk mencegah nilai pembagian dengan nol atau untuk memberikan pembobotan pada skor kemunculan bersama.

Untuk perhitungan *coherence score* UCI, dapat dilihat pada persamaan (4).

$$UCI = \frac{2}{N \cdot (N-1)} \sum_{i=1}^{N=1} \sum_{j=i+1}^{N} PMI(w_i, w_j) \tag{4}$$

Dimana :

- $\frac{2}{N \cdot (N-1)}$ : Faktor normalisasi untuk memastikan bahwa koherensi dihitung sebagai rata-rata dari semua pasangan kata dalam topik. N: Jumlah total kata dalam topik. Faktor normalisasi ini menyesuaikan hasil PMI agar sesuai dengan jumlah pasangan kata yang mungkin dalam topik.
- $\sum_{i=1}^{N=1} \sum_{j=i+1}^{N}$ : Simbol sum (penjumlahan) ganda yang mengindikasikan bahwa perhitungan UCI akan menghitung PMI untuk setiap pasangan kata (wi,wj) dalam topik, di mana i dan j adalah indeks kata. Penjumlahan pertama berjalan dari i = 1 hingga i = N−1. Penjumlahan kedua berjalan dari j = i+1 hingga j = N.

Untuk perhitungan *coherence score* Umass, dapat dilihat pada persamaan (5)

$$UMASS = \frac{2}{N \cdot (N-1)} \sum_{i=1}^{N=1} \sum_{j=i}^{i-1} \log \frac{P(w_i, w_j) + \varepsilon}{P(w_j)} \quad (5)$$

Dimana :

- $\frac{2}{N \cdot (N-1)}$ : Faktor normalisasi untuk memastikan bahwa koherensi dihitung sebagai rata-rata dari semua pasangan kata dalam topik. N: Jumlah total kata dalam topik. Faktor normalisasi ini menyesuaikan perhitungan koherensi dengan jumlah pasangan kata yang mungkin dalam topik.
- $\sum_{i=1}^{N=1} \sum_{j=i+1}^{N}$ : Simbol sum (penjumlahan) ganda yang mengindikasikan bahwa perhitungan Umass akan menghitung setiap pasangan kata (wi,wj) dalam topik, di mana i dan j adalah indeks kata. Penjumlahan pertama berjalan dari i = 1 hingga i = N−1. Penjumlahan kedua berjalan dari j = i+1 hingga j = N.
- $\log \frac{P(w_i, w_j) + \varepsilon}{P(w_j)}$ : Logaritma natural dari rasio antara probabilitas kemunculan bersama wi dan wj dengan probabilitas kemunculan wj saja. P(wi,wj) mengindikasikan probabilitas bahwa kata wi dan wj muncul bersama dalam dokumen yang sama. P(wj) adalah probabilitas kemunculan kata wj dalam dokumen. **ϵ**: Nilai kecil yang ditambahkan untuk menghindari pembagian dengan nol atau logaritma dari nol.

Untuk perhitungan *coherence score* NPMI, dapat dilihat pada persamaan (6)

$$NPMI(w_i, w_j) = \frac{\log \frac{P(w_i, w_j) + \varepsilon}{P(w_i).P(w_j)}}{-\log P(w_i).P(w_j) + \varepsilon} \quad (6)$$

Dimana :

- $P(w_i, w_j)$ : Probabilitas kemunculan bersama kata wi dan wj dalam dokumen yang sama. Dalam konteks rumus *Normalized Pointwise Mutual Information* (NPMI), wi dan wj adalah simbol yang digunakan untuk mewakili dua kata tertentu dalam sebuah topik.
- $P(w_i)$ : Probabilitas kemunculan kata wi dalam dokumen. Dimana wi merupakan kata pertama dalam pasangan kata yang sedang dianalisis.
- $P(w_j)$ : Probabilitas kemunculan kata wi dalam dokumen. Dimana wj merupakan kata kedua dalam pasangan kata yang sedang dianalisis.
- $\varepsilon$ : Nilai kecil yang ditambahkan untuk menghindari pembagian dengan nol atau logaritma dari nol, umumnya digunakan untuk stabilitas numerik.

*Perplexity* berguna untuk menetapkan nilai optimal dari beberapa topik yang didapatkan yang mana semakin rendah nilai *perplexity* maka semakin baik model LDA yang dihasilkan. Persamaan untuk menghitung nilai *Perplexity* dituliskan dalam persamaan (7).

$$Perplexity\ (T) = \ \mathrm{e}^{-\frac{1}{N}\sum_{i=1}^{N}\log P\,(\,\mathrm{wi}\mid \mathrm{w1},\mathrm{w2},\ldots,\mathrm{wi}-1\,)} \qquad (7)$$

Dimana :

- *Perplexity*(T) : Nilai *perplexity* dari model untuk kumpulan teks T. *Perplexity* adalah ukuran yang digunakan untuk menilai seberapa baik sebuah model probabilistik dalam memprediksi sebuah sampel data. Semakin rendah nilai perplexity, semakin baik model tersebut.
- Exp : Fungsi eksponensial (basis e), digunakan untuk mengembalikan hasil dari logaritma ke skala probabilitas asli.
- N : Jumlah total kata dalam *corpus* teks yang diuji.
- $\sum_{i=1}^{N}$ : Simbol sum (penjumlahan) dari i = 1 hingga i = N, yang menunjukkan bahwa perhitungan akan menjumlahkan nilai-nilai yang dihitung untuk setiap kata wi dalam *corpus* teks, dari kata pertama hingga kata ke-N.
- log : Fungsi logaritma natural (basis e). Logaritma digunakan untuk mengubah perkalian probabilitas menjadi penjumlahan untuk memudahkan perhitungan dan menangani nilai - nilai probabilitas yang sangat kecil.

- P(wi|w1,w2,…,wi−1): Probabilitas kemunculan kata ke-i (wi) diberikan kata-kata sebelumnya dalam urutan (w1,w2,…,wi-1). Ini adalah probabilitas kondisional yang dihitung oleh model probabilistik. Probabilitas ini menunjukkan seberapa besar kemungkinan kata wi muncul setelah urutan kata-kata sebelumnya.

## 4. HASIL DAN PEMBAHASAN

Penelitian ini diawali dengan membangun corpus dengan cara melakukan *crawling* dokumen naskah pidato Presiden Joko Widodo dari situs resmi Sekretariat Negara Republik Indonesia. Dari hasil *crawling* tersebut didapatkan data sebanyak 5988 naskah pidato presiden dari tanggal 20 Oktober 2014 sampai dengan 2 Maret 2024 seperti yang terlihat pada Tabel 1.

Tabel 1. Naskah Pidato Presiden RI

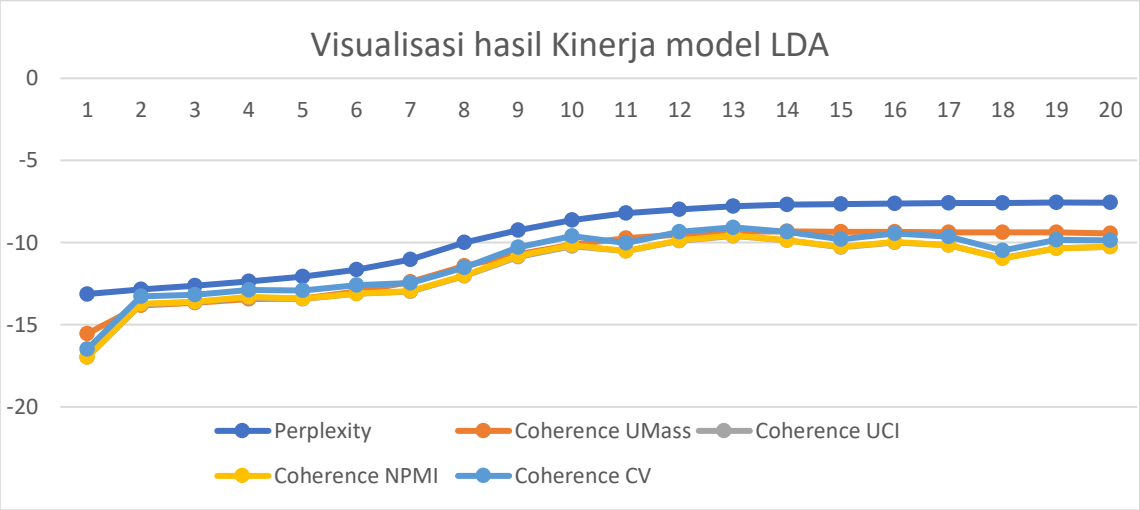
| No | Kategori | Jumlah |
|---|---|---|
| 1 | Amanat/Arahan | 134 |
| 2 | Berita | 28 |
| 3 | Dialog | 142 |
| 4 | Keterangan Pers | 1518 |
| 5 | Pengantar | 500 |
| 6 | Perkembangan Penanganan COVID-19 | 4 |
| 7 | Sambutan | 2041 |
| 8 | Transkrip Pidato | 1621 |
| **Total** | | **5988** |

Dalam melakukan pemodelan topik ini tahapan yang dilakukan dimulai dengan melakukan *preprocessing text* yaitu mulai dari tokenisasi, *Stopwords Removal*, dan *Stemming/Lemmatization*. Selanjutnya *corpus* dibuat dalam bentuk *words-vector* dan membuat *dictionary-term* untuk memetakkan kata unik ke numerik ID. *Corpora dictionary* Pidato Presiden RI yang dihasilkan selanjutnya diolah menggunakan LDA dengan memanfaatkan Pustaka Gensim yang merupakan bahasa Python untuk pemodelan topik, analisis bahasa, dan pemrosesan bahasa alami.

Adapun pemodelan topik dengan metode LDA ini menggunakan $\kappa$=2 s.d $\kappa$=22, selanjutnya model yang dihasilkan dalam setiap percobaan k-*n* dievaluasi menggunakan metrik *coherence* dan *perplexity*.

### *4.1. Topic Interpretation*

Dari beberapa hasil, penelitian ini menemukan hasil *cluster* topik dengan kata kunci yang dapat ditafsirkan dengan topik yang relevan dan diurutkan dari topik nilai terbaik ke terendah, kluster terbaik adalah yang k=16.




Gambar 3. Visualisasi Kinerja model

Berdasarkan data yang dijelaskan selanjutnya, berikut adalah beberapa poin utama yang dapat disimpulkan:

4.1.1. Koherensi Topik

Koherensi tertinggi (*Coherence_cv*) tercapai pada topik-topik dengan k=16, dengan nilai masing-masing 0.554. Ini menunjukkan bahwa topik-topik ini paling kohesif dan kemungkinan besar paling mudah diinterpretasikan.

4.1.2. *Perplexity*

Nilai *perplexity* terbaik (terendah) ada pada topik dengan k=21, dengan nilai -13.130. Ini menunjukkan bahwa model dengan 5 topik memiliki kemampuan prediksi terbaik pada data yang tidak terlihat.

4.1.3. Koherensi Kombinasi

Topik dengan K=16 ("Nasionalisme dan Nilai-nilai kebangsaan") memiliki nilai Coherence_cv tertinggi dan juga memiliki nilai Coherence_npmi terbaik, menunjukkan koherensi internal yang kuat. Topik dengan K=14 ("Pemerintahan Daerah") juga memiliki koherensi yang baik secara keseluruhan.

Table 2. topik interpretation $\kappa$ =21

| ***k*** | **Keyword** | **Topic Interpretation** |
|---|---|---|
| 2 | "orang", "kerja", "nama", "jam", "anakanak", "terima", "percaya", "anak", "pilih", "indonesia" | Pekerjaan dan Masyarakat |
| 3 | "nggak", "asia", "bendung", "paket", "indonesia", "world", "bencana", "air", "korban", "banjir" | Bencana dan Bantuan |
| 4 | "perintah", "rakyat", "masyarakat", "hukum", "tni", "layan", "negara", "polri", "aman", "lindung" | Pemerintahan dan Hukum |
| 5 | "sehat", "rumah", "sakit", "obat", "darurat", "dokter", "bpjs", "masyarakat", "layan", "bantu" | Kesehatan dan Layanan Medis |

| *k* | Keyword | Topic Interpretation |
|---|---|---|
| 6 | "menteri", "masuk", "kait", "ya", "lihat", "betul", "negara", "cepat", "urus", "perintah" | Pemerintahan dan Kebijakan |
| 7 | "bangun", "jalan", "tol", "selesai", "saing", "cepat", "kawasan", "bandara", "airport", "biaya" | Infrastruktur dan Pembangunan |
| 8 | "listrik", "jakarta", "kereta", "transportasi", "dki", "mobil", "jembatan", "mrt", "macet", "banten" | Transportasi dan Perkotaan |
| 9 | "juta", "sertifikat", "tanah", "bank", "rp", "ya", "beli", "desa", "pegang", "pakai" | Keuangan dan Properti |
| 10 | "presiden", "indonesia", "republik", "ya", "wartawan", "widodo", "joko", "tanggal", "partai", "kuala" | Kepresidenan dan Politik |
| 11 | "covid", "pandemi", "es", "sehat", "tangan", "odi", "anindito", "vaksin", "vaksinasi", "waralaba" | Pandemi COVID-19 |
| 12 | "ekonomi", "tumbuh", "persen", "uang", "tingkat", "investasi", "infrastruktur", "daerah", "belanja", "fokus" | Ekonomi dan Pertumbuhan |
| 13 | "indonesia", "asean", "kerja", "bidang", "malaysia", "negara", "bahas", "kawasan", "temu", "kunjung" | Hubungan Internasional |
| 14 | "kota", "provinsi", "gubernur", "kabupaten", "jawa", "bupati", "timur", "barat", "infrastruktur", "daerah" | Pemerintahan Daerah |
| 15 | "rp", "tani", "harga", "juta", "subsidi", "hutan", "beras", "jual", "tanam", "ribu" | Pertanian dan Subsidi |
| 16 | "negara", "bangsa", "indonesia", "hormat", "satu", "nilainilai", "wa", "pilih", "jaga", "saudara" | Nasionalisme dan Nilai-nilai kebangsaan |
| 17 | "industri", "usaha", "negara", "impor", "investasi", "pasar", "ekspor", "masuk", "produksi", "indonesia" | Industri dan Perdagangan |
| 18 | "presiden", "indonesia", "republik", "ya", "joko", "widodo", "iya", "terima", "", "kasih" | Kepresidenan |
| 19 | "negara", "indonesia", "bangsa", "hormat", "satu", "rakyat", "ketua", "jaga", "pimpin", "hadirin" | Kebangsaan dan Nasionalisme |
| 20 | "persen", "triliun", "ekonomi", "uang", "desa", "angka", "tumbuh", "fokus", "turun", "belanja" | Ekonomi Makro |
| 21 | "bangun", "indonesia", "ubah", "negara", "dunia", "didik", "cepat", "daya", "kembang", "sumber" | Pendidikan dan Pengembangan SDM |

Ini adalah hasil *clustering* yang pertama, Tabel 2 adalah model topik yang pertama kali dengan membuat kelompok secara acak namun dibuat 21 kelompok, setiap *corpus* akan mengisi sesuai *dictionary* setelah *corpus* dipecah dalam teknik *tokenizing* kedalam sebuah bentuk vektor. Setelah diamati dari keseluruhan topik yang ada, topik ke 16 adalah topik dengan perhitungan nilai *coherence* cv,npmi, umas, uci terbaik. Maka selanjutnya dibuat model cluster dengan (K) = 16, hasilnya seperti yang terlihat pada Tabel 3.

Tabel 3. topik interpretation K=16

| K | *Keyword* | Topic Interpretation |
|---|---|---|
| 1 | terima, presiden, anak-anak, ajar, sekolah, anak, sila, indonesia, kasih, kartu | Pendidikan dan Anak |
| 2 | presiden, indonesia, republik, tanya, joko, widodo, alur, wartawan, oke, baik | Statement Presiden |
| 3 | industri, usaha, ekspor, negara, impor, investasi, produk, hormat, peluang, negeri | Industri dan Ekonomi |
| 4 | bantu, sehat, cegah, tangan, covid, korban, belanda, batas, krisis, pandemi | Kesehatan dan Krisis |
| 5 | masuk, negara, menteri, ya, lihat, kerja, urus, nama, milik, desa | Pemerintahan dan Kerja |

| K | *Keyword* | Topic Interpretation |
|---|---|---|
| 6 | kota, provinsi, gubernur, kabupaten, wabarakatuh, warahmatullahi, timur, jawa, bupati, barat | Daerah dan Kepemimpinan |
| 7 | perintah, masyarakat, rakyat, layan, hukum, tni, aman, sistem, laksana, cepat | Pemerintahan dan Masyarakat |
| 8 | juta, sertifikat, tanah, bank, rp, beli, nggih, pinjam, pegang, pakai | Keuangan dan Properti |
| 9 | business, asia, one, economic, regional, indonesia, years, time, people, also | Ekonomi Regional dan Global |
| 10 | ubah, nggak, didik, guru, muda, cepat, kompetisi, sumber, bangun, teknologi | Pendidikan dan Teknologi |
| 11 | ekonomi, persen, tumbuh, uang, tingkat, triliun, anggar, daerah, investasi, fokus | Ekonomi dan Pertumbuhan |
| 12 | indonesia, kerja, dunia, negara, bangun, kuat, tingkat, bidang, asean, ekonomi | Pembangunan dan Kerja Sama Internasional |
| 13 | bangun, menteri, jalan, selesai, infrastruktur, listrik, kait, tol, labuh, cepat | Infrastruktur |
| 14 | rp, tani, harga, subsidi, juta, pasar, jual, beli, beras, ribu | Pertanian dan Harga |
| 15 | indonesia, negara, bangsa, hormat, saudara-saudara, satu, jaga, hadirin, pimpin, pilih | Nasionalisme dan Kesatuan |
| 16 | rumah, sehat, sakit, obat, bpjs, dokter, layan, dokter, masyarakat, sembuh | Kesehatan dan Rumah Sakit |

Topik yang pertama, Pendidikan dan Anak, mencakup berbagai inisiatif presiden dalam memajukan pendidikan anak-anak di Indonesia, termasuk program kartu Indonesia pintar yang bertujuan untuk membantu anak-anak dalam pendidikan. Topik kedua, Statement Presiden, berfokus pada berbagai pernyataan dan pidato yang disampaikan oleh Presiden Joko Widodo, serta interaksinya dengan wartawan. Selanjutnya, topik Industri dan Ekonomi menyoroti pentingnya industri, ekspor-impor, dan investasi dalam mendukung pertumbuhan ekonomi negara, peluang pengembangan ekonomi dan produk dalam negeri. Topik keempat, Kesehatan dan Krisis, mencakup isu-isu kesehatan, terutama pencegahan COVID-19, pembatasan sosial dan penanganan krisis pandemi, serta bantuan yang diberikan kepada korban.

Dalam topik Pemerintahan dan Kerja, pembahasan berfokus pada tugas-tugas pemerintahan, kerja menteri, dan administrasi negara, termasuk urusan desa. Topik Daerah dan Kepemimpinan membahas peran penting gubernur, bupati, dan kepala daerah lainnya dalam memimpin dan mengelola wilayah masing-masing, dalam konteks ini presiden menyoroti kinerja pemerintahan daerah di pulau jawa. Topik Pemerintahan dan Masyarakat terkait dengan layanan yang diberikan pemerintah kepada masyarakat, termasuk aspek hukum, keamanan yang dijaga oleh TNI, dan sistem administrasi publik yang dilaksanakan secara cepat dan efisien. Topik Keuangan dan Properti menyoroti isu-

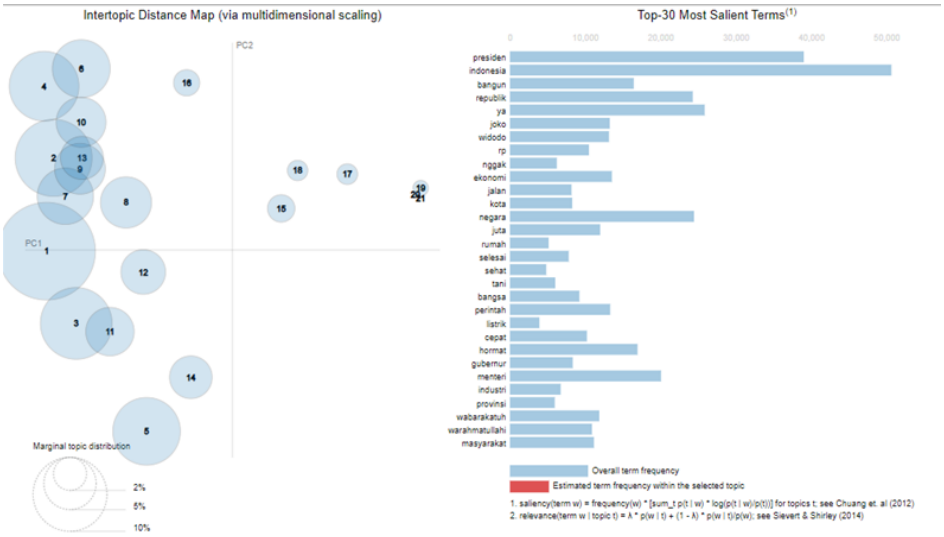
isu keuangan dan properti seperti sertifikat tanah, transaksi keuangan, dan pinjaman bank serta pengelolaan pemakaian dana yang efektif.

Topik Ekonomi Regional dan Global adalah topik yang melihat ekonomi dalam konteks regional dan global, termasuk hubungan bisnis di kawasan Asia dengan Indonesia untuk jangka waktu kedepan serta dampaknya kepada masyarakat. Sementara itu, topik Pendidikan dan Teknologi membahas peran pendidikan dan teknologi dalam membangun sumber daya manusia yang kompetitif dalam perubahan zaman dan peningkatan kompetisi kaum milenial, serta peran penting guru dan pendidikan yang cepat. Topik Ekonomi dan Pertumbuhan menyoroti berbagai aspek pertumbuhan ekonomi, investasi, tingkat pertumbuhan, tingkat inflasi mata uang maupun nilai jual beli rupiah, pengelolaan anggaran daerah serta fokus penggunaan anggaran. Sementara itu, Topik Pembangunan dan Kerja Sama Internasional berfokus pada upaya pembangunan nasional dan kerja sama internasional, termasuk kontribusi Indonesia dalam ASEAN, hubungan bilateral Indonesia dengan negara – negara di dunia.

Dalam topik Infrastruktur, pembahasan berfokus pada pembangunan infrastruktur, peran kementrian PUPR seperti jalan tol, listrik, dan proyek-proyek besar lainnya yang diselesaikan dengan cepat. Topik Pertanian dan Harga menekankan isu-isu penting dalam pertanian, harga pasar, subsidi pertanian maupun subsidi bibit tani, dan transaksi jual beli hasil tani seperti beras.

Topik Nasionalisme dan Kesatuan membahas pentingnya nasionalisme, kepemimpinan dan persatuan bangsa dalam sebuah pemilihan umum, serta peran presiden dan masyarakat dalam menjaga kesatuan dan keutuhan negara. Terakhir, topik Kesehatan dan Rumah Sakit mencakup layanan kesehatan dalam rumah sakit dan pelayanannya, ketersediaan stok obat, pengelolaan BPJS, peranan dokter, dan upaya penyembuhan masyarakat dari berbagai penyakit.

Topik yang dimodelkan dengan LDA divisualisasikan menggunakan pyLDAvis untuk melihat persebaran topik secara interaktif dan intuitif, eksplorasi topik, melihat distribusi kata, menganalisis kualitas model dan identifikasi topik overlap seperti yang ditujukan pada Gambar 4.




Gambar 4. Visualisasi persebaran topik dengan pyLDAvis

Dalam PyLDAvis, Setiap lingkaran pada plot utama mewakili sebuah topik. Ukuran lingkaran menunjukkan prevalensi relatif dari topik tersebut dalam *corpus*. Prevalensi topik menunjukkan proporsi dokumen yang membahas topik tersebut. Misalnya, jika topik tertentu memiliki prevalensi tinggi, berarti topik itu sering muncul atau banyak dibahas dalam *corpus*. Topik-topik yang terpisah jauh menunjukkan bahwa topik memiliki distribusi kata yang sangat berbeda, sementara topik-topik yang berdekatan menunjukkan distribusi kata yang lebih mirip. Dalam visualisasi tersebut, bar berwarna biru menunjukkan frekuensi kata di seluruh *corpus* yang terdapat dalam suatu topik. PyLDAvis memungkinkan pula untuk dapat mengontrol control relevansi, yakni keseimbangan antara eksklusivitas dan frekuensi kata dalam topik yang dipilih. Eksklusivitas kata menunjukkan seberapa spesifik sebuah kata untuk topik tertentu. Kata-kata dengan eksklusivitas tinggi adalah kata-kata yang muncul terutama dalam satu topik dan jarang muncul di topik lain. Nilai lambda yang tinggi (mendekati 1). menunjukkan kata-kata yang umum dalam *corpus* namun juga penting dalam topik yang dipilih. Nilai lambda yang rendah (mendekati 0) menunjukkan kata-kata yang unik dan sangat relevan untuk topik yang dipilih.

## 5. KESIMPULAN

Dari hasil analisis data dan temuan, dapat disimpulkan sebagai berikut:

1) Penelitian ini dapat mengidentifikasi topik utama secara efektif mengidentifikasi topik-topik utama yang sering dibahas dalam pidato, pernyataan, dan publikasi media Presiden Joko Widodo. Melalui penggunaan teknik pemodelan topik Latent

Dirichlet Allocation (LDA), penelitian ini mengungkapkan bahwa topik-topik utama mencakup berbagai bidang seperti Pendidikan dan Anak, Kesehatan dan Rumah Sakit, Industri dan Ekonomi, Pemerintahan dan Kerja, serta Nasionalisme dan Kesatuan. Identifikasi ini memberikan wawasan yang mendalam mengenai area fokus utama dalam komunikasi publik Presiden.

2) Metode LDA secara efektif dapat mengelompokkan kata-kata dalam teks ke dalam topik-topik yang relevan. Model yang menggunakan jumlah topik k=16 menunjukkan koherensi tertinggi, yaitu 0.554, yang menunjukkan konsistensi dan relevansi topik yang dihasilkan. Sebaliknya, model dengan jumlah topik k=21 menghasilkan perplexity terbaik, yaitu -13.130, menandakan kemampuan model dalam menangani ketidakpastian dalam data. Temuan ini menekankan pentingnya memilih jumlah topik yang sesuai untuk mencapai hasil analisis yang optimal.
3) Model Topik dapat mengungkapkan perubahan tematik dari waktu ke waktu bagaimana Presiden merespons isu-isu baru dan menyesuaikan kebijakan serta komunikasinya untuk menghadapi tantangan yang muncul. Analisis tren ini memberikan pemahaman tentang dinamika dan evolusi kebijakan yang diprioritaskan oleh Presiden.

## 6. DAFTAR PUSTAKA

[1] Y. O. Santoso *et al.*, "Pengelompokan jurnal ilmiah berdasarkan judul menggunakan lda 1,2," *Proxies*, vol. 3, no. 1, pp. 32–42, 2019.

[2] P. A. Telnoni and E. Rosely, "Pelabelan Data Dengan Latent Dirichlet Allocation dan K-Means Clustering pada Data Twitter Menggunakan Bahasa Indonesia Data Labeling using Latent Dirichlet Allocation and K-Means Clustering on Indonesian-Based Twitter," vol. 7, no. 2, pp. 885–892, 2020.

[3] I. M. Kusnanta, B. Putra, and P. Kusumawardani, "Analisis Topik Informasi Publik Media Sosial di Surabaya Menggunakan Pemodelan Latent Dirichlet Allocation ( LDA )," vol. 6, no. 2, pp. 4–9, 2017.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," vol. 3, pp. 993–1022, 2003.

[5] Y. Matira and I. Setiawan, "Pemodelan Topik pada Judul Berita Online Detikcom Menggunakan Latent Dirichlet Allocation," *Estimasi J. Stat. Its Appl.*, vol. 4, no. 1, pp. 2721–379, 2023, doi: 10.20956/ejsa.vi.24843.

[6] A. O. Widodo, F. Septiadi, and N. A. Rakhmawati, "Analisis Tren Konten Pada Vtuber Indonesia Menggunakan Latent Dirichlet Allocation," 2023. [Online]. Available: http://e-journal.stmiklombok.ac.id/index.php/jire

[7] 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, [Online]. Available: www.jurnal.unimed.ac.id

[8] R. Mitchel, *Web Scraping with Python*, vol. 53. Sebastopol: O'Reilly Media, Inc, 2018.

[9] Y. S. Emma Haddi, Xiaohui Liu, "The Role of Text Pre-processing in Sentiment Analysis," *Procedia Comput. Sci.*, vol. Volume 17, p. Pages 26-32, 2013.

[10] A. T. J. H, "Preprocessing Text untuk Meminimalisir Kata yang Tidak Berarti dalam Proses Text Mining," *J. Inform. UPGRIS*, vol. 1, pp. 1–9, 2015.

[11] G. Rosalinda, R. Santoso, and P. Kartikasari, "Pemodelan Topik Ulasan Aplikasi Netflix Pada Google Play Store Menggunakan Latent Dirichlet Allocation," *J. Gaussian*, vol. 11, no. 4, pp. 554–561, Feb. 2023, doi: 10.14710/j.gauss.11.4.554-561.