ANALISIS *TOPIC MODELLING* IBUKOTA NUSANTARA PADA *PLATFORM X* MENGGUNAKAN *LATENT DIRICHLET ALLOCATION*

TUGAS AKHIR

Program Studi

S1 Sistem Informasi

Oleh:

LUTHFI KRISNA BAYU

21410100005

FAKULTAS TEKNOLOGI DAN INFORMATIKA

UNIVERSITAS DINAMIKA

2025

ANALISIS *TOPIC MODELLING* IBUKOTA NUSANTARA PADA *PLATFORM X* MENGGUNAKAN *LATENT DIRICHLET ALLOCATION*

TUGAS AKHIR

Diajukan sebagai salah satu syarat untuk menyelesaikan
Program Sarjana

Oleh:

Nama	**: Luthfi Krisna Bayu**
NIM	**: 21410100005**
Program Studi	**: S1 Sistem Informasi**

FAKULTAS TEKNOLOGI DAN INFORMATIKA
UNIVERSITAS DINAMIKA
2025

Tugas Akhir

ANALISIS *TOPIC MODELLING* IBUKOTA NUSANTARA PADA *PLATFORM X* MENGGUNAKAN *LATENT DIRICHLET ALLOCATION*

Dipersiapkan dan disusun Oleh

Luthfi Krisna Bayu

NIM: 21410100005

Telah diperiksa, dibahas dan disetujui oleh Dewan Pembahas

Pada: 05 Maret 2025

Susunan Dewan Pembahas

Pembimbing

I. Tutut Wurijanto, M.Kom.
NIDN. 0703056702

II. Julianto Lemantara, S.Kom., M.Eng.
NIDN. 0722108601

Pembahas

I. Vivine Nurcahyawati, M.Kom.
NIDN. 0723018101

Digitally signed by Vivine Nurcahyawati
Date: 2025.03.07 09:27:38 +07'00'

Tugas Akhir ini telah diterima sebagai salah satu persyaratan

Untuk memperoleh gelar sarjana

Fakultas Teknologi dan Informatika
UNIVERSITAS Dinamika

Dr. Anjik Sukmaaji, S.Kom., M.Eng.

NIDN. 0731057301

Dekan Fakultas Teknologi dan Informatika

UNIVERSITAS DINAMIKA

"Selalu andalkan Tuhan dalam setiap langkah hidup,
maka Ia akan memampukan"

Luthfi Krisna Bayu

Laporan Tugas Akhir ini

Saya persembahkan kepada

Keluarga, Dosen Pembimbing, dan

Teman-teman yang saya kasihi

PERNYATAAN

PERSETUJUAN PUBLIKASI DAN KEASLIAN KARYA ILMIAH

Sebagai mahasiswa **Universitas Dinamika**, Saya :

Nama	: Luthfi Krisna Bayu
NIM	: 21410100005
Program Studi	: S1 Sistem Informasi
Fakultas	: Fakultas Teknologi dan Informatika
Jenis Karya	: Tugas Akhir
Judul Karya	: **ANALISIS *TOPIC MODELLING* IBUKOTA NUSANTARA PADA *PLATFORM X* MENGGUNAKAN *LATENT DIRICHLET ALLOCATION***

Menyatakan dengan sesungguhnya bahwa :

1. Demi pengembangan Ilmu Pengetahuan, Teknologi dan Seni, Saya menyetujui memberikan kepada **Universitas Dinamika** Hak Bebas Royalti Non-Eksklusif (***Non-Exclusive Royalty Free Right***) atas seluruh isi/sebagian karya ilmiah Saya tersebut diatas untuk disimpan, dialihmediakan, dan dikelola dalam bentuk pangkalan data (***database***) untuk selanjutnya didistribusikan atau dipublikasikan demi kepentingan akademis dengan tetap mencantumkan nama Saya sebagai penulis atau pencipta dan sebagai pemilik Hak Cipta.
2. Karya tersebut diatas adalah hasil karya asli Saya, bukan plagiat baik sebagian maupun keseluruhan. Kutipan, karya, atau pendapat orang lain yang ada dalam karya ilmiah ini semata-mata hanya sebagai rujukan yang dicantumkan dalam Daftar Pustaka Saya.
3. Apabila dikemudian hari ditemukan dan terbukti terdapat tindakan plagiasi pada karya ilmiah ini, maka Saya bersedia untuk menerima pencabutan terhadap gelar kesarjanaan yang telah diberikan kepada Saya.

Demikan surat pernyataan ini Saya buat dengan sebenar-benarnya.

Surabaya, 11 Februari 2025

Luthfi Krisna Bayu
NIM: 21410100005

ABSTRAK

Pemindahan Ibu Kota Negara/Nusantara (IKN) ke Kalimantan Timur menjadi isu strategis yang banyak diperbincangkan di media sosial, terutama di *platform X* (sebelumnya *Twitter*). Diskusi mencakup berbagai aspek, seperti pembangunan infrastruktur, transportasi, kebijakan pemerintah, serta opini publik. Lalu untuk memahami pola diskusi tersebut, penelitian ini menggunakan metode *topic modeling* dengan model *Latent Dirichlet Allocation* (LDA) yang cocok untuk mengelompokkan topik dari data teks dalam jumlah besar. Pendekatan CRISP-DM diterapkan sebagai metodologi penelitian, mencakup tahap *business understanding*, *data understanding*, *data preparation*, pemodelan, evaluasi menggunakan *coherence score*, serta *deployment* dalam bentuk visualisasi menggunakan pustaka *pyLDAvis* dan *wordcloud*. Hasil analisis menunjukkan bahwa model LDA berhasil mengidentifikasi lima topik utama yang mencerminkan fokus diskusi publik mengenai IKN, yaitu (1) konsep *smart city* dalam pembangunan ibu kota baru, (2) investasi dan pembangunan infrastruktur, (3) transportasi udara dan pengembangan bandara, (4) dampak sosial-ekonomi bagi masyarakat, serta (5) relokasi pusat pemerintahan dan ASN. Hasil penelitian ini memberikan wawasan bagi pemerintah dan pemangku kebijakan dalam memahami isu-isu utama terkait IKN untuk strategi komunikasi dan kebijakan yang lebih efektif. Selain itu, temuan ini juga memberikan gambaran bagi masyarakat mengenai berbagai aspek pembangunan IKN sehingga dapat meningkatkan partisipasi publik dalam diskusi kebijakan terkait IKN. Evaluasi model menggunakan *coherence score* menunjukkan nilai sebesar 0.429356, yang mengindikasikan tingkat keterpaduan yang cukup baik dalam menghasilkan topik yang bermakna.

Kata Kunci: Ibu Kota Nusantara, *Latent Dirichlet Allocation*, Media Sosial, *pyLDAvis*, *Topic Modeling*

KATA PENGANTAR

Puji syukur ke hadirat Tuhan YME yang telah memberikan berkat dan karunia-Nya sehingga penulis dapat menyelesaikan laporan tugas akhir ini dengan judul "ANALISIS *TOPIC MODELLING* IBUKOTA NUSANTARA PADA *PLATFORM X* MENGGUNAKAN *LATENT DIRICHLET ALLOCATION*" ini dengan baik dan lancar. Penyelesaian laporan Tugas Akhir ini sebagai syarat wajib untuk menyelesaikan program sarjana. Tidak terlepas dari bantuan dari pihak yang telah memberikan masukan, nasihat, saran, kritik kepada penulis. Oleh karena itu, pada kesempatan ini penulis ingin menyampaikan rasa terima kasih kepada:

1. Ayah dan Ibu tercinta, serta sanak saudara yang memberikan doa dan dukungan penuh kepada saya
2. Bapak Tutut Wurijanto, M.Kom. selaku Dosen Pembimbing I yang sudah memberikan bimbingan selama proses penyelesaian tugas akhir.
3. Ko Julianto Lemantara, S.Kom., M.Eng. selaku Dosen Pembimbing II yang sudah memberikan bimbingan selama proses penyelesaian tugas akhir.
4. Ibu Vivine Nurcahyawati, M.Kom. selaku Dosen Penguji yang telah menguji hasil tugas akhir.
5. Sahabat dan teman – teman perkuliahan di Universitas Dinamika Surabaya yang telah membantu dalam proses penyelesaian tugas akhir.

Penulis menyadari bahwa laporan ini masih jauh dari kata sempurna. Dengan demikian penulis mengharapkan kritik dan saran yang membangun dari pembaca untuk penyempurnaan dalam menyelesaikan laporan. Semoga laporan Kerja Praktik ini dapat bermanfaat untuk penulis sendiri, dan para pembaca.

Surabaya, 10 Februari 2025

Penulis

DAFTAR ISI

Halaman

ABSTRAK vii
KATA PENGANTAR viii
DAFTAR ISI ix
DAFTAR GAMBAR xi
DAFTAR TABEL xiii
DAFTAR LAMPIRAN xiv
BAB I PENDAHULUAN 1
1.1 Latar Belakang 1
1.2 Rumusan Masalah 3
1.3 Batasan Masalah 3
1.4 Tujuan 4
1.5 Manfaat 4
BAB II LANDASAN TEORI 5
2.1 Penelitian Terdahulu 5
2.2 Ibukota Nusantara (IKN) 6
2.3 Pemrograman *Python* 6
2.4 *Topic Modelling* 7
2.5 *Latent Dirichlet Allocation* (LDA) 8
2.6 *Natural Language Processing* (NLP) 10
2.7 *Coherence Score* 11
2.8 CRISP-DM 12
BAB III METODOLOGI PENELITIAN 14
3.1 *Business Understanding* 14
3.2 *Data Understanding* 14
3.3 *Data Preparation* 17
3.4 *Modelling* 19
3.5 *Evaluation* 22
3.6 *Deployment* 22

BAB IV HASIL DAN PEMBAHASAN 23
4.1 Hasil *Data Understanding* 23
4.1.1 *Crawling Data Tweet* 23
4.1.2 Hasil *Crawling* 25
4.1.3 EDA 25
4.2 *Data Preparation* 28
4.2.1 *Case Cleaning* 29
4.2.2 *Case Folding* 31
4.2.3 *Stopword Removal* 32
4.2.4 *Normalization* 33
4.2.5 *Stemming* 33
4.2.6 *Tokenization* 34
4.2.7 *Bag of Words* 35
4.3 *Modelling* 36
4.4 *Evaluation* 37
4.5 *Deployment* 40
4.5.1 *Wordcloud* 40
4.5.2 *pyLDAvis* 42
BAB V PENUTUP 49
5.1 Kesimpulan 49
5.2 Saran 49
DAFTAR PUSTAKA 50
LAMPIRAN 54

DAFTAR GAMBAR

Halaman

Gambar 2.1 Konsep LDA 9
Gambar 2.2 Metodologi CRISP-DM 12
Gambar 3.1 *Data Understanding* 14
Gambar 3.2 IPO *Data Preparation* 17
Gambar 3.3 *Modelling* 20
Gambar 4.1 *Tweet Harvesting step* 1 23
Gambar 4.2 *Tweet Harvesting step* 2 24
Gambar 4.3 Hasil *Crawling Data Tweet* 25
Gambar 4.4 *Load dataset* 26
Gambar 4.5 *Shape & info of dataframe* 27
Gambar 4.6 *Missing values & duplicated data* 28
Gambar 4.7 Seleksi Fitur 29
Gambar 4.8 *Case Cleaning* 1 29
Gambar 4.9 *Case Cleaning 2* 30
Gambar 4.10 *Case Folding* 31
Gambar 4.11 *Stopword Removal* 32
Gambar 4.12 *Normalization* 33
Gambar 4.13 *Stemming* 34
Gambar 4.14 *Tokenization* 35
Gambar 4.15 BoW 36
Gambar 4.16 *Modelling with* LDA 36
Gambar 4.17 Kode *Coherence Score* 38
Gambar 4.18 Hasil *Coherence Score* 39
Gambar 4.19 *Wordcloud* topik-topik terkait IKN 40
Gambar 4.20 *pyLDAvis* Topik 1 42
Gambar 4.21 *pyLDAvis* Topik 2 43
Gambar 4.22 *pyLDAvis* Topik 3 44
Gambar 4.23 *pyLDAvis* Topik 4 45
Gambar 4.24 *pyLDAvis* Topik 5 47

Gambar L1.1 Data *Tweet* Terkait IKN 54
Gambar L1.2 *Case Cleaning & Folding* 54
Gambar L1.3 *Stopwords Removal* 54
Gambar L1.4 *Stemming* 55
Gambar L1.5 *Tokenization* 55
Gambar L1.6 Membuat BoW (*Bag of Words*) 56
Gambar L1.7 Distribusi Topik (*Dirichlet Distribution*) 56
Gambar L1.8 Pemilihan Topik untuk Setiap Kata (*Multinomial Distribution*) 57
Gambar L1.9 *Word Distribution* 1 58
Gambar L1.10 *Word Distribution* 2 59
Gambar L1.11 *Word Distribution* 3 60
Gambar L1.12 Probabilitas Gabungan Dokumen 60
Gambar L1.13 *Insight* Hasil Probabilitas Gabungan 61
Gambar L1.14 Inferensi Topik 62
Gambar L1.15 *Insight* Hasil Inferensi Topik 62

DAFTAR TABEL

Halaman

Tabel 2.1 Penelitian Terdahulu 5
Tabel 4.1 Hasil *Coherence Score* 39
Tabel 4.2 Hasil *Wordcloud* IKN 41

DAFTAR LAMPIRAN

Halaman

Lampiran 1 Simulasi LDA 54
Lampiran 2 *Form* Bimbingan TA 63
Lampiran 3 Plagiasi 64
Lampiran 4 Biodata Penulis 65

BAB I
PENDAHULUAN

1.1 Latar Belakang

Pemindahan Ibukota Negara (IKN) dari Jakarta ke Ibukota Nusantara (IKN) telah menjadi isu nasional yang menyedot perhatian luas dari berbagai kalangan (Fristikawati dkk., 2022). Obyek penelitian ini adalah diskusi dan opini masyarakat terkait IKN yang disampaikan melalui *platform X*, salah satu platform media sosial yang sangat populer di Indonesia. Analisis terhadap data dari *platform X* dapat memberikan gambaran yang lebih jelas mengenai persepsi publik terhadap pemindahan ibukota ini, serta isu-isu yang dianggap penting oleh masyarakat.

Masyarakat memiliki peran yang sangat penting dalam proses pemindahan Ibukota Negara (IKN) ke Nusantara. Pemindahan ini menimbulkan beragam opini, kritik, dan masukan dari berbagai kalangan, yang disuarakan melalui media sosial seperti *platform X* (Narasi *Newsroom*, 2024). Bagi masyarakat, penelitian ini penting karena memungkinkan mereka memberikan masukan yang konstruktif serta mengontrol jalannya proyek. Penelitian analisis topik menggunakan metode *topic modelling*, opini publik dapat diorganisir secara objektif dan menyeluruh, sehingga isu-isu yang diprioritaskan oleh masyarakat bisa diidentifikasi dengan lebih jelas. Hal ini membantu masyarakat untuk mengantisipasi kekurangan yang mungkin muncul serta berperan sebagai penengah dalam perdebatan yang ada. Selain itu, masyarakat juga dapat mengidentifikasi hal-hal yang perlu diperbaiki dan memberikan apresiasi terhadap aspek-aspek yang dianggap berhasil dalam proses pemindahan IKN. Penelitian ini juga memberi kesempatan bagi masyarakat untuk menyuarakan kritik yang lebih berbobot, yang pada akhirnya dapat diolah menjadi masukan yang penting bagi pengambil keputusan.

Lalu dari sisi pemerintah dan para pemangku kepentingan, analisis *topic modelling* terhadap diskusi publik mengenai IKN memberikan manfaat yang sangat penting. Penelitian ini dapat membantu pemerintah dalam mengidentifikasi topik-topik yang menjadi perhatian utama masyarakat, sehingga bisa memprioritaskan hal-hal yang dianggap paling mendesak. Masukan dari masyarakat, baik dalam

bentuk kritik maupun apresiasi, dapat dijadikan sebagai bahan evaluasi untuk memperbaiki kebijakan atau strategi komunikasi yang diterapkan terkait proyek IKN (Anggraini, 2022). Selain itu, dengan memahami isu-isu yang berkembang, pemerintah dapat mengantisipasi potensi risiko atau kekhawatiran yang ada di kalangan masyarakat (Purnama & Chotib, 2023). Oleh karena itu, topik-topik yang terungkap dari analisis ini bukan hanya membantu pemerintah dalam mengelola komunikasi proyek dengan lebih efektif, tetapi juga mendukung terciptanya dialog yang lebih sehat antara masyarakat dan pemerintah, yang pada akhirnya memperbaiki kualitas keputusan dan kebijakan yang diambil terkait IKN.

Pada penelitian analisis teks, terdapat beberapa metode yang sering digunakan, antara lain analisis sentimen, analisis *social network*, dan *topic modelling* (Wanniarachchi dkk., 2023). Pada analisis teks yang kompleks seperti opini masyarakat terhadap pemindahan Ibukota Nusantara, *topic modelling* merupakan metode yang tepat karena mampu mengungkap tema-tema tersembunyi dari kumpulan data teks yang besar. Tidak seperti metode lain yang hanya berfokus pada sentimen / *social network*, *topic modelling* memberikan gambaran yang lebih mendalam tentang topik-topik utama yang sedang dibahas dalam teks. Metode ini tidak hanya mengidentifikasi kata-kata yang sering muncul, tetapi juga mengungkap hubungan antar kata dan bagaimana mereka membentuk topik-topik tertentu (Hardiyanti dkk., 2023). Menggunakan *topic modelling*, dapat secara otomatis mengelompokkan diskusi masyarakat menjadi beberapa topik berbeda, yang mempermudah pemahaman isu-isu utama terkait pemindahan IKN. Oleh karena itu, metode ini dipilih karena kemampuannya untuk memberikan hasil yang lebih kaya dan terstruktur dibandingkan metode lainnya, terutama dalam analisis data teks besar seperti data dari media sosial.

Topic modelling adalah metode yang bertujuan untuk menemukan struktur tersembunyi dalam kumpulan dokumen dengan mengidentifikasi sekumpulan topik yang tersembunyi (Cahyono & Astuti, 2023). Salah satu algoritma yang paling banyak digunakan dalam *topic modelling* adalah *Latent Dirichlet Allocation* (LDA). LDA bekerja dengan mengasumsikan bahwa setiap dokumen adalah campuran dari berbagai topik, dan setiap topik adalah distribusi dari kata-kata (Garg & Rangra, 2022).

Keuntungan penggunaan LDA adalah secara efektif mengelompokkan teks dalam jumlah besar ke dalam beberapa topik, yang sangat berguna saat menganalisis data media sosial (Erniyati dkk., 2023). Melalui *model* ini, opini publik yang kompleks dapat direduksi menjadi topik-topik utama yang mudah dianalisis. Namun LDA juga memiliki beberapa kekurangan, seperti perlunya menentukan jumlah topik sebelum dianalisis, sebab jika tidak tepat maka akan mempengaruhi keakuratan hasil (Baghmohammad dkk., 2021). Selain itu, LDA sangat bergantung pada kualitas *preprocessing data* dan memiliki keterbatasan ketika menangani topik yang tumpang tindih atau topik yang berubah seiring waktu (Been & Byeon, 2023).

Cara mengatasi kelemahan-kelemahan tersebut, penelitian ini akan mengadopsi beberapa langkah perbaikan. Pertama, penentuan jumlah topik akan dioptimalkan dengan menggunakan metode *Coherence Score*, yang dapat membantu memastikan bahwa jumlah topik yang dipilih adalah yang paling representatif dari data yang dianalisis. Pada tahap *preprocessing data*, dilakukan *prepocessing* standar dengan tambahan normalisasi.

Tujuan dari penelitian ini adalah untuk mengidentifikasi dan menganalisis topik-topik utama yang dibahas oleh masyarakat mengenai IKN di *Platform X*. Penelitian ini diharapkan dapat memberikan wawasan/*insight* yang lebih mendalam tentang persepsi dan isu-isu yang berkembang terkait pemindahan ibukota.

1.2 Rumusan Masalah

Berdasarkan uraian latar belakang di atas, maka rumusan masalah dari proposal Tugas Akhir (TA) ini yaitu :

Bagaimana metode *Latent Dirichlet Allocation* (LDA) dapat digunakan untuk mengidentifikasi dan mengelompokkan topik-topik utama yang muncul dalam diskusi masyarakat terkait pemindahan Ibukota Nusantara (IKN) di *Platform X*?

1.3 Batasan Masalah

Adapun batasan masalah dalam melakukan penelitian ini adalah sebagai berikut:

1. Penelitian ini hanya menggunakan data teks yang diambil dari Platform X selama periode tertentu yang mencakup diskusi terkait pemindahan Ibukota Nusantara (IKN). Rentang waktu yang digunakan adalah 1 Januari 2024 – 30 Agustus 2024.
2. Penelitian ini akan menggunakan optimasi jumlah topik dengan metode *Coherence Score*.
3. Data yang digunakan dalam penelitian kali ini menggunakan 1000 data *tweet* dengan *keyword* yaitu IKN.
4. Data yang diambil merupakan *tweet* berbahasa Indonesia
5. Kamus NLP yang digunakan menggunakan Sastrawi

1.4 Tujuan

Berdasarkan uraian rumusan masalah, maka tujuan yang ingin dicapai dalam penelitian ini adalah :

Mengaplikasikan model *Latent Dirichlet Allocation* (LDA) untuk mengidentifikasi dan mengelompokkan topik-topik utama yang muncul dalam diskusi masyarakat terkait pemindahan Ibukota Nusantara (IKN) di *Platform X*.

1.5 Manfaat

Penelitian ini dapat memberikan manfaat sebagai berikut :

1. Memperkaya literatur ilmiah dalam analisis teks, khususnya penerapan *Latent Dirichlet Allocation* (LDA) untuk *topic modelling* di media sosial, serta menjadi referensi bagi penelitian serupa di masa depan.
2. Memberikan wawasan bagi pemerintah dan pemangku kepentingan mengenai isu-isu utama yang dibahas masyarakat terkait pemindahan IKN, sehingga dapat digunakan untuk strategi komunikasi dan respons kebijakan yang lebih efektif.
3. Berkontribusi dalam pengembangan *big data analytics*, terutama dalam analisis teks skala besar, dengan optimasi model yang dapat dijadikan referensi bagi penelitian dan aplikasi analisis data yang lebih canggih kedepannya.

BAB II
LANDASAN TEORI

2.1 Penelitian Terdahulu

Tabel 2.1 Penelitian Terdahulu

Peneliti	Judul	Hasil	Persamaan & Perbedaan
Hery Oktafiandi	Implementasi LDA untuk Pengelompokan Topik *Twitter* Bertagar #Mypertamina	Penelitian ini mengeksplorasi Twitter dengan fokus pada tagar #Mypertamina. Data yang dianalisis berasal dari 149 *tweet* yang dikelompokkan menggunakan metode LDA. Temuan penelitian mengindikasikan adanya 3 kluster data dengan nilai koherensi tertinggi mencapai 0.468.	Persamaan : Menggunakan Model LDA untuk mengelompokkan topik. *Preprocessing* yang digunakan standar Menggunakan *pyLDAvis* sebagai visualisasi implementasi model LDA. Perbedaan : Tidak menggunakan BoW, melainkan menggunakan model *Bigram* dan *Trigram*.
Sajarwo Anggai, Tukiyat, Abu Khalid Rivai, Rafi Mahmud Zain	Ekstraksi Topik dalam *Dataset* Menggunakan Teknik Pemodelan Topik	Penelitian ini membahas tentang ekstraksi untuk menentukan topik-topik utama terkait pidato serta publikasi media terkait Presiden Joko Widodo. Hasil dari model LDA didapat 16 topik dengan nilai koheren 0.554, serta 21 topik dengan nilai *perplexity* -13.130	Persamaan : *Preprocessing* yang digunakan menggunakan *preprocessing standar* Implementasi visualisasi *model* menggunakan *pyLDAvis*. Perbedaan : tidak menggunakan BoW, melainkan menggunakan TF-IDF.
Yoga Sahria & Dhomas Hatta Fudholi	Analisis Topik Penelitian Kesehatan di Indonesia Menggunakan Metode *Topic Modeling* LDA (*Latent Dirichlet Allocation*)	Penelitian ini menghasilkan 2 topik dengan nilai koheren yaitu 0.57. Berdasarkan hasil tersebut, dilakukan analisis interpretasi topik dan hasilnya adalah terdapat 2 topik dominan yaitu topik umum dan topik penyakit.	Persamaan : Menggunakan *preprocessing* standar. . Menggunakan visualisasi *pyLDAvis*. Perbedaan : tidak menggunakan BoW, penelitian tersebut memakai TF-IDF. Data yang digunakan berasal dari jurnal yang membahas terkait kesehatan.

2.2 Ibukota Nusantara (IKN)

Pemindahan Ibukota Negara (IKN) dari Jakarta ke Nusantara merupakan proyek nasional yang mencerminkan langkah besar dalam upaya pemerintah Indonesia untuk mengatasi permasalahan seperti overpopulasi, polusi, dan ketimpangan pembangunan di Jakarta (Fristikawati dkk., 2022). Selain itu, pemindahan ini diharapkan dapat mempercepat pembangunan di wilayah Kalimantan, menciptakan pemerataan ekonomi, dan mendukung konsep pembangunan berkelanjutan yang meliputi pembangunan hijau dan *smart city* (Fristikawati & Adipradana, 2022).

Namun, rencana ini tidak lepas dari berbagai perdebatan, baik terkait pelaksanaan pembangunan maupun momen peresmiannya yang dianggap oleh sebagian pihak terlalu terburu-buru. Beberapa pihak memandang proyek ini sebagai *strategic move* dalam memajukan wilayah luar Jawa, sementara yang lain mempertanyakan efektivitas alokasi anggaran negara dan dampaknya terhadap lingkungan (Ma'mun, 2023). Diskusi mengenai IKN tidak hanya berlangsung di forum resmi, tetapi juga meluas di media sosial. Salah satunya adalah *platform X* sebagai media atau *platform* untuk beradu argumen dan opini, yang tentunya analisis teks menjadi alat penting untuk memahami opini masyarakat. Melalui pemanfaatan model algoritma seperti LDA, penelitian ini bertujuan menghasilkan topik pembicaraan publik yang nantinya dapat menjadi perhatian pemerintah serta mencerminkan kekhawatiran masyarakat terkait IKN.

2.3 Pemrograman *Python*

Python adalah salah satu bahasa pemrograman yang populer dan banyak digunakan di berbagai bidang, terutama untuk analisis data, pembelajaran mesin (*machine learning*), serta pengolahan bahasa alami (*natural language processing*) (Jalolov, 2023). *Python* dikenal karena sintaksnya yang sederhana dan intuitif, sehingga memudahkan para pemula sekaligus memberikan fleksibilitas bagi pengguna tingkat lanjut. Selain itu, *Python* memiliki ekosistem pustaka/*library* yang sangat kaya, seperti *NumPy*, *Pandas*, dan *Matplotlib* yang digunakan untuk manipulasi dan visualisasi data. Pada bidang *text mining*, *Python* juga menyediakan *library* seperti *NLTK* dan *spaCy*, yang memungkinkan proses pengolahan bahasa

alami (NLP), seperti *tokenized*, *stemming*, dan penghapusan *stopwords*, untuk membantu mempersiapkan data teks mentah menjadi lebih bersih dan siap dianalisis.

Pada konteks penelitian ini, *Python* digunakan untuk mengimplementasikan algoritma *topic modelling Latent Dirichlet Allocation* (LDA). *Library Gensim* adalah salah satu *library Python* yang dirancang khusus untuk analisis *topic modelling* dan akan digunakan untuk membangun *model* LDA yang mampu mengidentifikasi topik-topik tersembunyi dalam diskusi masyarakat di *Platform X*. Keunggulan lain *Python* adalah kemampuannya untuk menangani data dalam skala besar dengan mudah, termasuk pengolahan teks yang kompleks (Mahammadodilovich, 2023). Kombinasi antara pustaka-pustaka yang tersedia dan sintaks yang sederhana membuat *Python* menjadi alat yang ideal untuk melakukan analisis data teks yang dihasilkan dari platform media sosial dalam penelitian ini.

2.4 *Topic Modelling*

Topic modelling adalah teknik yang digunakan untuk menemukan struktur tersembunyi dalam sekumpulan dokumen (Cahyono & Astuti, 2023). Pada konteks teks, metode ini berfungsi untuk mengidentifikasi dan mengelompokkan berbagai topik yang terdapat dalam kumpulan data teks yang besar. *Topic modelling* sangat bermanfaat untuk mengelola dan menganalisis data teks yang tidak terstruktur, seperti data dari media sosial atau artikel berita, dengan tujuan untuk mengidentifikasi tema-tema utama yang sedang dibahas (Murshed dkk., 2023). Algoritma yang paling umum digunakan dalam *topic modelling* adalah *Latent Dirichlet Allocation* (LDA). Pada penelitian ini, *topic modelling* diterapkan untuk mengidentifikasi topik-topik utama yang muncul dalam diskusi masyarakat di *platform X* terkait dengan pemindahan Ibukota Nusantara (IKN). Teknik tersebut relevan karena opini masyarakat mengenai IKN sangat beragam, mencakup isu lingkungan, ekonomi, sosial, hingga pandangan politik. Oleh karena itu, dengan memanfaatkan *topic modelling*, penelitian ini mampu memberikan wawasan mendalam tentang persepsi publik terkait IKN yang tersebar di *platform* media sosial, khususnya pada *platform X*.

2.5 *Latent Dirichlet Allocation* (LDA)

Latent Dirichlet Allocation (LDA) adalah algoritma *topic modelling* yang paling populer. LDA bekerja dengan mengasumsikan bahwa setiap dokumen adalah campuran dari beberapa topik, dan setiap topik adalah distribusi dari kata-kata (Garg & Rangra, 2022). LDA menghasilkan sekumpulan topik yang masing-masing diwakili oleh sekumpulan kata-kata kunci yang dominan dalam dokumen tersebut. Keunggulan LDA terletak pada kemampuannya untuk menangani data teks yang besar dan mengidentifikasi topik secara otomatis tanpa supervisi (Pan & Xue, 2023). Tidak hanya itu saja, LDA juga tergolong dalam *Soft Clustering* yang di mana setiap objek/kata yang tersusun dapat memiliki lebih dari satu *cluster* topik (Pardede & Waskita, 2023). Namun, LDA memiliki kelemahan, seperti ketergantungan pada jumlah topik yang harus ditentukan di awal, yang memengaruhi akurasi hasil (Inoue dkk., 2023).

LDA menggunakan distribusi *Dirichlet*, yaitu distribusi yang digunakan untuk menentukan proporsi topik dalam dokumen serta proporsi kata dalam topik (Yu & Xiang, 2023). Distribusi tersebut memungkinkan algoritma untuk memodelkan variabilitas topik di antara dokumen secara fleksibel, tanpa harus menentukan *keyword* secara eksplisit. Pendekatan probabilistik yang digunakan oleh LDA membantu dalam menangani kumpulan dokumen teks yang besar secara otomatis dan efisien. Hal ini memungkinkan LDA untuk mengidentifikasi topik tersembunyi dalam teks secara otomatis, bahkan ketika teks tersebut kompleks dan heterogen.

Adapun langkah awal dari LDA yaitu penentuan Distribusi Topik θ (langkah awal untuk menentukan proporsi awal topik untuk dokumen). Distribusi topik dalam dokumen ditentukan berdasarkan distribusi *Dirichlet* (Blei dkk., 2003) :

$$\theta \sim Dir(\alpha) \quad (1)$$

Setelah distribusi topik diketahui, LDA menentukan topik Zn (topik ke-n) yang dialokasikan untuk setiap kata dalam dokumen menggunakan distribusi *Multinomial* berdasarkan θ (Blei dkk., 2003)

$$\mathcal{Z}n \sim Multinomial(\theta) \quad (2)$$

Selanjutnya, berdasarkan topik yang dipilih Zn, kata tertentu Wn (kata ke-n) dipilih dari distribusi kata β (Blei dkk., 2003)

$$\mathcal{W}n \sim p(\mathcal{W}n|\mathcal{Z}n, \beta) \quad (3)$$

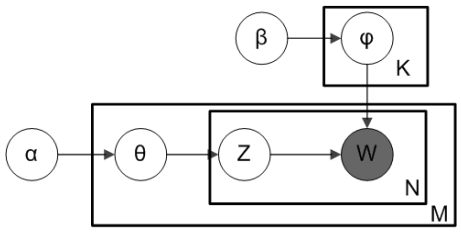
Secara matematis, LDA bekerja berdasarkan distribusi probabilitas dengan rumus dasar sebagai berikut (Blei dkk., 2003) :

$$p(\theta, Z, W|\alpha, \beta) = p(\theta|\alpha) \prod_{n=1}^{N} p\,(Zn|\theta) p(Wn|Zn, \beta) \quad (4)$$

Keterangan dari rumus tersebut adalah sebagai berikut :

1. θ adalah distribusi topik dalam dokumen.
2. Zn adalah topik yang dialokasikan untuk kata ke-n dalam dokumen.
3. Wn adalah kata ke-n dalam dokumen tersebut.
4. α adalah *hyperparameter* distribusi *Dirichlet* yang mengontrol seberapa tersebar topik dalam dokumen.
5. β adalah *hyperparameter* distribusi *Dirichlet* untuk distribusi kata dalam topik.

Distribusi *Dirichlet* $p(\theta \mid \alpha)$ mengatur proporsi topik yang ada dalam dokumen, sementara $p(Zn \mid \theta)$ menentukan probabilitas topik mana yang dipilih untuk setiap kata dalam dokumen. Setelah topik dipilih, $p(Wn \mid Zn, \beta)$ menentukan probabilitas memilih kata tertentu dalam suatu topik berdasarkan distribusi kata yang diatur oleh β. Proses iteratif ini menghasilkan distribusi topik yang optimal untuk setiap dokumen.



Gambar 2.1 Konsep LDA
(Sumber: Suparyati & Utami, 2022:3)

Penjelasan bagan :

1. M menandakan banyaknya dokumen
2. N menandakan banyaknya kata dalam 1 dokumen tertentu

3. α parameter *Dirichlet*. Mengontrol distribusi topik setiap dokumen
4. β parameter *Dirichlet*. Mengontrol distribusi kata untuk setiap topik
5. θ merupakan topik distribusi untuk dokumen M. Merepresentasikan probabilitas dokumen M
6. φ merupakan distribusi kata untuk topik tertentu. Merepresentasikan probabilitas topik ke-k, yang mengandung kata tertentu.
7. Z merupakan topik untuk kata ke-n dari suatu dokumen
8. W merupakan spesifik kata

Berdasarkan bagan pada gambar 1, α θ β φ adalah *latent* / *hidden variabel*, sedangkan W adalah variabel yang di observasi.

Penerapan LDA, khususnya dalam penelitian ini, beberapa *library* yang sering digunakan adalah *Gensim* dan *PyLDAvis*. *Gensim* menyediakan fungsi-fungsi yang efisien untuk pemrosesan data teks dalam jumlah besar, termasuk penerapan LDA dengan optimasi yang memungkinkan identifikasi topik secara akurat (Gomez dkk., 2023). Selain itu, *PyLDAvis* digunakan untuk memvisualisasikan hasil *topic modelling*, sehingga memudahkan pemahaman terhadap hubungan antara topik dan kata-kata yang menyusunnya (Naury dkk., 2021). Kedua *library* ini bekerja secara sinergis untuk mendukung analisis topik yang mendalam dan memvisualisasikan topik secara interaktif, yang membuat hasil analisis lebih mudah diinterpretasikan.

2.6 *Natural Language Processing* (NLP)

Natural Language Processing (NLP) adalah cabang dari kecerdasan buatan (AI) yang berfokus pada interaksi antara komputer dan bahasa manusia (Wang dkk., 2023). Tujuan utama NLP adalah untuk memungkinkan mesin memahami, menginterpretasikan, dan menghasilkan teks atau ucapan dalam bahasa alami. Pada tahap pengolahan teks, NLP mencakup berbagai tugas seperti tokenisasi, *stemming*, penghapusan *stopwords*, serta ekstraksi fitur teks (Sergii V. & Oleksandr V., 2023). Proses tersebut penting dalam *text mining* untuk membersihkan dan menyiapkan data sebelum dianalisis lebih lanjut. Teknik NLP juga digunakan dalam berbagai aplikasi seperti *Virtual Assistant*, *Neural Machine Translation*, *Chatbots*, dan sistem pencarian informasi (Patil dkk., 2023).

Pada konteks penelitian ini, NLP berperan penting dalam proses *preprocessing* teks dari *platform X* untuk meningkatkan kualitas data sebelum dianalisis menggunakan model *topic modelling*. Langkah-langkah *preprocessing* NLP digunakan untuk memastikan data teks menjadi lebih bersih dan terstruktur, sehingga algoritma *Latent Dirichlet Allocation* (LDA) dapat bekerja secara optimal. Melalui bantuan *library* seperti NLTK di *Python*, penelitian ini mampu mengidentifikasi topik utama dalam diskusi terkait pemindahan Ibukota Nusantara (IKN) di Platform X secara lebih akurat dan mendalam.

2.7 *Coherence Score*

Coherence Score merupakan salah satu metrik yang berfungsi untuk mengukur kualitas model *topic modelling*, seperti *Latent Dirichlet Allocation* (LDA). Metrik ini mengevaluasi seberapa konsisten atau relevan kata-kata kunci dalam setiap topik yang dihasilkan oleh model (Pardede & Waskita, 2023). Nilai *coherence score* yang lebih tinggi menunjukkan bahwa topik yang dihasilkan memiliki kualitas yang lebih baik, karena kata-kata dalam topik tersebut lebih saling berkaitan (Ajinaja dkk., 2023).

Pada penerapan LDA, *coherence score* dimanfaatkan untuk menentukan jumlah topik (κ) yang optimal. Caranya adalah dengan membandingkan *coherence score* pada berbagai nilai κ, dapat diketahui jumlah topik yang paling sesuai untuk mewakili data teks yang dianalisis. Perhitungan *coherence score* umumnya didasarkan pada hubungan *co-occurrence*, di mana pasangan kata dalam suatu topik dinilai berdasarkan kemunculan bersama mereka dalam dokumen (Weisser dkk., 2023). Perhitungan ini menggunakan distribusi kata dalam *corpus* dengan rumus berikut (Pardede & Waskita, 2023) :

$$score(vi, vj, \epsilon) = \log \frac{D(vi, vj) + \epsilon}{D(vj)} \quad (5)$$

Keterangan:

1. vi dan vj merupakan pasangan kata kunci yang dianalisis dalam topik
2. $D(vi, vj)$ merupakan jumlah dokumen yang mengandung kedua kata vi dan vj
3. $D(vj)$ merupakan jumlah dokumen yang mengandung kata vj

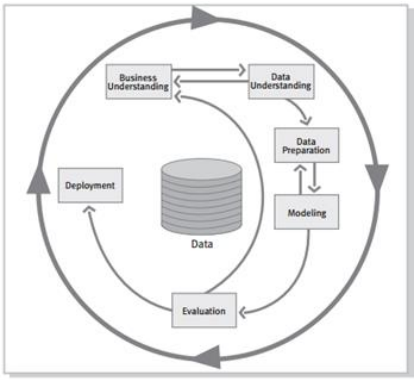
4. ϵ merupakan parameter *smoothing* kecil (biasanya mendekati nol) untuk menghindari kesalahan akibat pembagian dengan nol

Coherence score memberikan cara kuantitatif untuk mengevaluasi interpretasi topik yang dihasilkan. Pada penelitian ini, metode tersebut digunakan untuk memilih jumlah topik κ yang paling relevan terhadap diskusi masyarakat mengenai pemindahan Ibukota Nusantara (IKN) di *platform X*.

2.8 CRISP-DM

Data mining dan *data science* merupakan bidang yang penting dan terus berkembang pesat dalam era *big data*. Kedua disiplin ilmu tersebut memberikan kemampuan untuk mengolah, menganalisis, dan menafsirkan data dalam jumlah besar guna menghasilkan informasi yang bermakna dan dapat digunakan (Prastiwi dkk., 2022). Pada penerapannya, diperlukan pendekatan metodologis yang sistematis agar setiap tahap dapat dilaksanakan secara terstruktur dan mencapai hasil yang diharapkan.

CRISP-DM (*Cross Industry Standard Process for Data Mining*) adalah metodologi standar yang paling banyak digunakan dalam proyek *data mining* dan *data science*. Proses ini terdiri dari enam fase yang berurutan namun *iteratif*, yaitu *Business Understanding*, *Data Understanding*, *Data Preparation*, *Modelling*, *Evaluation*, dan *Deployment* (Martinez-Plumed dkk., 2021). CRISP-DM menawarkan pendekatan yang fleksibel dan mudah disesuaikan dengan berbagai tipe proyek analitik, sehingga membantu *data scientist* dalam merancang solusi analisis yang komprehensif dan terstruktur (Bokrantz dkk., 2023).



Gambar 2.2 Metodologi CRISP-DM
(Sumber: Halim dkk., 2023:2)

Tahapan pertama adalah *Business Understanding*, di mana pemahaman tentang tujuan bisnis menjadi prioritas (Pambudi, 2023). Pada fase ini, penting untuk merumuskan pertanyaan yang jelas terkait dengan persepsi publik mengenai IKN dan apa yang ingin dicapai dari analisis data tersebut, serta kriteria keberhasilan proyek. Selanjutnya adalah tahap *Data Understanding*, yang melibatkan pengumpulan data awal dan eksplorasi data untuk memahami karakteristik *dataset* (Pambudi, 2023). Eksplorasi ini bertujuan untuk memastikan bahwa data yang tersedia cukup representatif dan dapat diolah lebih lanjut untuk menjawab pertanyaan penelitian. Setelah memahami data, langkah berikutnya adalah *Data Preparation*, yang mencakup proses pembersihan dan transformasi data (Pambudi, 2023). Tahap ini bertujuan untuk mempersiapkan data yang bersih dan relevan agar model dapat bekerja secara optimal. Tahap keempat adalah *Modelling*, di mana algoritma *machine learning* / *data mining* diterapkan untuk membangun *model* (Christian & Qi, 2022). Pada kasus ini, *model* LDA (*Latent Dirichlet Allocation*) digunakan untuk analisis topik. Setelah *model* diterapkan, dilakukan fase *Evaluation* untuk menilai kinerja *model* yang telah dibangun (Pambudi, 2023). Evaluasi ini memastikan bahwa *model* siap untuk digunakan dalam langkah-langkah selanjutnya. Tahap terakhir melibatkan implementasi model di dunia nyata, baik dengan cara menghasilkan laporan / presentasi hasil untuk *stakeholder* atau bahkan dengan menerapkan *model* ke dalam sistem (Pambudi, 2023). Pada kasus ini, hasil visualisasi *model* LDA menggunakan *PyLDAvis* dapat dipresentasikan untuk memudahkan pemahaman mengenai topik-topik yang ditemukan dari diskusi publik terkait IKN.

BAB III
METODOLOGI PENELITIAN

Bab ini menjelaskan langkah-langkah yang dilakukan dalam penelitian ini berdasarkan pendekatan metodologi CRISP-DM. Tahapan-tahapan ini dirancang untuk memberikan kerangka kerja yang sistematis dalam menyelesaikan permasalahan penelitian, khususnya dalam analisis topik mengenai persepsi publik terhadap IKN (Ibu Kota Nusantara) pada *platform X* menggunakan model LDA.

3.1 *Business Understanding*

Pada tahap ini, tujuan utama adalah memahami masalah bisnis dan konteks yang melatarbelakangi penelitian, serta merumuskan solusi yang ingin dicapai melalui proyek penelitian ini. Pemindahan Ibu Kota Negara (IKN) ke Kalimantan Timur menjadi topik yang ramai diperbincangkan di *platform X*, mencakup berbagai perspektif seperti dukungan terhadap pembangunan, dampak ekonomi, hingga kritik dari masyarakat. Namun, karena volume percakapan (*tweet*) yang besar & data yang tidak terstruktur, sulit untuk memahami pola utama dalam diskusi ini secara menyeluruh. Oleh karena itu, penelitian ini menggunakan *topic modelling* dengan LDA untuk mengelompokkan topik secara otomatis, sehingga dapat mengungkap isu-isu dominan tanpa harus membaca setiap *tweet* secara manual.

Hasil analisis ini, diharapkan dapat memberikan *insight* bagi pemerintah dan masyarakat. Pemerintah dapat memahami opini publik, mengidentifikasi isu-isu krusial, dan menyusun strategi komunikasi yang lebih efektif. Sementara itu, masyarakat dapat memperoleh wawasan lebih terstruktur mengenai berbagai perspektif yang berkembang.

3.2 *Data Understanding*

Gambar 3.1 *Data Understanding*

Pada tahap ini, penelitian berfokus pada pengumpulan dan pemahaman data yang akan dianalisis. Data yang digunakan dalam penelitian ini diambil dari *Platform X* (sebelumnya dikenal sebagai *Twitter*), menggunakan teknik *tweet harvesting* untuk mengumpulkan opini dan diskusi publik terkait pemindahan Ibukota Nusantara (IKN)

. Data yang diperoleh sejumlah 1.000 data yang meliputi *tweet* yang memuat kata kunci yang relevan, seperti "Ibukota Nusantara," "IKN," serta berbahasa Indonesia `lang=*id*`, selama periode 1 Januari – 30 Agustus 2024. Teknik *tweet harvesting* memungkinkan pengambilan data langsung dari *platform* media sosial menggunakan *Auth Token* yang ada pada saat *login* ke dalam aplikasi *X*. Data ini diambil dalam bentuk CSV yang berisi berbagai informasi tentang setiap *tweet*, seperti konten teks, *metadata* pengguna, dan waktu publikasi. Setelah data berhasil di *crawling*, langkah selanjutnya adalah memahami struktur *dataset*. *Dataset* yang diperoleh terdiri dari 1.000 baris data dengan berbagai atribut/fitur/kolom yang menggambarkan karakteristik setiap *tweet*. Eksplorasi ini dilakukan untuk memastikan konsistensi data dan memvalidasi keberadaan atribut yang dibutuhkan. Berikut penjelasan dari atribut/fitur/kolom yang terdapat dalam *dataset* yang berhasil di *crawling* menggunakan *tweet harvest* :

a) ***conversation_id_str*** : ID unik untuk setiap percakapan atau *thread* di *Platform X*. Ini memungkinkan peneliti melacak seluruh rangkaian diskusi yang terkait dalam satu percakapan.
b) ***created_at*** : Waktu dan tanggal ketika *tweet* tersebut di *posting*. Formatnya mengikuti standar waktu UTC yang menyediakan informasi tentang kapan *tweet* tersebut dikirim.
c) ***favorite_count*** : Jumlah kali *tweet* tersebut disukai (*liked*) oleh pengguna lain di *platform*. Atribut ini menunjukkan tingkat popularitas *tweet* tertentu.
d) ***full_text*** : Isi lengkap dari *tweet*, yang memuat opini atau diskusi terkait IKN. Analisis dilakukan terhadap teks ini untuk mengidentifikasi topik dan isu utama.
e) ***id_str*** : ID unik yang diberikan untuk setiap *tweet*, yang dapat digunakan untuk identifikasi spesifik *tweet* dalam analisis atau pengambilan data lebih lanjut.
f) ***image_url*** : URL dari gambar yang diunggah bersama *tweet*

g) ***in_reply_to_screen_name*** : Menunjukkan nama pengguna yang dituju oleh *tweet* tersebut, jika *tweet* itu merupakan balasan terhadap *tweet* orang lain.

h) ***Lang*** : Bahasa yang digunakan dalam *tweet*, misalnya *in* (Indonesia), yang membantu dalam pemrosesan teks untuk mengelompokkan *tweet* berdasarkan bahasa.

i) ***Location*** : Lokasi geografis dari pengguna yang mengirimkan *tweet*, jika tersedia. Atribut ini memberikan konteks tambahan terkait persepsi masyarakat di wilayah tertentu.

j) ***quote_count*** : Jumlah kali *tweet* tersebut dikutip oleh pengguna lain. Ini menunjukkan seberapa sering *tweet* digunakan sebagai referensi dalam diskusi lain.

k) ***reply_count*** : Jumlah balasan yang diterima oleh *tweet*, menunjukkan tingkat interaksi langsung dengan *tweet* tersebut.

l) ***retweet_count*** : Jumlah kali *tweet* tersebut dibagikan atau di-*retweet* oleh pengguna lain, menunjukkan seberapa *viral* atau tersebar *tweet* tersebut.

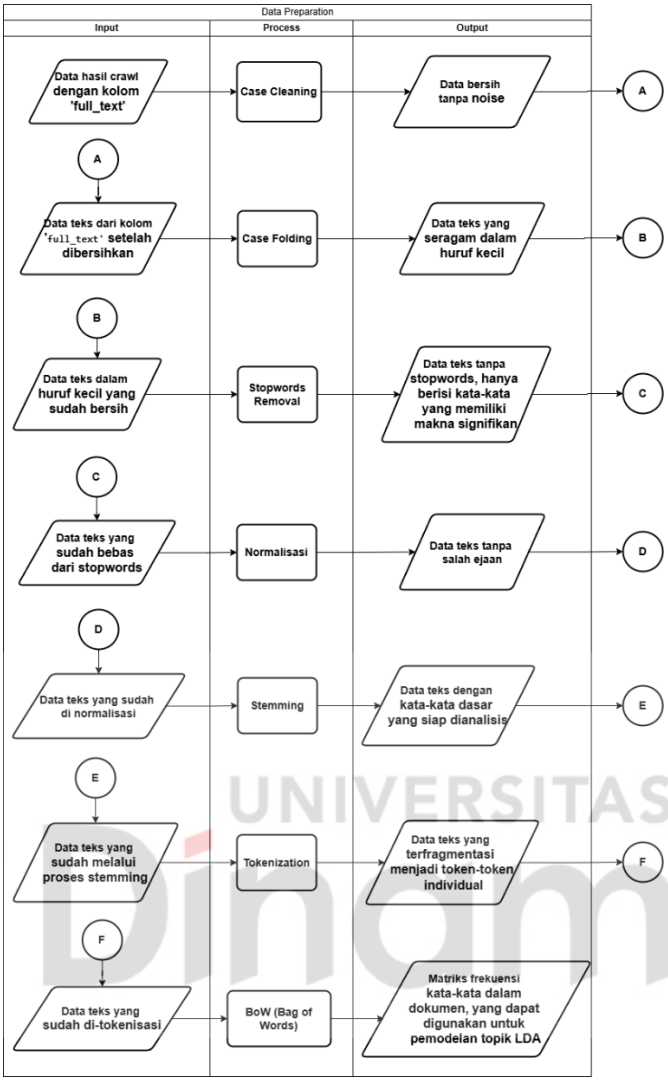
m) ***tweet_url*** : URL dari *tweet* tersebut, yang memungkinkan akses langsung ke *tweet* jika diperlukan.

n) ***user_id_str*** : ID unik dari pengguna yang mem*posting tweet*, yang memungkinkan identifikasi lebih lanjut mengenai perilaku dan preferensi pengguna.

o) ***Username*** : Nama pengguna yang terkait dengan akun yang mem*posting tweet*. Atribut ini sering digunakan untuk analisis perilaku pengguna atau untuk menyegmentasikan data berdasarkan pengirim.

Berdasarkan analisis di atas, maka atribut/fitur/kolom yang akan digunakan nantinya dalam proses *Data Preparation* dan seterusnya adalah kolom '*full_text*' yang memuat isi *tweet*, karena kolom ini mengandung informasi utama yang akan diolah untuk mengidentifikasi topik-topik dalam diskusi terkait IKN. Atribut ini penting dalam penerapan *Latent Dirichlet Allocation* (LDA), di mana teks *tweet* akan diolah melalui *preprocessing*, guna menghasilkan representasi teks yang siap untuk dianalisis.

3.3 *Data Preparation*



Gambar 3.2 IPO *Data Preparation*

Tahap ini adalah proses mempersiapkan data untuk dianalisis lebih lanjut. Beberapa teknik yang digunakan dalam tahap ini termasuk :

a) *Case Cleaning*: Menghilangkan data yang tidak relevan, duplikat, atau tidak valid, seperti *tweet* dengan teks kosong atau hanya berisi simbol.

b) *Case Folding*: Mengubah semua huruf dalam teks menjadi huruf kecil untuk menjaga konsistensi dalam analisis.

c) *Stopword Removal*: Menghapus kata-kata yang umum muncul tetapi tidak memiliki makna signifikan dalam analisis (seperti "dan", "di", "ke", dll.) menggunakan daftar *stopword* dalam bahasa Indonesia.

d) Normalisasi: Mengubah kata-kata yang terdapat salah ejaan atau salah ketik pada setiap dokumen menjadi kata-kata yang benar (sesuai KBBI).

e) *Stemming*: Mengubah setiap kata ke bentuk dasar atau akarnya. Misalnya, kata "berjalan" dan "berjalannya" akan diubah menjadi kata dasar "jalan". *Stemming* dalam bahasa Indonesia dilakukan menggunakan *library* Sastrawi.

f) *Tokenization*: Memecah teks menjadi kata-kata individual (*token*), yang merupakan langkah penting untuk analisis berbasis teks.

g) *BoW (Bag of Words)*: Menggunakan metode *Bag of Words* untuk merepresentasikan teks sebagai kumpulan kata-kata unik (*vocabulary*) dan menghitung frekuensi kemunculannya. Teknik ini menghasilkan matriks frekuensi yang berguna untuk analisis topik menggunakan model LDA.

Lalu untuk *pseudocode* dari *Data Preparation* adalah sebagai berikut :

Start Data Preparation

1. *Load dataset* dari *file CSV*

   Input : *dataset* yang berisikan kolom '*full_text*' saja

   Output : *loaded dataset*

2. *Clean data*

   For setiap baris dalam *dataset* :

   If baris kosong *OR* mengandung simbol :

   Remove baris

   Else If baris memiliki entri yang duplikat :

   Remove baris duplikat

   Else If baris mengandung *tag* yang tidak diperlukan (#, @, *URLs*)

   Remove these elements

   Else If baris mengandung *emoji OR* karakter khusus *OR number*

   Remove emoji OR karakter khusus OR number

   EndIf

   Output : *cleanded dataset*

3. *Perform case folding*

   For setiap teks dalam kolom '*full_text*' :

   Convert text to lowercase

   Output : Teks dengan huruf yang konsisten

4. *Remove stopwords*

   Load daftar *stopword* bahasa Indonesia

   For setiap teks dalam kolom '*full_text*' :

      Remove kata yang terdapat dalam daftar *stopword*

   Output : teks tanpa *stopwords*

5. *Perform normalization*

   Load dictionary normalisasi dari *file* eksternal

   For setiap kata dalam kolom '*full_text*' :

      Replace kata yang ditemukan dalam *dictionary* dengan kata yang benar

   Output : Teks yang sudah dinormalisasi

6. *Perform stemming*

   Load Indonesian stemming library (Sastrawi)

   For setiap kata dalam kolom '*full_text*' :

      Convert kata ke bentuk dasar

   Output : Teks dengan kata dasar

7. *Tokenize text*

   For setiap teks dalam kolom '*full_text*' :

      Split teks menjadi kata-kata individual (*token*)

   Output : teks yang sudah ditokenisasi

8. *Compute Bag of Words (BoW)*

   Input : teks yang telah di tokenisasi

   Output : matriks BoW

9. *Export prepared data*

   Save cleaned, stemmed, and tokenized data along with BoW matrix

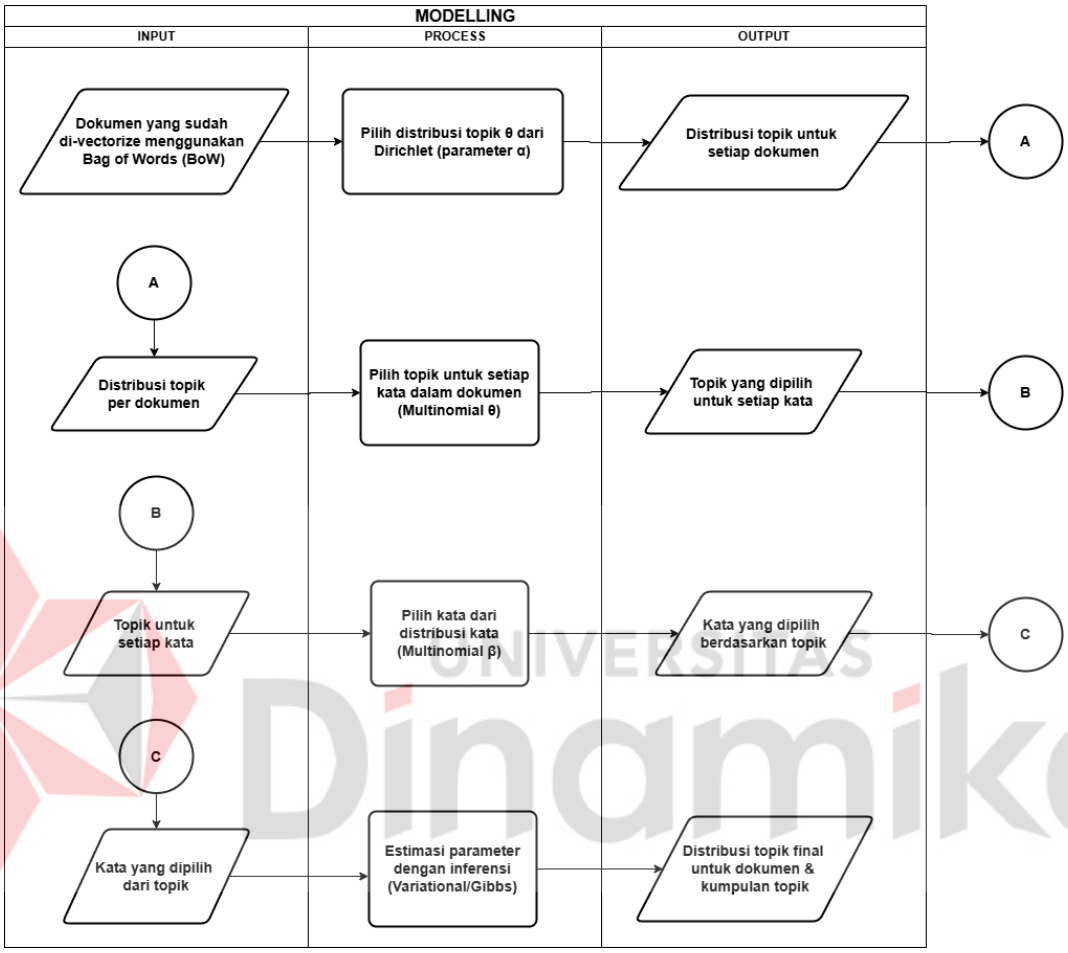
   Output : *dataset* siap untuk dimodelkan

END Data Preparation

3.4 *Modelling*

Model *Latent Dirichlet Allocation* (LDA) digunakan untuk mengidentifikasi dan memetakan topik utama dari kumpulan *tweet* terkait pemindahan Ibukota Nusantara (IKN) yang terdapat dalam kolom '*full_text*' yang telah dilakukan *preporcesing*. Pada proses pengerjaan ini, penulis menggunakan *library Gensim* di

Python, untuk melakukan implementasi LDA dalam analisis topik teks. Berikut adalah proses IPO yang menggambarkan tahapan-tahapan dari penerapan LDA dalam analisis data ini.



Gambar 3.3 *Modelling*

Setelah teks dokumen diubah menjadi vektor kata menggunakan *Bag of Words* (BoW), hasilnya akan menjadi *Input* 1 dalam proses *Latent Dirichlet Allocation* (LDA). Pada Proses 1, distribusi topik θ ditentukan untuk setiap dokumen berdasarkan distribusi *Dirichlet* dengan parameter α. Distribusi ini menunjukkan seberapa besar proporsi setiap topik dalam dokumen. Misalnya, jika ada dua topik, Dokumen 1 bisa memiliki distribusi topik $\theta = [0.6,0.4]$, yang berarti 60% dokumen tersebut terkait dengan Topik 1 dan 40% terkait dengan Topik 2 (contohnya dapat dilihat pada lampiran 1). Distribusi ini dihitung menggunakan rumus (1) / jika pada kode, semua proses LDA dijalankan menggunakan *library Gensim*.

Hasil dari proses ini adalah *Output* 1, yaitu distribusi topik untuk setiap dokumen yang menunjukkan seberapa besar topik tertentu muncul dalam dokumen tersebut. Selanjutnya, *Output* 1 menjadi *Input* 2 untuk Proses 2, di mana LDA menentukan topik $\mathcal{Z}n$ untuk setiap kata dalam dokumen. Proses ini menggunakan distribusi *Multinomial* yang didasarkan pada distribusi topik θ dari hasil sebelumnya. Sebagai contoh, kata pertama dalam Dokumen 1, "pemindahan", mungkin berhubungan dengan Topik 1, sedangkan kata kedua, "IKN", berhubungan dengan Topik 2. Pemilihan topik untuk setiap kata ini dilakukan menggunakan rumus (2).

Hasil dari proses ini adalah *Output* 2, yaitu topik yang telah dipilih untuk setiap kata dalam dokumen tersebut. *Output* 2 kemudian berfungsi sebagai *Input* 3 dalam Proses 3, di mana kata $\mathcal{W}n$ dipilih dari distribusi kata β, tergantung pada topik $\mathcal{Z}n$ yang telah dipilih sebelumnya untuk setiap kata. Jika kata tersebut terhubung dengan topik $\mathcal{Z}n$=Topik 1, kemungkinan kata yang akan dipilih adalah "ekonomi", "pembangunan", atau "pertumbuhan". Pemilihan kata berdasarkan topik ini dihitung dengan rumus (3).

Proses ini menghasilkan *Output* 3, yaitu kata yang dipilih dari distribusi kata untuk setiap topik terkait kata dalam dokumen tersebut. *Output* 3 kemudian digunakan sebagai *Input* 4 dalam Proses 4, di mana LDA melakukan estimasi parameter melalui metode inferensi seperti *Variational Inference* atau *Gibbs Sampling*. Pada tahap ini, distribusi topik dan kata diperbarui secara bertahap hingga ditemukan distribusi yang paling mungkin untuk dokumen dan kata dalam setiap topik. Probabilitas dokumen secara keseluruhan dihitung dengan rumus (4).

Proses ini menghasilkan *Output* 4, yaitu distribusi topik final untuk setiap dokumen, serta kumpulan topik yang berisi kata-kata yang paling relevan dengan setiap topik. Misalnya, Topik 1 mungkin berisi kata-kata seperti "pemindahan", "ekonomi", dan "pembangunan", sementara Topik 2 dapat terdiri dari kata-kata seperti "infrastruktur", "IKN", dan "proyek". Hasil akhirnya adalah kluster topik laten yang ditemukan dalam dokumen, yang membantu dalam mengidentifikasi tema/topik utama dalam kumpulan teks tersebut.

3.5 *Evaluation*

Setelah model dibuat, perlu dilakukan evaluasi untuk memastikan kualitas hasil yang diperoleh. Evaluasi dilakukan menggunakan *Coherence Score*, yang mengukur seberapa baik topik yang dihasilkan oleh model LDA. *Coherence score* memberikan indikasi apakah kata-kata dalam setiap topik memiliki hubungan yang bermakna. Rumus perhitungan *Coherence Score* dapat dilihat pada rumus (5).

Proses evaluasi dilakukan dengan menguji model LDA untuk berbagai jumlah topik, misalnya dari 2 hingga 10 topik. Setiap jumlah topik, *Coherence Score* dihitung menggunakan rumus di atas atau jika pada *python* menggunakan *c_v*, yang mengevaluasi kualitas topik berdasarkan relevansi kata-kata di dalamnya. Selain itu, grafik visualisasi *coherence score* juga digunakan untuk menentukan jumlah topik yang optimal. Grafik ini membantu mengidentifikasi jumlah topik dengan nilai *coherence* tertinggi, sehingga memudahkan dalam menentukan jumlah topik yang paling tepat untuk model.

3.6 *Deployment*

Tahap ini adalah implementasi model yang memungkinkan peneliti untuk menginterpretasikan topik-topik yang dibuat oleh LDA. Visualisasi hasil *topic modelling* dilakukan menggunakan *pyLDAvis*, yang memungkinkan visualisasi interaktif dari hasil LDA.

a) *pyLDAvis*: Merupakan *tool* visualisasi (hanya ada pada model LDA) yang memberikan representasi grafis dari distribusi topik dan kata-kata yang terkait dengan setiap topik.
b) Hasil: Hasil visualisasi tersebut dapat digunakan untuk menyampaikan topik-topik utama yang dibahas oleh masyarakat terkait pemindahan IKN, serta wawasan / *insight* dari penelitian ini dapat membantu pemerintah dan pihak-pihak terkait dalam memahami opini publik dan isu-isu yang perlu ditangani lebih lanjut.

BAB IV
HASIL DAN PEMBAHASAN

4.1 Hasil *Data Understanding*

Pada tahap pengambilan data teks, penelitian ini menggunakan *platform X* (sebelumnya dikenal sebagai *Twitter*) untuk mengumpulkan teks terkait opini dan diskusi publik mengenai pemindahan Ibu Kota Nusantara (IKN). Proses pengumpulan data (*crawling*) dilakukan melalui teknik *tweet harvesting*, yang memungkinkan penarikan data secara langsung dari *platform* dengan menggunakan *Auth Token* yang diperoleh saat *login* ke aplikasi *X*.

4.1.1 *Crawling Data Tweet*

Proses pengumpulan data dilakukan dengan metode *tweet harvesting*, yaitu teknik yang memungkinkan pengambilan data langsung dari *platform X* menggunakan *Auth Token* dari akun ketika *login* pada aplikasi *X*. Implementasi dengan kode program adalah sebagai berikut :

Twitter Auth Token

```
#@title Twitter Auth Token

twitter_auth_token = '[redacted]' # change this auth token
```

```
# Import required Python package
!pip install pandas

# Install Node.js (because tweet-harvest built using Node.js)
!sudo apt-get update
!sudo apt-get install -y ca-certificates curl gnupg
!sudo mkdir -p /etc/apt/keyrings
!curl -fsSL https://deb.nodesource.com/gpgkey/nodesource-repo.gpg.key | sudo gpg --dearmor -o /etc/apt/keyrings/nodesource.gpg

!NODE_MAJOR=20 && echo "deb [signed-by=/etc/apt/keyrings/nodesource.gpg] https://deb.nodesource.com/node_$NODE_MAJOR.x nodistro main" | sudo tee /etc/apt/sources.list.d/nodesourc

!sudo apt-get update
!sudo apt-get install nodejs -y

!node -v
```

Gambar 4.1 *Tweet Harvesting step* 1

Pada gambar 4.1, dilakukan langkah awal untuk mempersiapkan autentikasi dan instalasi *library* yang diperlukan dalam proses pengumpulan data (*crawling*) dari *platform X* (*Twitter*). Pertama, variabel *twitter_auth_token* digunakan untuk menyimpan *token* autentikasi API yang diperlukan agar aplikasi dapat mengakses data dari *platform* tersebut. *Token* ini bersifat unik dan harus diganti dengan *token autentikasi* yang valid milik pengguna.

Selanjutnya, *library Python pandas* di *install* menggunakan perintah *!pip install pandas*. *Library* ini akan digunakan untuk pengelolaan dan analisis data yang diambil, seperti membaca *file* CSV atau memproses *dataset*. Karena alat yang digunakan untuk *crawling*, yaitu *tweet-harvest*, dibangun menggunakan *Node.js*, maka diperlukan instalasi *Node.js* dalam *work environtment*.

Instalasi *Node.js* dimulai dengan memperbarui daftar paket pada sistem menggunakan *!sudo apt-get update*. Kemudian, dilakukan instalasi sertifikat keamanan, dengan *curl*, dan GnuPG untuk memastikan proses instalasi berjalan aman. Direktori khusus untuk menyimpan kunci keamanan juga dibuat, dan kunci keamanan *Node.js* diunduh dari sumber resmi untuk memastikan validitas *repository*. Setelah itu, *repository Node.js* versi 20 ditambahkan ke dalam daftar *repository* sistem, sehingga *Node.js* dapat di *install* langsung dari sumber resminya.

Setelah langkah-langkah tersebut, *Node.js* di *install* dengan perintah *!sudo apt-get install nodejs -y*. Untuk memastikan instalasi berhasil, versi *Node.js* yang terpasang diperiksa menggunakan *!node -v*. Dengan semua persiapan ini, langkah awal sudah siap untuk menjalankan *tweet-harvest* dan memulai proses pengumpulan data dari *platform X*.

```
# Crawl Data

filename = 'IKN.csv'
search_keyword = 'IKN since:2024-01-01 until:2024-08-30 lang:id'
limit = 1000

!npx -y tweet-harvest@2.6.1 -o "{filename}" -s "{search_keyword}" --tab "LATEST" -l {limit} --token {twitter_auth_token}
```

Gambar 4.2 *Tweet Harvesting step* 2

Kode di atas (gambar 4.2) digunakan untuk *crawling* data dari *platform X* menggunakan *tweet-harvest*. Proses ini mengambil 1.000 *tweet* berbahasa Indonesia terkait kata kunci "IKN", dengan rentang waktu 1 Januari – 30 Agustus 2024. *Crawling* dilakukan dengan perintah *!npx tweet-harvest*, menyimpan hasilnya dalam *file IKN.csv*. Opsi *--tab "LATEST"* memastikan hanya *tweet* terbaru yang diambil, sementara autentikasi dilakukan dengan *auth token* API. Data yang diperoleh akan digunakan untuk analisis lebih lanjut.

4.1.2 Hasil *Crawling*

	conversation_id_str	created_at	favorite_count	full_text	id_str	image_url	in_reply_to_screen_name	lang	location	quote_count	reply_count	retweet_co
0	1828993551813996936	Thu Aug 29 23:57:32 +0000 2024	2	@Aryprasetyo85 Kalau cara-cara seperti ini yg ...	1829307426627289397	NaN	Aryprasetyo85	in	Ciomas, Indonesia	0	0	
1	1829102996015460448	Thu Aug 29 23:51:45 +0000 2024	0	@Boediantar4 Ga tertarik lagi sma perkara d jk...	1829305974580236763	NaN	Boediantar4	in	NaN	0	0	
2	1829070888047882256	Thu Aug 29 23:50:14 +0000 2024	0	@Masfkr @ludovicusdwi @ImaginarySteady @jokowi...	1829305590449143953	NaN	Masfkr	in	NaN	0	1	
3	1828993551813996936	Thu Aug 29 23:48:27 +0000 2024	2	@Aryprasetyo85 dan banyak cebong bangga dg ikn	1829305142610723146	NaN	Aryprasetyo85	in	Depok, Indonesia	0	0	
4	1829295018278093151	Thu Aug 29 23:47:43 +0000 2024	0	@Hansunriko IKN Nusantara🇮🇩	1829304959818731958	NaN	Hansunriko	in	NaN	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	
995	1828932556471054642	Wed Aug 28 23:07:56 +0000 2024	0	Upacara di IKN Usai Basuki Sebut Pembangunan B...	1828932556471054642	NaN	NaN	in	NaN	0	0	
996	1828931584851939674	Wed Aug 28 23:04:04	0	Jokowi Ungkap Alasan Belum	1828931584851939674	NaN	NaN	in	indonesia	0	0	

Gambar 4.3 Hasil *Crawling Data Tweet*

Pada gambar 4.3, hasil *crawling* data menghasilkan 1.000 *tweet* terkait topik Ibukota Nusantara (IKN) dengan berbagai atribut/fitur/kolom, seperti *full_text* (isi teks *tweet*), *created_at* (waktu publikasi), *favorite_count* (jumlah *like*), *retweet_count* (jumlah *retweet*), dan ***lang*** (bahasa *tweet*). *Dataset* ini menyediakan informasi yang kaya untuk analisis opini publik mengenai pemindahan IKN.

4.1.3 EDA

Pada tahap ini, eksplorasi data awal (*Exploratory Data Analysis*) dilakukan untuk memahami karakteristik data hasil *crawling*. Eksplorasi ini bertujuan memastikan kualitas data dan menentukan kolom yang akan dianalisis lebih lanjut.

A. *Load Dataset*

Pada tahap ini, *dataset* hasil *crawling* dimuat ke dalam *work environment program* untuk diproses lebih lanjut. Proses ini merupakan langkah awal dalam eksplorasi data guna memastikan *dataset* siap digunakan dalam analisis berikutnya.

```
file_path = "/content/drive/MyDrive/Colab Notebooks/LDA for IKN/IKN.csv"
df = pd.read_csv(file_path)
```

```
df.head()
```

	conversation_id_str	created_at	favorite_count	full_text	id_str	image_url	in_reply_to_screen_name	lang	location
0	1828993551813996936	Thu Aug 29 23:57:32 +0000 2024	2	@Aryprasetyo85 Kalau cara-cara seperti ini yg ...	1829307426627289397	NaN	Aryprasetyo85	in	Ciomas, Indonesia
1	1829102996015460448	Thu Aug 29 23:51:45 +0000 2024	0	@Boediantar4 Ga tertarik lagi sma perkara d jk...	1829305974580236763	NaN	Boediantar4	in	NaN
2	1829070888047882256	Thu Aug 29 23:50:14 +0000 2024	0	@Masfkr @ludovicusdwi @ImaginarySteady @jokowi...	1829305590449143953	NaN	Masfkr	in	NaN
3	1828993551813996936	Thu Aug 29 23:48:27 +0000 2024	2	@Aryprasetyo85 dan banyak cebong bangga dg ikn	1829305142610723146	NaN	Aryprasetyo85	in	Depok, Indonesia
4	1829295018278093151	Thu Aug 29 23:47:43 +0000 2024	0	@Hansunriko IKN Nusantaraid	1829304959818731958	NaN	Hansunriko	in	NaN

Gambar 4.4 *Load dataset*

Kode pada Gambar 4.4 memuat *dataset* hasil *crawling* yang disimpan dalam *file* CSV bernama IKN.csv. Pada *cell* pertama, *file dataset* diakses menggunakan pustaka *pandas* dengan mendefinisikan lokasi *file* yang disimpan di *GDrive* dan memuatnya ke dalam sebuah *dataframe* untuk mempermudah pengolahan data. Selanjutnya, pada *cell* kedua, kode *df.head()* digunakan untuk menampilkan lima baris pertama *dataset*. *Output* yang ditampilkan menunjukkan struktur *dataset* yang berisi berbagai kolom, seperti *conversation_id_str* (ID unik percakapan), *created_at* (waktu *tweet* dipublikasikan), *favorite_count* (jumlah "*like*"), *full_text* (isi teks *tweet*), *lang* (bahasa yang digunakan), *location* (lokasi pengguna), *retweet_count* (jumlah *retweet*), dan *username* (nama pengguna). Informasi ini memberikan gambaran awal mengenai isi *dataset* dan memastikan data telah dimuat dengan benar untuk tahap analisis berikutnya.

B. *Explore Dataset*

Bagian ini bertujuan untuk mengeksplorasi *dataset* secara lebih mendalam, dengan menganalisis struktur serta karakteristik data yang telah dimuat sebelumnya. Hal tersebut dilakukan guna memastikan data sesuai dengan kebutuhan analisis.

```
1 df.shape
```

```
(1000, 15)
```

```
1 df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 15 columns):
 #   Column                  Non-Null Count  Dtype
---  ------                  --------------  -----
 0   conversation_id_str     1000 non-null   int64
 1   created_at              1000 non-null   object
 2   favorite_count          1000 non-null   int64
 3   full_text               1000 non-null   object
 4   id_str                  1000 non-null   int64
 5   image_url               388 non-null    object
 6   in_reply_to_screen_name 493 non-null    object
 7   lang                    1000 non-null   object
 8   location                488 non-null    object
 9   quote_count             1000 non-null   int64
 10  reply_count             1000 non-null   int64
 11  retweet_count           1000 non-null   int64
 12  tweet_url               1000 non-null   object
 13  user_id_str             1000 non-null   int64
 14  username                1000 non-null   object
dtypes: int64(7), object(8)
memory usage: 117.3+ KB
```

Gambar 4.5 *Shape & info of dataframe*

Gambar 4.5 tersebut menunjukkan hasil eksplorasi awal *dataset* untuk memahami struktur data. Pada *cell* pertama, kode *df.shape* digunakan untuk mengetahui ukuran *dataset*. *Output* (1000, 15) menunjukkan bahwa *dataset* memiliki 1.000 baris dan 15 kolom, menandakan bahwa data yang berhasil di-*crawl* terdiri dari 1.000 entri dengan 15 atribut atau fitur yang berbeda.

Pada *cell* kedua, kode *df.info()* digunakan untuk memberikan gambaran umum tentang tipe data dan jumlah nilai yang tersedia di setiap kolom. *Output* menunjukkan bahwa semua kolom memiliki tipe data yang sesuai, seperti *int64* untuk kolom numerik dan *object* untuk kolom teks. Selain itu, terdapat informasi mengenai kolom dengan data yang tidak lengkap, seperti *image_url*, *in_reply_to_screen_name*, dan *location*, yang masing-masing hanya memiliki 388, 493, dan 488 nilai *non-null*. Hal ini menunjukkan bahwa sebagian data pada kolom tersebut kosong (*null*). Dari hasil ini, penggunaan memori untuk menyimpan *dataset* juga tercatat sebesar 117,3 KB.

```
1 duplicate_rows = df.duplicated()
2 print("Jumlah baris duplikat:", duplicate_rows.sum())
```

```
Jumlah baris duplikat: 0
```

```
1 missing_values = df.isnull().sum()
2 print("Jumlah missing values per kolom:\n", missing_values)
```

```
Jumlah missing values per kolom:
 conversation_id_str          0
created_at                    0
favorite_count                0
full_text                     0
id_str                        0
image_url                   612
in_reply_to_screen_name     507
lang                          0
location                    512
quote_count                   0
reply_count                   0
retweet_count                 0
tweet_url                     0
user_id_str                   0
username                      0
dtype: int64
```

Gambar 4.6 *Missing values & duplicated data*

Gambar 4.6 menunjukkan hasil analisis untuk memeriksa keberadaan data duplikat dan nilai yang hilang (*missing values*) dalam *dataset*. Analisis ini penting untuk memastikan kualitas data yang akan digunakan dalam proses lebih lanjut. Pada *cell* pertama, kode *df.duplicated()* digunakan untuk mendeteksi baris yang terduplikasi dalam *dataset*. *Output* yang dihasilkan menunjukkan bahwa tidak ada baris duplikat dalam *dataset*. Hal ini memastikan bahwa setiap entri dalam *dataset* adalah unik dan tidak memerlukan tindakan lebih lanjut terkait duplikasi data. Pada *cell* kedua, kode *df.isnull().sum()* digunakan untuk menghitung jumlah nilai yang hilang di setiap kolom *dataset*. *Output* menunjukkan bahwa beberapa kolom memiliki nilai yang hilang.

4.2 *Data Preparation*

Pada tahap ini, data diproses terlebih dahulu untuk memastikan kualitasnya memadai dan sesuai dengan kebutuhan analisis. Proses ini mencakup berbagai langkah pembersihan dan transformasi data sehingga lebih terorganisir dan siap digunakan dalam analisis lanjutan. Salah satu langkah awal yang dilakukan adalah seleksi fitur, yaitu memilih kolom yang relevan untuk dianalisis.

```
1 dfIKN = df[['full_text']]
```

```
1 dfIKN.head()
```

	full_text
0	@Aryprasetyo85 Kalau cara-cara seperti ini yg ...
1	@Boediantar4 Ga tertarik lagi sma perkara d jk...
2	@Masfkr @ludovicusdwi @ImaginarySteady @jokowi...
3	@Aryprasetyo85 dan banyak cebong bangga dg ikn
4	@Hansunriko IKN Nusantara🇮🇩

```
1 dfIKN.shape
```

```
(1000, 1)
```

```
1 dfIKN.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1000 entries, 0 to 999
Data columns (total 1 columns):
 #   Column     Non-Null Count  Dtype
---  ------     --------------  -----
 0   full_text  1000 non-null   object
dtypes: object(1)
memory usage: 7.9+ KB
```

Gambar 4.7 Seleksi Fitur

Gambar 4.7 menjelaskan tahap seleksi fitur dengan memilih kolom *full_text* sebagai satu-satunya fitur yang digunakan dalam analisis topik dengan LDA. *dfIKN = df[['full_text']]* digunakan untuk membuat *dataframe* baru yang hanya berisi teks utama *tweet*. Pemilihan hanya kolom *full_text* dilakukan karena fokus analisis adalah menemukan pola topik dari teks *tweet*. Kolom lain, seperti *favorite_count* atau *location*, tidak relevan dalam proses ini.

4.2.1 *Case Cleaning*

Tahap ini dilakukan proses penyesuaian format teks agar seragam, seperti mengubah semua huruf menjadi huruf kecil. Langkah ini bertujuan untuk menghindari duplikasi yang disebabkan oleh perbedaan kapitalisasi dan mempermudah analisis data.

```
1 def remove_tweet_special(text):
2     # remove tab, new line, ans back slice
3     text = text.replace('\\t'," ").replace('\\n'," ").replace('\\u'," ").replace('\\',"")
4     # remove non ASCII (emoticon, chinese word, .etc)
5     text = text.encode('ascii', 'replace').decode('ascii')
6     # remove mention, link, hashtag
7     text = ' '.join(re.sub("([@#][A-Za-z0-9]+)|(\w+:\/\/\S+)"," ", text).split())
8     # remove incomplete URL
9     return text.replace("http://", " ").replace("https://", " ")
10
11 dfIKN['full_text'] = dfIKN['full_text'].apply(remove_tweet_special)
```

Gambar 4.8 *Case Cleaning* 1

Gambar 4.8 menunjukkan proses awal dalam tahap *Case Cleaning* untuk membersihkan teks pada kolom *full_text*. Proses ini dilakukan dengan menggunakan fungsi *remove_tweet_special*, yang bertujuan untuk menghapus elemen-elemen tidak relevan dari teks, seperti karakter khusus, *URL*, dan *mention*. Langkah pertama dalam fungsi ini adalah menghapus karakter seperti *tab (\t)*, baris baru *(\n)*, dan *backslash (\)* untuk memastikan tidak ada simbol atau karakter tak terlihat yang mengganggu analisis teks. Selanjutnya, teks di enkode ke format ASCII sehingga karakter non-ASCII seperti *emoticon* atau huruf non-Latin diganti dengan tanda ?, menjaga konsistensi teks hanya dengan huruf dan simbol standar. Fungsi ini juga menghapus *mention* seperti *@username*, *URL* seperti *https://link.com*, dan *hashtag* menggunakan ekspresi *reguler*. Selain itu, *URL* yang tidak lengkap juga dihapus untuk memastikan teks benar-benar bersih dari elemen yang tidak relevan. Setelah fungsi ini diterapkan, teks pada kolom *full_text* menjadi lebih bersih dan siap untuk langkah *preprocessing* berikutnya, yang penting untuk memastikan kualitas data dan fokus analisis pada isi utama teks.


```python
#remove number
def remove_number(text):
    return  re.sub(r"\d+", "", text)

dfIKN['full_text'] = dfIKN['full_text'].apply(remove_number)

#remove punctuation
def remove_punctuation(text):
    return text.translate(str.maketrans("","",string.punctuation))

dfIKN['full_text'] = dfIKN['full_text'].apply(remove_punctuation)

#remove whitespace leading & trailing
def remove_whitespace_LT(text):
    return text.strip()

dfIKN['full_text'] = dfIKN['full_text'].apply(remove_whitespace_LT)

#remove multiple whitespace into single whitespace
def remove_whitespace_multiple(text):
    return re.sub('\s+',' ',text)

dfIKN['full_text'] = dfIKN['full_text'].apply(remove_whitespace_multiple)

# remove single char
def remove_singl_char(text):
    return re.sub(r"\b[a-zA-Z]\b", "", text)

dfIKN['full_text'] = dfIKN['full_text'].apply(remove_singl_char)
```

Gambar 4.9 *Case Cleaning 2*

Pada Gambar 4.9, terdapat serangkaian proses *case cleaning* lanjutan yang bertujuan untuk membersihkan teks lebih lanjut. Pertama, angka dihapus menggunakan fungsi *remove_number* yang mengaplikasikan ekspresi *reguler* untuk

menghapus semua digit dalam teks, memastikan data teks tidak terkontaminasi informasi numerik yang tidak relevan. Selanjutnya, tanda baca dihilangkan dengan fungsi *remove_punctuation* yang menggunakan *modul string.punctuation* untuk menghapus semua tanda baca. Fungsi *remove_whitespace_LT* menghapus spasi berlebih di awal dan akhir teks menggunakan metode *strip()*, sementara fungsi *remove_whitespace_multiple* menggantikan spasi ganda dengan spasi tunggal. Terakhir, fungsi *remove_singl_char* menghapus huruf tunggal yang tidak memberikan makna signifikan, seperti "a", "b", atau "c", menggunakan ekspresi *reguler*. Proses-proses ini membantu merapikan teks, mengurangi kebisingan, dan memastikan data yang lebih bersih dan terstruktur, sehingga dapat meningkatkan kualitas analisis *topic modelling*.

4.2.2 *Case Folding*

Pada bagian ini, dilakukan *case folding*, yaitu proses untuk mengubah seluruh teks menjadi huruf kecil (*lowercase*). Tujuan dari tahap ini adalah untuk menghindari perbedaan antara huruf besar dan kecil yang dapat mengganggu kualitas data. Melalui cara tersebut, dapat dipastikan bahwa kata yang sama yang ditulis dengan variasi huruf besar/kecil dianggap sebagai satu entitas yang sama, sehingga meningkatkan akurasi dan konsistensi dalam pemrosesan teks lebih lanjut.

```
1 dfIKN['full_text'] = dfIKN['full_text'].str.lower()
2
3 print('Case Folding Result : \n')
4 print(dfIKN['full_text'].head())
```

```
Case Folding Result :

0    kalau caracara seperti ini yg dipakai bisa jad...
1    ga tertarik lagi sma perkara  jkarta gw lbih t...
2    katanya kan dateng aja ke ikn bang nanti juga ...
3                    dan banyak cebong bangga dg ikn
4                                        ikn nusantara
Name: full_text, dtype: object
```

Gambar 4.10 *Case Folding*

Gambar 4.10 menunjukkan proses *case folding* yang diterapkan pada *dataset*. Pada kode tersebut, fungsi *str.lower()* digunakan untuk mengubah seluruh teks dalam kolom *'full_text'* menjadi huruf kecil (*lowercase*). Hal ini bertujuan untuk

memastikan bahwa teks telah berhasil diubah ke dalam format huruf kecil secara konsisten.

4.2.3 *Stopword Removal*

Pada bagian ini, dilakukan penghapusan *stopwords*, yaitu kata-kata umum yang sering muncul dalam teks tetapi tidak memberikan informasi berarti dalam analisis seperti "dan", "atau", "dari", dan sebagainya. *Stopword removal* bertujuan untuk mengurangi kompleksitas teks dan fokus pada kata-kata yang lebih relevan untuk analisis lebih lanjut.

```
# Memuat daftar stopwords bahasa Indonesia dari NLTK
list_stopwords = stopwords.words('indonesian')

# Memuat stopwords tambahan dari file txt
with open('/content/drive/MyDrive/Colab Notebooks/LDA for IKN/stopwords_tambahan.txt', 'r',
          encoding='utf-8') as file:
    stopwords_tambahan = file.read().splitlines()

# Menambahkan stopwords tambahan ke daftar utama
list_stopwords.extend(stopwords_tambahan)

# Konversi ke set untuk optimasi pencarian
list_stopwords = set(list_stopwords)

# Fungsi untuk melakukan stopword removal
def stopwords_removal_direct(text):
    """
    Menghapus stopwords langsung dari teks.
    :param text: String teks
    :return: String teks tanpa stopwords
    """
    filtered_text = " ".join([word for word in text.split() if word not in list_stopwords])
    return filtered_text

# Menerapkan stopword removal pada kolom full_text
dfIKN['full_text_cleaned'] = dfIKN['full_text'].apply(stopwords_removal_direct)

# Menampilkan hasil
print('Stopword Removal Result: \n')
print(dfIKN[['full_text', 'full_text_cleaned']].head())
```

Gambar 4.11 *Stopword Removal*

Pada Gambar 4.11, Proses ini menghapus kata-kata umum yang tidak memiliki makna signifikan dalam analisis menggunakan daftar *stopwords* dari NLTK dan tambahan dari *file stopwords_tambahan.txt*. *File* eksternal ini berisi kata-kata umum dalam percakapan media sosial, seperti "yg", "gak", dan "kalo". Seluruh daftar *stopwords* dikonversi ke tipe set untuk efisiensi pencocokan. Fungsi *stopwords_removal_direct* digunakan untuk menghapus *stopwords* dari teks dengan memfilter hanya kata-kata yang bermakna. Hasilnya disimpan dalam kolom *full_text_cleaned*, sehingga teks yang dianalisis lebih bersih dan relevan untuk tahapan selanjutnya.

4.2.4 *Normalization*

Pada bagian ini, dilakukan normalisasi teks, yaitu proses mengganti kata-kata yang salah ejaan atau salah ketik dengan bentuk kata yang benar. Normalisasi bertujuan untuk memastikan konsistensi dalam teks sehingga kata-kata yang memiliki arti sama tetapi ditulis berbeda (misalnya, "jkarta" menjadi "jakarta) dapat dikenali sebagai satu entitas. Proses ini penting untuk meningkatkan kualitas data.

```python
# Membaca file normalization_dict.txt
normalization_dict = {}
with open('/content/drive/MyDrive/Colab Notebooks/LDA for IKN/normalization_dict.txt',
          'r', encoding='utf-8') as file:
    for line in file:
        key, value = line.strip().split(": ")
        normalization_dict[key] = value

# Fungsi untuk normalisasi teks
def normalize_text(text):
    """
    Mengganti kata-kata dalam teks berdasarkan normalization_dict.
    :param text: String teks
    :return: String teks setelah normalisasi
    """
    words = text.split()
    normalized_words = [normalization_dict.get(word, word) for word in words]
    return " ".join(normalized_words)

# Menerapkan normalisasi pada kolom 'full_text_cleaned'
dfIKN['full_text_cleaned'] = dfIKN['full_text_cleaned'].apply(normalize_text)

# Menampilkan hasil
print('Normalization result:\n')
print(dfIKN[['full_text', 'full_text_cleaned']].head())
```

Gambar 4.12 *Normalization*

Pada Gambar 4.12, dilakukan Proses normalisasi dilakukan dengan memanfaatkan *file* eksternal *normalization_dict.txt* yang berisi pasangan kata dalam format "kata_salah:kata_benar" (misal, "jkarta:jakarta"). *File* ini dibaca menggunakan *Pandas* dan dikonversi menjadi *dictionary* sebagai referensi normalisasi. Fungsi *normalize_text* mengganti kata-kata dalam teks berdasarkan *dictionary* tersebut. Prosesnya melibatkan tokenisasi teks, pengecekan dan penggantian kata sesuai referensi, lalu penyusunan ulang teks. Normalisasi diterapkan pada kolom *full_text_cleaned* menggunakan metode *.apply()*, menghasilkan teks yang lebih konsisten dan siap untuk analisis lebih lanjut.

4.2.5 *Stemming*

Pada bagian ini, dilakukan proses *stemming* untuk mentransformasi kata-kata ke bentuk dasar atau akar kata. Proses tersebut bertujuan untuk mengurangi variasi

kata yang memiliki makna dasar yang sama, sehingga teks menjadi lebih sederhana dan konsisten untuk analisis. *Stemming* dalam penelitian ini dilakukan menggunakan pustaka Sastrawi, yang secara khusus dirancang untuk menangani bahasa Indonesia.

```
# Membuat stemmer
factory = StemmerFactory()
stemmer = factory.create_stemmer()

# Fungsi untuk melakukan stemming
def stemming_text(text):
    """
    Melakukan stemming pada teks menggunakan Sastrawi.
    :param text: String teks
    :return: String teks setelah stemming
    """
    return stemmer.stem(text)

# Menerapkan stemming pada kolom 'full_text_cleaned'
dfIKN['ft_stemmed'] = dfIKN['full_text_cleaned'].apply(stemming_text)

# Menampilkan hasil
print('Hasil Stemming:\n')
print(dfIKN[['full_text_cleaned', 'ft_stemmed']].head())
```

Gambar 4.13 *Stemming*

Pada gambar 4.13, proses *stemming* dilakukan dengan menggunakan pustaka Sastrawi, yang merupakan alat *stemming* untuk bahasa Indonesia. Pertama, dibuat objek *stemmer* dengan memanfaatkan *StemmerFactory* dari pustaka Sastrawi. Fungsi *stemming_text* dirancang untuk menerima teks sebagai *input*, lalu menggunakan metode *stemmer.stem()* untuk mengubah setiap kata dalam teks tersebut ke bentuk dasarnya. Proses *stemming* ini diterapkan pada kolom *full_text_cleaned* di *dataframe dfIKN*, dengan hasilnya disimpan dalam kolom baru bernama *ft_stemmed*. Dengan demikian, kata-kata akan direduksi ke bentuk dasar yang lebih sederhana dan konsisten, sehingga mempermudah analisis lebih lanjut. Berdasarkan hal tersebut, kata-kata dalam teks yang terproses akan direduksi ke bentuk dasar yang lebih sederhana dan konsisten.

4.2.6 *Tokenization*

Pada bagian ini, dilakukan proses *tokenization* untuk memecah teks menjadi *unit-unit* terkecil yang disebut *token*. Pada tahap ini, setiap kata dalam teks akan

dipisahkan dan diubah menjadi elemen-elemen yang dapat digunakan dalam tahap analisis selanjutnya.

```python
# Fungsi untuk melakukan tokenisasi
def tokenize_text(text):
    """
    Melakukan tokenisasi pada teks.
    :param text: String teks
    :return: List token
    """
    return word_tokenize(text)

# Menerapkan tokenisasi pada kolom 'ft_stemmed'
dfIKN['ft_tokenized'] = dfIKN['ft_stemmed'].apply(tokenize_text)

# Menampilkan hasil
print('Hasil Tokenisasi:\n')
print(dfIKN[['ft_stemmed', 'ft_tokenized']].head())
```

Gambar 4.14 *Tokenization*

Pada Gambar 4.14, menggambarkan proses *tokenization*, yaitu tahap memecah teks menjadi elemen-elemen kata (*token*) yang lebih kecil. Proses ini dimulai dengan membuat fungsi *tokenize_text*, yang dirancang untuk menerima teks dalam bentuk *string* sebagai *input* dan menggunakan fungsi *word_tokenize* dari pustaka NLTK untuk memecah teks tersebut menjadi daftar *token*. Fungsi ini kemudian diterapkan pada kolom *ft_stemmed* di *dataframe dfIKN*. Hasil tokenisasi disimpan dalam kolom baru bernama *ft_tokenized*, yang berisi setiap teks dalam bentuk daftar *token*. Proses ini penting untuk mempersiapkan data dalam *format* yang lebih terstruktur dan memungkinkan analisis lebih lanjut.

4.2.7 *Bag of Words*

Pada tahap ini, metode *Bag of Words* (*BoW*) digunakan untuk merepresentasikan teks dalam bentuk vektor numerik. Teknik ini mengabaikan tata urutan kata dan hanya memperhitungkan kemunculan kata dalam dokumen, sehingga mempermudah proses analisis dan pemodelan data. Representasi *BoW* membantu mengubah data teks menjadi bentuk yang dapat diproses oleh model *machine learning*.

```python
# Membuat dictionary dari data tokenized
dictionary = corpora.Dictionary(dfIKN['ft_tokenized'])

# Konversi data ke format Bag of Words (BoW)
bow_corpus = [dictionary.doc2bow(text) for text in dfIKN['ft_tokenized']]

# Tampilkan contoh dictionary dan BoW
print("Dictionary:\n", dictionary.token2id)  # Menampilkan kata dan ID-nya
print("\nBoW Corpus Contoh:\n", bow_corpus[:2])  # Menampilkan dua contoh dokumen BoW
```

Gambar 4.15 BoW

Pada gambar 4.15 menampilkan kode program yang digunakan untuk mengonversi data *tokenized* ke dalam *format Bag of Words* (*BoW*). Pertama, kode tersebut membuat sebuah kamus (*dictionary*) dari data yang sudah melalui proses *tokenisasi* dengan menggunakan *corpora.Dictionary*, yang mengaitkan setiap kata dengan ID unik. Selanjutnya, data *tokenized* diubah menjadi representasi *BoW* menggunakan metode *doc2bow*, yang menghasilkan sebuah daftar pasangan (ID kata, frekuensi kemunculan kata) untuk setiap dokumen. Terakhir, kode tersebut menampilkan kamus yang berisi pasangan kata dan ID-nya melalui *dictionary.token2id*, serta contoh dua dokumen pertama dari *corpus BoW*. Proses ini bertujuan untuk mempersiapkan data teks dalam *format* numerik yang akan digunakan dalam penerapan model LDA.

4.3 *Modelling*

Bagian ini membahas tahap terkait pembangunan model untuk analisis data. Pada tahap ini, algoritma LDA digunakan untuk mengidentifikasi topik-topik tersembunyi dalam *dataset* yang telah diproses sebelumnya.

```python
# Menyusun model LDA
lda_model = LdaModel(bow_corpus,
                     num_topics=5,
                     id2word=dictionary,
                     passes=10,
                     alpha=0.1,  # Dokumen lebih spesifik ke beberapa topik utama
                     eta=0.01,   # Topik lebih spesifik dengan kata-kata unik
                     random_state=22
                     )

# Melihat topik yang ditemukan
topics = lda_model.print_topics(num_words=5)
for topic in topics:
    print(topic)
```

```
(0, '0.015*"smart" + 0.015*"city" + 0.014*"pindah" + 0.014*"bangun" + 0.013*"jakarta"')
(1, '0.026*"bangun" + 0.019*"indonesia" + 0.018*"investasi" + 0.017*"logistik" + 0.014*"dukung"')
(2, '0.021*"landas" + 0.020*"uji" + 0.020*"coba" + 0.013*"jalan" + 0.012*"rakyat"')
(3, '0.049*"masyarakat" + 0.046*"bangun" + 0.037*"nusantara" + 0.035*"beneran" + 0.034*"kalimantan"')
(4, '0.021*"pindah" + 0.020*"kantor" + 0.015*"jokowi" + 0.012*"jakarta" + 0.009*"asn"')
```

Gambar 4.16 *Modelling with* LDA

Pada Gambar 4.16, proses *modelling* dengan LDA ditampilkan. Model tersebut digunakan untuk mengidentifikasi lima topik utama dari data yang telah diproses sebelumnya dalam format *BoW*. Kode program dimulai dengan menyusun model LDA menggunakan fungsi *LdaModel* dari pustaka *Gensim*. Parameter yang digunakan meliputi:

1. *bow_corpus*: Data dalam format *BoW* sebagai *input* model.
2. *num_topics*=5: Menentukan jumlah topik yang ingin ditemukan, yaitu lima topik.
3. *id2word=dictionary*: Menghubungkan ID kata dalam *BoW* dengan kata aslinya.
4. *passes*=10: Menentukan jumlah iterasi untuk melatih model agar mendapatkan hasil yang lebih optimal.
5. *alpha*=0.1: Parameter ini mengontrol distribusi kata dalam topik, diatur lebih rendah agar setiap dokumen lebih spesifik ke beberapa topik utama.
6. *eta*=0.01: Parameter ini mengontrol distribusi kata dalam dokumen, yang diatur lebih kecil agar setiap topik memiliki kata-kata unik yang lebih spesifik.
7. *Random_state*=22: Parameter ini memastikan bahwa hasil yang diperoleh bersifat deterministik (konsisten).

Setelah model dilatih, topik yang ditemukan ditampilkan menggunakan metode *print_topics(num_words=5)*, yang menampilkan lima kata dengan bobot tertinggi dalam setiap topik. *Output* menunjukkan setiap topik dengan daftar kata-kata utama yang relevan serta bobotnya. Contohnya, pada topik keempat (indeks 3), kata "masyarakat" memiliki bobot tertinggi (0.049), diikuti oleh kata-kata seperti "bangun," "nusantara," "beneran," dan "kalimantan."

4.4 *Evaluation*

Bagian ini, membahas proses evaluasi terhadap model LDA yang telah dibangun. Evaluasi bertujuan untuk memastikan bahwa model yang dihasilkan mampu menggambarkan distribusi topik secara akurat dan relevan dengan data. Pada evaluasi kali ini, digunakan metrik *coherence score* untuk mengukur seberapa baik kata-kata dalam setiap topik saling berhubungan, sehingga dapat memberikan validasi terhadap kualitas model yang telah diterapkan.

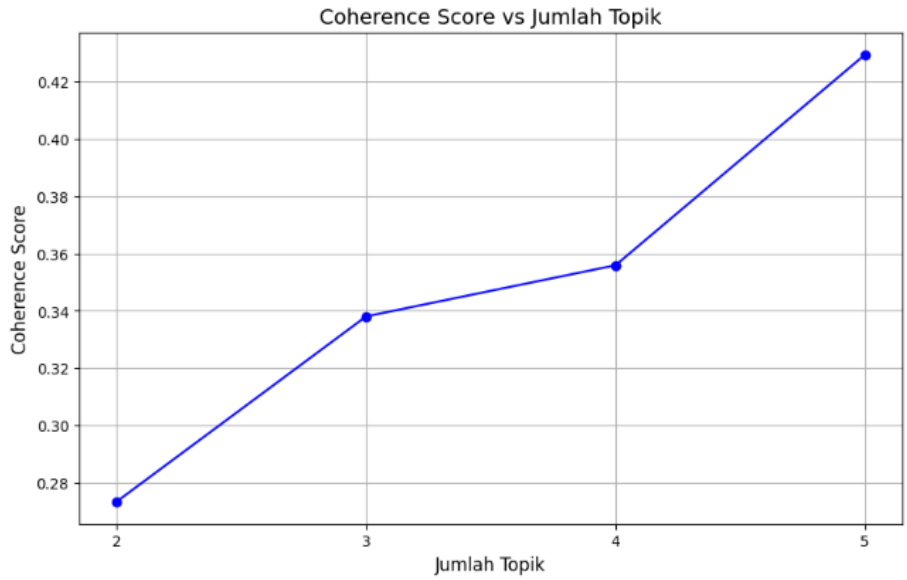
```
coherence_scores = []
# Menguji jumlah topik dari 2 hingga 10
for num_topics in range(2, 6):
    lda_model = LdaModel(
        bow_corpus,
        num_topics=num_topics,
        id2word=dictionary,
        passes=10,
        alpha=0.1,
        eta=0.01,
        random_state=22
    )
    coherence_model = CoherenceModel(
        model=lda_model,
        texts=dfIKN['ft_tokenized'],
        dictionary=dictionary,
        coherence='c_v'
    )
    score = coherence_model.get_coherence()
    coherence_scores.append((num_topics, score))

# Membuat DataFrame dari hasil untuk tabel
coherence_df = pd.DataFrame(coherence_scores, columns=['Num Topics', 'Coherence Score'])
# Menampilkan tabel hasil
print(coherence_df)
# Membuat grafik
plt.figure(figsize=(10, 6))
plt.plot(coherence_df['Num Topics'], coherence_df['Coherence Score'], marker='o', linestyle='-', color='b')
plt.title('Coherence Score vs Jumlah Topik', fontsize=14)
plt.xlabel('Jumlah Topik', fontsize=12)
plt.ylabel('Coherence Score', fontsize=12)
plt.xticks(coherence_df['Num Topics'])
plt.grid()
plt.show()
```

Gambar 4.17 Kode *Coherence Score*

Pada Gambar 4.17, merupakan kode proses perhitungan dan visualisasi *coherence score* untuk mengevaluasi model LDA berdasarkan jumlah topik yang diuji. Kode tersebut dimulai dengan membuat daftar kosong *coherence_scores* untuk menyimpan hasil evaluasi. Kemudian, model LDA dibangun dengan jumlah topik yang bervariasi, mulai dari 2 hingga 5, menggunakan data *BoW* dan *dictionary* yang telah dibuat sebelumnya. Lalu untuk setiap model LDA, metrik *coherence* dihitung menggunakan pustaka *CoherenceModel*, dengan mempertimbangkan teks yang sudah di-tokenisasi, kamus, dan nilai koherensi *c_v*.

Hasil perhitungan *coherence score* untuk setiap jumlah topik disimpan dalam bentuk tabel (*Dataframe*). Selain itu, hasil tersebut divisualisasikan dalam grafik untuk menunjukkan hubungan antara jumlah topik dan *coherence score*. Grafik ini mempermudah analisis untuk menentukan jumlah topik optimal, yaitu saat *coherence score* mencapai nilai tertinggi.

	Num Topics	Coherence Score
0	2	0.273475
1	3	0.338063
2	4	0.355902
3	5	0.429356



Gambar 4.18 Hasil *Coherence Score*

Pada Gambar 4.18, ditampilkan tabel hasil perhitungan & grafik *coherence score* untuk setiap jumlah topik yang diuji menggunakan model LDA. Kolom *Num Topics* (pada tabel) dan jumlah topik (pada grafik sumbu x) menunjukkan jumlah topik yang diuji dalam rentang 2 hingga 5, sedangkan kolom *Coherence Score* (pada tabel & grafik sumbu y) mencerminkan nilai koherensi yang diperoleh untuk setiap model.

Berdasarkan hasil tersebut, dapat diamati bahwa *coherence score* bervariasi berdasarkan jumlah topik. Berikut adalah interpretasi dari hasilnya:

Tabel 4.1 Hasil *Coherence Score*

Jumlah Topik	Hasil
2 topik	*coherence score* bernilai 0.273475, menunjukkan hubungan antar kata dalam topik masih kurang kuat.
3 topik	*coherence score* meningkat menjadi 0.338063, menunjukkan adanya peningkatan kualitas dalam pemisahan topik.
4 topik	*coherence score* meningkat lagi menjadi 0.355902, menandakan model semakin mampu mengelompokkan kata-kata yang lebih relevan.
5 topik	*coherence score* mencapai nilai tertinggi sebesar 0.429356, menunjukkan model ini memberikan kualitas topik terbaik dibanding jumlah topik lainnya.

Berdasarkan hasil ini, jumlah topik optimal adalah 5, karena memberikan *coherence score* tertinggi. Hal ini menunjukkan bahwa model LDA dengan 5 topik

mampu menghasilkan topik yang lebih relevan dan bermakna dibandingkan jumlah topik lainnya.

4.5 *Deployment*

Tahap *Deployment* merupakan langkah terakhir dalam proses analisis data yang bertujuan untuk menyajikan hasil model secara visual agar lebih mudah dipahami.

4.5.1 *Wordcloud*

Pada bagian ini, digunakan visualisasi *wordcloud* untuk menggambarkan kata-kata yang paling sering muncul dalam setiap topik yang dihasilkan oleh model LDA. Visualisasi ini mempermudah identifikasi kata-kata kunci utama yang membentuk topik, sehingga memberikan gambaran umum yang informatif kepada pembaca.

Gambar 4.19 *Wordcloud* topik-topik terkait IKN

Pada gambar 4.19, ditampilkan visualisasi *wordcloud* untuk masing-masing topik yang dihasilkan oleh model LDA. Visualisasi ini mempermudah identifikasi kata-kata kunci yang paling sering muncul dalam tiap topik, di mana ukuran *font* menunjukkan bobot atau frekuensi kemunculan kata dalam topik tertentu. Berikut adalah penjelasan berdasarkan hasil *wordcloud* untuk masing-masing topik:

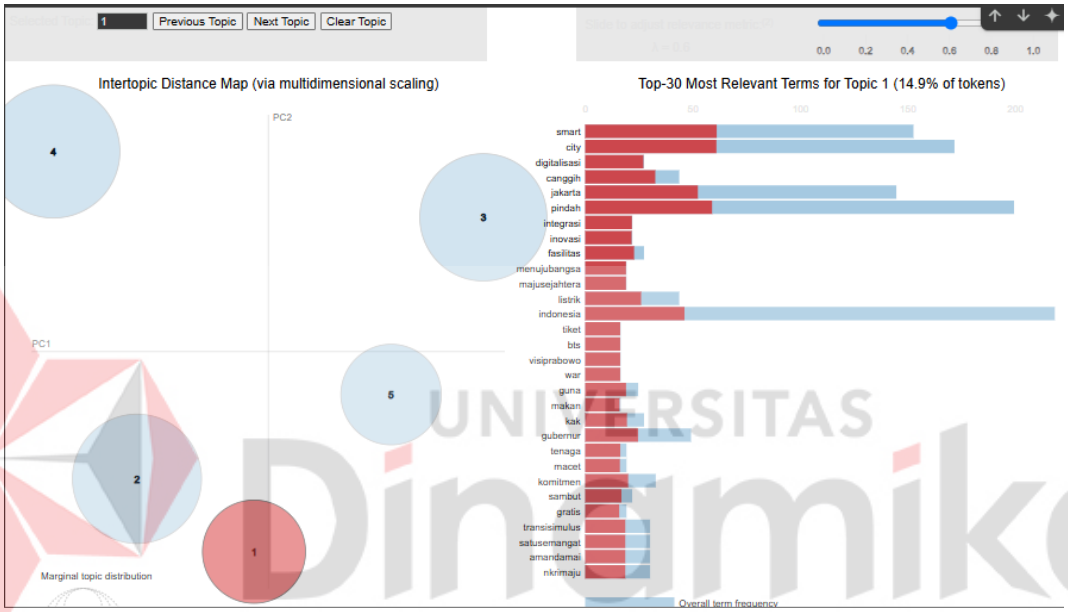
Tabel 4.2 Hasil *Wordcloud* IKN

Topik	Kata Kunci Utama	Penjelasan
1	('0.015*"*smart*" + 0.015*"*city*" + 0.014*"pindah" + 0.014*"bangun" + 0.013*"jakarta" + 0.011*"indonesia" + 0.008*"canggih" + 0.007*"kota" + 0.007*"investor" + 0.007*"asn"')	Topik ini membahas konsep *Smart City* di IKN. Kata "*smart*" dan "*city*" menegaskan bahwa ibu kota baru akan dirancang dengan teknologi canggih. Sementara itu, "pindah" dan "jakarta" menunjukkan fokus pada perpindahan pusat pemerintahan ke lokasi baru.
2	('0.026*"bangun" + 0.019*"indonesia" + 0.018*"investasi" + 0.017*"logistik" + 0.014*"dukung" + 0.013*"city" + 0.012*"infrastruktur" + 0.011*"smart" + 0.010*"investor" + 0.010*"nusantara"')	Topik ini berfokus pada investasi dan pembangunan infrastruktur. Kata "investasi", "logistik", dan "infrastruktur" menyoroti aspek pendanaan dan pengembangan fasilitas fisik di IKN. Kata "dukung" dan "nusantara" mengindikasikan adanya dukungan dari berbagai pihak terhadap proyek ini.
3	('0.021*"landas" + 0.020*"uji" + 0.020*"coba" + 0.013*"jalan" + 0.012*"rakyat" + 0.012*"lancar" + 0.012*"pacu" + 0.012*"bandara" + 0.012*"air" + 0.011*"mulus"')	Topik ini membahas transportasi dan infrastruktur mobilitas di IKN. Kata "landas", "uji", dan "coba" menunjukkan adanya pengujian fasilitas bandara dan jalan.
4	('0.049*"masyarakat" + 0.046*"bangun" + 0.037*"nusantara" + 0.035*"beneran" + 0.034*"kalimantan" + 0.034*"timur" + 0.031*"bawa" + 0.030*"ekonomi" + 0.029*"dampak" + 0.028*"positif"')	Topik ini menyoroti dampak pembangunan IKN bagi masyarakat dan ekonomi. Kata "masyarakat" dan "bangun" mencerminkan keterlibatan publik dalam proyek ini. Kata "ekonomi", "dampak", dan "positif" menunjukkan harapan akan pertumbuhan ekonomi di Kalimantan Timur akibat pembangunan ibu kota baru.
5	('0.021*"pindah" + 0.020*"kantor" + 0.015*"jokowi" + 0.012*"jakarta" + 0.009*"asn" + 0.009*"udara" + 0.009*"langsung" + 0.009*"banget" + 0.008*"negara" + 0.008*"september"')	Topik ini membahas relokasi pusat pemerintahan ke IKN. Kata "pindah" dan "kantor" mengindikasikan perpindahan instansi pemerintah, sementara "jokowi" dan "asn" menunjukkan peran pemerintah dan Aparatur Sipil Negara dalam kebijakan pemindahan ini.

4.5.2 *pyLDAvis*

Pada bagian ini, digunakan pustaka *pyLDAvis* untuk memvisualisasikan hasil model LDA secara interaktif. *pyLDAvis* menyediakan representasi dua dimensi dari distribusi topik, hubungan antar topik, serta kata-kata yang paling relevan dalam setiap topik. Visualisasi ini membantu dalam memahami struktur model topik secara lebih *detail*.

A. Topik 1



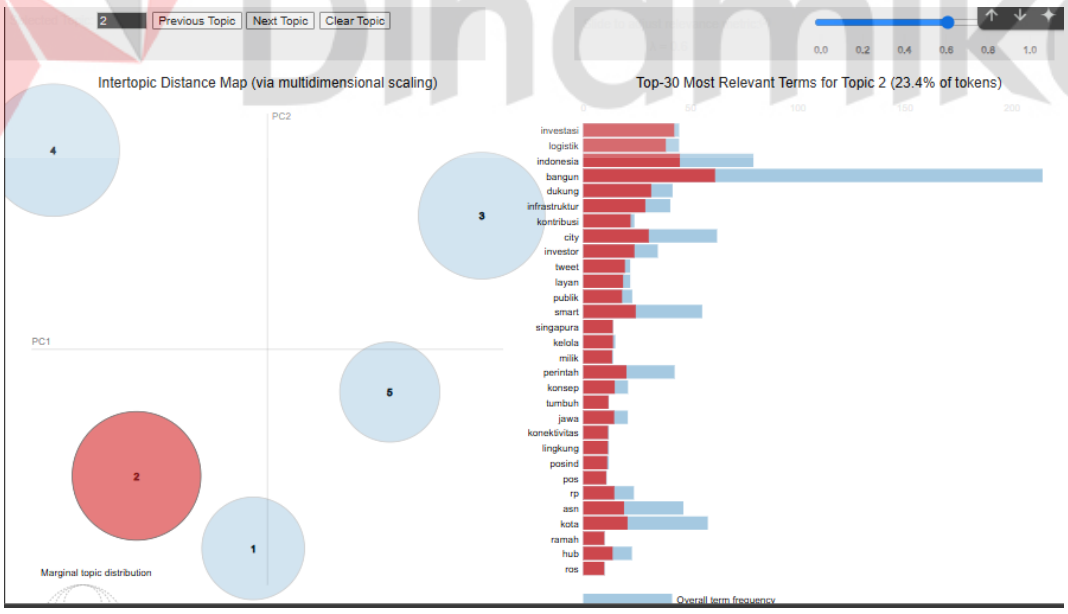
Gambar 4.20 *pyLDAvis* Topik 1

Pada gambar 4.20, visualisasi *pyLDAvis* ditampilkan untuk Topik 1 dengan pengaturan λ (*lambda*) sekitar 0.6. Berdasarkan *Intertopic Distance Map*, Topik 1 mencakup 14.9% dari total *token*. Pada bagian *Top-30 Most Relevant Terms (Bar Chart)*, kata-kata yang paling dominan dalam Topik 1 adalah "*smart*," "*city*," "digitalisasi," "canggih," "jakarta," "pindah," "integrasi," "inovasi," "fasilitas," "menuju," "bangsa," "maju," "sejahtera," "listrik," "indonesia," dan "tiket." Kata "*smart*" dan "*city*" memiliki batang merah lebih panjang dibandingkan batang biru, menunjukkan bahwa topik ini erat kaitannya dengan konsep *Smart City* di IKN. Selain itu, keberadaan kata "digitalisasi," "canggih," dan "inovasi" menunjukkan bahwa diskusi ini berfokus pada penerapan teknologi dan transformasi digital dalam pembangunan ibukota baru. Kata "jakarta" dan "pindah" juga muncul

sebagai kata yang cukup dominan, mengindikasikan bahwa diskusi mengenai perpindahan ibukota dari Jakarta ke IKN menjadi bagian penting dari pembahasan ini. Selain itu, kata "menuju," "bangsa," "maju," dan "sejahtera" menunjukkan optimisme terhadap pembangunan IKN sebagai kota masa depan yang lebih modern dan maju. Sementara itu, kata "listrik," "indonesia," dan "tiket" dapat dikaitkan dengan pembahasan mengenai pengelolaan energi serta aksesibilitas di ibukota baru.

Berdasarkan nilai lambda (λ) sekitar 0.6, visualisasi ini menunjukkan kombinasi antara kata-kata yang sering muncul dalam Topik 1 dan kata-kata yang lebih spesifik untuk topik tersebut. Kata-kata yang memiliki batang merah lebih panjang dibandingkan dengan batang biru merupakan kata-kata yang lebih khas dalam Topik 1 dibandingkan dengan topik lainnya. Secara keseluruhan, Topik 1 menggambarkan bahwa diskusi mengenai implementasi konsep *Smart City*, integrasi infrastruktur, serta dampak pembangunan terhadap kesejahteraan masyarakat dan kemajuan bangsa.

B. Topik 2

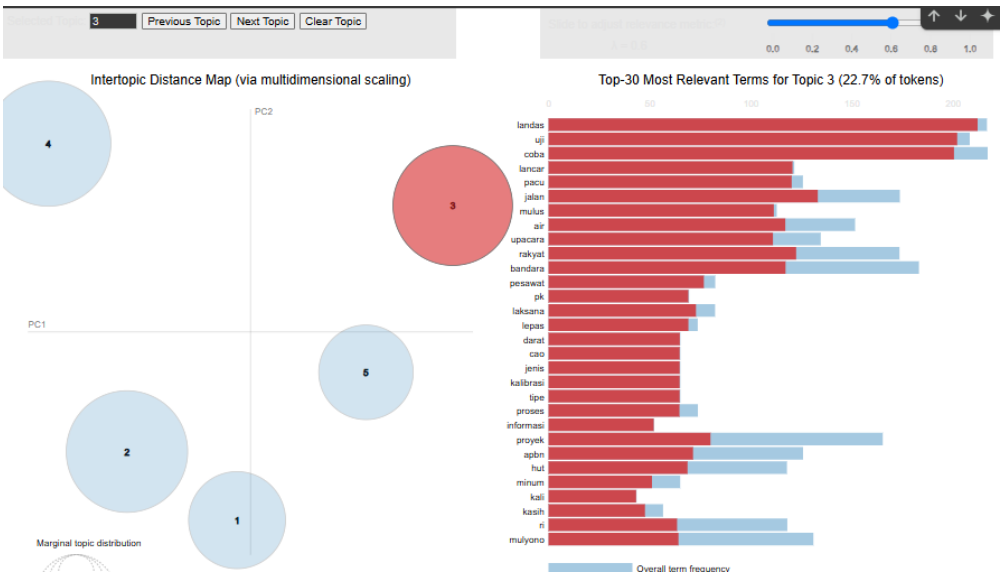


Gambar 4.21 *pyLDAvis* Topik 2

Pada gambar 4.21, visualisasi *pyLDAvis* ditampilkan untuk Topik 2 dengan pengaturan λ sekitar 0.6. Berdasarkan *Intertopic Distance Map*, terlihat bahwa

Topik 2 memiliki ukuran lingkaran sebesar 23.4% dari total *token*. Pada bagian *bar chart*, kata-kata yang paling dominan dalam Topik 2 adalah "investasi," "logistik," "indonesia," "bangun," "dukung," "infrastruktur," "kontribusi," "*city*," "investor," "singapura," "layanan," dan "publik." Keberadaan kata "investor" dan "singapura" menunjukkan bahwa topik ini membahas peran investasi asing dalam pembangunan IKN, terutama terkait dengan potensi keterlibatan Singapura. Kata "investasi" dan "logistik" memiliki batang merah lebih panjang dibandingkan batang biru, menunjukkan bahwa diskusi dalam topik ini berfokus pada investasi dan sektor logistik di IKN. Selain itu, kata "bangun," "dukung," dan "infrastruktur" menunjukkan bahwa pembangunan infrastruktur menjadi bagian utama dalam diskusi terkait IKN, yang juga melibatkan peran serta investor. Kata "kontribusi" dan "investor" semakin menguatkan bahwa pembahasan dalam topik ini menyoroti keterlibatan investor asing dalam mendukung pembangunan IKN. Keberadaan kata "singapura" menjadi indikasi bahwa negara tersebut memiliki ketertarikan khusus atau bahkan berperan dalam investasi atau kerja sama pembangunan IKN. Secara keseluruhan, Topik 2 menggambarkan bahwa diskusi mengenai IKN tidak hanya berfokus pada aspek pembangunan infrastruktur, tetapi juga menyoroti keterlibatan investasi asing, terutama dari Singapura, dalam mendukung pengembangan IKN.

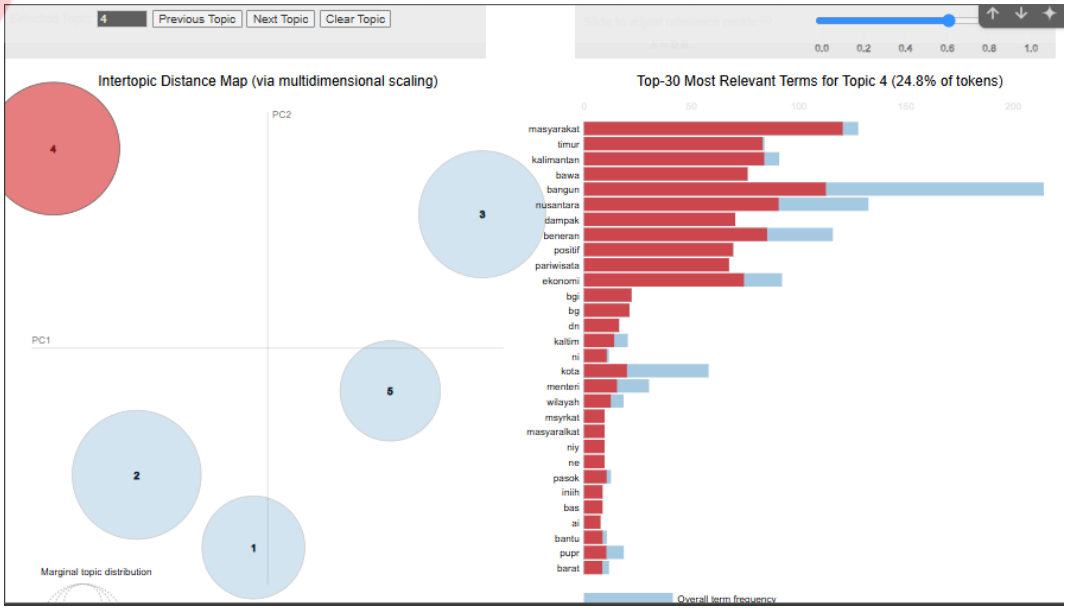
C. Topik 3



Gambar 4.22 *pyLDAvis* Topik 3

Pada gambar 4.22, visualisasi *pyLDAvis* ditampilkan untuk Topik 3 dengan pengaturan λ sekitar 0.6. Berdasarkan *Intertopic Distance Map*, Topik 3 mencakup 22.7% dari total *token*, menunjukkan bahwa topik ini cukup signifikan dalam *dataset*. Pada *bar chart* untuk Topik 3, kata-kata yang paling dominan antara lain "landas," "uji," "coba," "lancar," "pacu," "jalan," "mulus," "air," "upacara," "bandara," "pesawat," "lepas," dan "kalibrasi." Keberadaan kata "landas," "pacu," "pesawat," "lepas," dan "bandara" menunjukkan bahwa topik ini berkaitan erat dengan infrastruktur transportasi udara, khususnya pembangunan dan pengujian bandara di IKN. Kata "uji," "coba," "proses," dan "kalibrasi" menunjukkan adanya tahapan pengujian dan verifikasi, yang kemungkinan besar berkaitan dengan kesiapan operasional bandara dan transportasi udara. Selain itu, kata "jalan," "mulus," "air," dan "lancar" menunjukkan bahwa topik ini juga membahas infrastruktur jalan atau aspek lain dari transportasi di IKN. Secara keseluruhan, Topik 3 menggambarkan bahwa diskusi mengenai IKN tidak hanya berfokus pada aspek pembangunan infrastruktur secara umum, tetapi juga menyoroti pengujian kesiapan operasional transportasi, khususnya bandara dan jalan, sebagai bagian dari rencana strategis dalam pengembangan ibukota baru.

D. Topik 4



Gambar 4.23 *pyLDAvis* Topik 4

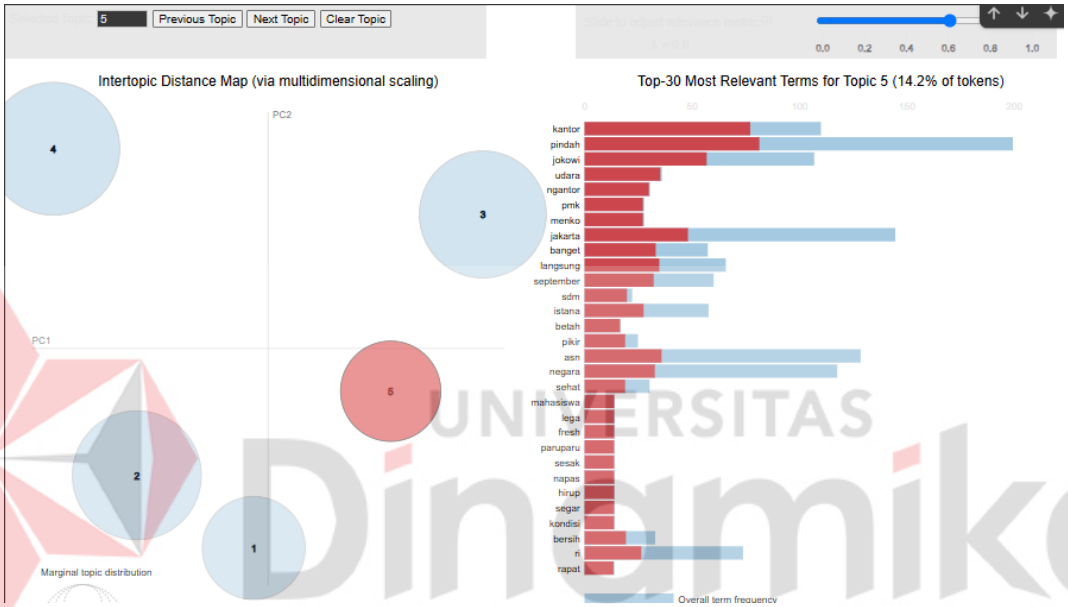
Pada gambar 4.23, visualisasi *pyLDAvis* ditampilkan untuk Topik 4 dengan λ sekitar 0.6. Berdasarkan *Intertopic Distance Map*, Topik 4 mencakup 24.8% dari total *token*, menjadikannya topik terbesar dan paling signifikan dalam *dataset*. Lalu, dari segi jarak antar topik, Topik 4 memiliki jarak yang cukup jauh dari topik-topik lainnya, menunjukkan bahwa tema dalam Topik 4 lebih unik dan tidak memiliki banyak tumpang tindih dengan topik lainnya.

Pada *bar chart* untuk Topik 4, kata-kata yang paling dominan antara lain "masyarakat," "timur," "kalimantan," "bawa," "bangun," "nusantara," "dampak," "beneran," "positif," "pariwisata," "ekonomi," dan "menteri." Keberadaan kata "masyarakat," "dampak," "positif," "ekonomi," dan "pariwisata" menunjukkan bahwa topik ini berfokus pada dampak pembangunan IKN terhadap masyarakat dan sektor ekonomi. Kata "kalimantan," "timur," dan "nusantara" memperkuat bahwa diskusi ini berkaitan erat dengan pemindahan ibukota ke Kalimantan Timur. Kata "bangun" menunjukkan adanya pembicaraan mengenai pembangunan dari IKN, sementara keberadaan kata "menteri" mengindikasikan bahwa topik ini juga berkaitan dengan kebijakan pemerintah dalam membangun dan mengelola IKN.

Berdasarkan interpretasi tema utama, Topik 4 dapat dikategorikan dalam tiga aspek utama. Pertama, dampak sosial pemindahan ibu kota terhadap masyarakat, yang ditunjukkan oleh kata-kata seperti "masyarakat," "dampak," "positif," dan "menteri." Diskusi kemungkinan mencakup pandangan positif maupun tantangan yang dihadapi masyarakat akibat perpindahan ibukota ke Kalimantan Timur. Kedua, peluang ekonomi dan pariwisata di IKN, yang tercermin dari kata-kata "ekonomi," "pariwisata," dan "positif." Ini menunjukkan bahwa pembangunan IKN bisa menciptakan peluang kerja baru, menarik investasi, dan meningkatkan kesejahteraan masyarakat sekitar. Selain itu, pariwisata di Kalimantan Timur juga menjadi salah satu aspek yang dibahas, terkait dengan pengembangan infrastruktur wisata di sekitar IKN. Ketiga, pembangunan dan implementasi kebijakan pemerintah, yang ditunjukkan oleh kata-kata "bangun," "bawa," dan "menteri." Keberadaan kata "menteri" menunjukkan bahwa keputusan terkait pembangunan ibukota melibatkan pejabat tinggi pemerintah dan kebijakan negara, sehingga pembahasan dalam topik ini kemungkinan mencakup pernyataan atau kebijakan resmi dari pemerintah mengenai masa depan IKN.

Secara keseluruhan, Topik 4 menggambarkan diskusi mengenai dampak sosial dan ekonomi dari pemindahan ibu kota ke IKN. Fokus utama diskusi adalah bagaimana pembangunan IKN memengaruhi masyarakat, baik dari sisi dampak langsung maupun potensi ekonomi dan pariwisata. Selain itu, terdapat pembahasan mengenai kebijakan pemerintah, dengan keterlibatan menteri dalam perencanaan dan implementasi pembangunan IKN.

E. Topik 5



Gambar 4.24 *pyLDAvis* Topik 5

Gambar 4.24, menunjukkan visualisasi *pyLDAvis* untuk Topik 5 dengan λ sekitar 0.6. Berdasarkan *Intertopic Distance Map*, Topik 5 mencakup 14.2% dari total *token*. Pada *bar chart* untuk Topik 5, kata-kata yang paling dominan dalam Topik 5 antara lain "kantor," "pindah," "jokowi," "udara," "pmk," "menko," "jakarta," "langsung," "september," "sdm," "istana," "belah," "rapat," "asn," "negara," "sehat," "mahasiswa," dan "*fresh*." Keberadaan kata "kantor," "pindah," dan "istana" menunjukkan bahwa topik ini membahas pemindahan kantor pemerintahan, termasuk Istana Kepresidenan dan lembaga negara lainnya ke IKN. Kata "jokowi," "menko," dan "pmk" mengindikasikan bahwa diskusi ini berkaitan dengan pernyataan atau kebijakan Presiden Joko Widodo serta peran kementerian dalam pemindahan pemerintahan ke IKN. Keberadaan kata "ASN" (Aparatur Sipil

Negara) menunjukkan bahwa topik ini juga mencakup perpindahan pegawai negeri ke ibukota baru. Selain itu, kata "sehat," "udara," dan "hirup" menunjukkan adanya diskusi mengenai kondisi lingkungan dan kualitas udara di lokasi baru dibandingkan dengan Jakarta.

Berdasarkan interpretasi tema utama, Topik 5 dapat dikategorikan dalam tiga aspek utama. Pertama, pemindahan kantor pemerintahan dan Aparatur Sipil Negara (ASN) ke IKN, yang tercermin dari kata-kata "kantor," "pindah," "istana," dan "ASN." Diskusi dalam topik ini mencakup kesiapan infrastruktur perkantoran, kesiapan ASN, serta proses transisi pemerintahan. Kedua, peran Presiden Jokowi dan kementerian dalam pemindahan, yang ditunjukkan oleh kata-kata "jokowi," "menko," dan "pmk." Keberadaan kata "menko" dan "pmk" menunjukkan bahwa kebijakan pemindahan ibukota dan kantor pemerintahan melibatkan pernyataan dari Presiden serta peran kementerian terkait. Diskusi ini juga mencakup bagaimana pemerintah merancang regulasi dan implementasi pemindahan ibukota. Ketiga, perbandingan kondisi lingkungan di Jakarta vs IKN, yang terlihat dari kata-kata "udara," "hirup," "sehat," dan "negara." Kemungkinan terdapat diskusi mengenai perbedaan kualitas udara dan kondisi lingkungan di IKN dibandingkan dengan Jakarta. Kata "*fresh*" bisa merujuk pada harapan bahwa IKN akan memberikan suasana baru yang lebih baik dibandingkan Jakarta, terutama dalam hal lingkungan kerja yang lebih sehat dan bebas polusi.

Secara keseluruhan, Topik 5 menggambarkan diskusi mengenai pemindahan kantor pemerintahan dan ASN ke IKN. Fokus utama diskusi adalah bagaimana perpindahan kantor dan aparatur sipil negara ke ibukota baru akan dilakukan. Peran Presiden Jokowi dan kementerian terkait juga menjadi bagian penting dalam pembahasan ini, khususnya dalam regulasi dan implementasi perpindahan. Selain itu, ada diskusi mengenai kualitas udara dan lingkungan di IKN dibandingkan dengan Jakarta, dengan harapan kondisi di IKN lebih sehat dan nyaman untuk bekerja.

BAB V
PENUTUP

5.1 Kesimpulan

Berdasarkan hasil penelitian yang telah dilakukan, dapat disimpulkan bahwa:

1. Model LDA berhasil mengelompokkan topik-topik utama yang muncul dalam diskusi publik mengenai IKN di *platform X*. Berdasarkan hasil analisis melalui pendekatan *wordcloud* dan *pyLDAvis*, diperoleh 5 topik utama, yaitu konsep *smart city*, investasi dan infrastruktur, transportasi dan mobilitas, dampak sosial-ekonomi, serta relokasi pemerintahan.
2. Topik yang dibentuk oleh LDA memiliki nilai *coherence score* yang cukup baik, yaitu 0.429356, yang menunjukkan bahwa pemodelan topik dapat menggambarkan pola diskusi publik secara jelas dan terstruktur.

5.2 Saran

Berdasarkan penelitian ini, saran yang dapat penulis berikan adalah:

1. Penyesuaian *hyperparameter* LDA, dapat dieksplorasi lebih lanjut dengan variasi yang lebih luas, guna memperoleh topik yang lebih representatif & berkualitas.
2. Menambahkan kamus bahasa lain selain Sastrawi untuk meningkatkan akurasi pemrosesan teks, terutama dalam menangani kata-kata tidak baku atau istilah spesifik dalam bahasa lain.
3. Dapat menggunakan *dataset* yang lebih besar untuk menangkap lebih banyak variasi topik yang muncul dalam percakapan publik terkait IKN.

DAFTAR PUSTAKA

Ajinaja, M. O., Adetunmbi, A. O., Ugwu, C. C., & Popoola, O. S. (2023). Semantic similarity measure for topic modeling using latent Dirichlet allocation and collapsed Gibbs sampling. *Iran Journal of Computer Science*, *6*(1). https://doi.org/10.1007/s42044-022-00124-7

Baghmohammad, M., Mansouri, A., & CheshmehSohrabi, M. (2021). Identification of topic development process of knowledge and information science field based on the topic modeling (LDA). *Iranian Journal of Information Processing and Management*, *36*(2).

Been, S., & Byeon, H. (2023). Analysis of Depression in News Articles Before and After the COVID-19 Pandemic Based on Unsupervised Learning and Latent Dirichlet Allocation Topic Modeling. *International Journal of Advanced Computer Science and Applications*, *14*(10). https://doi.org/10.14569/IJACSA.2023.0141018

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Blei et al., 2003 - Latent Dirichlet Allocation. Dalam *Journal of Machine Learning Research* (Vol. 3, Nomor 4/5).

Bokrantz, J., Subramaniyan, M., & Skoogh, A. (2023). Realising the promises of artificial intelligence in manufacturing by enhancing CRISP-DM. *Production Planning and Control*. https://doi.org/10.1080/09537287.2023.2234882

Cahyono, N., & Angga Reni Dwi Astuti. (2023). Analisis Topic Modelling Persepsi Pengguna Internet Menggunakan Metode Latent Dirichlet Allocation. *Indonesian Journal of Computer Science*, *12*(1). https://doi.org/10.33022/ijcs.v12i1.3155

Christian, Y., & Qi, K. O. Y. R. (2022). Penerapan K-Means pada Segmentasi Pasar untuk Riset Pemasaran pada Startup Early Stage dengan Menggunakan CRISP-DM. *JURIKOM (Jurnal Riset Komputer)*, *9*(4). https://doi.org/10.30865/jurikom.v9i4.4486

Erniyati, E., Harsani, P., Mulyati, M., & Fahriza, L. D. (2023). Topic Modeling LDA and SVM in Sentiment Analysis of Hotel Reviews. *Komputasi: Jurnal Ilmiah Ilmu Komputer dan Matematika*, *20*(2). https://doi.org/10.33751/komputasi.v20i2.7604

Fristikawati, Y., & Adipradana, N. (2022). Perlindungan Lingkungan, dan Pembangunan Ibukota Negara (IKN) Dalam Tinjauan Hukum. *Jurnal Justisia : Jurnal Ilmu Hukum, Perundang-undangan dan Pranata Sosial*, *7*(2). https://doi.org/10.22373/justisia.v7i2.15586

Fristikawati, Y., Alvander, R., & Wibowo, V. (2022). Pengaturan Dan Penerapan Sustainable Development Pada Pembangunan Ibukota Negara Nusantara. *Jurnal Komunitas Yustisia*, *5*(2).

Garg, M., & Rangra, P. (2022). Bibliometric Analysis of Latent Dirichlet Allocation. *DESIDOC Journal of Library and Information Technology*, *42*(2). https://doi.org/10.14429/djlit.42.2.17307

Gomez, M. J., Ruipérez-Valiente, J. A., & García Clemente, F. J. (2023). Exploring Technology- and Sensor-Driven Trends in Education: A Natural-Language-Processing-Enhanced Bibliometrics Study †. Dalam *Sensors* (Vol. 23, Nomor 23). https://doi.org/10.3390/s23239303

Hardiyanti, L., Anggraini, D., & Kurniawati, A. (2023). Identify Reviews of Pedulilindungi Applications using Topic Modeling with Latent Dirichlet Allocation Method. *IJCCS (Indonesian Journal of Computing and Cybernetics Systems)*, *17*(4). https://doi.org/10.22146/ijccs.86025

Inoue, M., Fukahori, H., Matsubara, M., Yoshinaga, N., & Tohira, H. (2023). Latent Dirichlet allocation topic modeling of free-text responses exploring the negative impact of the early COVID-19 pandemic on research in nursing. *Japan Journal of Nursing Science*, *20*(2). https://doi.org/10.1111/jjns.12520

Jalolov, T. S. (2023). Teaching the Basics of Python Programming. *International Multidisciplinary Journal for Research & Development (IMJRD)*, *10*(11).

Mahammadodilovich, S. S. (2023). Importance of Python Programming Language in Machine Learning. *International Bulletin of Engineering and Technology*, *3*(9).

Ma'mun, A. R. (2023). Problematika Komunikasi Politik Pendanaan Pembangunan Ibu Kota Negara Nusantara. *POLITICOS: Jurnal Politik dan Pemerintahan*, *3*(1). https://doi.org/10.22225/politicos.3.1.2023.1-16

Martinez-Plumed, F., Contreras-Ochando, L., Ferri, C., Hernandez-Orallo, J., Kull, M., Lachiche, N., Ramirez-Quintana, M. J., & Flach, P. (2021). CRISP-DM Twenty Years Later: From Data Mining Processes to Data Science Trajectories. *IEEE Transactions on Knowledge and Data Engineering*, *33*(8). https://doi.org/10.1109/TKDE.2019.2962680

Murshed, B. A. H., Mallappa, S., Abawajy, J., Saif, M. A. N., Al-ariki, H. D. E., & Abdulwahab, H. M. (2023). Short text topic modelling approaches in the context of big data: taxonomy, survey, and analysis. *Artificial Intelligence Review*, *56*(6). https://doi.org/10.1007/s10462-022-10254-w

Narasi Newsroom. (2024, Juli 16). Pembangunan IKN diwarnai banyak kritik. *Narasi Newsroom*.

Naury, C., Fudholi, D. H., & Hidayatullah, A. F. (2021). Topic Modelling pada Sentimen Terhadap Headline Berita Online Berbahasa Indonesia

Menggunakan LDA dan LSTM. *JURNAL MEDIA INFORMATIKA BUDIDARMA*, *5*(1). https://doi.org/10.30865/mib.v5i1.2556

Nova Anggraini. (2022, Maret 21). *Ingin Sumbang Saran Hingga Kritik Terkait Ibu Kota Baru? Bisa, Lewat Laman Resmi IKN Ini*. haluanpadang.com.

Pambudi, A. (2023). PENERAPAN CRISP-DM MENGGUNAKAN MLR K-FOLD PADA DATA SAHAM PT. TELKOM INDONESIA (PERSERO) TBK (TLKM) (STUDI KASUS: BURSA EFEK INDONESIA TAHUN 2015-2022). *Jurnal Data Mining dan Sistem Informasi*, *4*(1). https://doi.org/10.33365/jdmsi.v4i1.2462

Pan, X., & Xue, Y. (2023). Advancements of Artificial Intelligence Techniques in the Realm About Library and Information Subject - A Case Survey of Latent Dirichlet Allocation Method. *IEEE Access*, *11*. https://doi.org/10.1109/ACCESS.2023.3334619

Pardede, D. L. C., & Waskita, M. A. I. (2023). ANALISIS PEMODELAN TOPIK UNTUK ULASAN TENTANG PEDULI LINDUNGI. *Jurnal Ilmiah Informatika Komputer*, *28*(1). https://doi.org/10.35760/ik.2023.v28i1.7925

Patil, R., Boit, S., Gudivada, V., & Nandigam, J. (2023). A Survey of Text Representation and Embedding Techniques in NLP. *IEEE Access*, *11*. https://doi.org/10.1109/ACCESS.2023.3266377

Prastiwi, H., Pricilia, J., & Raswir, E. (2022). Jurnal Informatika Dan Rekayasa Komputer (JAKAKOM) Implementasi Data Mining Untuk Menentuksn Persediaan Stok Barang Di Mini Market Menggunakan Metode K-Means Clustering. *Informatika Dan Rekayasa Komputer (JAKAKOM)*, *1*(2).

Purnama, S. J., & Chotib, C. (2023). ANALISIS KEBIJAKAN PUBLIK PEMINDAHAN IBU KOTA NEGARA. *Jurnal Ekonomi dan Kebijakan Publik*, *13*(2). https://doi.org/10.22212/jekp.v13i2.3486

Sergii V., M., & Oleksandr V., N. (2023). Data preprocessing and tokenization techniquesfortechnical Ukrainian texts. *Applied Aspects of Information Technology*, *6*(3). https://doi.org/10.15276/aait.06.2023.22

Wang, J. K., Wang, S. K., Lee, E. B., & Chang, R. T. (2023). Natural Language Processing (NLP) in AI. Dalam *Digital Eye Care and Teleophthalmology: A Practical Guide to Applications*. https://doi.org/10.1007/978-3-031-24052-2_17

Wanniarachchi, V. U., Scogings, C., Susnjak, T., & Mathrani, A. (2023). Hate Speech Patterns in Social Media: A Methodological Framework and Fat Stigma Investigation Incorporating Sentiment Analysis, Topic Modelling and Discourse Analysis. *Australasian Journal of Information Systems*, *27*. https://doi.org/10.3127/ajis.v27i0.3929

Weisser, C., Gerloff, C., Thielmann, A., Python, A., Reuter, A., Kneib, T., & Säfken, B. (2023). Pseudo-document simulation for comparing LDA, GSDMM and GPM topic models on short and sparse text using Twitter data. *Computational Statistics*, *38*(2). https://doi.org/10.1007/s00180-022-01246-z

Yu, D., & Xiang, B. (2023). Discovering topics and trends in the field of Artificial Intelligence: Using LDA topic modeling. *Expert Systems with Applications*, *225*. https://doi.org/10.1016/j.eswa.2023.120114