BAB II
TINJAUAN PUSTAKA

2.1 Penelitian Yang Relevan

Penelitian yang relevan ini dilakukan untuk membandingkan penelitian sebelumnya dari jurnal dan paper yang sesuai sebagai referensi topik penelitian untuk memperkuat posisi penelitian ini sendiri. Penelitian ini akan membahas topik penelitian, metodologi yang digunakan, dan hasil penelitian.

Tabel 2. 1 Penelitian Terdahulu *Topic Modeling*

Nama Penulis	Roman Egger, Joanne Yu
Tahun Terbit	2022
Judul Penelitian	A *Topic Modeling* Comparison Between *LDA*, NMF, Top2Vec, and *BERTopic* to Demystify X Posts
Metode Penelitian	Penelitian ini membandingkan empat teknik pemodelan topik untuk menganalisis komentar di X terkait dengan COVID-19 dan perjalanan (#covidtravel). Empat metode yang dibandingkan adalah *LDA* (Latent Dirichlet Allocation), NMF (Non-negative Matrix Factorization), Top2Vec, dan *BERTopic*. Data yang digunakan berjumlah 31.800 komentar unik berbahasa Inggris yang dikumpulkan menggunakan alat ekstraksi data Phantombuster. Setiap model diproses dengan teknik pra-pemrosesan yang berbeda: *LDA* dan NMF melalui proses *Stopword*s removal, tokenisasi, *Stemming*, lemmatization, dan transformasi menggunakan TF-IDF. Sedangkan Top2Vec dan *BERTopic* mempertahankan teks asli dan menggunakan *embedding* dan clustering dengan algoritma UMAP dan HDBSCAN.
Hasil Penelitian	Hasil penelitian ini menunjukkan bahwa *BERTopic* menghasilkan topik yang paling jelas, spesifik, dan stabil, diikuti oleh NMF

	yang memiliki topik terpisah dengan baik. Top2Vec memberikan temuan yang unik namun cenderung menghasilkan topik yang tumpang tindih, sedangkan *LDA* sering kali menghasilkan topik yang sangat umum dan kurang relevan untuk data teks pendek seperti komentar. Evaluasi manual oleh peneliti mengindikasikan bahwa *BERTopic* adalah yang paling unggul, diikuti oleh NMF, Top2Vec, dan *LDA*.
Keterkaitan Penelitian	Penelitian ini sangat relevan dengan penelitian saya karena membandingkan *LDA* dan *BERTopic*, yang keduanya adalah teknik utama yang digunakan dalam analisis ulasan X terkait KPU. *BERTopic* terbukti lebih efektif dalam menganalisis teks pendek dan menghasilkan topik yang lebih koheren dan spesifik dibandingkan *LDA*, yang sering kali menghasilkan topik yang lebih umum.

Penelitian pertama yang dilakukan oleh (Egger & Yu, 2022) relevan dengan penelitian ini karena membandingkan kinerja *LDA* dan *BERTopic* dalam menganalisis teks pendek dari X, sehingga memberikan dasar teoritis bagi pemilihan kedua metode tersebut dalam konteks percakapan publik mengenai KPU. Namun, penelitian tersebut memiliki beberapa keterbatasan, seperti fokus pada isu global COVID-19 yang tidak mencakup konteks kelembagaan lokal, evaluasi yang lebih banyak dilakukan secara manual, serta belum mengaitkan hasil *Topic Modeling* dengan analisis persepsi publik. Penelitian ini berupaya melengkapi kekurangan tersebut dengan menerapkan evaluasi kuantitatif koherensi topik dan menganalisis topik yang terbentuk untuk memahami persepsi masyarakat terhadap KPU.

Tabel 2. 2 Penelitian *LDA* dan *BERTopic*

Nama Penulis	Gan et al.
Tahun Terbit	2025

Judul Penelitian	*Experimental Comparison of Three Topic Modeling Methods: LDA, Top2Vec, and BERTopicAllocation, Non-Negative Matrix Factorization and BERTopic*
Metode Penelitian	Penelitian ini membandingkan tiga metode *Topic Modeling*: *LDA*, Top2Vec, dan *BERTopic* dengan menggunakan data dari X terkait topik sosial dan politik. Penelitian ini mengandalkan evaluasi koherensi topik dan evaluasi manual yang dilakukan oleh peneliti untuk menilai kualitas topik yang dihasilkan oleh masing-masing model.*LDA* dan Top2Vec menggunakan TF-IDF untuk representasi kata, sementara *BERTopic* menggunakan teknik *embedding* dan clustering seperti UMAP dan HDBSCAN untuk memetakan topik.
Hasil Penelitian	Berdasarkan evaluasi yang dilakukan, *BERTopic* menghasilkan topik yang lebih terfokus dan koheren dibandingkan dengan *LDA* dan Top2Vec. *LDA* sering kali menghasilkan topik yang umum dan tidak relevan, sementara Top2Vec cenderung menghasilkan topik yang tumpang tindih. *BERTopic* menunjukkan keunggulan dalam menghasilkan topik yang lebih stabil dan lebih mudah dipahami, terutama untuk data pendek dan dinamis seperti komentar.
Keterkaitan Penelitian	Hasil dari penelitian ini memberikan dukungan kuat untuk pilihan *BERTopic* dalam penelitian Anda, karena *BERTopic* terbukti lebih baik dalam menangani data X yang pendek, seperti ulasan X terkait KPU, yang memerlukan pemetaan topik yang lebih relevan dan stabil

Penelitian yang dilakukan oleh (Gan et al., 2024) berjudul "Experimental Comparison of Three *Topic Modeling* Methods: *LDA*, Top2Vec, and *BERTopic*" membandingkan efektivitas tiga metode *Topic Modeling* (*LDA*, Top2Vec, dan *BERTopic*) relevan dengan penelitian ini karena membandingkan *LDA*, Top2Vec, dan *BERTopic* pada data X sehingga memperkuat dasar pemilihan metode dalam menganalisis percakapan publik mengenai KPU. Namun, penelitian tersebut belum diterapkan pada konteks isu kelembagaan seperti KPU dan tidak menambahkan analisis koherensi yang lebih

mendalam. Penelitian ini melengkapi kekurangan tersebut dengan fokus pada konteks KPU dan menggunakan evaluasi kuantitatif koherensi topik untuk menghasilkan analisis yang lebih terukur.

Tabel 2. 3 Perbandingan Metode *Topic Modeling*

Nama Penulis	Olusola Babalola, Bolanle Ojokoh, Olutayo Boyinbode
Tahun Terbit	2024
Judul Penelitian	*Comprehensive Evaluation of LDA, NMF, and BERTopic's Performance on News Headline Topic Modeling*
Metode Penelitian	Penelitian ini mengevaluasi tiga metode *Topic Modeling* (*LDA*, NMF, dan *BERTopic*) dalam pemodelan topik untuk judul berita. Setiap metode diterapkan pada dataset yang berisi artikel berita dari berbagai sektor dan diuji untuk memetakan topik dominan.*LDA* dan NMF mengandalkan TF-IDF dan teknik pra-pemrosesan yang ketat, sementara *BERTopic* mengandalkan *embedding* dan clustering.
Hasil Penelitian	*BERTopic* mengungguli *LDA* dan NMF dalam hal koherensi topik dan spesifikasi topik. *LDA* cenderung menghasilkan topik yang terlalu umum dan tidak cocok untuk data pendek, sedangkan NMF memberikan hasil yang cukup baik, namun tidak sebaik *BERTopic*.
Keterkaitan Penelitian	Penelitian ini sangat relevan dengan penelitian Anda karena *BERTopic* lebih unggul daripada *LDA* dalam analisis data yang pendek dan dinamis seperti komentar. Ini memperkuat keputusan saya untuk menggunakan *BERTopic* dalam menganalisis ulasan X terkait KPU

Penelitian yang dilakukan oleh (Babalola et al., 2024) berjudul "Comprehensive Evaluation of *LDA*, NMF, and *BERTopic*'s Performance on News Headline *Topic Modeling*" mereka membandingkan tiga metode *Topic Modeling* (*LDA*, NMF, dan *BERTopic*) relevan dengan penelitian ini karena membandingkan performa *LDA*, NMF, dan *BERTopic* serta menunjukkan keunggulan *BERTopic* dalam menghasilkan topik yang

lebih koheren, sehingga mendukung pemilihan metode dalam penelitian ini. Namun, penelitian tersebut hanya berfokus pada judul berita dan tidak diterapkan pada data teks pendek dari media sosial seperti X. Penelitian ini melengkapi kekurangan tersebut dengan menerapkan metode pada percakapan publik mengenai KPU serta menambahkan evaluasi koherensi topik yang lebih terukur dalam konteks isu politik lokal

Tabel 2. 4 Penerapan *LDA* dan *BERTopic* pada Berbagai Domain

Nama Penulis	Roman Egger
Tahun Terbit	2024
Judul Penelitian	*Topic Modeling in Tourism*
Metode Penelitian	Penelitian ini mengaplikasikan *LDA* dan *BERTopic* untuk menganalisis data pariwisata seperti ulasan wisatawan di platform media sosial. Setiap model diuji untuk menentukan topik dominan yang dibicarakan dalam data teks yang dikumpulkan.*LDA* menggunakan metode tokenisasi dan TF-IDF untuk representasi kata, sementara *BERTopic* menggunakan *embedding* untuk memetakan topik secara lebih otomatis.
Hasil Penelitian	*BERTopic* menghasilkan topik yang lebih terfokus dan spesifik dibandingkan dengan *LDA*, yang cenderung menghasilkan topik yang umum. *BERTopic* juga lebih efektif untuk analisis data pendek seperti komentar.
Keterkaitan Penelitian	Penelitian ini memberikan bukti lebih lanjut bahwa *BERTopic* lebih efektif dalam analisis data sosial media pendek dan dinamis. Hasil ini relevan untuk analisis ulasan X terkait KPU, di mana pemetaan topik yang terfokus dan stabil sangat dibutuhkan.

Penelitian yang dilakukan oleh (Egger, 2024) relevan dengan penelitian ini karena membandingkan *LDA* dan *BERTopic* pada data teks pendek dalam konteks pariwisata,

sehingga memperkuat dasar metodologis dalam penggunaan kedua metode untuk menganalisis percakapan publik mengenai KPU. Namun, penelitian tersebut terbatas pada domain pariwisata dan tidak membahas isu politik atau persepsi publik terhadap lembaga negara. Oleh karena itu, penelitian ini melengkapi kekurangan tersebut dengan menerapkan metode yang sama pada konteks politik, khususnya isu KPU, serta menambahkan evaluasi koherensi topik secara kuantitatif guna memperoleh hasil yang lebih terukur dan relevan.

Tabel 2. 5 Penelitian Terdahulu *Topic Modeling* Berbasis Embedding

Nama Penulis	Amandeeep Kaur, James R. Wallace
Tahun Terbit	2024
Judul Penelitian	*Moving Beyond LDA: A Comparison of Unsupervised Topic Modeling Techniques*
Metode Penelitian	Penelitian ini membandingkan *LDA*, *BERTopic*, dan Top2Vec dalam pemodelan topik tidak terawasi pada data sosial media. Setiap model diuji untuk mengidentifikasi topik-topik yang relevan dalam data yang sangat dinamis dan berbasis teks pendek. *LDA* menggunakan TF-IDF untuk representasi kata, sementara *BERTopic* menggunakan *embedding* dan teknik clustering.
Hasil Penelitian	*BERTopic* mengungguli *LDA* dalam menghasilkan topik yang terfokus dan lebih mudah dipahami. *LDA* sering kali menghasilkan topik yang lebih umum dan tidak mampu menangani data sosial media yang bersifat dinamis dengan baik.
Keterkaitan Penelitian	Penelitian ini mendukung penggunaan *BERTopic* dalam penelitian saya karena menunjukkan bahwa *BERTopic* lebih efektif dalam menganalisis ulasan X yang pendek dan dinamis, seperti ulasan tentang KPU.

Penelitian yang dilakukan oleh (Kaur & Wallace, 2024) relevan dengan tugas akhir ini karena sama-sama membahas teknik pemodelan topik pada data teks, khususnya teks pendek seperti media sosial. Jurnal tersebut memberikan dasar teori dan pembanding mengenai kelebihan dan kekurangan metode *LDA* serta model berbasis *embedding* seperti *BERTopic*. Namun, penelitian mereka tidak secara spesifik menguji performa model pada konteks bahasa Indonesia dan tidak diterapkan pada data dengan karakteristik tertentu seperti komentar publik atau ulasan yang menjadi objek penelitian dalam tugas akhir ini. Selain itu, jurnal tersebut hanya membandingkan kualitas topik secara umum, tanpa mengevaluasi aspek seperti kebutuhan *preprocessing* tambahan atau stabilitas model pada dataset kecil. Kekurangan inilah yang akan dilengkapi dalam tugas akhir ini melalui pengujian metode yang lebih spesifik pada data Indonesia serta analisis tambahan terkait efektivitas model dalam konteks penelitian yang lebih sempit.

Tabel 2. 6 *Topic Modeling* Multibahasa Menggunakan *BERTopic*

Nama Penulis	Darija Medvecki, Bojana Bašaragin, Adela Ljajić, Nikola Milošević
Tahun Terbit	2024
Judul Penelitian	*Multilingual Transformer and BERTopic for Short Text Topic Modeling: The Case of Serbian*
Metode Penelitian	Penelitian ini menggunakan *BERTopic* untuk menganalisis teks pendek dalam bahasa Serbia, dengan tujuan mengeksplorasi efektivitas model ini dalam menganalisis data teks yang lebih singkat dan multibahasa. Penelitian ini menggunakan *embedding* untuk mewakili kata-kata dalam vektor dan UMAP untuk mengurangi dimensi data. Proses HDBSCAN digunakan untuk mengelompokkan topik-topik yang dihasilkan. Evaluasi dilakukan dengan menggunakan koherensi topik dan interpretasi manual oleh peneliti untuk menilai kualitas topik yang dihasilkan.

Hasil Penelitian	*BERTopic* terbukti lebih efektif daripada *LDA* dalam menghasilkan topik yang terfokus dan koheren, bahkan dalam bahasa non-Inggris seperti Serbia. *LDA* sering kali menghasilkan topik yang lebih umum, tidak terfokus, dan kurang relevan dalam konteks teks pendek.
Keterkaitan Penelitian	Penelitian ini relevan untuk penelitian saya karena *BERTopic* juga dapat digunakan untuk bahasa selain bahasa Inggris, dan keunggulannya dalam menganalisis data teks pendek seperti komentar menjadikannya sangat cocok untuk analisis ulasan X terkait KPU yang bersifat dinamis dan cepat berubah.

Penelitian yang dilakukan oleh Darija Medvecki, Bojana Bašaragin, Adela Ljajić, dan Nikola Milošević pada tahun 2024 berjudul "Multilingual Transformer and *BERTopic* for Short Text *Topic Modeling*: The Case of Serbian" mengeksplorasi penggunaan *BERTopic* dalam menganalisis teks pendek dalam bahasa Serbia, dengan tujuan untuk mengevaluasi efektivitas model ini dalam menangani data teks pendek yang multibahasa, relevan dengan tugas akhir ini, karena sama-sama memanfaatkan *BERTopic* untuk menganalisis teks pendek, sehingga memberikan dasar empiris mengenai efektivitas model tersebut dibandingkan *LDA*. Namun, penelitian tersebut hanya berfokus pada bahasa Serbia dan tidak membahas konteks percakapan publik Indonesia, khususnya terkait isu kelembagaan seperti KPU. Selain itu, penelitian tersebut tidak menguji performa *BERTopic* terhadap dinamika percakapan publik di media sosial yang lebih heterogen. Oleh karena itu, tugas akhir ini melengkapi kekurangan tersebut dengan menerapkan *BERTopic* dan *LDA* pada data percakapan publik mengenai KPU di Indonesia, sehingga memberikan kontribusi baru pada penggunaan *Topic Modeling* dalam konteks sosial-politik nasional.

Tabel 2. 7 *Topic Modeling* pada Media Sosial X dengan Pendekatan *LDA*

Nama Penulis	T Ramamoorthy, V Kulothungan, B
Tahun Terbit	2024

Judul Penelitian	*Topic Modeling and Social Network Analysis Approach to X Data*
Metode Penelitian	Penelitian ini menggunakan kombinasi antara *LDA* untuk *Topic Modeling* dan analisis jaringan sosial untuk memetakan hubungan antara pengguna X dan topik yang mereka diskusikan. Model ini diterapkan pada dataset X yang mengandung berbagai diskusi tentang topik sosial-politik. *LDA* diterapkan untuk mengekstrak topik dari komentar, sementara analisis jaringan sosial digunakan untuk memahami hubungan antara pengguna dan bagaimana mereka terlibat dalam diskusi tentang topik tersebut. Evaluasi dilakukan dengan melihat keterkaitan antar pengguna dan relevansi topik yang dihasilkan oleh *LDA*.
Hasil Penelitian	Penelitian ini menunjukkan bahwa *LDA* efektif untuk mengekstraksi topik dari data X, meskipun hasilnya kurang terfokus untuk topik-topik yang lebih spesifik. Integrasi analisis jaringan sosial menambah nilai dalam memetakan hubungan antar pengguna X
Keterkaitan Penelitian	Penelitian ini relevan dengan penelitian saya karena analisis jaringan sosial dapat memberikan wawasan tambahan dalam menganalisis ulasan X terkait KPU, membantu untuk memetakan dinamika opini publik yang terlibat dalam diskusi seputar KPU. *BERTopic* bisa jadi lebih efektif daripada *LDA* untuk mendapatkan topik yang lebih terfokus, namun pendekatan analisis jaringan ini bisa sangat berguna dalam penelitian yang mengaitkan berbagai pengguna dengan topik yang dibicarakan.

Penelitian yang dilakukan oleh (Ramamoorthy et al., 2024) berjudul "*Topic Modeling* and Social Network Analysis Approach to X Data" menggabungkan *LDA* untuk *Topic Modeling* dan analisis jaringan sosial untuk memetakan hubungan antara pengguna X dan topik yang mereka diskusikan. Relevan dengan tugas akhir ini karena sama-sama menggunakan data X untuk mengidentifikasi topik diskusi publik, sehingga memberikan

gambaran mengenai kemampuan *LDA* dalam memetakan percakapan sosial-politik. Namun, penelitian tersebut hanya menggunakan *LDA* dan tidak membandingkannya dengan metode modern seperti *BERTopic* yang lebih unggul untuk teks pendek. Selain itu, kajian tersebut lebih menekankan analisis jejaring sosial daripada mengevaluasi kualitas topik. Oleh sebab itu, tugas akhir ini melengkapinya dengan membandingkan *LDA* dan *BERTopic* secara langsung serta mengevaluasi koherensi topik dalam konteks percakapan publik mengenai KPU di Indonesia, sehingga memberikan kontribusi metodologis yang lebih komprehensif.

Tabel 2. 8 Pengembangan dan Visualisasi *Topic Modeling* pada Data X

Nama Penulis	Andreas Buchmüller et al.
Tahun Terbit	2022
Judul Penelitian	*Twitmo: A X Data Topic Modeling and Visualization Package for R*
Metode Penelitian	Penelitian ini mengembangkan paket R yang disebut Twitmo, yang dirancang khusus untuk pemodelan topik dan visualisasi data X. Paket ini mendukung penggunaan *LDA* dan alat-alat visualisasi untuk menampilkan hasil model topik secara interaktif.Paket ini memungkinkan peneliti untuk mengonfigurasi jumlah topik dan mengimpor data X dalam format yang sesuai untuk analisis. Evaluasi dilakukan dengan membandingkan hasil visualisasi dan koherensi topik yang dihasilkan oleh *LDA*.
Hasil Penelitian	Twitmo memungkinkan pengguna untuk menganalisis data X dan menghasilkan visualisasi yang interaktif untuk topik-topik yang relevan. *LDA* menghasilkan topik yang lebih umum, tetapi paket ini memberikan kemudahan dalam memvisualisasi hasilnya sehingga dapat lebih mudah dipahami.

Keterkaitan Penelitian	Twitmo menyediakan alat yang sangat berguna dalam analisis ulasan X, karena memberikan visualisasi yang jelas untuk topik-topik yang diekstraksi. Hasil penelitian ini memperkuat pendekatan *LDA* untuk analisis ulasan X terkait KPU, meskipun *BERTopic* bisa lebih unggul untuk topik yang lebih terfokus.

Penelitian yang dilakukan oleh (Buchmüller et al., 2022). Relevan dengan tugas akhir ini karena sama-sama menggunakan data X dan menerapkan *LDA* untuk pemodelan topik, sehingga memberikan gambaran mengenai kemampuan *LDA* dalam memetakan percakapan publik. Namun, penelitian tersebut hanya berfokus pada pengembangan alat (Twitmo) dan tidak membandingkan *LDA* dengan metode pemodelan topik yang lebih modern seperti *BERTopic*. Selain itu, evaluasi yang dilakukan terbatas pada visualisasi dan koherensi topik tanpa analisis mendalam terhadap kualitas topik pada isu tertentu. Oleh karena itu, tugas akhir ini melengkapinya dengan membandingkan *LDA* dan *BERTopic* secara langsung serta mengevaluasi kinerjanya dalam konteks percakapan publik mengenai KPU, sehingga memberikan kontribusi yang lebih komprehensif secara metodologis dan substantif.

Tabel 2. 9 *Topic Modeling* dan Analisis Sentimen pada Media Sosial X

Nama Penulis	HP Suresha, KK Tiwari
Tahun Terbit	2023
Judul Penelitian	*Topic Modeling and Sentiment Analysis of Electric Vehicles of X Data*
Metode Penelitian	Penelitian ini menggabungkan *Topic Modeling* dan analisis sentimen untuk menganalisis komentar yang membahas tentang kendaraan listrik. Metode *LDA* (Latent Dirichlet Allocation) digunakan untuk mengekstraksi topik dominan dalam diskusi X tentang kendaraan listrik. Setelah topik diekstraksi, analisis sentimen dilakukan untuk memahami perasaan positif, negatif, atau netral yang terkait dengan

	kendaraan listrik dalam percakapan publik. *LDA* digunakan untuk memetakan berbagai topik yang sering muncul di X terkait kendaraan listrik. Analisis sentimen dilakukan dengan menggunakan algoritma machine learning untuk menilai sentimen di setiap komentar yang mengandung topik terkait kendaraan listrik.
Hasil Penelitian	*LDA* mampu mengidentifikasi beberapa topik dominan terkait Elon Musk, meskipun beberapa topik yang dihasilkan bersifat sangat umum dan tidak cukup spesifik. *LDA* mengidentifikasi tema-tema utama yang berkaitan dengan Musk, tetapi dengan keterbatasan dalam menangani komentar yang sangat pendek.
Keterkaitan Penelitian	Penelitian ini menunjukkan bahwa *LDA* mampu mengidentifikasi berbagai topik dominan terkait kendaraan listrik, seperti inovasi teknologi, kebijakan pemerintah, dan kesadaran lingkungan. Namun, beberapa topik yang dihasilkan masih terlalu umum dan kurang spesifik. Analisis sentimen mengungkapkan bahwa sentimen positif lebih dominan dalam diskusi tentang kendaraan listrik, meskipun ada juga sentimen negatif terkait biaya tinggi dan masalah infrastruktur yang masih ada. Penelitian ini memberikan wawasan penting mengenai persepsi publik terhadap kendaraan listrik melalui media sosial.

Penelitian yang dilakukan oleh (Suresha & Kumar Tiwari, 2021). Relevan karena sama-sama memanfaatkan data X dan menerapkan *LDA* untuk memetakan topik percakapan publik, sehingga memberikan gambaran mengenai kinerja *LDA* dalam konteks teks pendek. Namun, penelitian tersebut hanya menggunakan *LDA* tanpa membandingkannya dengan metode yang lebih mutakhir seperti *BERTopic*, serta lebih berfokus pada analisis sentimen daripada evaluasi kualitas topik secara mendalam. Oleh karena itu, tugas akhir ini melengkapinya dengan menghadirkan perbandingan langsung antara *LDA* dan *BERTopic* serta menilai koherensi topik dalam konteks percakapan publik

mengenai KPU, sehingga menghasilkan analisis topik yang lebih komprehensif dan sesuai dengan karakteristik data X di Indonesia.

Tabel 2. 10 Penerapan *BERTopic* pada Isu Politik di Media Sosial

Nama Penulis	M Mendonca, A Figueira
Tahun Terbit	2024
Judul Penelitian	Topic Extraction: *BERTopic*'s Insight into the 117th Congress
Metode Penelitian	Penelitian ini menggunakan *BERTopic* untuk mengekstraksi topik-topik dari data X terkait dengan Kongres AS ke-117. Data yang digunakan meliputi komentar yang terkait dengan kegiatan Kongres. *BERTopic* digunakan untuk mengidentifikasi dan mengelompokkan topik-topik yang relevan berdasarkan komentar yang diposting oleh anggota Kongres dan masyarakat.
Hasil Penelitian	*BERTopic* menunjukkan kemampuannya dalam menghasilkan topik yang lebih spesifik dan koheren dibandingkan dengan *LDA*. Topik-topik yang dihasilkan berfokus pada isu-isu tertentu yang dibicarakan di Kongres, dengan keterkaitan antar kata yang kuat.
Keterkaitan Penelitian	Penelitian ini sangat relevan dengan penelitian saya karena penggunaan *BERTopic* untuk analisis topik-topik politik di X memberikan panduan yang sangat berguna untuk analisis ulasan X terkait KPU, yang sering melibatkan wacana politik.

Penelitian yang dilakukan oleh (Xverse et al., 2024). Relevan karena memanfaatkan *BERTopic* untuk menganalisis percakapan politik di X, sehingga memberikan gambaran tentang kemampuan *BERTopic* dalam menghasilkan topik yang koheren pada data teks pendek. Namun, penelitian tersebut hanya berfokus pada penggunaan *BERTopic* tanpa membandingkannya dengan metode lain seperti *LDA*, dan tidak mengevaluasi performa model dalam konteks isu publik yang berbeda. Tugas akhir

ini melengkapinya dengan membandingkan *BERTopic* dan *LDA* pada data ulasan X terkait KPU, sehingga mampu menunjukkan perbedaan kinerja kedua model serta memberikan analisis topik yang lebih komprehensif.

Secara keseluruhan, hasil kajian terhadap penelitian-penelitian terdahulu menunjukkan adanya kecenderungan bahwa model berbasis *embedding* seperti *BERTopic* lebih mampu menghasilkan topik yang koheren dan spesifik pada data teks pendek, khususnya media sosial, dibandingkan dengan *LDA* yang cenderung menghasilkan topik yang lebih umum meskipun memiliki tingkat keberagaman yang baik. Berbagai studi juga menegaskan pentingnya evaluasi koherensi sebagai indikator kualitas model pemodelan topik. Namun demikian, masih terbatas penelitian yang secara khusus membandingkan kedua metode tersebut dalam konteks percakapan publik di Indonesia, terutama pada isu kelembagaan. Oleh karena itu, penelitian ini bertujuan untuk mengisi celah tersebut melalui analisis komparatif antara *LDA* dan *BERTopic* secara kuantitatif dalam konteks sosial-politik nasional.

2.2 Landasan Teori

Landasan teori membahas berbagai teori yang digunakan dalam penelitian berikut ini, teori-teori dihimpun dan didapatkan dari berbagai buku, jurnal dan artikel yang terpercaya yang membahas, mendukung dan menjadi referensi dari penelitian kali ini.

2.2.1 Machine learning

Machine Learning merupakan elemen kunci dalam dunia kecerdasan buatan yang bertujuan untuk menangani berbagai masalah. Ini adalah pengaplikasian kecerdasan buatan yang berfokus pada penciptaan sistem yang bisa belajar secara mandiri tanpa perlu diprogram berkali-kali. Machine Learning adalah area penelitian ilmiah yang mengeksplorasi algoritma dan model statistik yang digunakan oleh komputer untuk menyelesaikan tugas tertentu tanpa petunjuk yang jelas, melainkan dengan menggunakan pola dan penalaran sebagai alternatif. Tujuan penggunaan machine learning adalah untuk melatih mesin agar dapat mengolah data dengan lebih efisien. Dengan mengidentifikasi pola dalam kumpulan data, machine learning bisa memprediksi serta memahami sifat dari objek yang belum dikenali (Nurhalizah et al., 2024)

Machine learning memanfaatkan metode untuk mengelola data dalam jumlah besar secara cerdas untuk menghasilkan output yang akurat. Berdasarkan metode pembelajaran yang digunakan, jenis-jenis machine learning dapat dibagi menjadi pembelajaran terawasi (*supervised learning*), pembelajaran tidak terawasi (*unsupervised learning*), pembelajaran semi terawasi (semi *supervised learning*), dan pembelajaran penguatan (reinforcement learning). Pembelajaran terawasi adalah salah satu metode machine learning yang mengandalkan dataset yang telah diberi label untuk melatih mesin, sehingga mesin dapat mengenali label input dengan memanfaatkan fitur yang ada untuk kemudian melakukan prediksi atau klasifikasi. Sementara itu, pembelajaran tidak terawasi adalah pendekatan yang berfokus pada penarikan kesimpulan berdasarkan dataset yang berisi input tanpa respon yang dilabeli (E.Retnoningsih & R.Pramudita, 2020).

Konsep utama dari machine learning dimulai pada tahun 1950-an saat Alan Turing memperkenalkan ide mengenai "Turing Test" untuk mengukur kecerdasan mesin (Turing, 1950). Di tahun 1952, Arthur Samuel menciptakan sebuah program komputer yang bisa belajar untuk bermain permainan dam (checkers), yang sering dianggap sebagai salah satu contoh awal dari machine learning (Samuel, 1959). Sejalan dengan pertumbuhan data digital dan peningkatan daya komputasi, ML mengalami perkembangan yang signifikan. Ide Deep Learning menggunakan jaringan syaraf tiruan bertingkat telah dikembangkan lebih lanjut oleh Hinton dan rekan-rekannya pada tahun 2006. Sampai sekarang, ML diterapkan di berbagai sektor, termasuk pengenalan suara, pengolahan bahasa alami (NLP), analisis media sosial, serta pemilu digital. Dengan kemampuan menganalisis data besar secara efisien serta menghasilkan prediksi akurat, machine learning terus membuka peluang baru di berbagai sektor industri.

2.2.2 Data Mining

Data mining merupakan proses sistematis untuk menemukan pola, pengetahuan baru, atau informasi tersembunyi dari kumpulan data berukuran besar. Menurut (Hendrickx et al., 2015), data mining adalah tahap inti dari *Knowledge Discovery in Databases (KDD)* yang mencakup proses pembersihan data, transformasi, penambangan pola, hingga interpretasi hasil (Kusuma et al., 2025). Proses ini memanfaatkan teknik

statistik, pembelajaran mesin, dan kecerdasan buatan untuk melakukan berbagai tugas seperti klasifikasi, klasterisasi, prediksi, asosiasi, dan deteksi anomali (Ha et al., 2011). Data Mining juga dimanfaatkan untuk mendeteksi kejadian-kejadian yang ganjil seperti berbagai penerapan data mining yakni Analisa pasar dan manajemen, Analisa perusahaan dan manajemen resiko, Telekomunikasi, Keuangan, Asuransi, Olahraga dan Astronomi (Rifai et al., 2019) . Seiring munculnya big data dan NLP, data mining berevolusi ke arah analisis teks, termasuk penerapan *Topic Modeling* seperti Latent Dirichlet Allocation (*LDA*) oleh (Campbell et al., 2015) dan model berbasis transformer seperti *BERTopic* (Grootendorst, 2022).

2.2.3 Teks Preprocessing

Pengolahan awal data merupakan tahap krusial dalam menganalisis informasi, khususnya dalam pemrosesan bahasa alami dan pembelajaran mesin. (HaCohen-Kerner et al., 2020), tujuan utama dari pengolahan teks awal adalah untuk menggambarkan setiap dokumen dalam bentuk vektor fitur, yaitu dengan memisahkan teks menjadi kata-kata tunggal. Kalimat teks dipandang sebagai transaksi (Siswipraptini, 2023). Pemilihan kata kunci dengan melalui proses pemilihan fitur dan langkah-langkah utama dalam pengolahan teks awal sangat krusial untuk tujuan pengindeksan kalimat(Kadhim, 2018), Langkah-langkah umum dalam *pre-processing* data mencakup:

1. Tokenisasi: Memecah teks menjadi unit-unit kecil seperti kata atau frasa.
2. Penghapusan Stop Words: Menghilangkan kata-kata umum yang tidak memberikan makna signifikan, seperti "dan", "atau", dll.
3. Penghapusan URL: Menghapus tautan atau URL (misalnya, https://...) yang sering muncul dalam komentar
4. Penghapusan spasi ganda: Menghapus spasi berlebihan yang ada di antara kata-kata dalam komentar agar teks lebih bersih dan terstruktur.
5. Pengahpusan Mention: Menghapus mention (@username) yang merujuk pada pengguna lain, karena tidak berkontribusi pada analisis topik atau sentimen dalam konteks komentar.
6. *Stemming/Lemmatization*: Mengubah bentuk inflektif suatu kata ke bentuk dasarnya untuk mengurangi variasi kata yang sama.

Sejarah pengolahan data awal dapat ditelusuri ke perkembangan awal teknik pengolahan informasi dalam dunia komputer pada tahun 1950-an dan 1960-an, saat para peneliti mulai memahami pentingnya mempersiapkan data sebelum melakukan analisis yang lebih mendalam. Dengan kemajuan dalam teknologi informasi serta pertumbuhan jumlah dan kerumitan data, metode pre-processing semakin maju untuk meningkatkan kualitas analisis yang dihasilkan.

2.2.4 Ekstraksi Fitur

Para peneliti dalam menyebutkan bahwa BoW yang digunakan pada metode *LDA* merupakan model multiguna yang dapat digunakan sebagai algoritma pemilihan fitur, dan klasifikasi dokumen dan gambar. Dalam klasifikasi dokumen, BoW adalah vektor jumlah kemunculan kata, yang disebut juga histogram dokumen tersebut. Beberapa kata yang tidak informatif, seperti; a, an, the, dan, dll. akan dihapus dari kamus setelah menghitung semua kata dari kamus yang muncul di dokumen. Misalkan *V* adalah kosakata kata-kata unik di seluruh korpus. Misalkan n adalah jumlah kata unik dalam kosa kata. Misalkan d_i adalah kata ke-i di *V*. Misalkan f_{di} adalah frekuensi data d_i pada dokumen D. *Bag of Word*s dimisalkan *BoW*(D) dari dokumen D adalah vektor dengan panjang n, di mana setiap elemen mewakili frekuensi kata yang terkait dalam kosakata (Arifiandy & Fahmi, 2024).

$$F(d_i) = \big(f(\omega_{1,} d_i), f(\omega_{2,} d_i), \dots f(\omega_{n,} d_i) \tag{2.1}$$

$$BoW(d_i) = (F_{d1}) + (F_{d2}), \dots (F_{dn}) \tag{2.2}$$

d_i merupakan dokumen ke-i hasil *preprocessing*,
w_j merupakan kata ke-j dalam kosakata (*vocabulary*),
n menyatakan jumlah kata unik dalam dataset,
$f(w_j, d_i)$ menyatakan frekuensi kemunculan kata $w_j$pada dokumen d_i.

Class-Based TF-IDF (*c-TF-IDF*) merupakan pengembangan dari metode TF-IDF yang menggeneralisasi proses perhitungan ke tingkat kluster dokumen. Pada metode ini, seluruh dokumen dalam satu kluster digabung dan diperlakukan sebagai satu dokumen besar, sehingga memungkinkan identifikasi kata yang paling merepresentasikan kluster tersebut (Grootendorst, 2022). *c-TF-IDF* bekerja dengan menghitung frekuensi kata pada kluster dan menyesuaikan nilai tersebut dengan frekuensi kata yang muncul pada seluruh kluster. Rumus *c-TF-IDF* (Listyawan et al., 2024) ditunjukkan sebagai berikut:

$$W_{t,c} = tf_{t,c} \, \mathrm{X} \log\left(1 + \frac{A}{tf_t}\right) \tag{2.3}$$

$tf_{t,c}$ = Frekuensi kata t pada kluster c

tf_{tc} = Frekuensi kata t pada semua kluster

A = Jumlah rata – rata kata per kelas A

BERTopic menggunakan model *embedding* (seperti BERT, Sentence-BERT, atau model berbasis Transformer lainnya) untuk mengonversi setiap dokumen atau kalimat menjadi vektor berdimensi tinggi. Proses ini menghasilkan representasi numerik dari teks.

$$v_i = "Embed"(d_i) \tag{2.4}$$

v_i adalah vektor *embedding* yang mewakili dokumen ke- i

d_i adalah dokumen ke- i dalam kumpulan data.

Embed adalah model *embedding* yang mengubah teks menjadi vektor

2.2.5 Latent Dirichlet Allocation

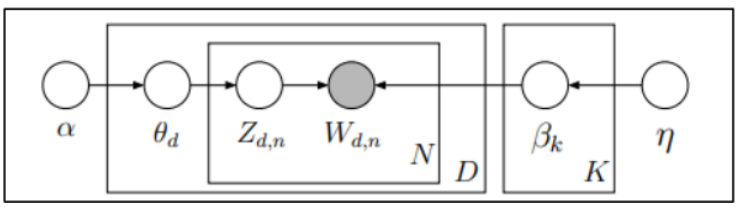
Metode *LDA*, adalah teknik analisis topik yang menggunakan statistik probabilitas untuk mengidentifikasi dan mengelompokkan topik-topik utama dalam kumpulan dokumen (Djunaidi et al., 2025). Metode ini menganalisis pola kemunculan kata untuk mengungkap struktur tematik yang mendasari teks (Penerapan et al., 2025).

LDA pertama kali diperkenalkan pada tahun 2003 sebagai model probabilistik generatif untuk kumpulan dokumen (Blei et al., 2003). Selanjutnya, metode ini banyak digunakan dan dikembangkan dalam berbagai penelitian, salah satunya oleh (Campbell et al., 2015) . *LDA* merupakan metode probabilistik generatif dari kumpulan corpus, Ide dasar *LDA* merupakan model unsupervised untuk pemodelan topik yang digunakan untuk menemukan informasi tersembunyi dan hubungan antar kata, *LDA* mengelompokan topik pada kumpulan teks berdasarkan kesamaan data, Pada *LDA* dokumen merupakan campuran topik yang belum diketahui dan setiap topik terdiri dari distribusi kata (Mutmainah et al., 2023). Dan dibawah ini adalah rumus dari metode *LDA* yang di buat oleh (Hikmah et al., 2020)

$$p(\theta \mid \alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1} \qquad (2.5)$$

1. Probabilitas distribusi topik θ*theta* yang diberikan parameter α*alpha*.
2. Fungsi Gamma, digunakan untuk normalisasi dalam distribusi Dirichlet.
3. Penjumlahan parameter α*alpha*, mengontrol sebaran topik.
4. Produk fungsi Gamma untuk setiap topik, berfungsi untuk normalisasi menentukan kontribusi topik, berdasarkan parameter α*alpha*.

Distribusi Dirichlet digunakan dalam *LDA* untuk memastikan bahwa nilai probabilitas topik bersifat non-negatif dan memiliki total probabilitas sama dengan satu. Nilai parameter αberpengaruh terhadap karakteristik dokumen, di mana nilai αyang kecil cenderung menghasilkan dokumen dengan satu topik dominan, sedangkan nilai αyang besar menghasilkan dokumen dengan distribusi topik yang lebih merata (Blei et al., 2003).



Gambar 2. 1 Model Grafik *LDA* (Oktaviana et al., 2022)

Grafik di atas menunjukkan struktur model Latent Dirichlet Allocation (*LDA*) yang digunakan untuk *Topic Modeling*. Dalam model ini, α mengontrol distribusi topik dalam dokumen, sedangkan θ*d* menggambarkan distribusi topik untuk dokumen *d.* $Z_{d,n}$ adalah indeks topik untuk kata ke-n dalam dokumen d_i $W_{d,n}$ adalah kata ke-n dalam dokumen tersebut. B_k mengontrol distribusi kata dalam topik k_i sementara η adalah parameter distribusi global topik dalam dokumen. Secara keseluruhan, model ini menggambarkan bagaimana *LDA* menghasilkan distribusi topik berdasarkan parameter-parameter tersebut untuk menganalisis teks.

2.2.6 *BERTopic*

BERTopic merupakan model *Topic Modeling* berbasis *embedding* yang dibangun di atas BERT, yang dapat menghasilkan representasi vektor kata dan kalimat dengan properti semantik, sehingga memungkinkan pemahaman secara kontekstual (Kukushkin et al., 2022). *BERTopic* diperkenalkan pada tahun 2020, adalah model topik yang menggunakan transformer BERT, teknik clustering, dan variasi *c-TF-IDF* (class-basedTF-IDF) untuk menghasilkan representasi topik yang koheren. Model ini dirancang untuk mengatasi keterbatasan model lain yang kurang memperhitungkan hubungan semantik antara kata. *BERTopic* unggul dalam berbagai benchmark berkat kinerjanya yang dapat diskalakan seiring dengan perkembangan model bahasa baru, fleksibilitasnya dalam memisahkan proses clustering dari representasi topik, serta kemampuannya untuk memodelkan evolusi topik melalui distribusi kata (Nursyahrina et al., 2024)

BERTopic memiliki 3 langkah utama untuk menghasilkan topik yang telah diolah langkah pertama yaitu document *embedding* yang mana pada tahap ini akan dilakukan konversi document atau biasanya teks menjadi bentuk vektor. Langkah selanjutnya adalah Document Clustering, pada tahap ini akan dilakukan clustering atau pengelompokkan data berdasarkan jarak terdekat. Langkah terakhir adalah Topic Representation, pada tahap ini akan ditentukan sebuah kalimat atau kata sebagai perwakilan kluster yang mana penentuan perwakilan kluster tersebut dilakukan dengan prosedur TF-IDF (Listyawan et al., 2024)

BERTopic menggunakan model *embedding* (seperti BERT, Sentence-BERT, atau model berbasis Transformer lainnya) untuk mengonversi setiap dokumen atau kalimat menjadi vektor berdimensi tinggi. Proses ini menghasilkan representasi numerik dari teks.

$$v_i = "Embed" (d_i) \tag{2.6}$$

Di mana:

v_i adalah vektor *embedding* yang mewakili dokumen ke- i

d_i adalah dokumen ke- i dalam kumpulan data.

Embed adalah model *embedding* yang mengubah teks menjadi vektor.

Setelah mendapatkan *embedding* untuk setiap dokumen, *BERTopic* menggunakan UMAP (Uniform Manifold Approximation and Projection) untuk mereduksi dimensi dan HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) untuk mengelompokkan vektor-vektor dokumen tersebut menjadi topik-topik yang relevan.

1. UMAP: Mereduksi dimensi *embedding* untuk menjaga struktur data.
2. HDBSCAN:Mengelompokkan dokumen berdasarkan kedekatan *embedding* dalam ruang dimensi rendah.

   HDBSCAN menghasilkan topik-topik berdasarkan densitas klaster, dengan persamaan umum:

$$Cluster(d_i) = \arg\max\big(\text{Density}(d_i)\big)$$

Selanjutnya *BERTopic* menggunakan teknik *c-TF-IDF* (class-based TF-IDF) untuk menilai pentingnya kata-kata dalam setiap topik. Rumusnya adalah:

$$c\text{-}TF\text{-}IDF(t,k) = \frac{TF(t,k)}{DF(t,k)} \times \log \frac{N}{df(t)} \tag{2.7}$$

Di mana:

1. $TF(t,k)$ adalah frekuensi kata t dalam topik k1
2. $DF(t,k)$ adalah jumlah dokumen yang mengandung kata t dalam topik k1
3. N adalah jumlah total dokumen,
4. $df(t)$ adalah jumlah dokumen yang mengandung kata t

2.2.7 *Topic Modeling*

Pemodelan topik adalah metode untuk menemukan topik dan tren dari corpus. Pemodelan topik telah diterapkan di beberapa bidang. Metode pemodelan topik umumnya digunakan untuk secara otomatis mengatur, memahami, mencari, dan meringkas arsip elektronik yang besar. "Topik" menandakan hubungan variabel yang tersembunyi, untuk diperkirakan, yang menghubungkan kata-kata dalam kosa kata dan kemunculannya dalam dokumen (Sahria et al., 2020).

Topic Modeling sendiri merupakan bentuk dari model statistic yang digunakan pada machine learning dan natural language processing (NLP) untuk mengidentifikasi struktur tersembunya dari kumpulan data atau teks. *Topic Modeling* memiliki fungsi untuk mengelompokkan kumpulan data atau teks menjadi beberapa topik. Hal ini dapat dimanfaatkan untuk segmentasi produk dan pembuatan rekomendasi produk untuk stakeholder yang akan memberikan keuntungan jangka Panjang (Listyawan et al., 2024). *Topic Modeling* dapat dibuat dengan berbagai metode seperti Latent Dirichlet Allocation (*LDA*), Non-negative Matrix Factorization (NMF), Top2VEC, dan *BERTopic* (Egger & Yu, 2022).

2.2.8 Python

Python adalah bahasa pemrograman tingkat tinggi yang berorientasi objek dan dibuat oleh Guido van Rossum. Python adalah bahasa lintas fungsi yang ditafsirkan secara maksimal yang memiliki banyak fungsi. Bahasa pemrograman berorientasi objek ini biasanya digunakan untuk merampingkan kumpulan data kompleks yang besar. Fitur lain yang membuat Python ini terkenal di kalangan ilmuwan dan penganalisis data adalah fleksibilitasnya. Karena fleksibilitasnya, bahasa ini memungkinkan untuk membuat data model, mensistemasikan kumpulan data, membuat algoritma yang didukung ML (Machine Learning), layanan web, dan

menerapkan penambangan data untuk menyelesaikan berbagai tugas dalam waktu singkat (Junaidi et al., 2023).

Python memiliki perkembangan popularitas yang luar biasa dalam komunitas scientific computing. Hal ini dikarenakan banyaknya library machine learning dan deep learning menggunakan python sebagai bahasa utama mereka. Beberapa contoh library yang menerapkan machine learning adalah Bidirectional Encoder Representations from Transformer atau BERT yang dikenal sebagai dasar dari beberapa library yang digunakan untuk menganalisis data seperti PyABSA dan *BERTopic* (Listyawan et al., 2024).

2.2.9 Coherence Value

Coherence Value adalah salah satu metrik yang digunakan untuk menilai kualitas topik yang dihasilkan oleh model *Topic Modeling*, seperti Latent Dirichlet Allocation (*LDA*) dan *BERTopic*. Metrik ini mengukur seberapa saling terkait kata-kata dalam suatu topik, yang artinya semakin tinggi Coherence Value, semakin baik kualitas topik tersebut. Dengan menggunakan Coherence Value, kita dapat mengidentifikasi seberapa koheren topik-topik yang dihasilkan, yang sangat penting dalam *unsupervised learning* untuk memastikan bahwa topik yang dihasilkan tidak hanya acak tetapi juga bermakna dan relevan dengan data yang dianalisis (Mimno et al., 2009).

Definisi *coherence value* yaitu mengukur tingkat kesamaan semantik antara katakata utama dalam satu topik. Digunakan untuk mengevaluasi dan membandingkan hasil *Topic Modeling*, serta memilih jumlah topik yang optimal. Nilai *coherence* yang tinggi menunjukkan bahwa topik-topik yang dihasilkan lebih mudah diinterpretasi oleh manusia (Fahlevvi & SN, 2022).

Dalam penelitian ini, jenis coherence value yang digunakan adalah C_v coherence. C_v coherence merupakan metrik evaluasi topik yang mengombinasikan informasi kemunculan bersama (*co-occurrence*) kata dalam dokumen dengan pendekatan kesamaan semantik, sehingga menghasilkan pengukuran koherensi topik yang lebih stabil dan representatif. Pendekatan ini diperkenalkan dan dikembangkan oleh (Röder et al., 2015). Sebagai metode evaluasi topik yang memiliki korelasi tinggi.
Secara matematis, C_v coherence dirumuskan oleh (Röder et al., 2015). Sebagai berikut:

$$C_v = \frac{T}{|T|} \sum_{t \epsilon T} C(t) \tag{2.8}$$

Di mana $|T|$merupakan jumlah topik yang dihasilkan oleh model, dan $C(t)$ adalah nilai koherensi untuk setiap topik ke-t, yang dihitung berdasarkan rata-rata tingkat keterkaitan antar pasangan kata dalam topik tersebut. Nilai koherensi antar kata diperoleh dari pola kemunculan kata dalam kumpulan dokumen menggunakan pendekatan probabilistik dan normalisasi tertentu.

2.2.10 Topik Diversity

Evaluasi model merupakan tahapan penting dalam *Topic Modeling* untuk menilai kualitas topik yang dihasilkan oleh suatu algoritma. Menurut (Blei & Jordan, 2003), model topik yang baik tidak hanya mampu mengelompokkan kata secara probabilistik, tetapi juga harus menghasilkan topik yang bermakna, mudah diinterpretasikan, dan tidak saling tumpang tindih. Oleh karena itu, selain *topic coherence* yang mengukur keterkaitan kata dalam satu topik, diperlukan metrik lain yang mampu menilai perbedaan antar topik, yaitu *topic diversity* (Dieng et al., 2020).

Topic diversity didefinisikan sebagai ukuran yang menunjukkan sejauh mana topik-topik yang dihasilkan oleh model bersifat beragam dan tidak redundan. (Dieng et al., 2020) dalam penelitiannya yang berjudul *Topic Modeling in Embedding Spaces* mengusulkan pengukuran *topic diversity* sebagai proporsi kata unik yang muncul pada kumpulan kata teratas dari seluruh topik. Pendekatan ini digunakan untuk mengukur tingkat variasi antar topik yang dihasilkan oleh model, di mana semakin tinggi proporsi kata unik, maka semakin kecil tingkat kemiripan antar topik.

rumus *topic diversity* yang biasa digunakan oleh Dieng et al. (2020) dan dirumuskan sebagai berikut:

$$TD = \frac{Unique\left(w_1 w_{2,\ldots\ldots} w_{k \text{ x } m}\right)}{K \text{ x } m} \tag{2.9}$$

TD adalah nilai *topic diversity*,

K adalah jumlah topik yang dihasilkan oleh model,

m adalah jumlah kata teratas (*top-m words*) yang diambil dari setiap topik,

Unique(·) adalah jumlah kata unik dari seluruh kata teratas pada semua topik.

2.2.11 *CRISP-DM*

CRISP-DM adalah standar industri yang banyak digunakan dalam proses Data Mining. Model ini terdiri dari enam tahapan utama (Ncr & Clinton, 1999).

1. *Business Understanding*

Tahap pertama yaitu *Business Understanding* berfokus pada pemahaman mendalam terhadap tujuan dan kebutuhan bisnis dari proyek data mining. Dalam fase ini, dilakukan identifikasi masalah bisnis, penetapan tujuan, serta perencanaan proyek yang sesuai dengan sasaran bisnis tersebut. Menurut (Wirth & Hipp, 2000) yang dikutip di jurnal, "tahap awal ini berfokus pada pemahaman tujuan dan persyaratan proyek dari perspektif bisnis, dan kemudian mengubah pengetahuan ini menjadi definisi masalah penambangan data, dan rencana proyek awal yang dirancang untuk mencapai tujuan tersebut". Hal ini menegaskan pentingnya merumuskan masalah bisnis ke permasalahan data mining dan membuat rencana proyek sebagai langkah awal.

2. *Data Understanding*

Tahap pemahaman data dimulai dengan pengumpulan data awal dan dilanjutkan dengan berbagai aktivitas untuk mengenal data, mengidentifikasi masalah kualitas data, menemukan wawasanpertama ke dalam data, atau mendeteksi subset menarik untuk membentuk hipotesis atas informasi tersembunyi (Wirth & Hipp, 2000).

3. *Data Preparation*

Data preparation adalah tahap penting dalam proses analisis data dan data mining yang melibatkan pembersihan, transformasi, dan pengorganisasian data mentah agar siap digunakan dalam proses pemodelan. Tahap ini memastikan bahwa data memiliki kualitas tinggi, bebas dari kesalahan, dan terstruktur secara

tepat untuk menghasilkan model yang akurat (Suad A. Alasadi & Wesem S. Bhaya, 2017). Aktivitas kunci dalam fase ini meliputi pembersihan data, transformasi data, dan pemilihan fitur.

4. *Modeling*

Tahap modeling merupakan proses pemilihan teknik atau algoritma, pengujian model dengan membagi data ke set pelatihan dan pengujian, membangun model serta, serta mengevaluasi performanya (Goud et al., 2024). Tahapan ini sering dilalui secara iteratif untuk mendapatkan model terbaik.

5. *Evaluation*

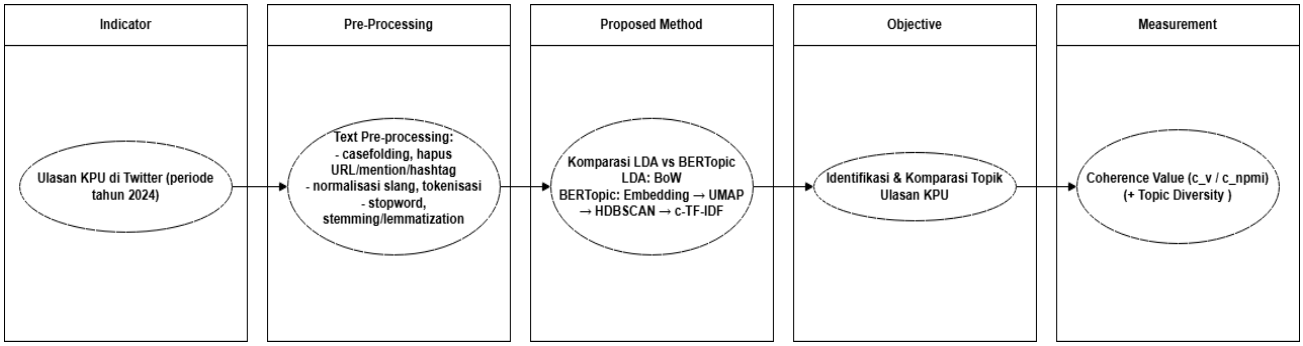
Pada tahap ini, model yang telah dibangun dinilai memiliki kualitas yang baik dari perspektif analisis data. Namun, sebelum dilakukan implementasi akhir, diperlukan evaluasi menyeluruh untuk memastikan bahwa model telah memenuhi tujuan penelitian serta tidak mengabaikan aspek penting yang relevan. Pada akhir tahap ini, harus ditetapkan keputusan mengenai kelayakan dan pemanfaatan hasil data mining yang diperoleh (Brzozowska et al., 2023).

6. Deployment

Tahap Deployment merupakan fase penerapan model ke lingkungan operasional agar hasil analisis dapat digunakan oleh stakeholder atau pihak terkait. Kegiatan dalam tahap ini sangat beragam, mulai dari penyusunan laporan hingga penerapan sistem monitoring yang berjalan secara otomatis. Pada fase ini, model diintegrasikan ke dalam proses nyata untuk membantu pengambilan keputusan strategis (Karmila et al., 2024).

2.3 Kerangka Pemikiran

Penelitian kuantitatif membutuhkan suatu kerangka konseptual atau biasa pula disebut Kerangka Pemikiran, yang berisi operasionalisasi teori yang relevan dan menjadi dasar utama dari penelitian yang akan dilakukan. Memuat pernyataan proposisional atau hubungan antar-konsep penelitian, sehingga dapat digambarkan dalam suatu bagan alur dan menjadi pedoman dalam penyusunan penelitian (Dan et al., 2021).



Gambar 2. 2 Kerangka Pemikiran

1. *Indicator*: Pada tahap ini, fokus utama adalah pada ulasan/percakapan publik tentang KPU yang diambil dari sosial media yaitu X. Data ini menjadi sumber informasi awal untuk analisis lebih lanjut. Data dikumpulkan melalui *scraping* data menggunakan ekstensi Chrome yaitu Web Scraper.
2. *Pre-Preprocessing*: Selanjutnya, data teks yang telah dikumpulkan akan melalui tahap *text pre-processing*. Proses ini bertujuan untuk membersihkan data dari elemen-elemen yang tidak relevan, seperti tanda baca, angka, dan *Stopwords* (kata umum yang diabaikan), serta melakukan normalisasi kata agar data menjadi seragam dan siap untuk dianalisis lebih lanjut.
3. *Proposed Method*: Penelitian ini membandingkan dua teknik *Topic Modeling*: Latent Dirichlet Allocation (*LDA*) dan *BERTopic*. *LDA* adalah model generatif-probabilistik berbasis *bag-of-words* dengan *prior* Dirichlet (α, β) untuk memodelkan distribusi topik per dokumen dan kata per topik. *BERTopic* menggunakan transformer *embedding*s untuk menangkap makna kalimat, mengelompokkan dokumen secara semantik, lalu mengekstrak kata kunci topik dengan *c-TF-IDF*. Perbandingan difokuskan pada kualitas topik (koherensi dan keterjelasan) pada teks pendek X terkait KPU.
4. Tujuan utama adalah mengidentifikasi peta topik percakapan tentang KPU di X dan membandingkan kualitas topik yang dihasilkan oleh *LDA* dan *BERTopic* dalam hal koherensi, spesifisitas, dan keterjelasan (interpretability) serta menyajikan insight isu publik paling dominan dan sensitif waktu.
5. *Measurement*: Pada tahap akhir, dilakukan evaluasi untuk menilai kualitas topik yang dihasilkan oleh model. Salah satu metrik utama yang digunakan

adalah Nilai Koherensi (*Coherence Value*). Nilai ini mengukur seberapa sering kata-kata dalam satu topik muncul bersama dalam dokumen.