BAB II

LANDASAN TEORI

2.1 Penelitian Terdahulu

Tabel 2. 1 Penelitian terkait dalam lima tahun terakhir

No	Penulis, Tahun, Judul, dan Jurnal	Metodologi dan Hasil	Analisis Kritis
1	**Yogie Oktavianus Sihombing dkk. (2024) [30].** ***Optimizing IndoRoBERTa Model for Multi-Class Classification of Sentiment & Emotion on Indonesian Twitter*****. 2024 IEEE 10th Information Technology International Seminar (ITIS).**	Model IndoRoBERTa-Base memperoleh akurasi tertinggi sebesar 97,3% dan F1-score sebesar 97,0% pada klasifikasi sentimen. Sementara itu, pada klasifikasi emosi, model memperoleh akurasi sebesar 82,2% dan F1-score sebesar 82,7%.	Kelebihan penelitian ini adalah penggunaan IndoRoBERTa pada data Twitter berbahasa Indonesia sehingga relevan dengan karakteristik teks media sosial. Namun, penelitian ini masih berfokus pada klasifikasi sentimen dan emosi, belum menggabungkan hasil sentimen dengan pemodelan topik untuk mengetahui isu pembahasan utama.
2	**M. Usama dkk. (2023) [31]**. *RoBERTa-GRU: A Hybrid Deep Learning Model for Enhanced Sentiment*	Hasil penelitian menunjukkan akurasi sebesar 94,63% pada IMDb, 89,59% pada Sentiment140, dan 91,52% pada Twitter US Airline Sentiment.	Kelebihan penelitian ini adalah penggunaan kombinasi RoBERTa dan GRU yang mampu menghasilkan performa tinggi pada beberapa dataset sentimen. Namun, penelitian ini belum

	Analysis. **Applied Sciences.**		berfokus pada data berbahasa Indonesia dan tidak membahas pemodelan topik setelah proses analisis sentimen.
3	***Lazuardy Syahrul Darfiansa dkk. (2025) [32]***. *Optimizing IndoBERT for Revised Bloom's Taxonomy Question Classification Using Neural Network Classifier.* ***Journal of Information Systems Engineering and Business Intelligence.***	Penelitian ini mengoptimalkan IndoBERT untuk klasifikasi pertanyaan berdasarkan Revised Bloom's Taxonomy. Model IndoBERT-LSTM memperoleh akurasi tertinggi sebesar 88,75%.	Penelitian ini menunjukkan bahwa IndoBERT dapat dikombinasikan dengan arsitektur neural network lain untuk meningkatkan performa klasifikasi. Namun, cakupan model masih terbatas pada bahasa Indonesia dan domain klasifikasi pertanyaan pendidikan.
4	***Alya Fitri Nurhaliza dkk. (2024) [33]***. *Penerapan Pemodelan Topik Menggunakan Metode Latent Dirichlet Allocation terhadap Pembahasan Pemilu Indonesia Tahun 2024 di Twitter.* ***Jurnal Pengembangan Teknologi Informasi dan Ilmu Komputer.***	Penelitian ini menggunakan LDA yang diintegrasikan dengan bigram untuk memodelkan topik pembahasan Pemilu 2024 di Twitter. Jumlah topik optimal diperoleh sebanyak 7 atau 8 topik dengan nilai coherence score sebesar 0,5826.	Penelitian ini relevan karena menggunakan data Twitter dan pemodelan topik. Namun, metode LDA masih memiliki keterbatasan dalam memahami konteks semantik dan data media sosial yang mengandung noise.
5	***Dziky Ridhwanullah (2022) [34]***.	Penelitian ini menggunakan LDA untuk memodelkan	Penelitian ini menunjukkan bahwa Twitter dapat digunakan

	Pemodelan Topik pada Cuitan tentang Penyakit Tropis di Indonesia dengan Metode Latent Dirichlet Allocation. ***Jurnal Ilmiah SINUS / Tesis Universitas Islam Indonesia***.	topik dari 2.737 cuitan mengenai penyakit tropis di Indonesia. Hasil penelitian menemukan 5 topik tren pembicaraan dengan nilai koherensi 0,564049.	untuk melihat tren pembicaraan publik. Namun, periode pengumpulan data hanya satu bulan sehingga cakupan temporalnya masih terbatas.
6	***Maria Ulfa Yanuar dan Wahyu Wibowo (2025/2026) [35]***. *Topic Modeling and Sentiment Analysis on Trans Jatim Application User Reviews.* ***MALCOM: Indonesian Journal of Machine Learning and Computer Science / ITS Repository***.	Penelitian ini menggunakan LDA untuk menghasilkan 5 topik ulasan pengguna aplikasi Trans Jatim dan IndoBERT untuk analisis sentimen. Model IndoBERT menghasilkan akurasi sebesar 88,2%.	Penelitian ini sangat relevan karena sama-sama membahas transportasi publik dan menggunakan IndoBERT. Namun, objek data yang digunakan berasal dari ulasan aplikasi, sedangkan penelitian ini menggunakan percakapan Twitter/X.
7	***Lenggo Geni dkk. (2023) [36]***. *Sentiment Analysis of Tweets Before the 2024 Elections in Indonesia Using IndoBERT Language Models.* ***Jurnal Ilmiah Teknik Elektro Komputer***	Penelitian ini melakukan fine-tuning IndoBERT pada data tweet menjelang Pemilu 2024. Model IndoBERT large-p1 memperoleh akurasi sebesar 83,5% dan mengungguli TextBlob, SVM, serta Naive Bayes.	Penelitian ini membuktikan bahwa IndoBERT efektif for analisis sentimen Twitter berbahasa Indonesia. Namun, dataset yang digunakan tidak seimbang karena didominasi oleh kelas netral.

	dan Informatika (JITEKI).		
8	***Aditya Saiful Rizky dan Erwin Yudi Hidayat (2025) [37]***. *Emotion Classification in Indonesian Text Using IndoBERT.* ***Computer Engineering and Applications***.	Penelitian ini menggunakan IndoBERT untuk mengklasifikasikan lima kelas emosi pada 5.079 tweet. Konfigurasi terbaik diperoleh pada rasio data 80:10:10, learning rate 1e-6, dan batch size 8.	Penelitian ini menunjukkan kemampuan IndoBERT dalam memahami ekspresi emosional pada teks Indonesia. Namun, model masih mengalami indikasi overfitting dan kesulitan membaca kalimat dengan emosi campuran.
9	***Agry Alfiah (2026) [38]***. *Sentiment Analysis of the TJ: TransJakarta Application Using the IndoBERT Model.* ***International Journal of Humanities, Social Sciences and Business (INJOSS)***.	Penelitian ini melakukan fine-tuning IndoBERT-base-p1 pada 3.431 ulasan aplikasi TJ: TransJakarta. Model memperoleh akurasi sebesar 87,76%.	Penelitian ini relevan karena membahas TransJakarta dan menggunakan IndoBERT. Namun, data berasal dari ulasan aplikasi, sedangkan penelitian ini berfokus pada opini publik di Twitter/X yang lebih terbuka dan real-time.
10	***Ardiansyah dkk. (2023) [39]***. *Analisis Sentimen terhadap Pelayanan Kesehatan berdasarkan Ulasan Google Maps menggunakan BERT.* ***Jurnal FASILKOM***.	Penelitian ini menggunakan model BERT untuk menganalisis 4.228 ulasan Google Maps terkait pelayanan kesehatan. Pada tahap pengujian, model memperoleh akurasi sebesar 86%.	Penelitian ini menunjukkan bahwa BERT dapat digunakan for analisis sentimen layanan publik. Namun, data mengalami ketidakseimbangan kelas sehingga performa pada kelas netral dan negatif kurang optimal.

11	***Dimas Eko Putro dkk. (2025) [40]***. *Analisis Sentimen Publik terhadap "Save Raja Ampat" di Media Sosial Menggunakan Model IndoBERT.* ***Bulletin of Computer Science Research***.	Penelitian ini menggunakan IndoBERT untuk menganalisis sentimen publik pada komentar media sosial terkait kampanye Save Raja Ampat. Dataset berjumlah 10.000 komentar dari TikTok dan YouTube.	Penelitian ini menunjukkan relevansi IndoBERT for isu publik di media sosial. Namun, terdapat indikasi overfitting setelah epoch tertentu dan distribusi kelas yang tidak seimbang.
12	***Al Farhan Irhanusa Ridat dan Auliya Rahman Isnain (2025) [41]***. *Pemodelan Topik dengan LDA untuk Memahami Persepsi Publik terhadap Apple Vision Pro melalui Analisis Berita Twitter.* ***JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)***.	Penelitian ini menggunakan LDA untuk menghasilkan 6 kelompok topik dengan nilai coherence score optimum sebesar 0,4646.	Penelitian ini relevan karena membahas persepsi publik di Twitter. Namun, LDA memiliki keterbatasan for memahami konteks bahasa slang, sarkasme, dan dinamika perubahan topik.
13	***Nathanael Abel Adrielvino dan Afifah Trista Ayunda (2026) [42]***. *Penerapan IndoBERT dan BERTopic dalam ABSA untuk Evaluasi*	Penelitian ini menggunakan IndoBERT dan BERTopic for melakukan aspect-based sentiment analysis terhadap 7.000 ulasan aplikasi e-government. Topik dipetakan	Penelitian ini relevan karena menggabungkan IndoBERT dan BERTopic. Namun, model masih kurang optimal dalam mengenali sentimen netral dan kalimat yang mengandung opini positif

	Kualitas Aplikasi E-Government Indonesia. ***RABIT: Jurnal Teknologi dan Sistem Informasi Univrab***.	berdasarkan 4 dimensi E-S-QUAL.	serta negatif secara bersamaan.
14	**Jennieta Neysa Anggiyanti dan Frederik Samuel Papilaya (2026) [43]**. *Implementasi Pemodelan Topik BERTopic dan Analisis Sentimen BERT terhadap Ulasan Aplikasi SATUSEHAT Berbasis Natural Language Processing Lanjutan*. **JIPI (Jurnal Ilmiah Penelitian dan Pembelajaran Informatika)**.	Penelitian ini menggabungkan BERT for klasifikasi sentimen dan BERTopic for pemodelan topik. Model BERT memperoleh akurasi 95%, sedangkan BERTopic menghasilkan 10 topik dengan coherence score 0,6101.	Penelitian ini menunjukkan bahwa kombinasi analisis sentimen dan BERTopic dapat menghasilkan pemahaman yang lebih lengkap. Namun, sebagian topik masih saling bertumpang tindih sehingga interpretasinya perlu kehati-hatian.
15	***Evi Yulianti dan Nuzulul Khairu Nissa (2024) [44]***. *ABSA of Indonesian Customer Reviews using IndoBERT: Single-sentence and Sentence-pair Classification Approaches.* ***Bulletin***	Penelitian ini mengevaluasi beberapa pendekatan IndoBERT for aspect-based sentiment analysis pada ulasan pelanggan berbahasa Indonesia. Pendekatan sentence-pair dengan auxiliary	Penelitian ini menunjukkan bahwa IndoBERT efektif for memahami hubungan aspek dan sentimen. Namun, pendekatan auxiliary sentences memiliki waktu komputasi

	of Electrical Engineering and Informatics.	sentences menghasilkan performa paling efektif.	jauh lebih tinggi sehingga kurang efisien.
16	***Rafi Albar Kurniawan dkk. (2026) [45]***. *Pemodelan Topik Ulasan Pengguna Honkai Star Rail Menggunakan BERTopic Berbasis IndoBERT.* ***Blend Sains Jurnal Teknik***.	Penelitian ini menggunakan BERTopic berbasis IndoBERT for mengelompokkan ulasan pengguna Play Store. Model memperoleh rata-rata coherence score C_V sebesar 0,665.	Penelitian ini menunjukkan bahwa BERTopic berbasis IndoBERT dapat menghasilkan topik yang cukup koheren. Namun, jumlah outlier sangat besar sehingga model belum sepenuhnya merepresentasikan keseluruhan korpus.
17	***Ramdhania, K. F., Hidayat, D. F., & Salkiawati, R. (2024). Implementasi Metode Naïve Bayes dan Support Vector Machine (SVM) untuk Menganalisis Sentimen Pengguna Twitter terhadap Transjakarta. Jurnal Matematika dan Pendidikan Matematika, vol. 9, no. 1, pp. 1–14.***	Penelitian ini membandingkan performa Naive Bayes dan SVM berbasis TF-IDF dalam mengklasifikasikan sentimen tweet pengguna terhadap layanan TransJakarta. Dimana disini Naïve Bayes mendapatkan Accuracy 76.43%, sedangkan SVM mendapatkan hasil 84.95%	Kelebihan penelitian ini adalah objek dan sumber data yang identik dengan penelitian ini (TransJakarta, Twitter/X), sehingga menjadi pembanding metode yang sangat relevan. Namun, penelitian ini belum dilengkapi dengan pemodelan topik, sehingga belum dapat menjelaskan isu spesifik yang melatarbelakangi setiap kelas sentimen.
18	***Ruriyanto, D. A., Raya, F. A., & Siswoyo, R. (2026). Analisis Sentimen Ulasan Publik pada***	Penelitian ini menggunakan Naive Bayes untuk mengklasifikasikan sentimen ulasan publik	Penelitian ini relevan karena tujuan akhirnya sejalan dengan penelitian ini, yaitu memberikan masukan untuk

	X Menggunakan Naive Bayes untuk Meningkatkan Kualitas Layanan TransJakarta. Journal of Research and Publication Innovation, 2026.	terhadap TransJakarta di media sosial X guna mendukung peningkatan kualitas layanan.	peningkatan kualitas layanan TransJakarta berdasarkan opini publik. Namun, penelitian ini hanya berhenti pada klasifikasi sentimen tanpa pemodelan topik, sehingga rekomendasi yang dihasilkan masih bersifat umum dan belum mengarah pada aspek layanan tertentu secara spesifik.

Mempelajari penelitian terdahulu dilakukan untuk memahami perkembangan metode analisis sentimen dan pemodelan topik pada data teks berbahasa Indonesia. Berdasarkan Tabel 2.1, pendekatan berbasis *transformer* seperti RoBERTa dan IndoBERT semakin banyak digunakan dalam berbagai tugas NLP, mulai dari klasifikasi sentimen, klasifikasi emosi, klasifikasi pertanyaan, hingga *aspect-based sentiment analysis*. Beberapa penelitian menunjukkan bahwa model berbasis *transformer* mampu menghasilkan performa yang cukup tinggi, seperti akurasi 97,3% dan F1-score 97,0% pada klasifikasi sentimen Twitter berbahasa Indonesia menggunakan IndoRoBERTa [30], akurasi 88,75% pada klasifikasi taksonomi pertanyaan menggunakan IndoBERT [32], 88,2% pada analisis sentimen aplikasi Trans Jatim [35], 83,5% pada sentimen *tweet* Pemilu 2024 [36], dan 87,76% pada aplikasi TJ: TransJakarta [38].

Hasil penelitian terdahulu juga menunjukkan bahwa IndoBERT relevan untuk data berbahasa Indonesia yang bersumber dari media sosial. Pada penelitian Geni dkk., IndoBERT digunakan untuk menganalisis sentimen *tweet* menjelang Pemilu 2024 dan mampu mengungguli beberapa metode konvensional seperti TextBlob, SVM, dan Naive Bayes [36]. Penelitian Rizky dan Hidayat juga menunjukkan bahwa IndoBERT dapat digunakan untuk memahami ekspresi emosi pada *tweet* berbahasa Indonesia, meskipun model masih mengalami kendala ketika

menghadapi kalimat dengan emosi campuran [37]. Hal ini menunjukkan bahwa IndoBERT memiliki kemampuan kontekstual yang kuat, tetapi tetap memerlukan penyesuaian terhadap karakteristik data media sosial.

Dalam konteks layanan publik dan transportasi, beberapa penelitian menunjukkan bahwa analisis sentimen dapat digunakan untuk memahami respons pengguna terhadap layanan yang diberikan. Penelitian Yanuar dan Wibowo menggunakan IndoBERT untuk menganalisis sentimen ulasan aplikasi Trans Jatim dan menghasilkan akurasi sebesar 88,2% [35]. Penelitian Alfiah juga menggunakan IndoBERT pada ulasan aplikasi TJ: TransJakarta dan memperoleh akurasi sebesar 87,76% [38]. Kedua penelitian tersebut relevan dengan penelitian ini karena sama-sama membahas layanan transportasi publik. Namun, penelitian ini memiliki perbedaan karena menggunakan data Twitter/X yang bersifat lebih terbuka, *real-time*, dan mencerminkan percakapan publik secara langsung

Pada sisi pemodelan topik, beberapa penelitian terdahulu masih menggunakan Latent Dirichlet Allocation. LDA digunakan untuk menganalisis topik pembahasan Pemilu 2024 di Twitter dengan *coherence score* 0,5826 [33], cuitan penyakit tropis dengan *coherence score* 0,564049 [34], dan persepsi publik terhadap Apple Vision Pro dengan *coherence score* 0,4646 [41]. Meskipun LDA masih banyak digunakan, beberapa penelitian menunjukkan bahwa metode tersebut memiliki keterbatasan dalam menangani *noise*, bahasa informal, sarkasme, serta konteks semantik pada data media sosial.

Sebagai alternatif, BERTopic mulai banyak digunakan karena memanfaatkan representasi *embedding* berbasis *transformer*. Penelitian Adrielvino dan Ayunda menggabungkan IndoBERT dan BERTopic untuk evaluasi kualitas aplikasi *e-government* [42]. Penelitian Anggiyanti dan Papilaya menggunakan BERT dan BERTopic untuk menganalisis ulasan aplikasi SATUSEHAT, dengan hasil 10 topik dan *coherence score* 0,6101 [43]. Penelitian Kurniawan dkk. juga menggunakan BERTopic berbasis IndoBERT dan memperoleh rata-rata *coherence score* C_V sebesar 0,665 [45]. Hal ini menunjukkan bahwa BERTopic dapat menjadi metode yang relevan untuk mengekstraksi topik dari data teks yang bersifat kontekstual.

Berdasarkan penelitian terdahulu tersebut, penelitian ini memiliki posisi yang berbeda karena berfokus pada analisis sentimen dan pemodelan topik mengenai layanan TransJakarta berdasarkan data Twitter/X. Objek penelitian ini dibatasi pada layanan TransJakarta, sedangkan metode yang digunakan adalah Indonesian RoBERTa untuk *auto labeling*, IndoBERT untuk analisis sentimen, dan BERTopic untuk pemodelan topik. Dengan demikian, penelitian ini tidak hanya mengklasifikasikan sentimen pengguna, tetapi juga mengidentifikasi isu utama yang muncul dalam percakapan masyarakat mengenai layanan TransJakarta.

2.2 Teori yang berkaitan

2.2.1 Analisis Sentimen

Analisis sentimen merupakan proses komputasional yang digunakan untuk mengidentifikasi opini, sikap, emosi, atau penilaian seseorang terhadap suatu objek tertentu melalui data teks. Objek tersebut dapat berupa produk, layanan, kebijakan, organisasi, maupun isu publik yang sedang dibicarakan masyarakat. Dalam konteks penelitian ini, analisis sentimen digunakan untuk mengetahui kecenderungan opini publik terhadap layanan transportasi umum di DKI Jakarta berdasarkan percakapan pada media sosial Twitter/X [46].

Klasifikasi sentimen umumnya dilakukan dengan mengelompokkan teks ke dalam kelas tertentu, seperti positif, negatif, atau netral. Sentimen positif dapat menunjukkan adanya apresiasi, kepuasan, atau pengalaman baik terhadap layanan transportasi umum. Sentimen negatif dapat menunjukkan keluhan, kritik, atau ketidakpuasan terhadap aspek tertentu dalam layanan. Sementara itu, sentimen netral dapat menunjukkan informasi, pertanyaan, atau pernyataan yang tidak secara jelas mengandung opini positif maupun negatif [47].

Dalam penelitian layanan publik, analisis sentimen dapat membantu memahami respons masyarakat terhadap kualitas layanan yang diberikan. Hasil sentimen dapat digunakan untuk melihat bagaimana masyarakat menilai layanan TransJakarta berdasarkan pengalaman yang

mereka sampaikan di Twitter/X. Dengan demikian, analisis sentimen dapat menjadi salah satu pendekatan berbasis data untuk mengevaluasi persepsi publik terhadap layanan transportasi umum.

2.2.2 Pemodelan Topik

Pemodelan topik merupakan pendekatan dalam Natural Language Processing yang digunakan untuk menemukan tema atau topik tersembunyi dalam kumpulan dokumen teks. Metode ini bekerja dengan mengelompokkan kata-kata atau dokumen yang memiliki keterkaitan makna sehingga dapat menggambarkan isu utama yang sering muncul dalam suatu kumpulan data. Dalam penelitian ini, pemodelan topik digunakan untuk mengetahui isu atau pembahasan utama yang muncul dalam percakapan masyarakat mengenai transportasi umum di DKI Jakarta [48].

Pemodelan topik penting karena analisis sentimen hanya menunjukkan kecenderungan opini, sedangkan topik dapat menjelaskan aspek apa yang sedang dibicarakan oleh masyarakat. Misalnya, apabila suatu tweet memiliki sentimen negatif, pemodelan topik dapat membantu mengetahui apakah sentimen negatif tersebut berkaitan dengan keterlambatan, kepadatan penumpang, fasilitas, aksesibilitas, keamanan, atau gangguan operasional. Dengan demikian, hasil pemodelan topik dapat memberikan pemahaman yang lebih rinci terhadap penyebab munculnya sentimen tertentu.

Pada penelitian ini, pemodelan topik dilakukan menggunakan BERTopic. BERTopic memanfaatkan representasi embedding berbasis transformer untuk menangkap makna semantik dari dokumen, kemudian melakukan pengelompokan dokumen berdasarkan kedekatan makna. Topik yang terbentuk kemudian direpresentasikan menggunakan pendekatan class-based TF-IDF agar kata-kata yang paling mencerminkan suatu topik dapat diidentifikasi [29].

2.2.3 Pre-processing

Pre-processing merupakan tahap awal dalam pengolahan data teks yang bertujuan untuk membersihkan dan menyiapkan data sebelum digunakan dalam proses analisis. Data yang berasal dari media sosial umumnya tidak terstruktur karena mengandung mention, hashtag, URL, emoji, tanda baca, singkatan, bahasa tidak baku, serta kemungkinan duplikasi. Oleh karena itu, pre-processing diperlukan agar data lebih bersih, konsisten, dan sesuai untuk digunakan dalam analisis sentimen maupun pemodelan topik [49].

Dalam penelitian ini, pre-processing disesuaikan dengan karakteristik data Twitter/X berbahasa Indonesia. Karena model yang digunakan adalah IndoBERT, data tidak perlu diterjemahkan ke bahasa Inggris. Hal ini berbeda dengan pendekatan yang menggunakan leksikon berbahasa Inggris, karena IndoBERT memang dikembangkan untuk pemrosesan bahasa Indonesia. Dengan demikian, tahapan pre-processing difokuskan pada pembersihan teks, normalisasi, dan penyesuaian format data agar dapat diproses oleh model IndoBERT dan BERTopic [28].

1. *Text filtering.* Text filtering merupakan proses membersihkan teks dari elemen yang tidak diperlukan dalam analisis. Pada data Twitter/X, elemen yang biasanya dihapus meliputi URL, mention pengguna, simbol, angka tertentu, karakter khusus, tanda baca berlebihan, dan teks kosong. Tahap ini bertujuan agar data yang dianalisis lebih berfokus pada isi opini atau pernyataan pengguna.

2. Case folding. Case folding merupakan proses mengubah seluruh huruf dalam teks menjadi huruf kecil. Tahap ini dilakukan untuk menyeragamkan bentuk kata agar kata yang sama tidak dianggap berbeda hanya karena perbedaan huruf kapital. Misalnya, kata "TransJakarta", "transjakarta", dan "TRANSJAKARTA" dapat diseragamkan menjadi "transjakarta".

3. Normalization. Normalization merupakan proses mengubah kata tidak baku, singkatan, atau slang menjadi bentuk yang lebih standar.

Tahap ini penting pada data Twitter/X karena pengguna sering menggunakan bahasa informal, seperti “gak”, “ga”, “nggak”, “tdk”, atau “bgt”. Normalisasi dapat membantu model dalam memahami teks secara lebih konsisten, terutama pada konteks bahasa Indonesia di media sosial.

4. Tokenization. Tokenization merupakan proses memecah teks menjadi unit-unit yang lebih kecil, seperti kata, subword, atau token. Pada model berbasis transformer seperti IndoBERT, tokenisasi dilakukan menggunakan tokenizer yang sesuai dengan model. Tokenizer ini dapat memecah kata menjadi subword tertentu sehingga kata yang jarang muncul tetap dapat diproses oleh model [50].

5. Remove duplicate data. Penghapusan data duplikat dilakukan untuk menghindari kemunculan data yang sama secara berulang. Pada data Twitter/X, duplikasi dapat muncul karena retweet, tweet dengan isi yang sama, atau hasil scraping yang mengambil data berulang. Tahap ini penting agar distribusi data tidak bias terhadap opini tertentu.

6. Remove Stopwords. Stopwords merupakan kata-kata umum yang sering muncul tetapi tidak selalu memiliki makna penting dalam analisis, seperti “yang”, “di”, “ke”, “dan”, atau “dari”. Penghapusan stopwords dapat membantu proses pemodelan topik karena kata-kata yang terlalu umum dapat mengganggu pembentukan topik. Namun, dalam analisis sentimen berbasis transformer, penghapusan stopwords perlu dipertimbangkan dengan hati-hati karena beberapa kata umum tetap dapat memengaruhi konteks kalimat.

7. Handling emoji and symbol. Emoji dan simbol sering digunakan dalam media sosial untuk mengekspresikan emosi. Pada beberapa kasus, emoji dapat memberikan informasi tambahan mengenai sentimen pengguna. Namun, emoji juga dapat dihapus apabila penelitian hanya berfokus pada teks. Dalam penelitian ini, penanganan emoji disesuaikan

dengan kebutuhan preprocessing agar data tetap konsisten untuk analisis sentimen dan pemodelan topik.

8. Filtering data relevan. Filtering data relevan dilakukan untuk memastikan bahwa tweet yang dianalisis benar-benar berkaitan dengan transportasi umum di DKI Jakarta. Tahap ini diperlukan karena kata kunci atau mention tertentu dapat menghasilkan tweet yang tidak sesuai dengan fokus penelitian. Dengan melakukan filtering, data yang digunakan dapat lebih merepresentasikan percakapan publik mengenai TransJakarta.

2.3 Framework/Algoritma yang digunakan

2.3.1 Knowledge Discovery in Database (KDD)

Dalam penelitian berbasis *data mining*, terdapat beberapa *framework* yang dapat digunakan untuk mengatur tahapan penelitian agar lebih sistematis. Beberapa *framework* yang umum digunakan adalah *Cross-Industry Standard Process for Data Mining* (CRISP-DM), *Sample, Explore, Modify, Model, Assess* (SEMMA), dan *Knowledge Discovery in Database* (KDD). Pada penelitian ini, *framework* yang digunakan adalah KDD karena alurnya sesuai untuk proses penemuan pengetahuan dari data yang belum terstruktur, mulai dari pemilihan data, pembersihan data, transformasi data, proses *data mining*, hingga interpretasi hasil [50].

KDD merupakan proses untuk menemukan pola, hubungan, atau informasi yang bernilai dari kumpulan data. Dalam konteks penelitian ini, KDD digunakan untuk mengolah data percakapan Twitter/X mengenai transportasi umum di DKI Jakarta agar dapat menghasilkan informasi berupa sentimen publik dan topik pembahasan utama. Alur KDD dinilai sesuai karena data Twitter/X memiliki karakteristik tidak terstruktur, sehingga perlu melalui tahapan pembersihan dan transformasi sebelum dianalisis menggunakan model IndoBERT dan BERTopic.

Tahapan KDD dalam penelitian ini terdiri dari lima tahap utama. Tahap pertama adalah *data selection*, yaitu proses pemilihan data Twitter/X yang berkaitan dengan TransJakarta. Tahap kedua adalah *data cleaning*, yaitu proses pembersihan data dari unsur yang tidak diperlukan seperti *URL*, *mention*, simbol, data duplikat, dan data tidak relevan. Tahap ketiga adalah *data transformation*, yaitu proses menyesuaikan data teks agar dapat digunakan oleh model. Tahap keempat adalah *data mining*, yaitu proses analisis sentimen menggunakan IndoBERT dan pemodelan topik menggunakan BERTopic. Tahap terakhir adalah *interpretation/evaluation*, yaitu proses mengevaluasi hasil model dan menginterpretasikan sentimen serta topik yang ditemukan.

Berdasarkan alur tersebut, KDD menjadi *framework* yang sesuai untuk penelitian ini karena dapat memandu proses analisis data secara bertahap. Penelitian ini tidak hanya melakukan klasifikasi sentimen, tetapi juga menggali topik pembahasan dari data Twitter/X. Oleh karena itu, penggunaan KDD membantu memastikan setiap tahapan penelitian dilakukan secara terstruktur dan dapat dijelaskan secara sistematis.

2.3.2 BERT

Bidirectional Encoder Representations from Transformers atau BERT merupakan model bahasa berbasis *transformer* yang dikembangkan untuk memahami konteks kata secara dua arah. Berbeda dengan pendekatan satu arah, BERT membaca konteks kata dari sisi kiri dan kanan kalimat secara bersamaan. Kemampuan ini membuat BERT dapat memahami hubungan antar kata secara lebih baik, terutama pada kalimat yang maknanya bergantung pada susunan dan konteks [27].

BERT dibangun menggunakan arsitektur *transformer encoder* yang memanfaatkan mekanisme *self-attention*. Mekanisme ini memungkinkan model memberikan bobot perhatian pada kata-kata tertentu dalam sebuah kalimat sesuai dengan kontribusinya terhadap makna keseluruhan. Dengan cara tersebut, BERT dapat menghasilkan representasi teks yang

lebih kontekstual dibandingkan pendekatan berbasis frekuensi kata atau metode konvensional.

Dalam prosesnya, BERT menggunakan *tokenizer* untuk mengubah teks menjadi bentuk *token*. Token tersebut kemudian direpresentasikan dalam bentuk *embedding* yang dapat diproses oleh model. BERT juga menggunakan token khusus seperti *[CLS]* dan *[SEP]*. Token *[CLS]* umumnya digunakan sebagai representasi keseluruhan kalimat untuk tugas klasifikasi, sedangkan token *[SEP]* digunakan sebagai penanda pemisah antar kalimat atau akhir kalimat [27].

Dalam penelitian ini, BERT menjadi landasan konseptual dari IndoBERT. IndoBERT sendiri merupakan model yang dikembangkan berdasarkan arsitektur BERT, tetapi dilatih menggunakan korpus Bahasa Indonesia. Oleh karena itu, pemahaman terhadap BERT diperlukan untuk menjelaskan dasar dari model IndoBERT yang digunakan dalam analisis sentimen.

2.3.3 IndoBERT

IndoBERT merupakan model bahasa pra-latih berbasis BERT yang dikembangkan khusus untuk Bahasa Indonesia. Model ini diperkenalkan dalam penelitian IndoLEM sebagai bagian dari upaya membangun *benchmark* dan model bahasa untuk berbagai tugas *Natural Language Processing* Bahasa Indonesia. IndoBERT dilatih menggunakan korpus Bahasa Indonesia sehingga lebih sesuai untuk memahami struktur bahasa, kosakata, dan konteks kalimat dalam Bahasa Indonesia [28].

Penggunaan IndoBERT dalam penelitian ini didasarkan pada karakteristik data yang berupa tweet berbahasa Indonesia. Data Twitter/X umumnya berisi kalimat pendek, bahasa informal, singkatan, serta konteks yang tidak selalu eksplisit. Dengan kemampuan representasi kontekstual dua arah, IndoBERT dapat membantu model memahami makna kalimat secara lebih baik dibandingkan metode yang hanya bergantung pada kata-kata tertentu.

Dalam analisis sentimen, IndoBERT digunakan untuk mengklasifikasikan tweet ke dalam kelas sentimen tertentu. Tweet yang telah melalui tahap *pre-processing* akan diproses oleh *tokenizer* IndoBERT, kemudian dimasukkan ke dalam model untuk menghasilkan prediksi sentimen. Hasil prediksi tersebut digunakan untuk mengetahui kecenderungan persepsi publik terhadap layanan TransJakarta.

Pemilihan IndoBERT juga relevan karena penelitian ini berfokus pada teks berbahasa Indonesia. Penggunaan model berbahasa Indonesia diharapkan dapat mengurangi keterbatasan yang mungkin muncul apabila menggunakan model bahasa umum atau model yang dilatih pada korpus bahasa asing. Dengan demikian, IndoBERT menjadi algoritma utama dalam tahap analisis sentimen pada penelitian ini.

2.3.4 BERTopic

BERTopic merupakan metode *topic modeling* yang memanfaatkan representasi semantik berbasis *transformer embedding*. Berbeda dengan metode pemodelan topik tradisional yang banyak bergantung pada distribusi kata, BERTopic menggunakan representasi vektor dari dokumen untuk menangkap kedekatan makna antar teks. Pendekatan ini membuat BERTopic relevan untuk data media sosial yang cenderung singkat, tidak terstruktur, dan memiliki variasi bahasa yang tinggi [29].

Secara umum, BERTopic memiliki beberapa tahapan utama. Tahap pertama adalah pembuatan *document embedding*, yaitu proses mengubah setiap dokumen atau tweet menjadi representasi vektor. Tahap kedua adalah *dimension reduction*, yaitu proses mengurangi dimensi vektor agar lebih mudah dikelompokkan. Tahap ketiga adalah *clustering*, yaitu proses mengelompokkan dokumen yang memiliki kedekatan makna. Tahap terakhir adalah pembuatan representasi topik menggunakan *class-based TF-IDF* atau c-TF-IDF [29].

Dalam penelitian ini, BERTopic digunakan untuk mengidentifikasi topik pembahasan publik mengenai transportasi umum di DKI Jakarta.

Setelah tweet diklasifikasikan berdasarkan sentimen menggunakan IndoBERT, data dapat dianalisis lebih lanjut untuk melihat tema pembahasan pada setiap kelompok sentimen. Dengan demikian, BERTopic membantu menjelaskan isu yang paling sering muncul, seperti keterlambatan, kepadatan, fasilitas, aksesibilitas, integrasi antarmoda, atau gangguan operasional.

Penggunaan BERTopic dalam penelitian ini memberikan nilai tambah karena hasil penelitian tidak hanya berhenti pada klasifikasi sentimen. Sentimen dapat menunjukkan kecenderungan opini publik, sedangkan topik dapat menjelaskan aspek layanan yang menjadi perhatian masyarakat. Oleh karena itu, kombinasi IndoBERT dan BERTopic dapat memberikan pemahaman yang lebih lengkap terhadap persepsi publik mengenai transportasi umum di DKI Jakarta.

2.3.5 RoBERTa

RoBERTa (*Robustly Optimized BERT Pretraining Approach*) merupakan model berbasis *transformer* yang dikembangkan sebagai pengembangan dari BERT dengan optimasi pada proses *pre-training*. Dalam analisis sentimen, RoBERTa mampu menghasilkan representasi teks kontekstual sehingga dapat membantu model memahami hubungan antarkata dalam kalimat, termasuk pada data media sosial yang bersifat pendek dan tidak terstruktur. Penelitian Sihombing dkk. menunjukkan bahwa IndoRoBERTa-Base memperoleh akurasi sebesar 97,3% dan *F1-score* sebesar 97,0% pada klasifikasi sentimen Twitter berbahasa Indonesia [30]. Selain itu, penelitian Usama dkk. menunjukkan bahwa model *hybrid* RoBERTa-GRU memperoleh akurasi sebesar 94,63% pada dataset IMDb, 89,59% pada Sentiment140, dan 91,52% pada Twitter US Airline Sentiment [31]. Hasil tersebut menunjukkan bahwa RoBERTa memiliki kemampuan yang baik dalam tugas analisis sentimen, terutama untuk menghasilkan representasi teks yang kuat. Dalam penelitian ini, RoBERTa digunakan pada tahap *auto labeling* untuk memberikan label

awal sentimen negatif, netral, dan positif sebelum data digunakan dalam proses *fine-tuning* IndoBERT.

2.3.6 UMAP

Uniform Manifold Approximation and Projection atau UMAP merupakan metode *dimension reduction* yang digunakan untuk mengurangi dimensi data berdimensi tinggi ke dalam dimensi yang lebih rendah. Dalam konteks BERTopic, UMAP digunakan untuk mereduksi dimensi *embedding* dokumen sebelum proses *clustering*. Tahap ini penting karena representasi *embedding* umumnya memiliki dimensi yang besar, sehingga perlu disederhanakan agar proses pengelompokan dokumen dapat berjalan lebih efektif [50].

UMAP bekerja dengan mempertahankan struktur kedekatan data sehingga dokumen yang memiliki makna serupa tetap berada dalam posisi yang berdekatan pada ruang dimensi rendah. Pada penelitian ini, UMAP digunakan sebagai bagian dari alur BERTopic untuk membantu mengelompokkan tweet berdasarkan kedekatan semantik. Dengan demikian, tweet yang membahas isu serupa mengenai transportasi umum dapat lebih mudah dikelompokkan ke dalam topik tertentu.

2.3.7 HDBSCAN

Hierarchical Density-Based Spatial Clustering of Applications with Noise atau HDBSCAN merupakan metode *clustering* berbasis kepadatan yang dapat mengelompokkan data tanpa harus menentukan jumlah klaster sejak awal. HDBSCAN bekerja dengan melihat kepadatan data dan membentuk kelompok berdasarkan area yang memiliki kepadatan tinggi. Metode ini juga dapat mengidentifikasi data yang tidak masuk ke dalam kelompok tertentu sebagai *noise* atau *outlier* [51].

Dalam BERTopic, HDBSCAN digunakan setelah tahap reduksi dimensi menggunakan UMAP. Data tweet yang telah direpresentasikan dalam bentuk vektor dan direduksi dimensinya akan dikelompokkan berdasarkan kedekatan semantik. Keunggulan HDBSCAN adalah

kemampuannya menemukan jumlah topik secara otomatis berdasarkan struktur data. Hal ini relevan untuk penelitian ini karena jumlah topik pembahasan publik mengenai transportasi umum belum diketahui sejak awal.

Selain membentuk kelompok topik, penggunaan HDBSCAN dalam BERTopic juga memungkinkan munculnya dokumen yang tidak masuk ke dalam *cluster* tertentu. Kondisi ini penting untuk dijelaskan karena data Twitter/X memiliki karakteristik yang pendek, informal, dan bervariasi, sehingga tidak semua tweet dapat dikelompokkan secara kuat ke dalam topik yang sama.

Dalam BERTopic, dokumen yang tidak memiliki kemiripan makna yang cukup kuat dengan topik tertentu akan dimasukkan ke dalam topik -1 atau disebut sebagai *outlier*. *Outlier* cukup umum ditemukan pada data media sosial karena *tweet* biasanya pendek, menggunakan bahasa informal, dan membahas banyak hal yang berbeda. *Outlier* tidak selalu berarti bahwa data tersebut salah, tetapi menunjukkan bahwa dokumen belum dapat dikelompokkan dengan jelas ke dalam salah satu topik. Untuk mengurangi jumlah *outlier*, BERTopic memiliki proses *outlier reduction* yang dapat memetakan ulang sebagian dokumen dari topik -1 ke topik yang paling mendekati berdasarkan nilai probabilitas atau kemiripan dokumen. Pada penelitian ini, *outlier reduction* digunakan agar *tweet* yang sebelumnya masuk ke dalam topik -1 tetap dapat dipertimbangkan dalam interpretasi topik tanpa menghapus data dari dataset.

2.3.8 Class-based TF-IDF

Class-based TF-IDF atau c-TF-IDF merupakan pendekatan yang digunakan dalam BERTopic untuk membentuk representasi topik. Pada pendekatan ini, dokumen-dokumen yang berada dalam satu klaster dianggap sebagai satu kelas, kemudian kata-kata penting pada kelas tersebut dihitung berdasarkan bobot frekuensi dan keunikannya

dibandingkan kelas lain. Dengan cara ini, BERTopic dapat menampilkan kata-kata yang paling merepresentasikan suatu topik [29].

Dalam penelitian ini, c-TF-IDF digunakan untuk membantu menginterpretasikan hasil topik dari tweet mengenai transportasi umum. Misalnya, apabila satu klaster memiliki kata dominan seperti "padat", "halte", "antre", dan "penuh", maka topik tersebut dapat diinterpretasikan sebagai pembahasan mengenai kepadatan atau antrean penumpang. Dengan demikian, c-TF-IDF membantu proses pemberian makna terhadap topik yang terbentuk.

2.3.9 Confusion Matrix

Confusion matrix merupakan tabel evaluasi yang digunakan untuk melihat performa model klasifikasi dengan membandingkan hasil prediksi model terhadap label sebenarnya. Dalam penelitian ini, *confusion matrix* digunakan untuk mengevaluasi performa model IndoBERT dalam mengklasifikasikan sentimen tweet. Evaluasi ini penting untuk mengetahui seberapa baik model dalam membedakan sentimen positif, negatif, dan netral.

Pada *confusion matrix*, terdapat beberapa komponen utama, yaitu *true positive* (TP), *true negative* (TN), *false positive* (FP), dan *false negative* (FN). TP menunjukkan data positif yang berhasil diprediksi sebagai positif, sedangkan TN menunjukkan data negatif yang berhasil diprediksi sebagai negatif. FP menunjukkan data yang salah diprediksi sebagai positif, sedangkan FN menunjukkan data positif yang gagal dikenali oleh model. Komponen-komponen tersebut digunakan untuk menghitung metrik evaluasi seperti *accuracy*, *precision*, *recall*, dan *F1-score* [52].

1. *Accuracy* merupakan rasio jumlah prediksi benar terhadap seluruh data yang diuji. Metrik ini menunjukkan tingkat ketepatan model secara keseluruhan. Namun, *accuracy* perlu diinterpretasikan

dengan hati-hati apabila dataset memiliki distribusi kelas yang tidak seimbang.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

2.*Precision* merupakan rasio prediksi benar pada suatu kelas terhadap seluruh data yang diprediksi sebagai kelas tersebut. Metrik ini menunjukkan seberapa tepat model ketika memberikan prediksi pada suatu kelas.

$$\text{Precision} = \frac{TP}{TP + FP}$$

3. *Recall* merupakan rasio prediksi benar pada suatu kelas terhadap seluruh data aktual pada kelas tersebut. Metrik ini menunjukkan kemampuan model dalam menemukan seluruh data yang seharusnya masuk ke kelas tertentu.

$$\text{Recall} = \frac{TP}{TP + FN}$$

4. *F1-score* merupakan rata-rata harmonik dari precision dan recall. Metrik ini digunakan untuk melihat keseimbangan antara ketepatan prediksi dan kemampuan model dalam menemukan data pada suatu kelas

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

2.3.10 Coherence Score

Coherence score merupakan metrik evaluasi yang digunakan untuk menilai kualitas topik yang dihasilkan oleh metode *topic modeling*. Metrik ini mengukur sejauh mana kata-kata utama dalam suatu topik memiliki keterkaitan makna. Semakin tinggi nilai *coherence score*, semakin baik keterhubungan antar kata dalam topik tersebut, sehingga topik lebih mudah diinterpretasikan oleh manusia [53].

Dalam penelitian ini, *coherence score* digunakan untuk mengevaluasi hasil pemodelan topik dari BERTopic. Evaluasi ini penting karena hasil topik tidak hanya perlu terbentuk secara teknis, tetapi juga harus memiliki makna yang dapat dipahami. Jika kata-kata dalam satu topik memiliki hubungan yang kuat, maka topik tersebut dapat dianggap lebih koheren dan lebih layak digunakan dalam interpretasi hasil.

Salah satu jenis *coherence score* yang sering digunakan adalah C_V *coherence*. C_V *coherence* menilai keterkaitan antar kata dalam suatu topik dengan mempertimbangkan kedekatan semantik dan statistik kemunculan kata. Pada penelitian ini, C_V *coherence* dapat digunakan untuk melihat kualitas topik yang dihasilkan BERTopic pada data tweet transportasi umum di DKI Jakarta.

2.4 Tools penelitian

2.4.1 Google Colab

Google Colab merupakan layanan *hosted Jupyter Notebook* yang dapat digunakan untuk menulis dan menjalankan kode Python melalui *browser*. Layanan ini memungkinkan pengguna menjalankan kode pada server Google tanpa perlu melakukan instalasi lingkungan pemrograman secara lokal. Google Colab juga menyediakan akses ke sumber daya komputasi seperti GPU dan TPU, sehingga sesuai digunakan untuk kebutuhan *machine learning*, *data science*, dan pemrosesan model berbasis *deep learning* [54].

Dalam penelitian ini, Google Colab digunakan sebagai lingkungan kerja untuk melakukan pengolahan data, *pre-processing*, pelatihan model IndoBERT, evaluasi model, serta pemodelan topik menggunakan BERTopic. Penggunaan Google Colab membantu proses eksperimen karena mendukung instalasi berbagai *library* Python yang dibutuhkan, seperti *pandas*, *transformers*, *scikit-learn*, dan *bertopic*. Selain itu, integrasi dengan Google Drive dapat memudahkan penyimpanan dataset, model, dan hasil analisis.

Pemilihan Google Colab dalam penelitian ini didasarkan pada kemudahan akses dan dukungan komputasi yang memadai. Model berbasis *transformer* seperti IndoBERT membutuhkan sumber daya komputasi yang cukup besar, terutama pada proses *fine-tuning*. Dengan adanya dukungan GPU, proses pelatihan dan pengujian model dapat dilakukan lebih efisien dibandingkan hanya menggunakan perangkat lokal dengan spesifikasi terbatas.