BAB II

TINJAUAN PUSTAKA

2.1. Studi Sebelumnya

Beberapa penelitian sebelumnya telah mengkaji kebijakan pembatasan usia pengguna media sosial dari berbagai perspektif. Stevens dkk. [33] dalam sebuah artikel pandangan menelaah secara kritis kebijakan pembatasan usia 16 tahun di Australia dengan merujuk preseden kebijakan serupa di Tiongkok dan Korea Selatan, dan menyimpulkan bahwa pembatasan usia saja tidak cukup tanpa regulasi pendamping terkait desain platform, transparansi algoritma, serta pembatasan *profiling* data anak. Hal ini menunjukkan perlunya penelitian lebih lanjut tentang bagaimana publik merespons kebijakan serupa di Indonesia. Mar'atusholihat dkk. [30] melakukan analisis sentimen terhadap kebijakan pembatasan penggunaan media sosial anak menggunakan algoritma Decision Tree dan Naïve Bayes pada komentar TikTok, dengan Naïve Bayes menghasilkan akurasi lebih tinggi sebesar 97,14 persen dibandingkan Decision Tree sebesar 91,90 persen. Namun, penelitian tersebut belum menerapkan pemodelan topik dan belum menangani ketidakseimbangan kelas secara optimal.

Sejumlah penelitian juga telah mengkaji sentimen publik terhadap kebijakan pemerintah menggunakan platform YouTube sebagai sumber data. Nandini & Rahardi [26] membandingkan beberapa algoritma *machine learning* untuk menganalisis komentar YouTube terkait kebijakan ekonomi, dengan SVM berbasis TF-IDF bigram sebagai model terbaik yang mencapai akurasi 96,08 persen, meskipun sentimen netral masih menjadi tantangan karena sifatnya yang ambigu. Setiawan dkk. [34] menggunakan IndoBERT untuk menganalisis 23.796 komentar YouTube terhadap putusan Mahkamah Konstitusi (MK) terkait batas usia minimum calon wakil presiden dan mencapai akurasi 95 persen, mengungguli BERT, SVM, dan Random Forest. Temuan ini memperlihatkan keunggulan IndoBERT dalam menangkap nuansa kontekstual bahasa Indonesia pada komentar kebijakan di YouTube, meskipun analisis pemodelan topik belum disertakan.

Selain YouTube, beberapa penelitian menganalisis sentimen kebijakan pemerintah dari berbagai platform sekaligus. Manoppo dkk. [27] menggunakan IndoBERT pada data dari platform X, Instagram, dan TikTok untuk menganalisis sentimen

terhadap kebijakan kenaikan PPN 12%, dengan akurasi 84,94 persen dan dominasi sentimen negatif. Khaqqi dkk. [32] menggabungkan IndoBERT dan BERTopic untuk menganalisis sentimen terhadap kebijakan tarif impor AS-Tiongkok dari komentar di empat platform, mencapai akurasi 93,03 persen dan menemukan korelasi positif lemah hingga sedang yang signifikan antara sentimen publik dan harga kedelai global sebagai satu-satunya indikator ekonomi yang signifikan, meskipun periode observasi yang singkat membatasi generalisasi temuan tersebut. Penelitian-penelitian multiplatform tersebut cenderung menggabungkan data lintas platform tanpa membandingkan secara khusus perbedaan pola sentimen antarplatform.

Beberapa peneliti juga menggabungkan analisis sentimen dan pemodelan topik dalam satu kerangka penelitian. Sihombing & Widiyaningtyas [31] mengintegrasikan IndoBERTweet dan BERTopic untuk menganalisis isu kesehatan publik dari platform X, menghasilkan *F1-score* 0,7822 dan 44 topik diskusi yang beragam. Ishak dkk. [35] menggunakan IndoBERT dan LDA untuk menganalisis sentimen terhadap pemerintahan Prabowo–Gibran pada artikel berita daring, dengan dominasi sentimen negatif sebesar 80 persen, dan menunjukkan bahwa kombinasi keduanya dapat memberikan gambaran yang lebih lengkap terhadap dinamika opini publik. Nur dkk. [36] menerapkan IndoBERT dan BERTopic pada opini konsumen tentang merek teknologi di platform X, menghasilkan akurasi 79,8 persen dan 16 klaster topik. Berbagai penelitian telah menerapkan pendekatan gabungan analisis sentimen dan pemodelan topik. Namun, belum ditemukan penelitian yang secara khusus mengkaji sentimen dan topik pada kebijakan pembatasan usia pengguna media sosial di Indonesia menggunakan IndoBERT dan BERTopic dengan data komentar dari YouTube dan Instagram.

Tabel 2.1. Penelitian Sebelumnya

Penulis	Tahun	Tujuan	Metode	Data	Hasil
Stevens dkk. [33]	2026	Menelaah kebijakan pembatasan usia 16 tahun di Australia	Artikel pandangan (telaah kritis dengan preseden Tiongkok dan Korea Selatan)	Kajian kebijakan	Pembatasan usia saja tidak cukup tanpa regulasi desain platform, transparansi algoritma, dan pembatasan profiling data anak
Mar'atusholihat dkk. [30]	2025	Analisis sentimen kebijakan pembatasan media sosial anak	Decision Tree dan Naïve Bayes (*Bag of Words*)	Komentar TikTok	Naïve Bayes 97,14% mengungguli Decision Tree 91,90%
Nandini & Rahardi [26]	2025	Analisis sentimen komentar kebijakan ekonomi	Perbandingan algoritma *machine learning* dan ensemble (TF-IDF)	Komentar YouTube	SVM TF-IDF bigram terbaik (96,08%); sentimen netral menjadi tantangan
Setiawan dkk. [34]	2025	Analisis sentimen putusan MK tentang batas usia calon wakil presiden	IndoBERT (dibandingkan BERT, SVM, Random Forest)	23.796 komentar YouTube	IndoBERT akurasi 95%, mengungguli model lain

Penulis	Tahun	Tujuan	Metode	Data	Hasil
Manoppo dkk. [27]	2025	Analisis sentimen kebijakan kenaikan PPN 12%	Fine-tuning IndoBERT (*class weighting*)	X, Instagram, TikTok	Akurasi 84,94%; sentimen negatif dominan
Khaqqi dkk. [32]	2025	Analisis sentimen kebijakan tarif impor AS-Tiongkok dan kaitannya dengan indikator ekonomi	IndoBERT + BERTopic + korelasi Spearman	Komentar empat platform (TikTok, Instagram, YouTube, X)	Akurasi 93,03%; korelasi positif lemah–sedang signifikan dengan harga kedelai global
Sihombing & Widiyaningtyas [31]	2025	Analisis sentimen dan pemodelan topik isu kesehatan publik	IndoBERTweet + BERTopic	Platform X	F1-Score 0,7822; 44 topik diskusi
Ishak dkk. [35]	2025	Analisis sentimen dan topik pemerintahan Prabowo–Gibran	IndoBERT + LDA	Artikel berita daring (Antara, Detik, Tempo)	Sentimen negatif 80%; kombinasi memberi gambaran opini lebih lengkap
Nur dkk. [36]	2025	Analisis sentimen dan topik opini konsumen merek teknologi	*Fine-tuning* IndoBERT + BERTopic	Platform X	Akurasi 79,8%; 16 klaster topik

2.2. Dasar Teori

2.2.1. Media Sosial

Media sosial merupakan platform berbasis internet yang memungkinkan pengguna membuat, berbagi, dan berinteraksi melalui konten digital [2], [5]. Berbeda dengan media tradisional yang bersifat satu arah, media sosial memungkinkan komunikasi dua arah sehingga setiap individu dapat sekaligus berperan sebagai konsumen dan produsen informasi [2], [5]. Dalam penelitian ini, platform yang digunakan sebagai sumber data adalah YouTube dan Instagram, dua platform yang banyak digunakan masyarakat Indonesia [4]. Keduanya menyediakan fitur kolom komentar sebagai ruang diskusi publik yang aktif, sehingga komentar penggunanya dapat dianalisis untuk memahami dinamika opini publik terhadap suatu kebijakan [34].

2.2.2. Kebijakan Pembatasan Usia Pengguna Media Sosial

Kebijakan pembatasan usia pengguna media sosial merupakan regulasi yang ditetapkan oleh pemerintah untuk melindungi anak dan remaja dari dampak negatif penggunaan media sosial [21]. Tren regulasi ini telah berkembang secara global. Beberapa negara, seperti Australia, bahkan telah lebih dahulu menerapkan kebijakan serupa dengan melarang pengguna di bawah 16 tahun mengakses platform media sosial [33]. Di Indonesia, kebijakan pembatasan usia pengguna media sosial dituangkan dalam Peraturan Pemerintah Nomor 17 Tahun 2025 tentang Tata Kelola Penyelenggaraan Sistem Elektronik dalam Pelindungan Anak atau yang dikenal sebagai PP Tunas [17], [18]. Regulasi ini mewajibkan Penyelenggara Sistem Elektronik untuk menyaring konten berbahaya, menyediakan mekanisme pelaporan, menerapkan verifikasi usia pengguna, serta melarang *profiling* data anak untuk kepentingan komersial [17], [18]. Sebagai implementasinya, Kementerian Komunikasi dan Digital secara resmi memberlakukan kebijakan pembatasan usia minimum 16 tahun bagi pengguna media sosial terhitung sejak 28 Maret 2026 [20].

Sebagai aturan pelaksana dari PP Tunas, Kementerian Komunikasi dan Digital menerbitkan Peraturan Menteri Komunikasi dan Digital Nomor 9 Tahun 2026, yang mengatur mekanisme teknis pembatasan usia secara lebih rinci [19]. Peraturan ini mewajibkan setiap Penyelenggara Sistem Elektronik untuk mencantumkan batasan usia minimum penggunaan layanannya, yang dikelompokkan ke dalam lima kategori rentang

usia, yaitu 3–5 tahun, 6–9 tahun, 10–12 tahun, 13–15 tahun, dan 16 tahun hingga belum berusia 18 tahun [19]. Selain itu, setiap platform diwajibkan melakukan penilaian mandiri untuk menentukan tingkat risiko layanannya terhadap anak, yang diklasifikasikan ke dalam dua kategori, yaitu profil risiko tinggi dan profil risiko rendah, berdasarkan aspek seperti potensi kontak dengan orang tidak dikenal, paparan konten berbahaya, eksploitasi anak sebagai konsumen, keamanan data pribadi, potensi adiksi, serta gangguan psikologis dan fisiologis anak [19]. Secara khusus, layanan jejaring dan media sosial dikategorikan sebagai layanan dengan profil risiko tinggi, kecuali ditentukan lain berdasarkan hasil penilaian mandiri, sehingga penyelenggara layanan tersebut diwajibkan menonaktifkan akun pengguna anak yang berusia di bawah 16 tahun [19]. Ketentuan inilah yang menjadi dasar hukum langsung bagi kebijakan pembatasan usia minimum 16 tahun bagi pengguna media sosial di Indonesia.

2.2.3. Opini Publik

Opini publik merupakan kumpulan pandangan, sikap, dan penilaian yang disampaikan oleh masyarakat terhadap suatu isu atau kebijakan yang dianggap memiliki kepentingan bersama [24]. Opini ini memiliki peran penting dalam evaluasi kebijakan karena mencerminkan tingkat penerimaan serta potensi permasalahan yang mungkin muncul dalam implementasinya [24]. Perkembangan media sosial telah mengubah cara masyarakat mengungkapkan opini, menjadikan platform digital sebagai ruang utama bagi publik untuk menyampaikan pandangan mereka secara langsung dan terbuka [23], [24]. Komentar pengguna di platform seperti YouTube dan Instagram merupakan salah satu bentuk ekspresi opini publik yang mencerminkan pandangan dan kekhawatiran masyarakat terhadap suatu kebijakan, sehingga dapat dianalisis untuk memahami dinamika respons publik secara mendalam [25].

2.2.4. Analisis Sentimen

Analisis sentimen merupakan bidang penelitian dalam pemrosesan bahasa alami (*Natural Language Processing*/NLP) yang bertujuan untuk mengidentifikasi, mengekstraksi, dan mengklasifikasikan opini atau perasaan yang terkandung dalam suatu teks [29], [31]. Klasifikasi sentimen umumnya dibagi ke dalam tiga kategori: positif untuk teks yang mengandung dukungan atau kepuasan, negatif untuk teks yang mengandung penolakan atau ketidakpuasan, dan netral untuk teks yang tidak menunjukkan kecenderungan emosi tertentu [29], [31]. Pendekatan analisis sentimen

telah berkembang dari metode berbasis leksikon yang mengandalkan kamus kata berbobot sentimen, kemudian metode *machine learning* klasik seperti Naïve Bayes dan Support Vector Machine, hingga pendekatan *deep learning* berbasis *transformer* yang mampu memahami konteks kalimat secara lebih mendalam [34], [36]. Dalam konteks teks media sosial berbahasa Indonesia, analisis sentimen menghadapi tantangan tersendiri akibat penggunaan bahasa informal, singkatan, *slang*, dan percampuran bahasa yang umum ditemukan dalam komentar pengguna, sehingga diperlukan model yang secara khusus dilatih menggunakan korpus bahasa Indonesia untuk menghasilkan klasifikasi yang akurat [29], [34].

2.2.5. Prapemrosesan Teks

Prapemrosesan teks merupakan tahapan awal dalam pengolahan teks yang bertujuan untuk membersihkan dan mengubah data teks mentah menjadi format yang terstruktur dan siap diproses oleh model *machine learning* [29]. Tahapan ini umumnya meliputi beberapa proses yang diterapkan secara berurutan. Pertama, *case folding* mengubah seluruh huruf dalam teks menjadi huruf kecil untuk menyeragamkan format penulisan [29], [31]. Kedua, *cleansing* menghapus elemen yang tidak relevan seperti URL, *mention*, *hashtag*, angka, simbol, dan karakter khusus yang tidak memiliki makna dalam analisis [31]. Ketiga, normalisasi mengubah kata-kata tidak baku, singkatan, dan *slang* menjadi bentuk standarnya menggunakan kamus normalisasi bahasa Indonesia [27], [31]. Keempat, *stopword removal* menghapus kata-kata umum yang tidak memiliki makna signifikan dalam analisis seperti kata ”dan”, “atau”, dan “yang” [29]. Terakhir, tokenisasi memecah teks menjadi unit kata atau token yang lebih kecil untuk memudahkan pemrosesan selanjutnya [31]. Penerapan tahapan prapemrosesan teks yang tepat sangat penting untuk meningkatkan kualitas data dan akurasi model, khususnya pada teks media sosial berbahasa Indonesia yang kaya akan variasi bahasa informal [27], [32].

2.2.6. BERT

BERT (Bidirectional Encoder Representations from Transformers) merupakan model bahasa berbasis arsitektur *transformer* yang dikembangkan oleh Google dan diperkenalkan oleh Devlin dkk. pada tahun 2019 [37]. Berbeda dengan model bahasa sebelumnya yang memproses teks secara satu arah, BERT menggunakan pendekatan dua arah (*bidirectional*) yang memungkinkan model memahami konteks sebuah kata

berdasarkan seluruh kata yang ada di sekitarnya secara bersamaan [37]. BERT dilatih menggunakan dua tugas utama: Masked Language Modeling (MLM) untuk memprediksi kata yang disembunyikan secara acak dalam sebuah kalimat, dan Next Sentence Prediction (NSP) untuk memahami hubungan logis antara dua kalimat [37]. Model ini tersedia dalam dua konfigurasi: BERT-Base dengan 12 lapisan *transformer*, 768 *hidden size*, dan 12 *attention heads*, serta BERT-Large dengan 24 lapisan *transformer*, 1024 *hidden size*, dan 16 *attention heads* [37]. Kemampuan dua arah ini menjadikan BERT unggul dalam berbagai tugas NLP termasuk klasifikasi teks dan analisis sentimen, yang dibuktikan dengan pencapaian *state-of-the-art* pada *benchmark* GLUE [37].

2.2.7. IndoBERT

IndoBERT merupakan adaptasi dari arsitektur BERT yang dikembangkan secara khusus untuk memproses teks berbahasa Indonesia [38]. Model ini dilatih menggunakan korpus Indo4B yang terdiri atas sekitar 4 miliar kata, dikumpulkan dari berbagai sumber teks bahasa Indonesia termasuk media sosial, berita daring, Wikipedia, dan berbagai korpus web lainnya, sehingga mencakup ragam bahasa baik formal maupun informal [38]. IndoBERT tersedia dalam dua varian: IndoBERT-base dengan 12 lapisan *transformer*, 768 *hidden size*, dan 12 *attention heads*, serta IndoBERT-large dengan 24 lapisan *transformer*, 1024 *hidden size*, dan 16 *attention heads* [38]. Dalam klasifikasi sentimen, IndoBERT menggunakan proses *fine-tuning*, yaitu penyesuaian bobot model yang telah dilatih sebelumnya menggunakan *dataset* berlabel sesuai tugas target, sehingga model dapat menghasilkan klasifikasi yang akurat pada domain dan konteks tertentu [27], [34]. IndoBERT terbukti mengungguli model *multilingual* seperti mBERT dalam berbagai tugas NLP berbahasa Indonesia termasuk analisis sentimen, karena pelatihan dengan korpus khusus bahasa Indonesia menghasilkan representasi linguistik yang lebih sesuai dengan karakteristik bahasa tersebut [38].

Pemilihan IndoBERT dalam penelitian ini didasarkan pada beberapa pertimbangan. Pertama, IndoBERT dilatih khusus pada korpus berbahasa Indonesia berskala besar yang mencakup ragam bahasa formal maupun informal [38], sehingga lebih sesuai untuk memproses teks berbahasa Indonesia dibandingkan model multilingual yang dilatih pada campuran banyak bahasa sekaligus dengan proporsi data Bahasa Indonesia yang relatif kecil. Kedua, karakteristik ini menjadikan IndoBERT lebih mampu memahami bahasa informal, singkatan, dan *slang* yang umum dijumpai pada

komentar media sosial, yang menjadi jenis data utama dalam penelitian ini. Ketiga keunggulan IndoBERT dibandingkan model lain juga telah dibuktikan pada penelitian sejenis. Setiawan dkk. [34] menunjukkan bahwa IndoBERT mencapai akurasi 95 persen serta mengungguli BERT, SVM, dan Random Forest dalam analisis sentimen komentar berbahasa Indonesia. Berdasarkan pertimbangan tersebut, IndoBERT dipilih sebagai model analisis sentimen pada penelitian ini.

2.2.8. Pemodelan Topik

Pemodelan topik (*topic modeling*) merupakan metode pembelajaran mesin tanpa pengawasan (*unsupervised learning*) yang digunakan untuk menemukan topik tersembunyi dalam sekumpulan dokumen teks tanpa memerlukan pelabelan data sebelumnya [31]. Pemodelan topik bekerja dengan cara menganalisis pola kemunculan kata dalam dokumen untuk mengelompokkan kata-kata yang sering muncul bersama ke dalam klaster topik yang merepresentasikan tema tertentu [34], [36]. Salah satu metode pemodelan topik yang paling banyak digunakan adalah Latent Dirichlet Allocation (LDA), yang mengasumsikan bahwa setiap dokumen merupakan campuran dari beberapa topik dan setiap topik merupakan distribusi probabilitas atas kata-kata dalam kosakata [35]. Meskipun LDA telah banyak digunakan dalam berbagai penelitian, metode ini memiliki keterbatasan dalam menangkap makna semantik kata karena hanya mengandalkan frekuensi kemunculan kata tanpa mempertimbangkan konteks kalimat [31], [35]. Perkembangan model berbasis *transformer* kemudian mendorong munculnya pendekatan pemodelan topik yang lebih mutakhir seperti BERTopic, yang memanfaatkan representasi semantik dari model bahasa untuk menghasilkan topik yang lebih koheren dan kontekstual dibandingkan dengan LDA [34], [36].

2.2.9. BERTopic

BERTopic merupakan metode pemodelan topik berbasis *transformer* yang dikembangkan oleh Grootendorst pada tahun 2022 dan dirancang untuk menghasilkan topik yang lebih koheren dan bermakna secara semantik dibandingkan metode konvensional seperti LDA [39]. BERTopic bekerja melalui tiga tahapan utama. Pertama, setiap dokumen diubah menjadi representasi vektor menggunakan model *embedding* berbasis *transformer* melalui *framework* Sentence-BERT [39]. Pendekatan ini mampu menangkap makna semantik teks secara kontekstual sehingga dokumen dengan topik yang serupa akan memiliki representasi vektor yang berdekatan dalam ruang vektor

[39]. Kedua, dimensi vektor yang dihasilkan direduksi menggunakan metode Uniform Manifold Approximation and Projection (UMAP) untuk menyederhanakan representasi dokumen tanpa kehilangan informasi semantik yang penting [39]. UMAP dipilih karena terbukti mampu mempertahankan fitur lokal dan global dari data berdimensi tinggi dengan lebih baik dibandingkan metode lain seperti PCA dan t-SNE [39]. Ketiga, dokumen yang telah direduksi dimensinya dikelompokkan menggunakan algoritma Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), yaitu perluasan dari DBSCAN yang mampu mendeteksi klaster dalam berbagai bentuk dan ukuran secara otomatis serta memodelkan dokumen yang tidak relevan sebagai *outlier* [39]. Setelah klaster terbentuk, BERTopic menggunakan metode class-based Term Frequency-Inverse Document Frequency (c-TF-IDF) untuk mengekstraksi kata-kata kunci yang paling representatif dari setiap topik dengan cara memperlakukan seluruh dokumen dalam satu klaster sebagai satu dokumen tunggal, sehingga pentingnya sebuah kata dapat diukur berdasarkan kontribusinya terhadap topik, bukan terhadap dokumen individual [39]. Kelebihan utama BERTopic dibandingkan LDA adalah kemampuannya dalam memahami makna semantik teks secara kontekstual melalui representasi *transformer*, menghasilkan topik yang lebih koheren, serta tidak memerlukan penentuan jumlah topik secara manual sebelum proses pengelompokan dilakukan [34], [36].

Pemilihan BERTopic dalam penelitian ini didasarkan pada beberapa pertimbangan. Pertama, berbeda dengan metode pemodelan topik konvensional seperti LDA yang mengandalkan frekuensi kemunculan kata tanpa mempertimbangkan konteks kalimat, BERTopic memanfaatkan representasi semantic dari model *transformer* sehingga mampu menangkap makna kontekstual pada teks pendek dan informal seperti komentar media sosial [39]. Kedua, BERTopic tidak mensyaratkan penentuan jumlah topik secara manual di awal proses, sehingga sesuai untuk penelitian ini yang belum memiliki asumsi awal mengenai jumlah tema yang mungkin muncul dari komentar publik terhadap kebijakan. Ketiga, pendekatan gabungan IndoBERT dan BERTopic telah terbukti efektif pada penelitian sejenis. Khaqqi dkk. [32] dan Nur dkk. [36] berhasil menerapkan kombinasi kedua metode tersebut untuk analisis sentimen dan pemodelan topik pada data media sosial berbahasa Indonesia, sehingga mampu menghasilkan informasi yang lebih komprehensif mengenai persepsi publik.

Berdasarkan pertimbangan tersebut, BERTopic dipilih sebagai metode pemodelan topik pada penelitian ini.