

BAB 1

PENDAHULUAN

1.1. Latar Belakang

Media sosial mengalami perkembangan dengan menjadi salah satu ruang publik digital yang berperan penting dalam membentuk opini masyarakat. Di Indonesia, platform X (sebelumnya Twitter) menjadi salah satu media sosial paling aktif dengan karakteristik utama berupa cuitan singkat (Fazri & Voutama, 2025). Setiap individu dapat mengekspresikan pendapat, kritik, ataupun dukungan terhadap isu tertentu dalam waktu yang sangat cepat. Hal ini menjadikan X sebagai sumber data penting dalam berbagai penelitian, khususnya yang berkaitan dengan analisis sentimen publik (Gregorius et al., 2025). Salah satu isu kebijakan publik yang memicu reaksi masyarakat adalah pembentukan Badan Pengelola Investasi Daya Anagata Nusantara (Danantara). Kebijakan ini bertujuan untuk mengelola investasi strategis nasional dan mendorong pembangunan ekonomi jangka panjang. Namun, seperti kebijakan pemerintah pada umumnya, kehadiran Danantara menuai respons yang beragam. Sebagian masyarakat menilai kebijakan ini sebagai langkah strategis, sementara sebagian lain menilainya penuh risiko, kurang transparan, atau tidak berpihak pada kepentingan publik (Siregar et al., 2025). Berdasarkan laporan IDN Times (2025), peluncuran Danantara juga menimbulkan pro dan kontra di kalangan masyarakat dan ekonom. Ekonom dari Universitas Gadjah Mada menilai bahwa keberadaan lembaga ini perlu diiringi dengan transparansi dan pengawasan ketat agar pengelolaan investasi publik dapat berjalan akuntabel dan tidak menimbulkan ketidakpercayaan masyarakat (Sintasari, 2025). Hal ini menunjukkan bahwa persepsi publik terhadap kebijakan pemerintah perlu dipahami secara komprehensif melalui analisis sentimen.

Analisis sentimen merupakan pendekatan dalam bidang *Natural Language Processing (NLP)* yang berfungsi untuk mengidentifikasi dan mengklasifikasikan opini publik menjadi kategori positif dan negatif. Pendekatan ini memungkinkan peneliti memahami persepsi masyarakat terhadap suatu isu atau kebijakan secara objektif berdasarkan data teks dari media sosial. Dalam konteks kebijakan publik, analisis sentimen membantu menggambarkan tingkat dukungan dan penolakan

masyarakat terhadap kebijakan yang diterapkan pemerintah. Melalui penerapan metode berbasis *machine learning* dan *deep learning*, analisis sentimen mampu mengubah opini yang bersifat kualitatif menjadi informasi kuantitatif yang dapat digunakan untuk mendukung pengambilan keputusan dan evaluasi kebijakan publik (Ningsih et al., 2024).

Dalam penerapan analisis sentimen, terdapat beberapa aspek penting yang harus dipersiapkan agar hasil analisis akurat dan representatif. Pertama, tahap pengumpulan data (*data collection*) dari media sosial seperti X (Twitter) perlu dilakukan secara sistematis menggunakan teknik *web crawling* atau *API* agar data yang diperoleh relevan dengan topik penelitian. Kedua, data yang diperoleh umumnya masih mentah dan perlu melalui proses *text preprocessing* seperti *cleaning*, *case folding*, dan *normalization* untuk menghilangkan noise dan menstandarkan teks (Siswoyo et al., 2025). Ketiga, diperlukan proses *labelling* atau pelabelan data untuk menentukan polaritas sentimen pada setiap teks, pada penelitian ini pelabelan dilakukan secara manual. Keempat, peneliti harus menyiapkan strategi penanganan ketidakseimbangan data (*imbalanced dataset*), pada penelitian ini menggunakan SMOTE agar model tidak bias terhadap kelas mayoritas (Ahmad et al., 2024). Terakhir, pemilihan model yang sesuai misalnya IndoBERT menjadi langkah krusial karena model berbasis *Transformer* memiliki kemampuan memahami konteks kalimat secara dua arah (*bidirectional*), sehingga mampu menghasilkan klasifikasi sentimen yang lebih akurat dan kontekstual (A'la, 2025).

Objek yang diteliti pada penelitian ini yaitu opini publik di media sosial X mengenai kebijakan Danantara. Cara kerja objek tersebut dapat dijelaskan dari komunikasi publik pada media sosial X, di mana pengguna menulis cuitan, kemudian cuitan tersebut tersebar dalam jaringan sosial, dan setiap cuitan dapat dianalisis untuk melihat sentimen dari pernyataan tersebut. Tujuan dari analisis sentimen adalah untuk memetakan persepsi publik dari teks atau opini yang bersifat positif, dan negatif (Ningsih et al., 2024).

Saat ini media sosial menyediakan data yang melimpah, namun terdapat sejumlah masalah dalam alur pengolahan data opini publik. Banyaknya jumlah cuitan menyulitkan analisis manual karena memerlukan waktu dan tenaga besar.

Cuitan di platform X sering kali memiliki variasi bahasa seperti singkatan, campuran bahasa asing, bahasa gaul, serta emotikon yang sulit diproses secara otomatis. Panjang teks yang terbatas hingga 280 karakter kerap menghasilkan kalimat yang tidak lengkap dan dapat memengaruhi maknanya. Beberapa kata memiliki makna ganda tergantung konteks kalimat. Misalnya, kata “bisa” dapat berarti “dapat” atau “racun”. Data yang diperoleh sering kali mengandung noise seperti cuitan yang tidak relevan, mengandung spam, atau tidak memuat opini yang jelas. Metode analisis sentimen konvensional seperti Naïve Bayes masih memiliki keterbatasan karena kesulitan menangkap hubungan kompleks antar kata, sehingga hasil analisis sering kali kurang akurat (Siswoyo et al., 2025).

Keterbatasan metode analisis sentimen konvensional dalam memahami konteks bahasa di media sosial X sering kali menjadi masalah. Hal ini penting karena konteks kalimat sering menentukan makna sesungguhnya dari opini masyarakat. Sebagai contoh, kalimat “Danantara tidak buruk juga” dan “Danantara buruk sekali” memiliki kata kunci yang sama, yaitu “buruk”. Namun, kalimat pertama seharusnya positif, sedangkan kalimat kedua jelas negatif. Metode berbasis TF-IDF atau Word2Vec mungkin memberi bobot besar pada kata “buruk”, sehingga kedua kalimat tersebut diprediksi negatif. Padahal, kalimat pertama mengandung sentimen positif. Masalah ini muncul karena metode konvensional tidak memperhatikan hubungan antar kata dalam kalimat (Ahmad et al., 2024). Oleh karena itu, diperlukan metode yang mampu menangkap konteks, struktur kalimat, dan nuansa bahasa agar hasil analisis sentimen menjadi lebih akurat dalam kerangka kerja yang sistematis.

Selain tantangan dalam memahami konteks bahasa, permasalahan lain yang sering muncul pada analisis sentimen adalah ketidakseimbangan jumlah data antar kelas sentimen. Misalnya, jumlah cuitan bernuansa positif terhadap kebijakan Danantara bisa jauh lebih banyak dibandingkan cuitan positif atau negatif. Ketidakseimbangan ini dapat menurunkan performa model karena algoritma cenderung bias terhadap kelas mayoritas. Untuk mengatasi masalah tersebut, dapat menggunakan metode *oversampling* SMOTE (*Synthetic Minority Oversampling Technique*). Penggunaan SMOTE dikarenakan teknik ini tidak hanya menduplikasi data minoritas, tetapi menghasilkan sampel sintetis baru berdasarkan kedekatan

antar data sehingga distribusi kelas menjadi lebih seimbang. Keunggulan SMOTE terletak pada kemampuannya mempertahankan informasi pada kelas mayoritas tanpa penghapusan data, meningkatkan representasi kelas minoritas, serta membuat model lebih sensitif terhadap pola yang jarang muncul. Berbagai penelitian menunjukkan bahwa SMOTE mampu meningkatkan stabilitas, sensitivitas, dan akurasi model pada dataset yang sangat tidak seimbang. Dengan demikian, penggunaan SMOTE diharapkan mampu mengurangi bias model dan menghasilkan prediksi sentimen yang lebih adil dan akurat (A'la, 2025).

IndoBERT adalah model pra-latih (pre-trained) yang mengadopsi arsitektur BERT (*Bidirectional Encoder Representations from Transformers*) dan dirancang khusus untuk pemrosesan Bahasa Indonesia. Model ini menggunakan mekanisme self-attention untuk menangkap keterkaitan antar kata secara dua arah, baik dari konteks sebelah kiri maupun kanan, sehingga representasi makna kata yang dihasilkan menjadi lebih tepat sesuai konteks kalimat (Pratama Putra, 2025). IndoBERT dilatih pada korpus berbahasa Indonesia dalam skala besar, seperti Wikipedia, Kompas, dan Tempo, yang membuat model memiliki pemahaman kontekstual yang kuat terhadap pola dan struktur bahasa Indonesia (Jayadianti et al., 2022). Sejumlah penelitian terdahulu juga melaporkan bahwa IndoBERT cenderung menghasilkan performa yang lebih baik dibandingkan metode klasik seperti SVM dan Naïve Bayes dalam tugas analisis sentimen (Alfatah, 2024).

Model IndoBERT memiliki performa yang cukup kuat walaupun tanpa proses balancing data secara khusus. Hal ini dibuktikan dalam jurnal yang dikaji, di mana distribusi kelas sentimen sudah berada pada proporsi yang relatif seimbang dan model tetap mampu mencapai akurasi sebesar 95% dengan nilai precision, recall, serta F1-score di atas 90% (Setiawan et al., 2025). Capaian tersebut menunjukkan bahwa kemampuan representasi kontekstual IndoBERT yang berbasis arsitektur transformer sudah memadai dalam menangkap pola sentimen secara optimal, sehingga kinerja klasifikasi tetap stabil dan tinggi meskipun tidak disertai perlakuan penyeimbangan data tambahan.

Pada penelitian ini, IndoBERT digunakan sebagai model utama untuk mengklasifikasikan sentimen publik terhadap kebijakan Danantara, memanfaatkan kemampuan representasi kontekstual bahasa Indonesia yang menjadi keunggulan

utama arsitektur Transformer pre-trained. Sebagai analisis, diterapkan SMOTE pada data latih untuk menguji sejauh mana penanganan ketidakseimbangan distribusi kelas dapat mempengaruhi performa model pada berbagai kondisi rasio data. Evaluasi dilakukan menggunakan metrik *accuracy*, *precision*, *recall*, dan *F1-score* untuk mengukur performa secara komprehensif (Randy Suryono, 2025).

Urgensi penelitian ini terletak pada kebutuhan memahami persepsi publik terhadap kebijakan Danantara melalui analisis sentimen yang akurat di platform X. Tantangan utama pada data media sosial adalah variasi bahasa yang tidak baku serta ketidakseimbangan jumlah data antar kelas sentimen yang dapat memicu bias model pada kelas mayoritas. Karena itu, penelitian ini memanfaatkan IndoBERT untuk klasifikasi sentimen biner (positif dan negatif) serta menguji dampak penerapan SMOTE pada berbagai tingkat ketidakseimbangan data latih melalui beberapa skenario rasio distribusi kelas. Dengan rancangan skenario yang terkontrol, penelitian ini memberikan bukti empiris kapan penerapan SMOTE membantu dan kapan efeknya tidak konsisten terhadap kemampuan generalisasi model.

1.2. Perumusan Masalah

1.2.1. Permasalahan

Analisis sentimen terhadap kebijakan Badan Pengelola Investasi Daya Anagata Nusantara (Danantara) menjadi penting untuk mengetahui bagaimana persepsi publik terbentuk di media sosial, khususnya pada platform X (Twitter). Media sosial menjadi ruang ekspresi masyarakat dalam menyampaikan opini, baik berupa dukungan maupun kritik terhadap kebijakan pemerintah. Melalui analisis sentimen, opini-opini tersebut dapat diolah menjadi informasi yang menggambarkan tingkat penerimaan masyarakat terhadap kebijakan Danantara.

Namun, dalam praktiknya, penelitian analisis sentimen terhadap kebijakan Danantara menghadapi dua permasalahan utama. Pertama, ketidakseimbangan jumlah data antar kelas sentimen (positif, dan negatif) yang sering kali membuat model cenderung bias terhadap kelas mayoritas. Akibatnya, hasil klasifikasi menjadi kurang representatif karena model lebih sering memprediksi kelas yang dominan, seperti sentimen positif, dan mengabaikan kelas minoritas (Ningsih et al., 2024). Kedua, pemilihan metode penyeimbangan data yang tepat menjadi faktor

penting untuk meningkatkan kinerja model analisis sentimen. Pada penelitian ini menggunakan *Synthetic Minority Oversampling Technique* (SMOTE) yang menambah data sintesis pada kelas minoritas.

1.2.2. Metode Usulan

Penelitian ini menerapkan IndoBERT sebagai pendekatan utama untuk melakukan klasifikasi sentimen biner (positif/negatif) pada cuitan berbahasa Indonesia. Kemampuan IndoBERT dalam menangkap konteks semantik bahasa Indonesia melalui mekanisme *self-attention* menjadikannya model yang kuat bahkan pada kondisi data yang tidak seimbang. Sebagai analisis tambahan, digunakan SMOTE pada data latih untuk menguji sejauh mana augmentasi berbasis TF-IDF dapat berkontribusi pada kemampuan IndoBERT mengenali kelas minoritas di berbagai kondisi distribusi, mulai dari *imbalance* ringan hingga ekstrem.

Evaluasi dilakukan tidak hanya pada akurasi keseluruhan, tetapi terutama pada *F1-score* agar penilaian performa mencerminkan kemampuan model secara proporsional di kedua kelas sentimen. Penelitian ini bertujuan mengidentifikasi kondisi distribusi data di mana IndoBERT sudah cukup kuat secara mandiri, serta kondisi ketidakseimbangan ekstrem di mana augmentasi SMOTE berpotensi memberikan manfaat tambahan. Dengan pendekatan evaluasi berbasis skenario ini, hasil penelitian diharapkan memberikan panduan praktis mengenai kapan SMOTE relevan untuk diterapkan bersama model IndoBERT.

1.2.3. Pertanyaan Penelitian

Berdasarkan latar belakang dan permasalahan yang telah dijelaskan, maka pertanyaan penelitian dalam studi ini dirumuskan sebagai berikut:

1. Seberapa akurat model IndoBERT dalam mengklasifikasikan sentimen publik terhadap kebijakan Danantara di media sosial X?
2. Bagaimana pengaruh SMOTE terhadap performa model IndoBERT dalam mengatasi ketidakseimbangan data pada analisis sentimen publik mengenai kebijakan Danantara?

1.3. Tujuan dan Manfaat Penelitian

1.3.1. Tujuan Penelitian

Tujuan dari penelitian ini adalah untuk:

1. Menghasilkan model analisis sentimen berbasis IndoBERT yang mampu mengklasifikasikan opini publik terhadap kebijakan Danantara di media sosial X secara lebih akurat, kontekstual, dan representatif.
2. Mengevaluasi pengaruh penerapan metode SMOTE terhadap performa model IndoBERT dalam menangani ketidakseimbangan data pada analisis sentimen publik mengenai kebijakan Danantara.

1.3.2. Manfaat Penelitian

Penelitian ini diharapkan memberikan manfaat sebagai berikut:

1. Manfaat Akademis

Penelitian ini dapat menjadi kontribusi dalam pengembangan ilmu di bidang *Natural Language Processing* (NLP), khususnya analisis sentimen berbahasa Indonesia. Dengan mengkaji penerapan IndoBERT, penelitian ini dapat menunjukkan bagaimana model berbasis Transformer mampu mengatasi keterbatasan metode konvensional yang tidak kontekstual. Selain itu, mengevaluasi efektivitas SMOTE untuk mengatasi ketidakseimbangan data.

2. Manfaat Praktis

Penelitian ini dapat memberikan gambaran empiris mengenai persepsi publik terhadap kebijakan Danantara di media sosial X. Hasil penelitian diharapkan dapat dijadikan masukan bagi pemerintah maupun pemangku kebijakan dalam mengevaluasi kebijakan serta menyusun strategi komunikasi publik yang lebih efektif.

1.4. Batasan Penelitian

Supaya penelitian ini terarah dan fokus, ruang lingkup penelitian ini dibatasi pada hal-hal berikut:

1. Objek penelitian dibatasi pada analisis sentimen publik terhadap kebijakan Badan Pengelola Investasi Daya Anagata Nusantara (Danantara) yang dibahas pada platform media sosial X (Twitter).
2. Data penelitian yang digunakan hanya berupa teks cuitan yang berbahasa Indonesia. Cuitan yang menggunakan bahasa asing, campuran, atau mengandung simbol/emoji yang tidak relevan tidak menjadi fokus utama.
3. Periode pengambilan data dibatasi pada rentang waktu Mei 2025 sampai Agustus 2025.

4. Kategorisasi sentimen dibatasi menjadi dua kelas, yaitu positif dan negatif.

Penelitian ini tidak menggunakan data dari platform media sosial lain seperti Instagram atau TikTok. Alasan utamanya adalah karena kedua platform tersebut lebih berfokus pada konten visual berupa gambar dan video, sehingga data teks yang tersedia relatif terbatas dan kurang sesuai untuk analisis sentimen berbasis teks. Komentar pengguna pada kedua platform tersebut juga cenderung bersifat singkat, ambigu, serta banyak mengandung emoji atau bahasa nonformal yang sulit diolah menggunakan pendekatan *Natural Language Processing* (NLP).

1.5. Metodologi Penulisan Skripsi

Struktur penulisan dalam skripsi ini disusun berdasarkan format Skripsi Program Studi Sistem Informasi Fakultas Teknik Universitas Trunodjoyo Madura, dengan rincian sebagai berikut:

BAB I. PENDAHULUAN

Membahas mengenai latar belakang permasalahan, perumusan masalah, tujuan dilakukannya penelitian, manfaat yang diharapkan dari penelitian, ruang lingkup masalah, metode penelitian yang digunakan, serta susunan penulisan secara sistematis.

BAB II. TINJAUAN PUSTAKA

Bab ini menyajikan kajian teori dan penelitian terdahulu yang berkaitan dengan penelitian, meliputi konsep analisis sentimen, *text preprocessing*, metode penyeimbang data, serta model *pre-trained language model* khususnya IndoBERT. Selain itu, bab ini juga membahas penelitian relevan sebelumnya mengenai analisis sentimen di media sosial dan identifikasi gap penelitian yang menjadi dasar penelitian ini.

BAB III. METODOLOGI PENELITIAN

Bab ini menjelaskan metodologi dan tahapan penelitian yang dilakukan, meliputi proses pengumpulan data (*crawling data* dari media sosial X), tahapan *text preprocessing*, proses pelabelan data, penggunaan model IndoBERT untuk klasifikasi sentimen, serta skenario pengujian performa model. Bab ini juga memuat rancangan alur penelitian secara menyeluruh.

BAB IV. HASIL DAN PEMBAHASAN

Bab ini menyajikan hasil pelaksanaan seluruh tahapan penelitian secara rinci, mulai dari proses pengumpulan data, pelabelan, text preprocessing, tokenisasi, balancing data, hingga pemodelan dan evaluasi. Pembahasan mencakup perbandingan performa 10 skenario pengujian menggunakan metrik akurasi, presisi, *recall*, dan F1-score, disertai analisis pengaruh penerapan SMOTE pada setiap kondisi ketidakseimbangan kelas. Selain itu, bab ini juga membahas hasil deployment model terbaik pada aplikasi web berbasis Streamlit sebagai implementasi praktis dari model yang telah dibangun.

BAB V. PENUTUP

Bab ini menyajikan kesimpulan dari seluruh rangkaian penelitian yang telah dilakukan, mencakup temuan utama mengenai performa model IndoBERT dan pengaruh SMOTE terhadap berbagai skenario ketidakseimbangan data. Kesimpulan disusun untuk menjawab pertanyaan penelitian yang telah dirumuskan pada BAB I. Selain itu, bab ini juga menyampaikan saran untuk pengembangan penelitian lebih lanjut guna mengatasi keterbatasan yang ada pada penelitian ini.