BAB III
METODOLOGI PENELITIAN

3.1 Jenis Penelitian

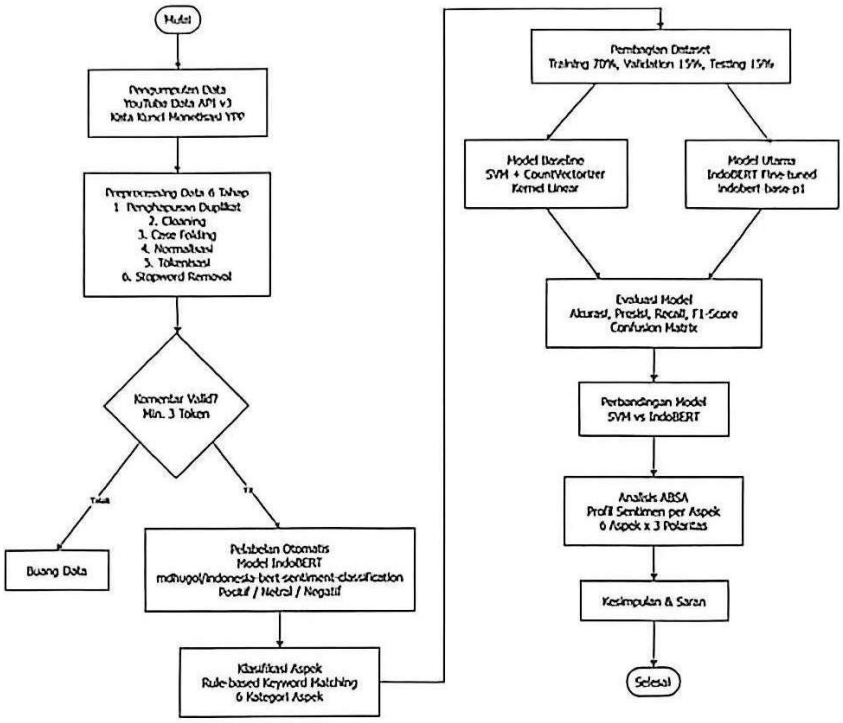
Penelitian ini menggunakan pendekatan kuantitatif dengan desain eksperimental komputasional, di mana performa dua model klasifikasi IndoBERT dan SVM diukur secara objektif menggunakan metrik evaluasi yang terstandarisasi (akurasi, presisi, recall, dan F1-Score weighted average) dan dibandingkan hasilnya. Pendekatan ini dipilih karena selaras dengan tujuan penelitian yang bersifat terukur dan dapat direproduksi. Data yang digunakan berupa komentar YouTube yang dikumpulkan dari video-video yang membahas kebijakan monetisasi YPP.

Seluruh tahapan penelitian diorganisasikan secara sistematis meliputi: (1) Pengumpulan Data menggunakan YouTube Data API v3; (2) Preprocessing Data dalam 6 tahap berurutan; (3) Pelabelan Data secara otomatis berbasis model IndoBERT; (4) Klasifikasi Aspek ke dalam 6 kategori; (5) Pemodelan dengan SVM dan IndoBERT; (6) Evaluasi menggunakan metrik standar NLP; dan (7) Analisis ABSA untuk menghasilkan profil sentimen per aspek. Implementasi dilakukan menggunakan Python 3.12 dengan library utama PyTorch, Hugging Face Transformers, dan Scikit-learn.

3.2 Diagram Alur Penelitian

Diagram alur berikut menggambarkan keseluruhan tahapan penelitian secara sistematis, dari pengumpulan komentar YouTube hingga penarikan kesimpulan berdasarkan hasil evaluasi model:

Dipindai dengan CamScanner



Gambar 3.1 Alur Penelitian

3.3 Pengumpulan Data

Data penelitian berupa komentar publik yang dikumpulkan dari video YouTube yang membahas kebijakan monetisasi YouTube Partner Program menggunakan YouTube Data API v3. Setiap komentar dikumpulkan beserta empat atribut: nama akun penulis (author), isi komentar (comment), jumlah likes, dan waktu publikasi (published_at). Kata kunci pencarian meliputi: "monetisasi YouTube", "syarat YPP", "cara monetisasi channel", "YouTube Partner Program Indonesia", dan "demonetisasi YouTube". Data dikumpulkan dalam rentang waktu Agustus 2024 hingga April 2026 dan disimpan dalam format .xlsx.

Tabel 3. 1 Detail Pengumpulan Data

Keterangan	Detail
Sumber data	Komentar publik video YouTube berbahasa Indonesia
Metode pengumpulan	YouTube Data API v3 (web scraping terprogram)

Dipindai dengan CamScanner

Kata kunci pencarian	"monetisasi YouTube", "syarat YPP", "cara monetisasi channel", dll.
Atribut yang dikumpulkan	author, comment, likes, published_at
Rentang waktu data	Agustus 2024 – April 2026
Format penyimpanan	Microsoft Excel (.xlsx) → dataMonetisasi.xlsx

3.4 Preprocessing Data

Preprocessing data merupakan tahap kritis yang menentukan kualitas representasi teks. Mengingat karakteristik komentar YouTube berbahasa Indonesia yang sangat informal mengandung slang, singkatan, emoji, dan URL preprocessing dilakukan dalam enam tahap berurutan:

1. Penghapusan Duplikat: Komentar yang identik secara teks dihapus untuk mencegah bias dalam distribusi kelas. Contoh: komentar promosi seperti "Gabung ke WA Komunitas Youtube disini 👉 https://..." yang muncul berulang di banyak video dihapus, hanya satu yang dipertahankan.
2. Cleaning (Pembersihan Teks): Menghapus semua elemen non-tekstual menggunakan regular expression, meliputi URL, mention (@username), hashtag (#), emoji, karakter HTML, dan simbol berulang. Contoh: "Mantap..sangat bermanfaat 👍👍👍" → "Mantap sangat bermanfaat".
3. Case Folding: Mengubah seluruh teks menjadi huruf kecil (lowercase). Contoh: "YouTube", "youtube", dan "YOUTUBE" diperlakukan identik.
4. Normalisasi: Mengubah kata tidak baku, singkatan, dan slang ke bentuk baku menggunakan kamus normalisasi Bahasa Indonesia yang diunduh dari repositori GitHub analysisdatasentiment/kamus_kata_baku (file kamuskatabaku.xlsx). Contoh: "yg" → "yang", "jdi" → "jadi", "gmn" → "bagaimana".

Dipindai dengan CamScanner

5. Tokenisasi: Memecah teks menjadi (token) individual menggunakan metode pemisahan berbasis spasi (.split()). contoh: "channel saya sepi penonton" → ["channel", "saya", "sepi", "penonton"].
6. Stopword Removal: Menghapus kata-kata umum yang tidak bermakna menggunakan daftar stopword Bahasa Indonesia dari library NLTK (nltk.corpus.stopwords, bahasa 'indonesian'). Contoh: dari ["channel", "saya", "sepi", "penonton"] → kata "saya" dihapus → ["channel", "sepi", "penonton"]. Komentar yang setelah proses ini memiliki kurang dari 3 token bermakna tidak diikutsertakan dalam analisis.

Tabel 3. 2 Tahapan Preprocessing

Tahap	Proses	Keterangan
1	Penghapusan Duplikat	Menghapus komentar yang identik secara teks
2	Cleaning	Hapus URL, emoji, simbol, karakter tidak relevan
3	Case Folding	Mengubah seluruh teks ke huruf kecil
4	Normalisasi Slang	Mengganti kata tidak baku ke bentuk baku
5	Tokenisasi	Pemecahan token
6	Stopword Removal	Penghapusan stopword

Dipindai dengan CamScanner

3.5 Pelabelan Data

Pelabelan data dilakukan secara otomatis menggunakan model IndoBERT untuk mengklasifikasikan komentar valid ke dalam tiga kategori sentimen. Pendekatan ini dipilih untuk menyeimbangkan akurasi label dengan efisiensi proses pelabelan pada dataset berskala ribuan komentar.

3.5.1 Skema Label Sentimen

Tiga kategori sentimen yang digunakan adalah:

1. Positif: komentar yang mengekspresikan dukungan, apresiasi, semangat, atau kepuasan. Contoh: "Mantap sangat bermanfaat, semangat terus abang!".
2. Negatif: komentar yang mengekspresikan keluhan, kritik, frustrasi, atau ketidakpuasan. Contoh: "Kenapa channel saya sepi penonton ya, sudah lama tapi belum monetisasi".
3. Netral: komentar yang bersifat informatif, bertanya, atau tidak memiliki kecenderungan polaritas. Contoh: "Cara daftar YPP bagaimana ya subscriber saya sudah 1000".

3.5.2 Prosedur Pelabelan Otomatis

Pelabelan dilakukan secara otomatis menggunakan model IndoBERT yang telah dilatih khusus untuk klasifikasi sentimen Bahasa Indonesia, yaitu mdhugol/indonesia-bert-sentiment-classification yang diakses melalui Hugging Face pipeline. Model ini menghasilkan tiga label sentimen: LABEL_0 (Positif) komentar mengekspresikan dukungan, apresiasi atau kepuasan, LABEL_1 (Netral) komentar bersifat informatif atau tidak memiliki polaritas jelas, dan LABEL_2 (Negatif) komentar mengekspresikan keluhan, kritik atau ketidakpuasan. Proses pelabelan dilakukan dalam batch berukuran 32 komentar dengan parameter truncation=True untuk menangani komentar yang melebihi batas panjang token model. Seluruh komentar hasil preprocessing yang valid diproses secara otomatis tanpa verifikasi manual mengingat skala dataset yang besar (Wilie, et al., 2020) (Koto, Rahimi, Lau, & Baldwin, 2020).

Perlu dicatat bahwa model yang digunakan untuk pelabelan (mdhugol/indonesia-bert-sentiment-classification) berasal dari keluarga arsitektur

Dipindai dengan CamScanner

yang sama (BERT) dengan model utama yang dievaluasi dalam penelitian ini (indobert-base-p1). Kesamaan arsitektur ini berpotensi menimbulkan bias, di mana model utama cenderung memiliki keselarasan representasi linguistik dengan model pelabel, sehingga keunggulan performanya terhadap SVM tidak murni mencerminkan kualitas model semata, melainkan juga dipengaruhi oleh kedekatan arsitektural dengan proses pembentukan ground truth. Keterbatasan ini menjadi salah satu catatan validitas penelitian yang dibahas lebih lanjut pada Bab V.

3.6 Klasifikasi Aspek

Klasifikasi aspek dilakukan menggunakan pendekatan berbasis kata kunci (rule-based keyword matching) terhadap enam kategori aspek yang diidentifikasi dari analisis konten aktual komentar dataset. Setiap komentar diklasifikasikan ke dalam satu kategori aspek yang paling dominan berdasarkan kata-kata kunci yang terkandung:

Tabel 3. 3 Kategori Aspek dan Kata Kunci

No	Aspek	Kata Kunci Utama
1	Dukungan Komunitas & Semangat	semangat, sukses, mantap, keren, hadir, menyimak, amin
2	Syarat & Kebijakan Monetisasi	monetisasi, subscriber, jam tayang, ypp, syarat, daftar
3	Apresiasi Konten Edukasi	bermanfaat, ilmu, terima kasih, info, edukasi, tutorial
4	Channel & Performa Konten	channel, konten, video, upload, penonton, views
5	Engagement Penonton	hadir, nonton, like, subscribe, penonton, komen

Dipindai dengan CamScanner

6	Bantuan Teknis & Akun	akun, gmail, masalah, error, bantuan, verifikasi

3.7 Pengembangan Model

Penelitian ini mengembangkan dan membandingkan dua model klasifikasi sentimen dengan pendekatan yang berbeda secara fundamental:

Tabel 3. 4 Perbandingan Konfigurasi Model SVM vs IndoBERT

Dimensi	SVM (Model Baseline)	IndoBERT (Model Utama)
Paradigma	Machine learning konvensional (Generasi 2)	Deep learning berbasis Transformer (Generasi 3)
Representasi Teks	CountVectorizer: vektor frekuensi kata statis	Representasi kontekstual bidireksional per token
Pemahaman Konteks	Tidak ada; kata direpresentasikan secara independen	Ada; makna kata bergantung konteks kalimat penuh
Penanganan Bahasa Informal	Terbatas; slang tidak dikenal dianggap kata asing	Baik; dilatih pada teks informal Bahasa Indonesia
Parameter Utama	Kernel Linear, C=1.0, CountVectorizer	LR=2e-5, batch=8, epoch=3, max token=128, dropout=0.3

3.7.1 Model Baseline: SVM dengan CountVectorizer

Model SVM diimplementasikan menggunakan library Scikit-learn dengan konfigurasi kernel Linear dan parameter C=1.0 (default). Representasi teks menggunakan CountVectorizer yang menghitung frekuensi kemunculan kata (term

Dipindai dengan CamScanner

frequency) tanpa pembobotan IDF, dengan n-gram range (1,2) untuk menangkap frasa dua kata yang sering bermakna sebagai satu unit. SVM berfungsi sebagai tolok ukur bawah yang harus dapat dilampaui oleh IndoBERT untuk memvalidasi keunggulan pendekatan deep learning pada domain ini Faisal et al. (2025).

3.7.2 Model Utama: IndoBERT Fine-tuned

Model IndoBERT diimplementasikan menggunakan Hugging Face Transformers library dengan model dasar indobenchmark/indobert-base-p1. Fine-tuning dilakukan dengan menambahkan lapisan klasifikasi di atas token [CLS]: lapisan Dense dengan aktivasi Softmax untuk 3 kelas (Positif, Netral, Negatif). Konfigurasi hyperparameter: optimizer AdamW, learning rate=2e-5, batch size=8 (per_device_train_batch_size=8, per_device_eval_batch_size=8), panjang sekuens maksimum=128 token (truncation=True), epoch=3, dan weight decay=0,01. Model terbaik dipilih berdasarkan metrik akurasi pada validation set (load_best_model_at_end=True). Proses fine-tuning dilakukan menggunakan GPU pada Google Colab (Wilie, et al., 2020) Devlin et al. (2019).

3.8 Pembagian Dataset

Dataset komentar berlabel dibagi menggunakan dua tahap split dengan rasio 70:15:15 (training:validation:testing). Pembagian dilakukan menggunakan train_test_split dengan random_state=42 untuk memastikan reprodusibilitas. Tahap pertama membagi data menjadi 70% training dan 30% sementara, kemudian tahap kedua membagi data sementara menjadi 15% validation dan 15% testing. Validation set digunakan untuk memantau performa model selama pelatihan IndoBERT dan memilih model terbaik (load_best_model_at_end), sedangkan testing set digunakan untuk evaluasi final kedua model.

Dipindai dengan CamScanner

Tabel 3. 5 Pembagian Dataset Training dan Testing

Subset	Persentase	Fungsi
Training Set	70%	Melatih parameter model SVM dan IndoBERT
Validation Set	15%	Memantau performa IndoBERT selama training dan memilih model terbaik
Testing Set	15%	Evaluasi final performa kedua model

3.9 Metrik Evaluasi

Performa kedua model dievaluasi menggunakan empat metrik standar klasifikasi multiclass: (1) Akurasi proporsi prediksi benar dari seluruh data testing, digunakan sebagai metrik utama perbandingan; (2) Presisi Weighted Average rata-rata presisi per kelas yang dibobot berdasarkan jumlah sampel; (3) Recall Weighted Average rata-rata recall per kelas yang dibobot; dan (4) F1-Score Weighted Average rata-rata harmonik dari presisi dan recall yang dibobot. Selain itu, Confusion Matrix divisualisasikan untuk menganalisis pola kesalahan klasifikasi antar kelas secara mendetail.

3.10 Analisis Sentimen Berbasis Aspek

Setelah model terbaik ditentukan, hasil klasifikasi sentimen dikombinasikan dengan hasil klasifikasi aspek untuk menghasilkan analisis ABSA. Setiap komentar memiliki dua label: label sentimen (Positif/Netral/Negatif) dan label aspek (salah satu dari 6 kategori). Kombinasi keduanya menghasilkan tabel distribusi sentimen per aspek (6×3 = 18 kombinasi) yang memungkinkan identifikasi aspek dengan sentimen positif tertinggi maupun aspek dengan sentimen negatif tertinggi.

3.11 Lingkungan Komputasi dan Alat Penelitian

Implementasi seluruh tahapan penelitian menggunakan spesifikasi teknis berikut: bahasa pemrograman Python 3.12, framework deep learning PyTorch, serta

Dipindai dengan CamScanner

akses model IndoBERT melalui Hugging Face Transformers. Terdapat dua model IndoBERT yang digunakan dengan fungsi berbeda, yaitu mdhugol/indonesia-bert-sentiment-classification yang telah di-fine-tune sebelumnya (pre-fine-tuned) dan digunakan untuk proses pelabelan data secara otomatis, serta indobenchmark/indobert-base-p1 yang digunakan sebagai model dasar (base model) untuk proses fine-tuning oleh peneliti sebagai model utama penelitian ini. Pengumpulan data dilakukan menggunakan Google API Python Client (YouTube Data API v3), manipulasi data menggunakan Pandas dan NumPy, implementasi SVM dan metrik evaluasi menggunakan library Scikit-learn, pemrosesan teks dan stopword Bahasa Indonesia menggunakan NLTK (Natural Language Toolkit), manajemen dataset untuk fine-tuning IndoBERT menggunakan library datasets (Hugging Face), serta visualisasi menggunakan Matplotlib dan Seaborn. Proses fine-tuning IndoBERT dilakukan pada Google Colab. Seluruh kode penelitian diimplementasikan dalam format Jupyter Notebook (.ipynb) untuk memastikan transparansi dan reprodusibilitas.

Dipindai dengan CamScanner