

DATA SCIENCE • PEMULA

Belajar Sentiment Analysis untuk Data Science Pemula

Dari konsep dasar hingga tiga pendekatan modern:
Rule-Based, Machine Learning, dan Large Language Models

Apa yang akan kita pelajari?

01 Apa itu Data Science?

02 Apa itu Sentiment Analysis?

03 Mengapa ini penting?

04 Pendekatan Rule-Based

05 Pendekatan Machine Learning

06 Pendekatan LLM (AI Modern)

07 Perbandingan Ketiga Pendekatan

08 Referensi

Apa itu Data Science?

Data Science adalah disiplin ilmu yang menggabungkan statistik, pemrograman, dan domain knowledge untuk mengekstrak wawasan bermakna dari data.

Provost & Fawcett, 2013 — Data Science for Business

Matematika & Statistik

Probabilitas, distribusi, uji hipotesis

Pemrograman & Komputasi

Python, SQL, manipulasi data

Domain Knowledge

Konteks bisnis atau keilmuan

Apa itu Sentiment Analysis?

Sentiment Analysis (Opinion Mining) adalah proses komputasi untuk mengidentifikasi dan mengekstrak opini subjektif dari teks — mengklasifikasikannya menjadi Positif, Negatif, atau Netral.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Morgan & Claypool Publishers.

POSITIF

"Pelayanannya sangat ramah!"

NETRAL

"Kantor buka jam 8 pagi."

NEGATIF

"Jalannya rusak dan gelap."

Mengapa Sentiment Analysis Penting?

2.5

Quintillion

byte data dibuat setiap hari di internet

80%

teks tidak terstruktur

dari seluruh data yang ada di dunia

4×

lebih cepat

pengambilan keputusan berbasis data vs intuisi

IDC Data Age Report, 2020 • McKinsey Global Institute, 2021

Pemda & Kebijakan

Evaluasi respons publik terhadap program pemerintah

Bisnis & Brand

Pantau reputasi produk dari ulasan pelanggan

Riset & Akademik

Analisis opini publik untuk studi sosial & politik

Media & Jurnalisme

Deteksi sentimen berita dan tren topik

Bagaimana Proses Sentiment Analysis?

1 Kumpulkan Data

Teks dari media sosial, ulasan, komentar, berita

2 Preprocessing

Bersihkan teks: hapus tanda baca, normalisasi, tokenisasi

3 Representasi Teks

Ubah kata menjadi angka (bag-of-words, TF-IDF, embedding)

4 Klasifikasi

Tentukan sentimen: Rule-Based / ML / LLM

5 Evaluasi

Ukur akurasi, precision, recall, F1-score

Rule-Based

Menggunakan daftar kata kunci, aturan linguistik, dan leksikon sentimen yang dibuat secara manual. Contoh: VADER, SentiWordNet, AFINN.

Kelebihan

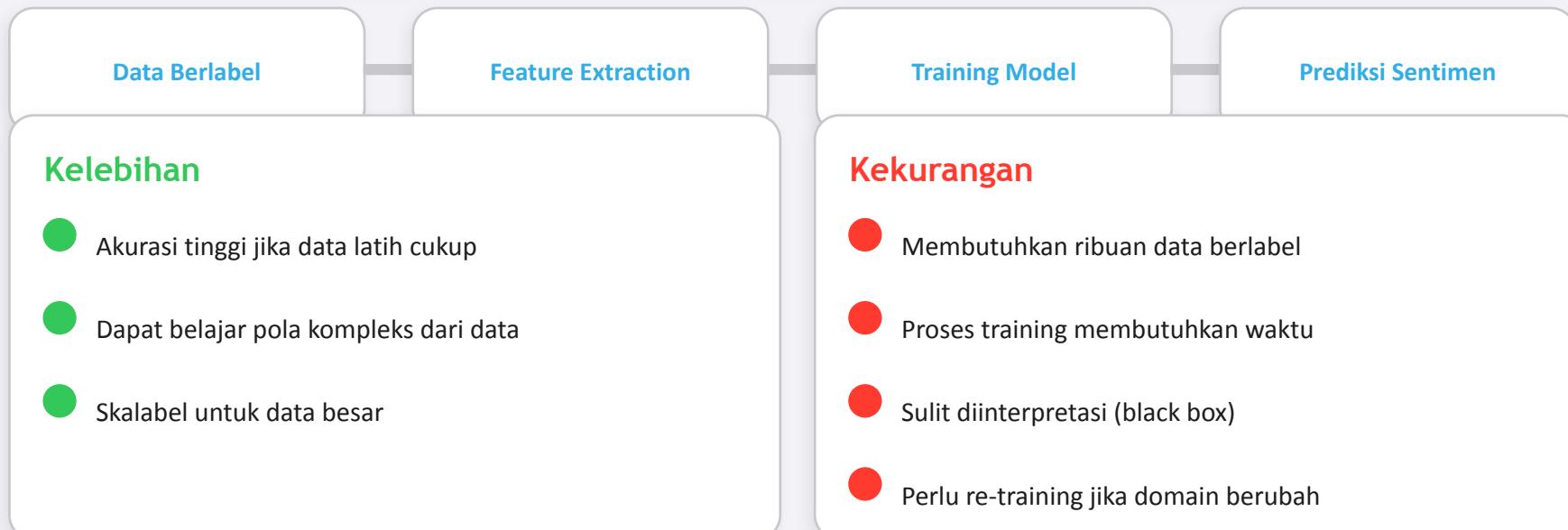
- Mudah dipahami dan dijelaskan
- Tidak butuh data latih (training data)
- Cepat diimplementasikan
- Transparan — aturannya terlihat jelas

Kekurangan

- Akurasi rendah untuk bahasa informal
- Tidak memahami sarkasme dan konteks
- Perlu update manual jika bahasa berubah
- Sulit menangani bahasa campuran (code-switching)

Machine Learning

Model belajar dari data berlabel menggunakan algoritma seperti Naive Bayes, SVM, atau Neural Network. Representasi teks: TF-IDF, Word2Vec, atau BERT.



Large Language Models

Model bahasa raksasa yang dilatih pada miliaran teks. Mampu memahami konteks, sarkasme, dan nuansa bahasa tanpa fine-tuning khusus.

Contoh: GPT-4 (OpenAI), Gemini (Google), Claude (Anthropic).

Kelebihan

- Memahami sarkasme, ironi, dan konteks
- Tidak perlu data latih (zero-shot)
- Mendukung banyak bahasa sekaligus
- Akurasi sangat tinggi pada teks kompleks

Kekurangan

- Biaya API bisa mahal untuk skala besar
- Privasi data: teks dikirim ke server eksternal
- Quota terbatas di tier gratis
- Latency lebih lambat dibanding rule-based

Pilih Pendekatan yang Tepat

Kriteria	Rule-Based	Machine Learning	LLM / AI Modern
Akurasi	Sedang	Tinggi	Sangat Tinggi
Data Latih	Tidak perlu	Ribuan label	Tidak perlu
Kecepatan	Sangat cepat	Sedang	Lebih lambat
Sarkasme	Tidak bisa	Terbatas	Mampu
Biaya	Gratis	Gratis/murah	Berbayar API
Cocok untuk	Prototipe	Produksi besar	Konteks kompleks

Referensi

1 Liu, B. (2012). Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

2 Provost, F., & Fawcett, T. (2013). Data Science for Business. O'Reilly Media.

3 Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. NAACL-HLT 2019.

4 Brown, T. B., et al. (2020). Language Models are Few-Shot Learners (GPT-3). Advances in Neural Information Processing Systems (NeurIPS).

5 Hutto, C. J., & Gilbert, E. (2014). VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text. ICWSM.

6 IDC. (2020). Data Age 2025: The Digitization of the World. International Data Corporation.

7 McKinsey Global Institute. (2021). The Age of AI: And Our Human Future. McKinsey & Company.

**Data adalah
bahan bakar,
analisis adalah
mesinnya.**

Sentiment Analysis membantu kita mendengar ribuan suara sekaligus — secara terstruktur dan terukur.