

Data Mining (Introduction)

Pelatihan Teknis Substantif
Jabatan Fungsional

Analisis Data Ilmiah (ADI)

Batch 1

7 Mei 2026

Dimas Sony Dewantara S.Kom., M.Kom.

Analisis Data Ilmiah- Fungsi Data dan Business Intelligence
Pusat Data dan Informasi - BRIN

Fokus:

- Analisis – Visualisasi Data – Pengembangan Dashboard
- Computer Vision
- Natural Language Processing (NLP)

Tools:

- Python
- Tableau - Superset

LinkedIn: [linkedin.com/in/dimas-sony-dewantara/](https://www.linkedin.com/in/dimas-sony-dewantara/)

GitHub: github.com/jimbotake



“May the **Insight** be with you”



Outline

- Pemahaman Data Mining
- Tahapan dalam Data Mining
- Metode Data Mining
- Implementasi Data Mining
- Sharing Pengalaman Implementasi
- Praktikum – Google Colab





“Kenapa platform digital bisa tahu apa yang kita suka?”

Contoh:

- 🎬 Netflix → rekomendasi film
- 🛒 E-commerce → produk yang “pas banget”
- 📱 Instagram → iklan yang relevan

Rekomendasi -> Pola -> Data Mining (Big Data)





Apa itu Data Mining?

“Proses mengekstraksi pola, hubungan, dan pengetahuan dari kumpulan data besar menggunakan teknik statistik dan komputasi.”

Tujuan Utama

- Menemukan pola tersembunyi
- Menghasilkan insight yang bermakna
- Mendukung pengambilan keputusan berbasis data



<https://usabilitygeek.com/big-data-small-insight/>



Data → Informasi → Insight → Keputusan

! Kenapa Data Mining Penting?

Ledakan Data (Data Explosion)

Data terus bertambah dari berbagai sumber (sensor, aplikasi, transaksi)

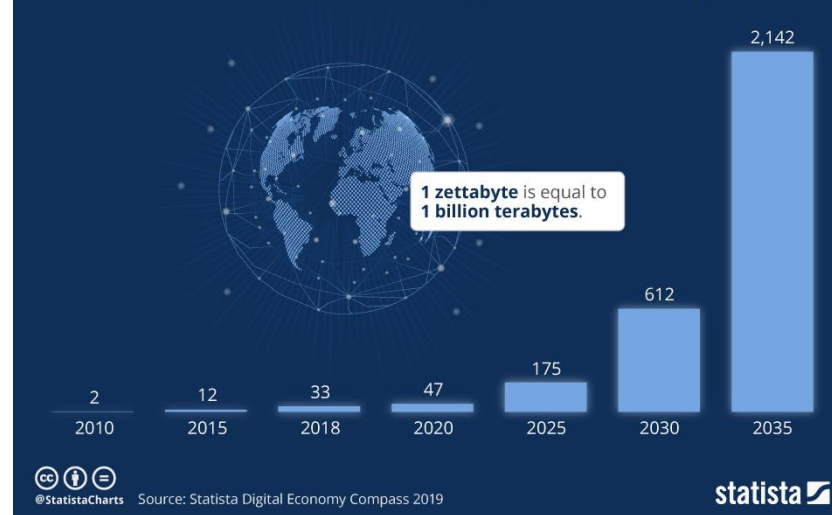
Keterbatasan Manusia

Analisis manual tidak mampu menangani data skala besar & kompleks

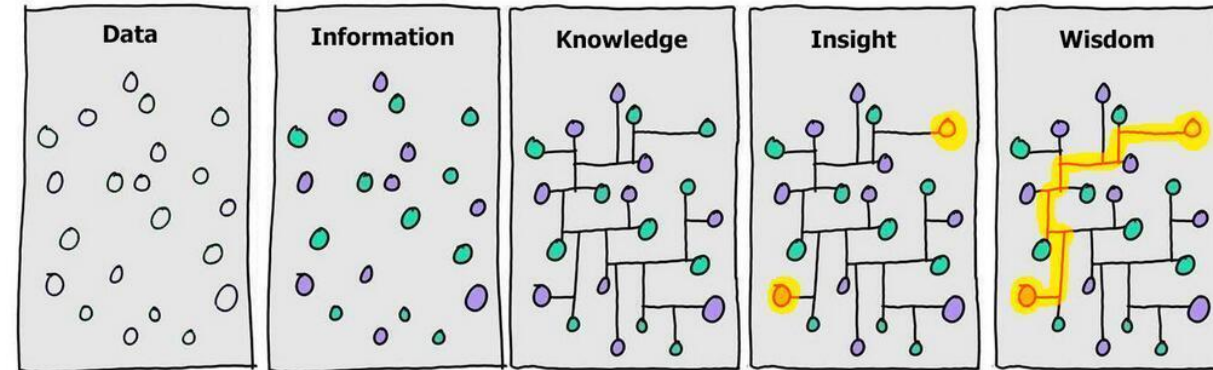
Pola Tersembunyi

Banyak insight penting tidak terlihat tanpa metode analitik

Global Data Creation is About to Explode
Actual and forecast amount of data created worldwide 2010-2035 (in zettabytes)



<https://lightedge.com/resources/the-data-explosion-and-hidden-data-storage-costs-in-the-cloud-could-object-storage-be-the-answer/>



<https://www.thefactory-lab.com/turning-data-into-wisdom-63>





Data vs Informasi

Data

Fakta mentah

Belum memiliki makna jelas

Berupa angka, teks, catatan

Contoh: 23, 27, 30

Informasi

Hasil interpretasi

Sudah memberi pemahaman

Menjawab pertanyaan

Suhu meningkat tiap jam



Transformasi Nilai

Data → Pengolahan → Analisis → Informasi

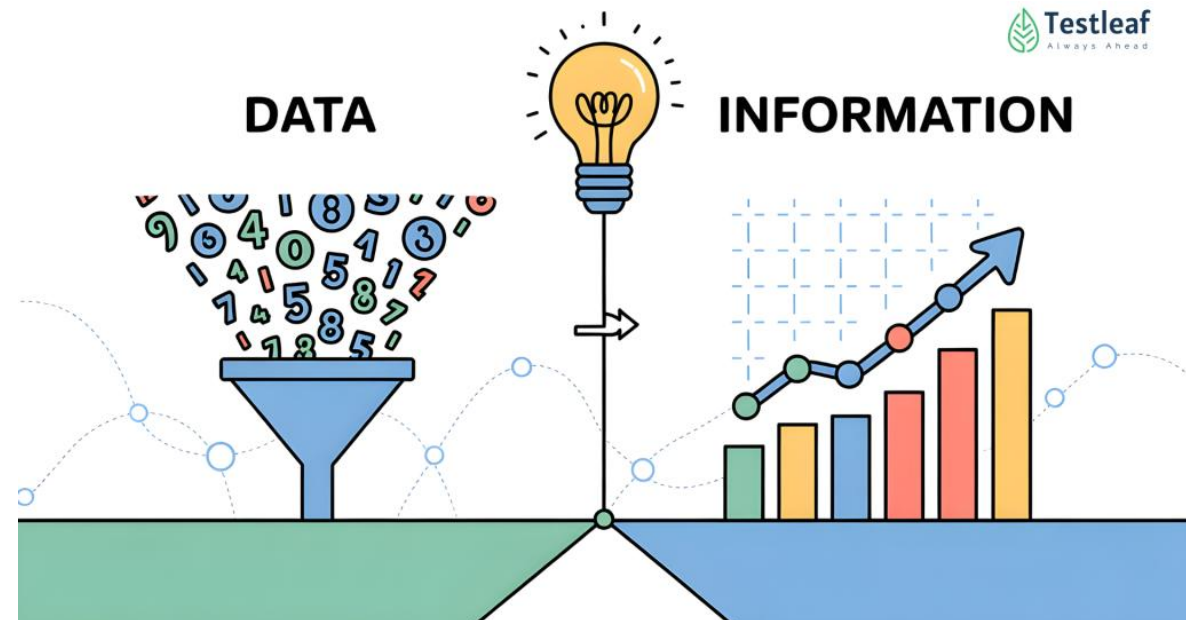



Contoh Nyata

Data: 100, 120, 150, 180






Informasi:

Penjualan meningkat selama 4 bulan berturut-turut.



 Analogi -> Data Mining = Menambang Emas dari Pasir

Pasir bercampur emas -> Disaring & dibersihkan ->
Emas ditemukan -> Bernilai tinggi

-  Makna Analogi
-  Pasir = data mentah, berantakan, banyak noise
-  Menyaring & mencuci = cleaning + preprocessing
-  Emas = insight / pola penting
-  Nilai ekonomi / ilmiah = keputusan & pengetahuan



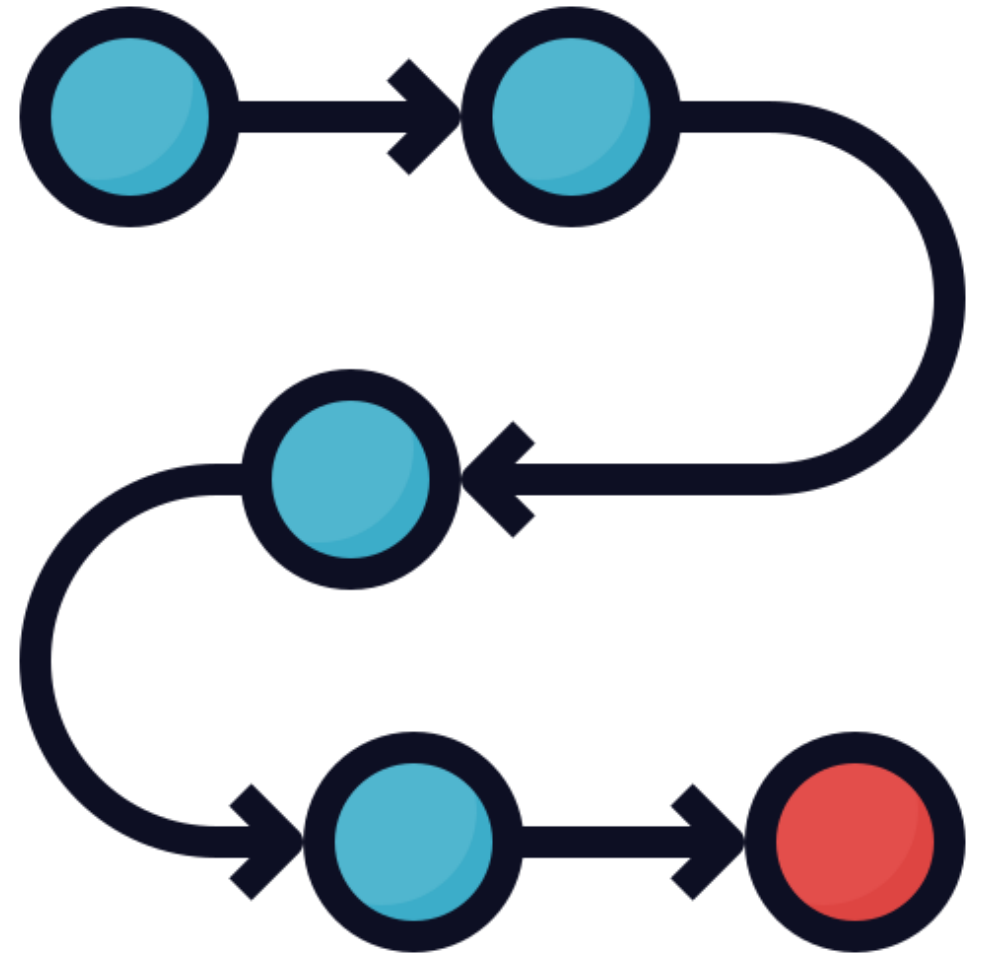


Tahapan Data Mining

1. Data Collection
- ↓
2. Data Cleaning
- ↓
3. Data Transformation
- ↓
4. Data Mining
- ↓
5. Evaluation & Interpretation

Penjelasan Singkat

- ◆ **1. Data Collection** Mengumpulkan data dari database, file, sensor, API, survei, sosmed, dll.
- ◆ **2. Data Cleaning** Mengatasi data kosong, duplikat, salah format, outlier.
- ◆ **3. Data Transformation** Mengubah data agar siap dianalisis (normalisasi, encoding, agregasi).
- ◆ **4. Data Mining** Menerapkan algoritma untuk menemukan pola.
- ◆ **5. Evaluation** Menilai apakah hasil valid, akurat, dan relevan.



Metode Data Mining

Statistik (Pendekatan Matematis)

- Analisis Deskriptif: Meringkas data melalui rata-rata, tren, dan distribusi untuk memahami kondisi saat ini.
- Analisis Inferensial: Mengambil kesimpulan besar dari sampel data kecil dengan tingkat kepercayaan tertentu.

Pendekatan Prediktif (Machine Learning)

- Klasifikasi: Mengelompokkan data ke kategori yang sudah ada (Contoh: Menentukan nasabah Layak/Tidak Layak kredit).
- Regresi: Memprediksi nilai angka pasti (Contoh: Prediksi harga properti atau omzet penjualan).
- Clustering: Mencari kelompok alami dari data yang belum berlabel (Contoh: Segmentasi perilaku konsumen).

Aturan Asosiasi (Pendekatan Relasional)

Pola Kebiasaan: Menemukan hubungan "Jika-Maka" antar item (Contoh: Pembeli popok bayi cenderung membeli bir di supermarket).

Deteksi Anomali (Pendekatan Keamanan)

Penyimpangan Data: Mengidentifikasi data yang sangat berbeda dari pola umum.
Contoh: Deteksi transaksi kartu kredit yang mencurigakan atau serangan siber.





Apa itu Klasifikasi?

“Metode untuk **memprediksi kategori atau kelas** berdasarkan data historis yang sudah memiliki label.”



Data Berlabel (**Input + Label**)



Model Belajar **Pola**



Prediksi Data Baru



Contoh Penggunaan



Email → **Spam / Bukan Spam**



Diagnosis → **Positif / Negatif**



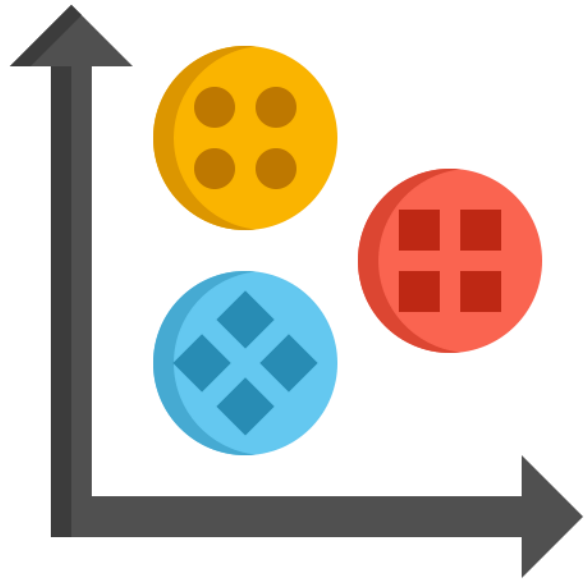
Transaksi → **Fraud / Normal**






Mahasiswa → **Lulus Tepat Waktu / Tidak**

● Apa itu Clustering?





“Metode untuk **mengelompokkan data berdasarkan kemiripan karakteristik**, tanpa label yang sudah ditentukan sebelumnya.”



Karakteristik Utama

-  Tidak membutuhkan label
-  Mencari struktur alami dalam data
-  Cocok untuk eksplorasi data awal

Contoh Penggunaan

-  Mengelompokkan spesimen biologis
-  Segmentasi pelanggan berdasarkan perilaku
-  Mengelompokkan wilayah berdasarkan iklim
-  Menemukan pola genetik yang mirip

Data Tanpa Label



Hitung Kemiripan



Bentuk Kelompok
(Cluster 1, 2, 3, dst)










Apa itu Regression?

“Metode untuk **memprediksi nilai numerik kontinu** berdasarkan hubungan antar variabel.”



Contoh Penggunaan

-  Prediksi suhu besok
-  Prediksi harga rumah
-  Prediksi hasil eksperimen
-  Prediksi permintaan produk
-  Prediksi penjualan bulanan

Input Variabel X
(luas rumah, waktu, dosis, dll)



Model Mempelajari Hubungan




Prediksi Nilai Y
(harga, suhu, hasil, dll)





Komparasi Metode

Teknik	Tujuan	Output	Contoh
Klasifikasi	Prediksi kategori	Label	Spam / Bukan Spam
Clustering	Mengelompokkan data	Cluster	Segmen pelanggan
Regression	Prediksi angka	Nilai numerik	Harga rumah

 **Pertanyaan Kunci**
Bukan “**algoritma mana paling hebat?**” tetapi, “**metode mana yang cocok untuk masalah yang dihadapi?**”





Tools Populer

Tools

Python

R

RapidMiner

Kelebihan

Fleksibel, ekosistem besar

Sangat kuat di statistik

Drag & drop, visual

Penggunaan

Analisis data, machine learning

Riset & analisis ilmiah

Pemula & prototyping



Pilihan Umum untuk Analisis Ilmiah



Python



R



python™



Mengapa Tools Sangat Penting dalam Data Mining?

Tools membantu kita untuk:

- ✓ Mengolah data dalam jumlah besar
- ✓ Menjalankan algoritma kompleks secara cepat
- ✓ Membuat visualisasi yang mudah dipahami
- ✓ Mengotomatisasi proses berulang
- ✓ Mengurangi human error

Perbandingan Sederhana

Manual Analysis

Lambat • Rentan salah • Sulit skala besar

Dengan Tools

Cepat • Konsisten • Skalabel





Penerapan Data Mining di E-commerce

Digunakan untuk:

- ✓ Sistem rekomendasi produk
- ✓ Analisis perilaku pengguna
- ✓ Prediksi pembelian berikutnya
- ✓ Personalisasi promo & iklan
- ✓ Segmentasi pelanggan

Contoh Nyata

- “Produk serupa yang mungkin Anda sukai”
- “Sering dibeli bersama”
- “Diskon khusus untuk Anda”
- “Lanjutkan belanja produk yang dilihat sebelumnya”

Cara Kerja Sederhana

Riwayat Klik + Pencarian + Pembelian



Analisis Pola Perilaku



Rekomendasi Produk yang Relevan



tokopedia





Penerapan Data Mining di Bidang Kesehatan

Digunakan untuk:

- ✓ Prediksi penyakit lebih dini
- ✓ Analisis rekam medis pasien
- ✓ Deteksi pola gejala
- ✓ Prediksi risiko pasien
- ✓ Mendukung keputusan klinis

Contoh Nyata

- Prediksi risiko diabetes
- Deteksi dini penyakit jantung
- Analisis pola rawat inap
- Rekomendasi tindakan lanjutan
- Identifikasi pasien berisiko tinggi

Cara Kerja Sederhana

Data Pasien + Riwayat Medis + Hasil Lab



Analisis Pola & Risiko



Prediksi / Rekomendasi Klinis





Kesimpulan



Inti Nilai Data Mining

Data Mentah



Analisis



Insight



Pengetahuan



Keputusan

- ✓ Data mining adalah proses menemukan insight dari data
- ✓ Membutuhkan data yang baik dan pemahaman konteks
- ✓ Digunakan di banyak bidang: bisnis, kesehatan, riset
- ✓ Membantu pengambilan keputusan berbasis data





Sharing Pengalaman Implementasi

Domain Sosial Media

Analisis Media Sosial Twitter dengan kata kunci “BRIN”

Pengumpulan Data: 1 – 27 Juli 2022

Tim Analisis Media Sosial

Organisasi Riset Ilmu Pengetahuan Sosial dan Humaniora

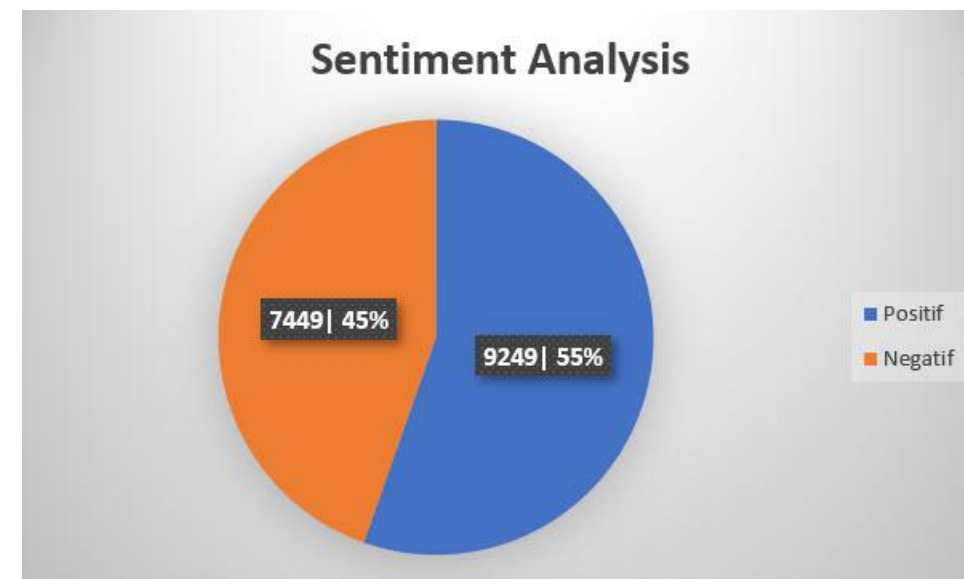
BRIN



Text Classification

Sentiment Analysis

Sentiment Analysis pada data tweet yang dikumpulkan menghasilkan perbandingan kumpulan tweet bersentimen positif sebesar 55% sedangkan tweet bersentimen negatif sebesar 45%. Analisis sentimen ini dilakukan menggunakan algoritma *deep learning CNN+LSTM* dengan akurasi pada data *training* sebesar 0,77%.



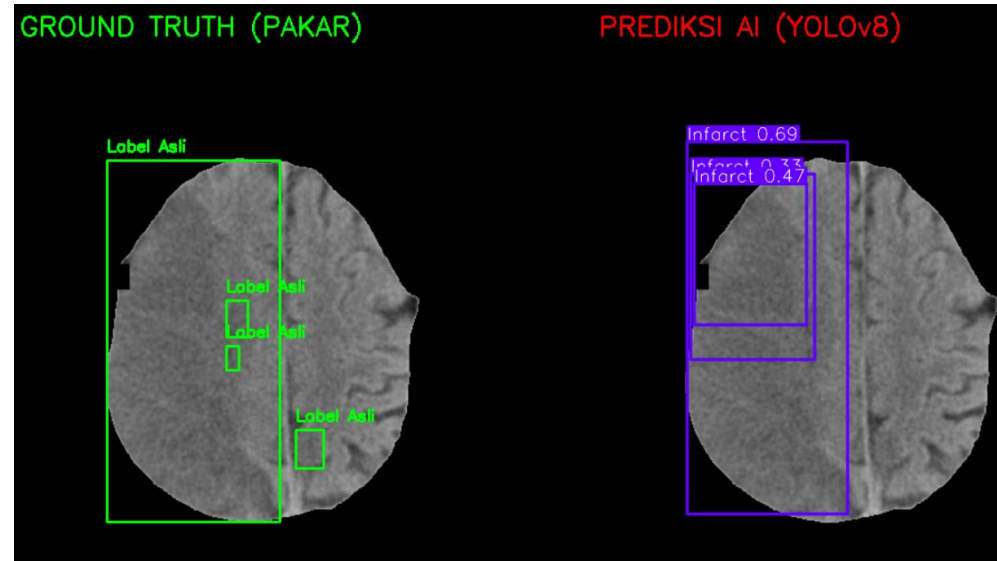
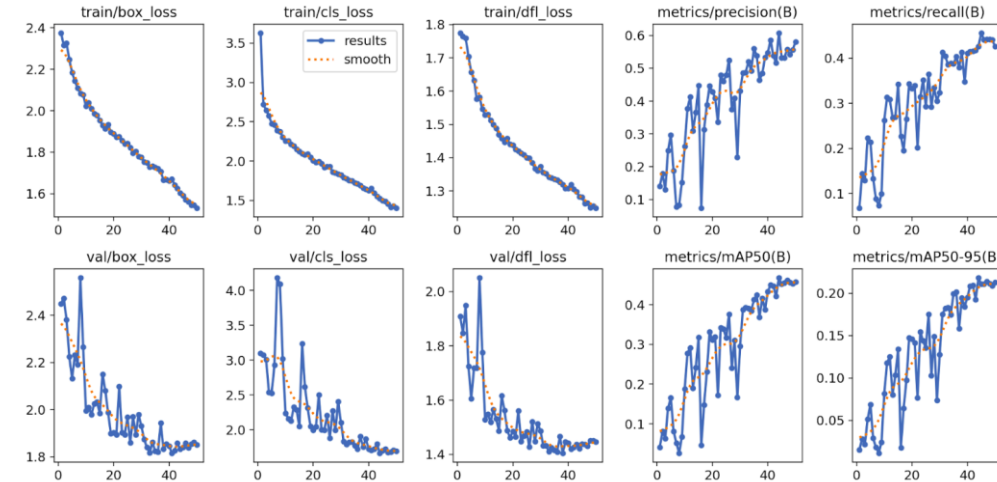
status_id	screen_name	text	tweet_prep	pred
x154285911283624	Torrez_20_06	Peneliti Utama Badan Riset dan Inovasi Nasional (BRIN) Prof Siti Zuhro mengakui ketokohan Ketua Umum DPP Partai Golkar @airlangga_hrt sebagai sosok yang pintar. #AirlanggaSolusiIndonesia https://t.co/TnXu2UUwNo	peneliti utama badan riset inovasi nasional brin prof siti zuhro mengakui ketokohan ketua dpp partai golkar airlangga_hrt sosok pintar airlanggasolusiindonesia http co tnXu2uuwno	1
x154286194957413	Torrez_20_06	"Diusung KIB sebagai capres 2024, Peneliti Utama Badan Riset dan Inovasi Nasional (BRIN) Prof Siti Zuhro mengakui ketokohan Ketua Umum DPP Partai Golkar @airlangga_hrt sebagai sosok yang pintar,????#AirlanggaSolusiIndonesiahttps://t.co/qDVLCAMpTd https://t.co/Msq1Sx6vns"	diusung kib capres 2024 peneliti utama badan riset inovasi nasional brin prof siti zuhro mengakui ketokohan ketua dpp partai golkar airlangga_hrt sosok pintar airlanggasolusiindonesiahttps co qdvlcamptd http co msq1sx6vns	1
x154286557610773	Zapline3	"Ketika 33 Lembaga Riset Kementerian dan Lembaga Negara dilebur ke BRIN, lebih dari 2.000 karyawan honorer termasuk peneliti eijkman diPHK. Dan kebanyakan muslim itu.Kagak ada yg bela koar2 tuh..Giliran tempat mabok dibubarin para kodok berisik semua.Krok krok"	33 lembaga riset kementerian lembaga negara dilebur brin 000 karyawan honorer peneliti eijkman diphk kebanyakan muslim kagak yg bela koar2 tuh giliran mabok dibubarin kodok berisik krok krok	0
x154286614807658	akhmadrifai_ID	"@BosTemlen Ketika 33 Lembaga Riset Kementerian dan Lembaga Negara dilebur ke BRIN, lebih dari 2.000 karyawan honorer termasuk peneliti eijkman diPHK. Dan kebanyakan muslim itu.Kagak ada yg bela koar2 tuh..Giliran tempat mabok dibubarin para kodok berisik semua.Krok krok"	bostemlen 33 lembaga riset kementerian lembaga negara dilebur brin 000 karyawan honorer peneliti eijkman diphk kebanyakan muslim kagak yg bela koar2 tuh giliran mabok dibubarin kodok berisik krok krok	0

Deteksi infarct stroke pada citra Non-Contrast CT (NCCT) menggunakan pendekatan YOLO (You Only Look Once)

“Merupakan metode berbasis *deep learning* dalam domain **Computer Vision** dan **Medical Imaging** yang bertujuan untuk mengidentifikasi area infark secara otomatis dan cepat.”

Domain Medis

Pada citra NCCT, infarct stroke umumnya ditandai dengan area hipodens (lebih gelap) yang seringkali sulit dibedakan pada fase awal. Dengan menggunakan YOLO, model dilatih untuk mendeteksi lokasi lesi tersebut. YOLO bekerja dengan membagi citra menjadi grid, kemudian memprediksi *bounding box*, kelas objek, dan tingkat kepercayaan (*confidence score*) untuk setiap area yang berpotensi mengandung infark.



CNN with multi stage image data augmentation methods for indonesia rare and protected orchids classification

2020 International Conference on Computer Science and Its Application in Agriculture, ICOSICA 2020 • Conference Paper • 2020 • DOI: 10.1109/ICOSICA49951.2020.9243174

Dewantara, Dimas Sony^{a, b}; Hidayat, Rachmat^{a, b}; Susanto, Heru^c; Arymurthy, Aniat Murni^a

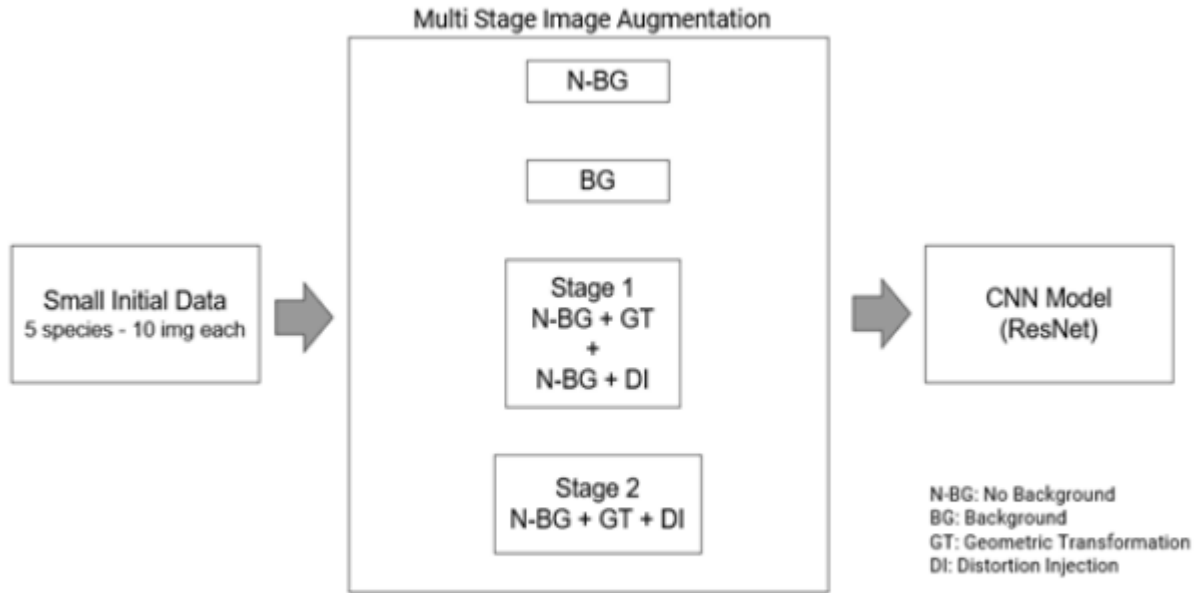


Fig. 4. Multi stage image augmentation scheme

Domain Konservasi

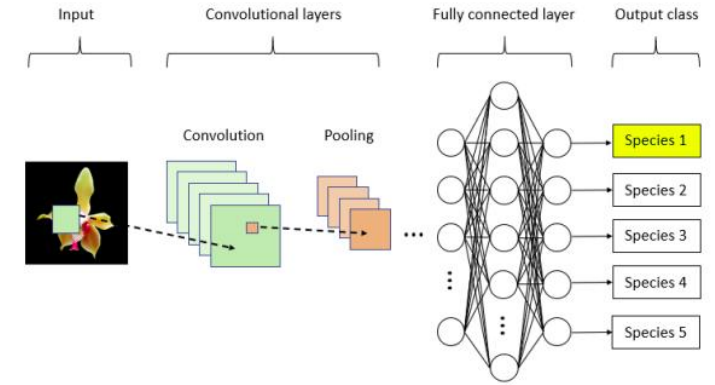


Fig. 5. CNN Architecture

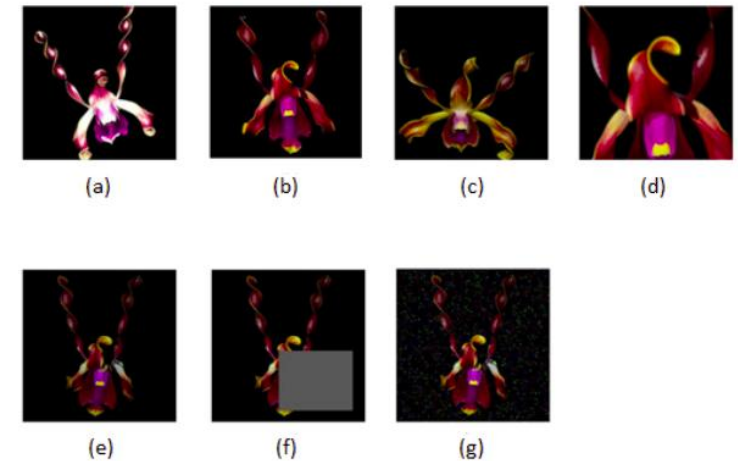


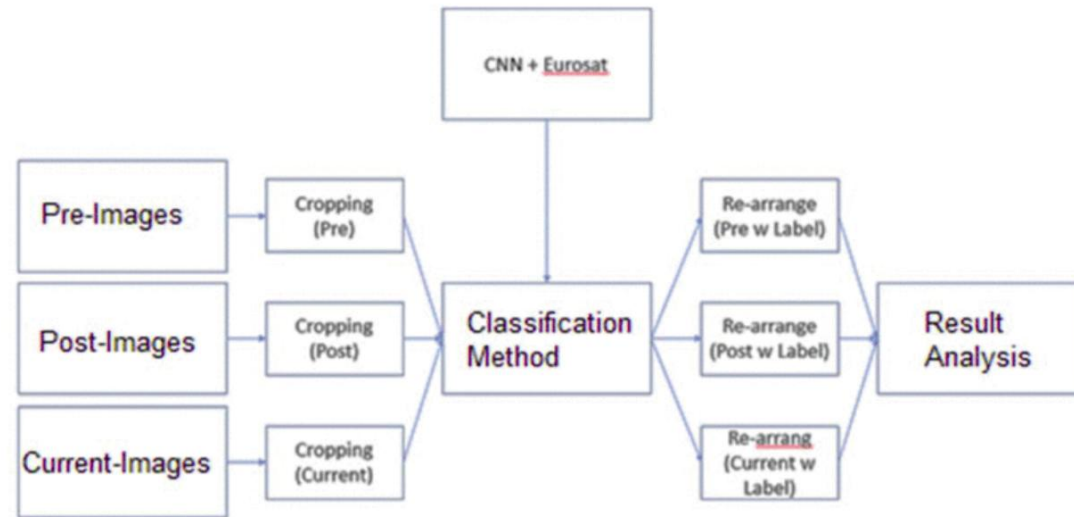
Fig. 3. Example of image augmentation results - Dendrobium lasianthera, (a) rotation, (b) horizontal flipping, (c) translation, (d) scaling, (e) brightness adjustment, (f) random eraser, (g) noise injection

Disaster Impact Analysis Uses Land Cover Classification, Case study: Petobo Liquefaction

2020 3rd International Conference on Computer and Informatics Engineering, IC2IE 2020 • Conference Paper • 2020 •

DOI: 10.1109/IC2IE50715.2020.9274573 

Hidayat, Rachmat^a ; Arymurthy, Aniaty Murni^a ; Dewantara, Dimas Sony^b 



Domain Kebencanaan

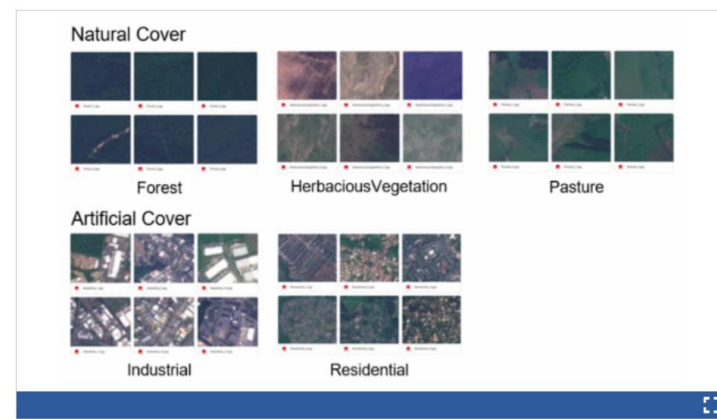


Fig. 2.
Eurosat Dataset with Natural Cover and Artificial Cover



Fig. 3.
Satellite Images of Areas Affected by the Petobo Liquefaction, Before a Disaster, Shortly After a Disaster, and Lastest States After a Disaster

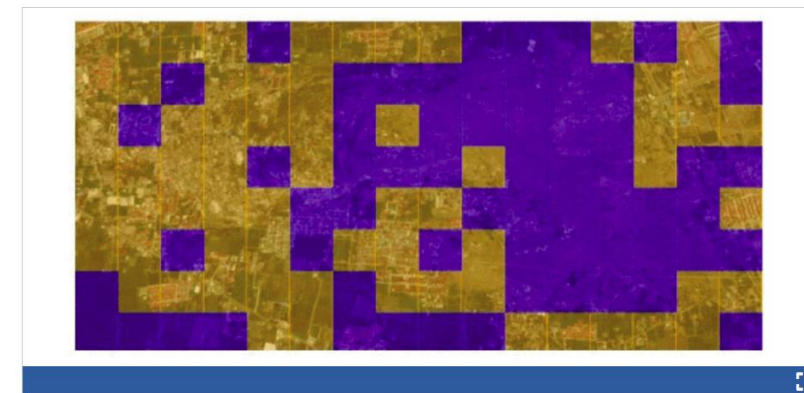


Fig. 10.
ResNet Result in Second Phase

Praktek – Youtube Comment Analysis

Pipeline

- **YouTube API Crawling**
- **Raw Comment Dataset**
- **Text Preprocessing (Cleaning + Stopword)**
- **Visualisasi - Wordcloud**
- Tokenization -> Word Embedding
- CNN-LSTM Training -> Labeled Data Training
- Saved Model -> Load Model
- Predict Crawled Comments
- Sentiment Result (Negative / Neutral / Positive)

Tools: Google Colab

Source: <https://s.id/HkeRs>





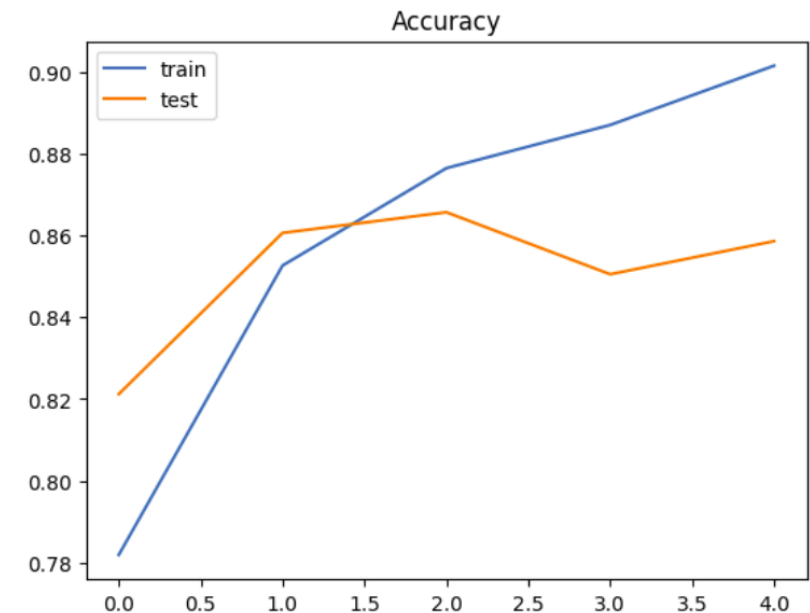
Praktek – Youtube Comment Sentiment Analysis

Data Training

	text1	label	text
1	warung ini dimiliki oleh pengusaha pabrik tahu...	positive	warung dimiliki pengusaha pabrik puluhan terke...
2	mohon ulama lurus dan k212 mmberi hujjah partai...	neutral	mohon ulama lurus k212 mmberi hujjah partai diw...
3	lokasi strategis di jalan sumatera bandung . t...	positive	lokasi strategis jalan sumatera bandung nyaman...
4	betapa bahagia nya diri ini saat unboxing pake...	positive	betapa bahagia unbox paket barang bagus meneta...
5	duh . jadi mahasiswa jangan sombong dong . kas...	negative	duh mahasiswa sombong kasih kartu kuning belaj...
...
10996	tidak kecewa	positive	kecewa
10997	enak rasa masakan nya apalagi kepiting yang me...	positive	enak masakan kepiting menyenangkan memilih kep...
10998	hormati partai-partai yang telah berkoalisi	neutral	hormati partai partai berkoalisi
10999	pagi pagi di tol pasteur sudah macet parah , b...	negative	pagi pagi tol pasteur macet parah bikin jengkel
11000	meskipun sering belanja ke yogya di riau junct...	positive	belanja yogya riau junction lihat foodlife lua...

11000 rows × 3 columns

Performa Model CNN + LSTM + Word Embedding





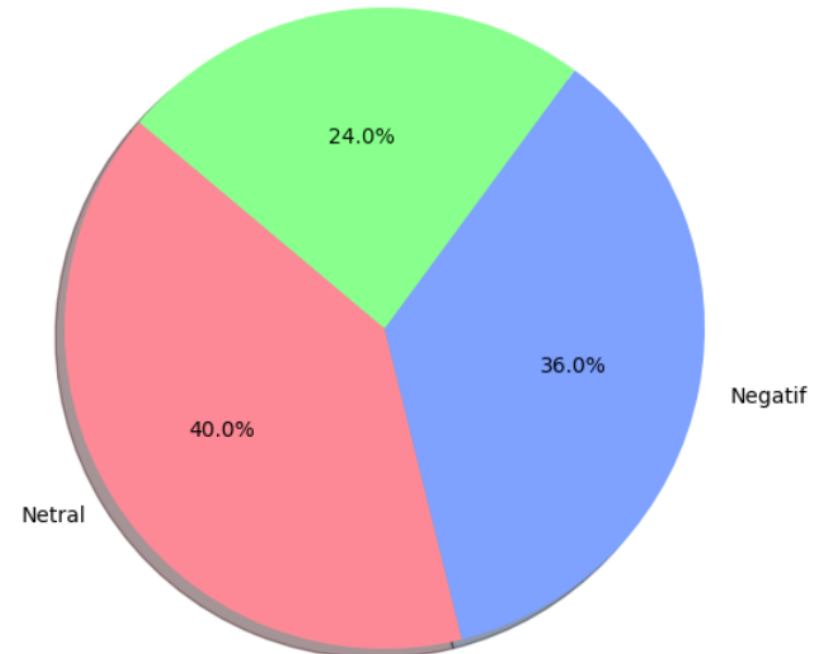
Praktek – Youtube Comment Sentiment Analysis

Comment Prediksi Sentiment

Konten sekeren ini belum ada komentar	konten keren komentar	2
Ga ada gambarnya	gambar	0
6.7 viewers tp gaada yg comment so here we go.....	viewer gaada comment so here we go	0
karang sambung kab mana tuh	karang sambung kab	1
2025	2025	1
Yang nonton 6,4 jt tetapi yang komen nihil. Bi...	nonton jt komen nihil bingung ngomong	0
Menurut cerita orang tua dulu dulu sekalih bah...	cerita orang tua kalih krsambung dulunya laut	0
Aku pernah kesana dan bagus bukit bukit nya	kesana bagus bukit bukit	2
Bagus sekali. Sangat informatif, disajikan den...	bagus informatif saji tarik mudah paham rekam ...	2
7tahun yang lalu penonton 6juta tapi kolom kom...	7tahun tonton 6juta kolom komentar kosong bayuh	0

0 = Negatif, 1 = Netral, 2 = Positif

Distribusi Hasil Prediksi Sentiment Comment Positif





Terima Kasih

LinkedIn: [linkedin.com/in/dimas-sony-dewantara/](https://www.linkedin.com/in/dimas-sony-dewantara/)

GitHub: github.com/jimbotake