Pengantar NLP

Gina Khayatun Nufus, M.Kom

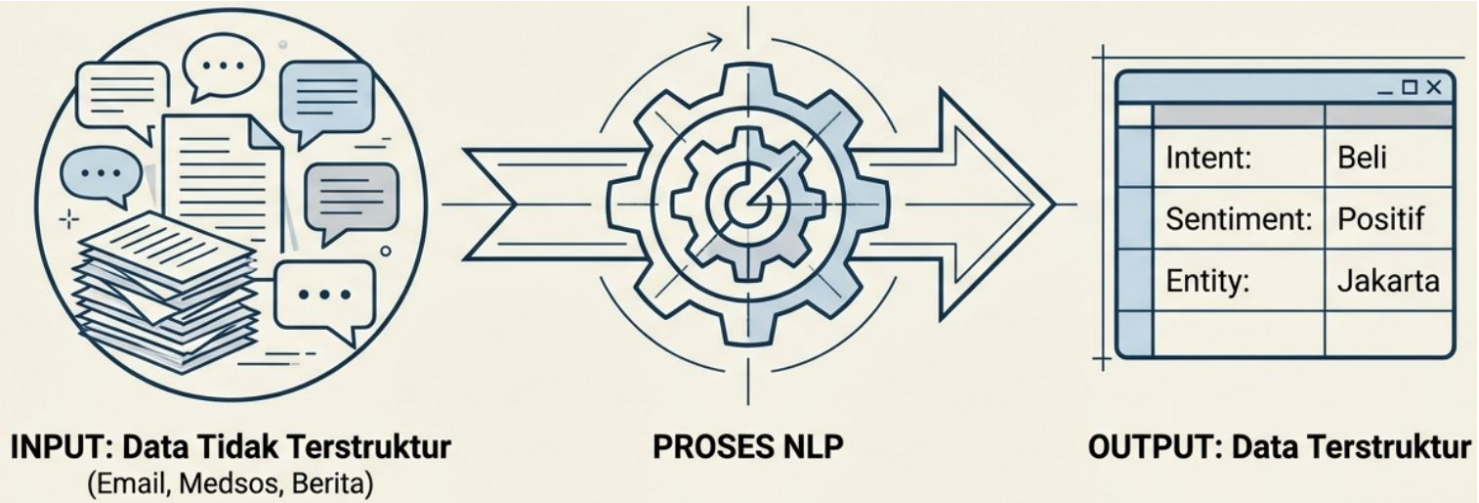
Overview

1. Pengenalan NLP. Aplikasi NLP dalam kehidupan sehari-hari
2. Tantangan dalam memproses bahasa natural (ambiguitas, konteks, variasi bahasa)
3. Jenis-jenis task NLP

Pengenalan NLP dan Aplikasi Dalam Kehidupan Sehari-hari

Definisi NLP

Natural Language Processing (NLP) merupakan cabang AI yang memungkinkan komputer memahami, memproses dan menghasilkan bahasa manusia.



Tujuan Utama : Agar mesin bisa **"membaca"** untuk melakukan tugas spesifik.

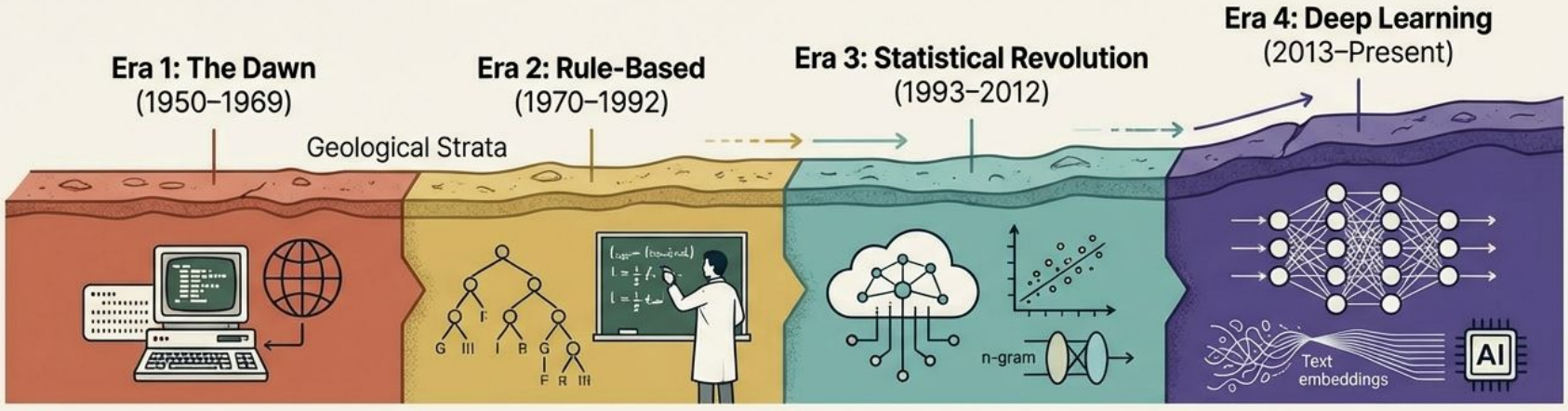
Aplikasi dalam kehidupan sehari-hari

Kontekstual Translation

Bank = “Financial Institution” , NOT “Riverbank”

Di balik aplikasi yang praktis, ada sistem yang bekerja keras memahami makna, tujuan (intent), dan konteks bahasa kita.

Four Eras of Innovation



Machine Translation & Code Breaking. Context: Cold War.

Hand-crafted Knowledge. Linguists writing grammar rules.

Machine Learning. N-Grams & Probability. "Let the data speak."

Representation Learning, Word Embeddings, & Large Language Models.

Era 1 : Periode Awal (Si Penerjemah Kaku)

Contoh: Georgetown-IBM Demo (1954).

Karakteristik : Dimulai dari penelitian mesin terjemahan

Motivasi: Perang Dingin (Menerjemahkan pesan Rusia ke Inggris).

Metode: *Look-up Dictionary* (Kata per kata).

Limitasi: Tidak tahu grammar. "Saya makan hati" diterjemahkan "I eat liver".

Idenya dari pemecah kode Perang Dunia II. Komputer cuma kayak kamus elektronik. Kalau urutan katanya beda sedikit, dia bingung.

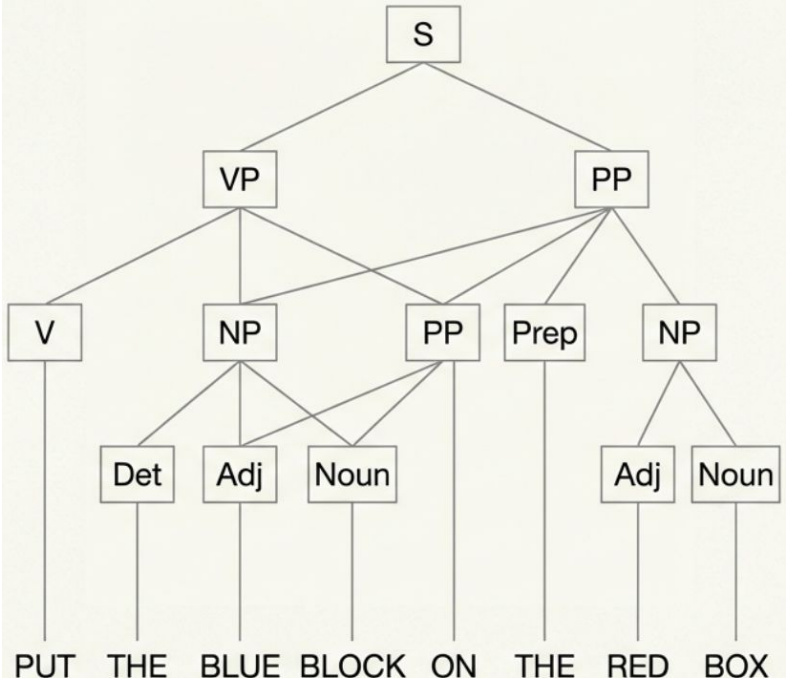
Era 2 : Rule-Based Systems (Dosen Grammar Galak)

Pendekatan: *Hand-crafted Rules* (Dibuat manual oleh ahli bahasa).

Masalah: *Brittle* (Rapuh). Ketemu kata gaul dikit, sistem error. Mahal & Lama.

Di era ini, Ahli Linguistik adalah raja. Mereka mencoba memprogram seluruh aturan tata bahasa ke dalam komputer. Berhasil? Ya, tapi hanya untuk lingkup sangat kecil (seperti dunia blok mainan).

Contoh: SHRDLU (Sistem tanya jawab dunia blok).



Limitasi di Era 2. Mengapa Kita berubah>

Poin Utama:

- **Brittle (Rapuh):** Gagal total jika input di luar aturan baku.
- **Tidak Scalable:** Butuh terlalu banyak tenaga ahli untuk menulis aturan baru.
- **Solusi:** Kita butuh data, bukan aturan manual!

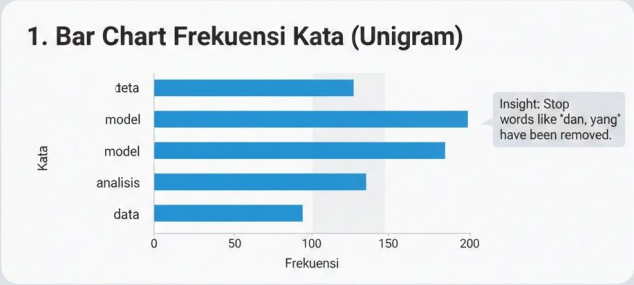
"Sistem ini gagal saat bertemu bahasa gaul atau kalimat non-baku. Kita tidak mungkin menulis aturan untuk setiap kemungkinan kalimat di dunia. Harus ada cara baru."

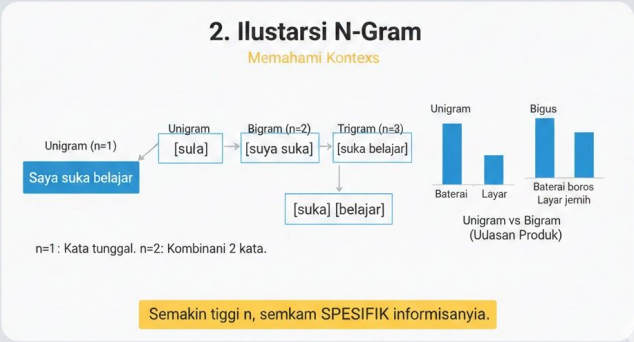
Era 3 (Fase 1) - Machine Learning Empiris (1993-1998)

Paradigma Baru: *"Let the data speak"* (Biarkan data bicara).
Metode: Statistik Murni (Counting).
Konsep: N-Grams & Probabilitas.

- Contoh: Peluang kata "Makan" diikuti "Nasi" > "Batu".

Masalah: *Sparsity* (Banyak kombinasi kata valid yang tidak muncul di data latih).





"Era 90-an internet mulai tumbuh, data teks melimpah. Kita berhenti mengajari komputer grammar, kita suruh dia menghitung probabilitas kemunculan kata saja."

Era 3 : Fase (2) - Supervised Learning Era (1998-2012)

Fokus: Klasifikasi dengan *Feature Engineering*.
Workflow:

1. Manusia memberi label (Annotation).
2. Komputer mempelajari pola dari fitur (kapitalisasi, sufiks, dll).

Algoritma: SVM, Hidden Markov Model (HMM).

"Menghitung saja tidak cukup. Kita masuk ke era di mana manusia menjadi 'guru' yang memberi kunci jawaban (label) pada ribuan data agar komputer bisa belajar pola (Supervised Learning)."

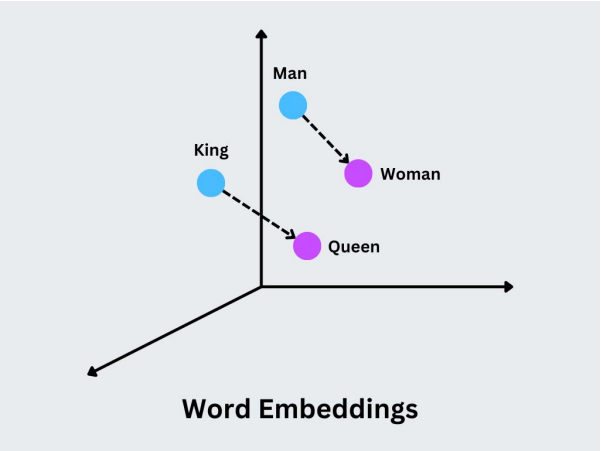
Era 4 (Fase 1) - Deep Learning Revolution (2013-2018)

Word2Vec (2013): Kata diubah menjadi vektor angka.

- king - man + woman ≈ queen
- Paris - France + Italy ≈ Rome

Keunggulan: Komputer memahami *makna* (semantik), bukan hanya mencocokkan karakter.
Model: RNN & LSTM (Mampu menangkap konteks kalimat lebih panjang).

Representasi Terdistribusi (Word Embeddings)



"Ini adalah *game changer*. Komputer tidak lagi melihat kata 'Kucing' dan 'Anjing' sebagai simbol berbeda, tapi sebagai dua vektor yang posisinya berdekatan karena sama-sama hewan peliharaan."

Era 4 (Fase 2) - Large Language Models (2018-Sekarang

Poin Utama:

- **Pergeseran Paradigma:** *Pre-training* + *Fine-tuning*.
- **Inovasi:** Transformer Architecture (BERT, GPT).
- **Self-Supervised Learning:**
  - Belajar dari data tak berlabel (internet).
  - Tugas: Menebak kata yang hilang (Masked LM) atau kata selanjutnya (Next Token Prediction).

"Alih-alih melabeli data satu per satu (mahal), kita biarkan model membaca seluruh internet dan belajar memprediksi kata sendiri. Ini melahirkan model raksasa yang punya pengetahuan umum sangat luas."

Analogi: Supervised vs Self-Supervised

Supervised: Kursus Privat (Belajar satu skill spesifik dengan guru).

- *Mahal, tidak bisa ditransfer ke skill lain.*

Self-Supervised: Membaca Seluruh Buku di Perpustakaan.

- *Punya pengetahuan umum, mudah belajar skill baru apa saja.*

"Dulu kalau mau bikin alat penerjemah, kita latih khusus terjemahan. Mau bikin chatbot, latih lagi dari nol. Sekarang, satu model pintar (Foundation Model) bisa melakukan semuanya hanya dengan instruksi berbeda."

Analisis Komparatif

Aspek	Era 3 : 1993-2012	Era 4: Deep Learning (2013-2018)	Era 4: Self-Superviside (2018-Sekarang)
Data	10-100M words	100M-1B words	10B-100B+ words
Representasi	Sparse vectors (one-hot)	Dense embeddings (100-300d)	Contextual embeddings (768-1024d)
Learning	Supervised per task	Supervised per task	Self-supervised + fine-tuning/prompting
Generalisasi	Limited (symbol-based)	Better (similarity-based)	Excellent (knowledge-based)
Feature Engineering	Manual & critical	Automatic	Automatic + emergent
Context Window	±5 words	±5 words	512-8000+ tokens
Training Cost	Low-Medium	Medium	Very High
Deployment Cost	Low	Medium	Medium-High
Task Adaptation	Train from scratch	Train from scratch	Few examples or just prompts!

Task NLP yang Relevan di Indonesia

Task NLP yang relevan di Indonesia

- Sentiment analysis ulasan Tokopedia/Google Maps: “Makanannya enak, tapi pelayanannya parah” → sistem harus menangkap sentimen campuran, bukan sekadar positif/negatif per kalimat.
- Moderasi komentar di media sosial: platform perlu mendeteksi ujaran kebencian, spam, atau hoaks dalam Bahasa Indonesia dan bahasa daerah (Jawa, Sunda, Minang), yang penuh slang dan variasi ejaan.

Tantangan Bahasa Indonesia dan Slang

- Ambiguitas makna: kata “bisa” (racun ular vs mampu) atau “tarik” (aksi keuangan vs gerakan fisik) butuh konteks kalimat untuk dipahami; model yang hanya melihat kata per kata akan mudah salah.
- Slang dan singkatan: “gk”, “ga”, “nggak”, “engga” → semua berarti “tidak”; “bgt”, “bgttt” → “banget”; jika dibiarkan apa adanya, model akan menganggap itu sebagai kata yang berbeda-beda, vocab jadi besar dan data makin jarang.

Contoh Bahasa Manusia yang akan diberikan ke Mesin

- Untuk manusia, kalimat “Pelayanannya parah, tapi makanannya enak bgt sih 😍” jelas artinya (sentimen campuran, nada bercanda).
- Untuk komputer, itu cuma rangkaian karakter dengan: emoji, huruf berulang “bgt”, campuran formal-informal, dan mungkin typo.

Kalau kamu diminta membuat program yang menghitung apakah review itu positif atau negatif, apa saja yang mengganggu?

Kesalahan Umum yang Harus di Hindari

- Mengabaikan konteks budaya/bahasa: Jangan asumsikan model Inggris langsung kerja untuk teks Indonesia (misalnya, slang "mantul" vs "excellent"); selalu tes dengan dataset lokal.
- Langsung lompat ke kode kompleks: Pemula sering skip teori, menyebabkan kebingungan saat debug; pahami dulu "mengapa" task NLP sulit (ambigu bahasa, seperti "bank" = sungai/uang).
- Overlook etika awal: Hindari bias data (misalnya, dataset dominan satu gender), yang bisa propagate ke model akhir; diskusikan kasus nyata seperti diskriminasi di sentiment analysis.

Kesimpulan

- Jenis Task NLP Dasar : Termasuk klasifikasi teks (misalnya, spam detection), sentiment analysis (positif/negatif/netral dari review), named entity recognition (NER, identifikasi nama orang/tempat), dan part-of-speech tagging (POS, label kata sebagai noun/verb). Mulai dengan task sederhana seperti menghitung frekuensi kata untuk memahami pola bahasa.
- **Pendekatan Tradisional vs Modern**: Tradisional menggunakan aturan linguistik (rule-based, seperti regex untuk pola) dan statistik (seperti n-grams untuk probabilitas kata berikutnya). Modern bergeser ke machine learning (ML) dan deep learning (DL) yang belajar pola dari data besar, lebih fleksibel untuk bahasa seperti Indonesia yang kompleks secara morfologi.
- **Komponen Utama**: Input (teks mentah), processing (tokenisasi awal), output (prediksi/hasil), dievaluasi dengan metrik seperti accuracy atau precision-recall.

Tugas Pertemuan 1:

1. Individu : Buat ringkasan perbedaan 4 era NLP dalam bentuk tabel/infografis.
2. Kelompok: Diskusikan satu sistem dari Era 2 dan presentasikan di pertemuan berikutnya. 1 Kelompok terdiri dari 2-3 orang. Laporan & PPT.