

Intro to Natural Language Processing

Indonesia AI

Proprietary document of Indonesia AI 2023



OBJECTIVE & OUTLINE

Proprietary document of Indonesia AI 2023



Intro to Natural Language Processing

Objektif: Memahami konsep dasar dari Natural Language Processing (NLP) atau Pemrosesan Bahasa Alami.

Outline:

1. Apa itu Natural Language Processing?
2. Sejarah NLP
3. Trend dalam Penerapan NLP
4. Kaidah Dasar dalam NLP
5. Aspek dasar dalam Ilmu Linguistik

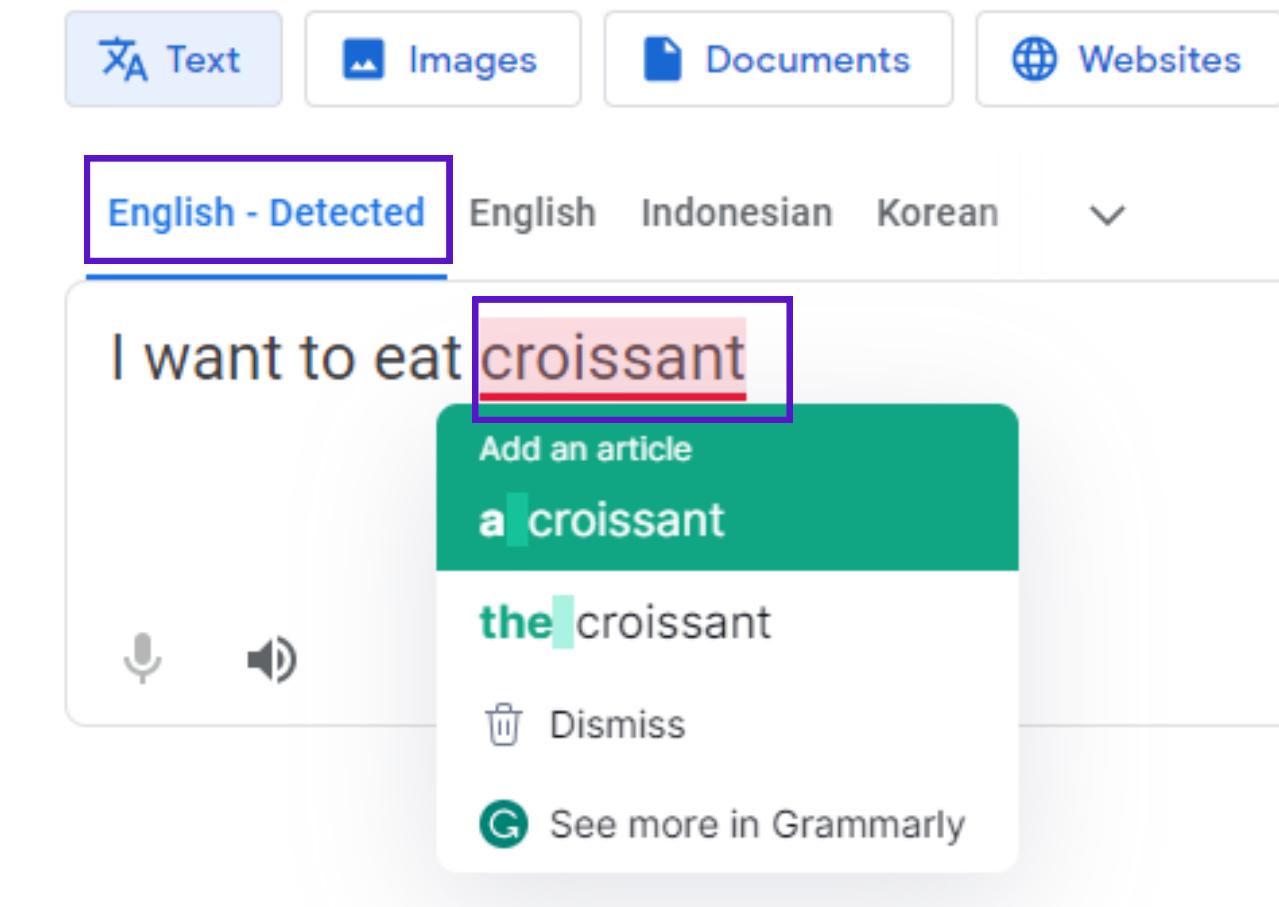
Natural Language Processing, What is it?

DEFINISI

Proprietary document of Indonesia AI 2023

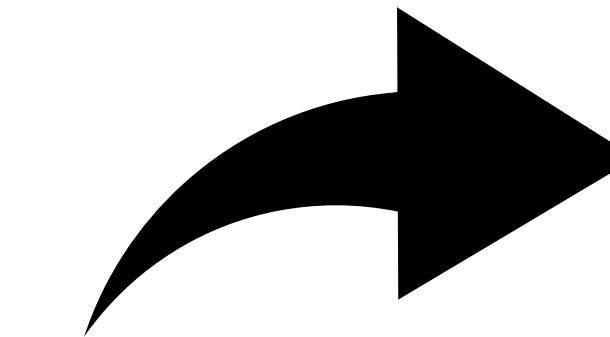
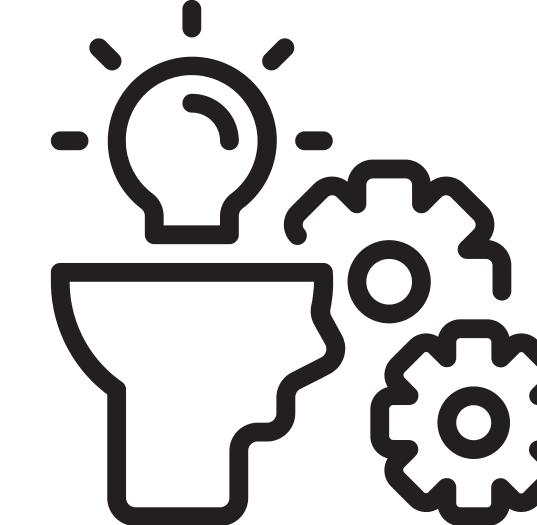
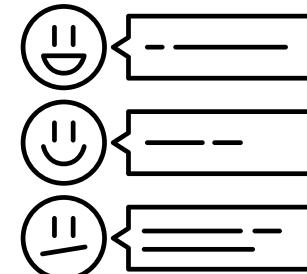
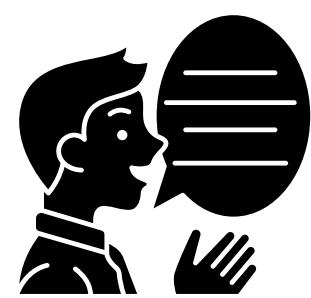
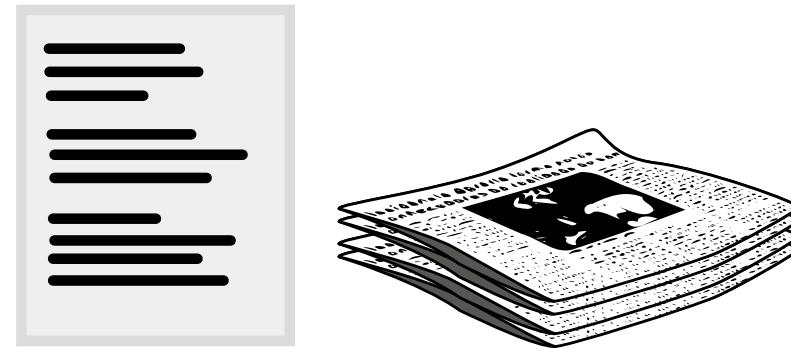
Merupakan cabang bidang AI yang terfokus pada yang berkaitan dengan memberi komputer kemampuan untuk **memahami teks** dan **kata-kata yang diucapkan** dengan cara yang sama seperti yang **mampu dilakukan oleh kita sebagai manusia** yang memiliki kecerdasan **linguistik**.

- International Business Machines (IBM) Corp.



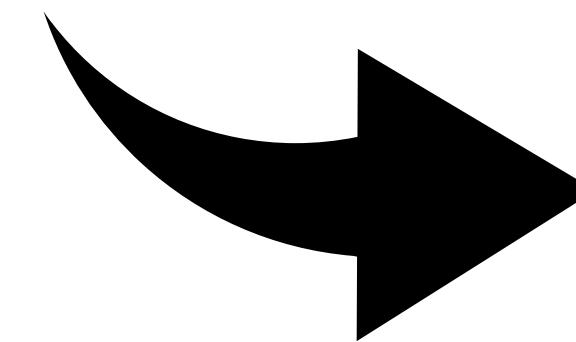
PENERAPAN NLP

Proprietary document of Indonesia AI 2023



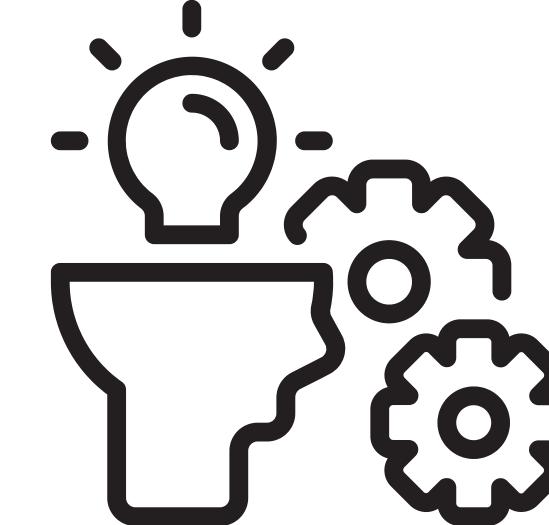
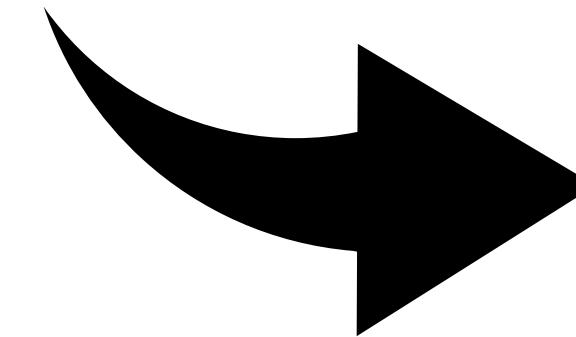
insights

NLP

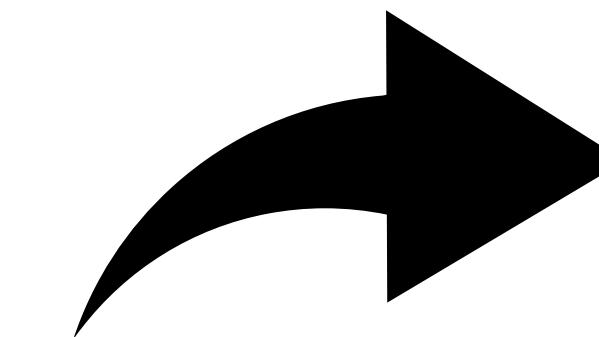


PENERAPAN NLP

Proprietary document of Indonesia AI 2023



NLP



insights

PENERAPAN NLP

Proprietary document of Indonesia AI 2023

Kemampuan NLP semakin dibutuhkan agar stakeholder, mulai dari organisasi, pemerintahan, bisnis, dan perorangan dapat mengambil insights dari data teks tersebut dan memanfaatkannya untuk **pengambilan keputusan yang lebih baik**.

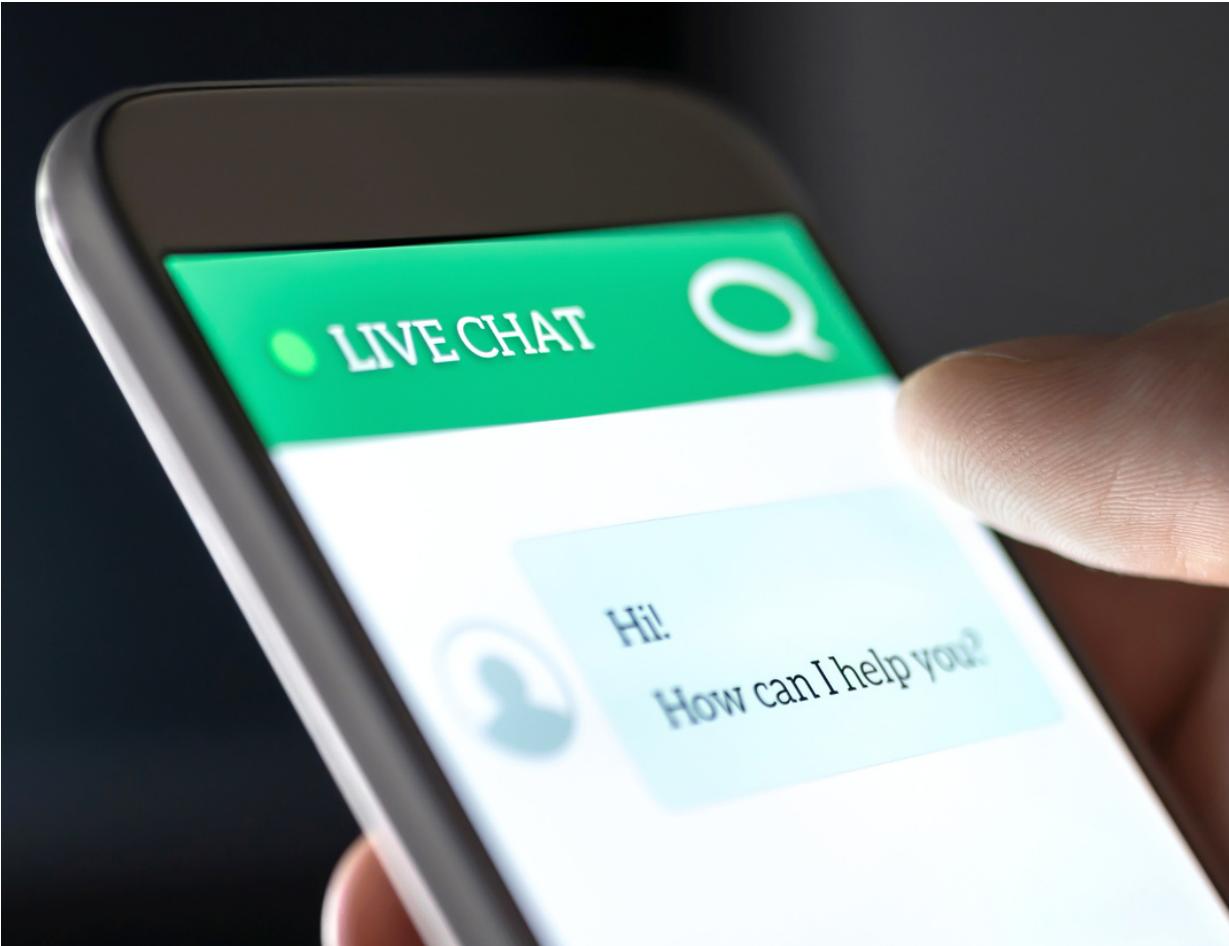
Data teks semacam ini mengandung informasi berharga yang dapat memberikan **wawasan tentang tren, opini publik, preferensi pelanggan**, dan masih banyak lagi.

NLP DALAM BERBAGAI DOMAIN (1)

Proprietary document of Indonesia AI 2023



News



Technology



Education

NLP DALAM BERBAГAI DOMAIN (2)

Proprietary document of Indonesia AI 2023



Finance



Healthcare

TANTANGAN DALAM NLP

Proprietary document of Indonesia AI 2023

Tantangan dalam NLP sangat beragam, dari kompleksitas struktur bahasa hingga variasi makna dalam konteks yang berbeda, hal ini **mempengaruhi keakuratan dan pemahaman mesin terhadap teks.**

1. Ambiguitas
2. Slang
3. Kesalahan pengejaan dan pelafalan
4. Irony and Sarcasm
5. Pengetahuan umum
6. Kreativitas
7. Training Data
8. Multilingual

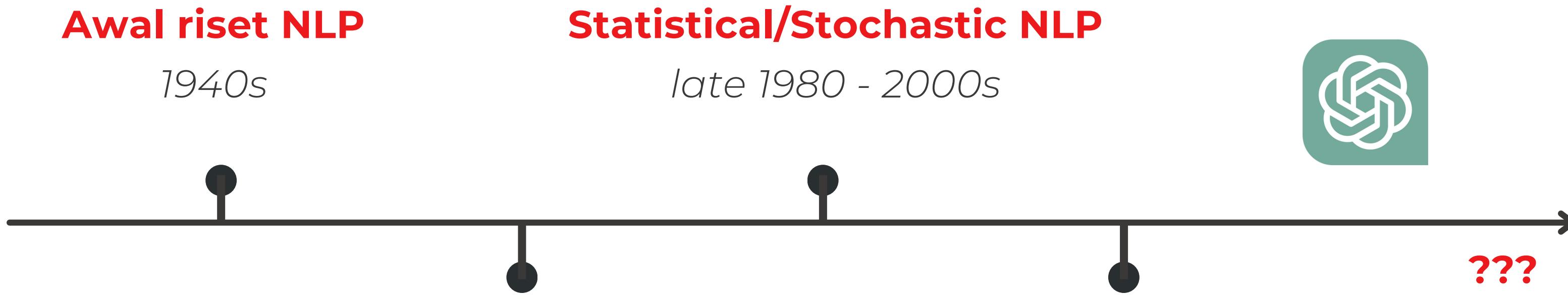


Any question guys ~

— Sejarah NLP

SEJARAH NLP

Proprietary document of Indonesia AI 2023



AWAL RISET

Proprietary document of Indonesia AI 2023

Bidang NLP sudah mulai diriset pada tahun sekitar 1940, setelah perang dunia ke-2.

- Di waktu itu, para ilmuwan mulai memahami **pentingnya** melakukan **penerjemahan** dari satu bahasa ke bahasa lainnya.

Awal riset NLP

1940s

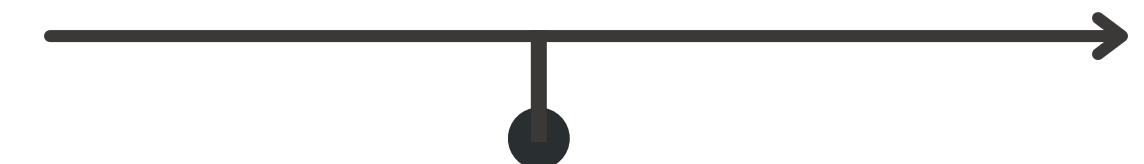


SYMBOLIC NLP

Proprietary document of Indonesia AI 2023

Idenya adalah menggunakan **aturan** dan **konsep ontologi**.

- Pada tahun 1950-an, dilakukan penerjemahan kalimat Rusia ke bahasa Inggris secara otomatis.
- Pada tahun 1960-an, dikembangkan sistem pemrosesan bahasa alami seperti SHRDLU.
- Pada tahun 1970-an, programmer mulai membuat ontologi konseptual untuk mengorganisir informasi dunia nyata dalam bentuk data komputer.
- Pada tahun 1980-an dan awal 1990-an, fokus pada bidang seperti rule-based parsing, morfologi, semantik, dan pemahaman bahasa alami.



Symbolic NLP

1950s - early 1990s

STATISTICAL NLP

Proprietary document of Indonesia AI 2023

Ditandai dengan **penggunaan algoritma machine learning** dan penekanan pada **analisis statistik**.

- Pada akhir 1980-an, terjadi revolusi dalam NLP dengan diperkenalkannya algoritma machine learning.
- Pada tahun 1990-an, terdapat keberhasilan dalam metode statistik khususnya dalam terjemahan mesin.
- Pada tahun 2000-an, penelitian semakin fokus pada unsupervised dan semi-supervised dengan adanya ketersediaan data bahasa mentah yang belum diannotasi.

Statistical/Stochastic NLP

late 1980 - 2000s

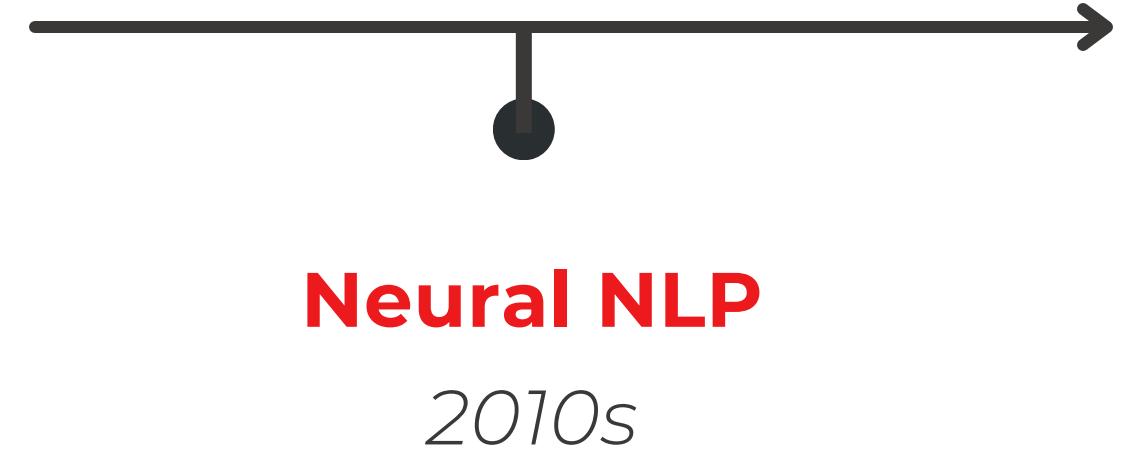


NEURAL NLP

Proprietary document of Indonesia AI 2023

Pada tahun 2010-an, terjadi kebangkitan pendekatan berbasis **neural network** dalam NLP.

- Kemampuan neural network untuk mengatasi **kompleksitas** dan variasi bahasa manusia menjadi penekanan utama.
- Penggunaan data besar dan komputasi yang canggih membantu kemajuan dalam penggunaan neural network dalam NLP.



Any question guys ~

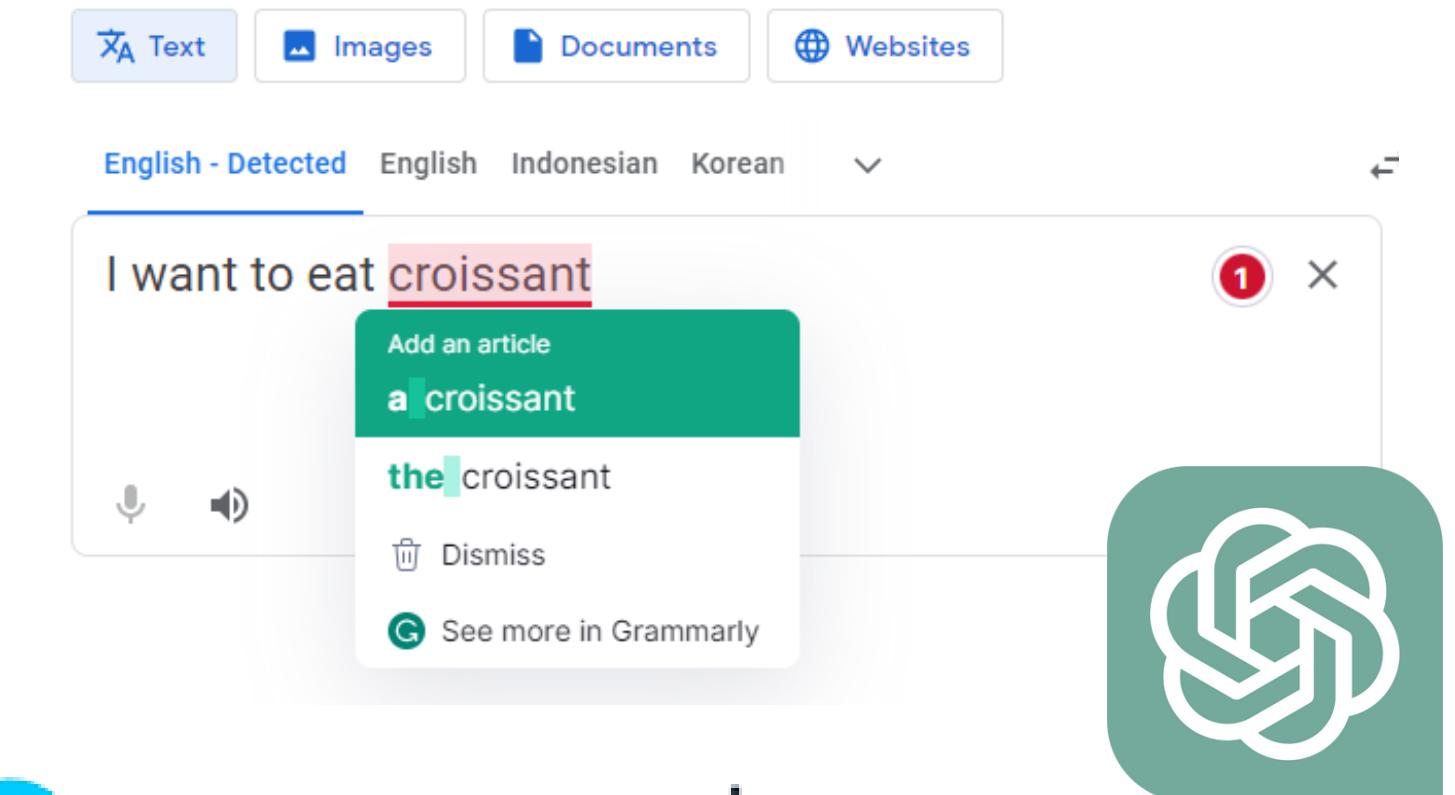
Trend dalam Penerapan NLP

APLIKASI NLP POPULER

Proprietary document of Indonesia AI 2023

Penerapan NLP yang populer secara global:

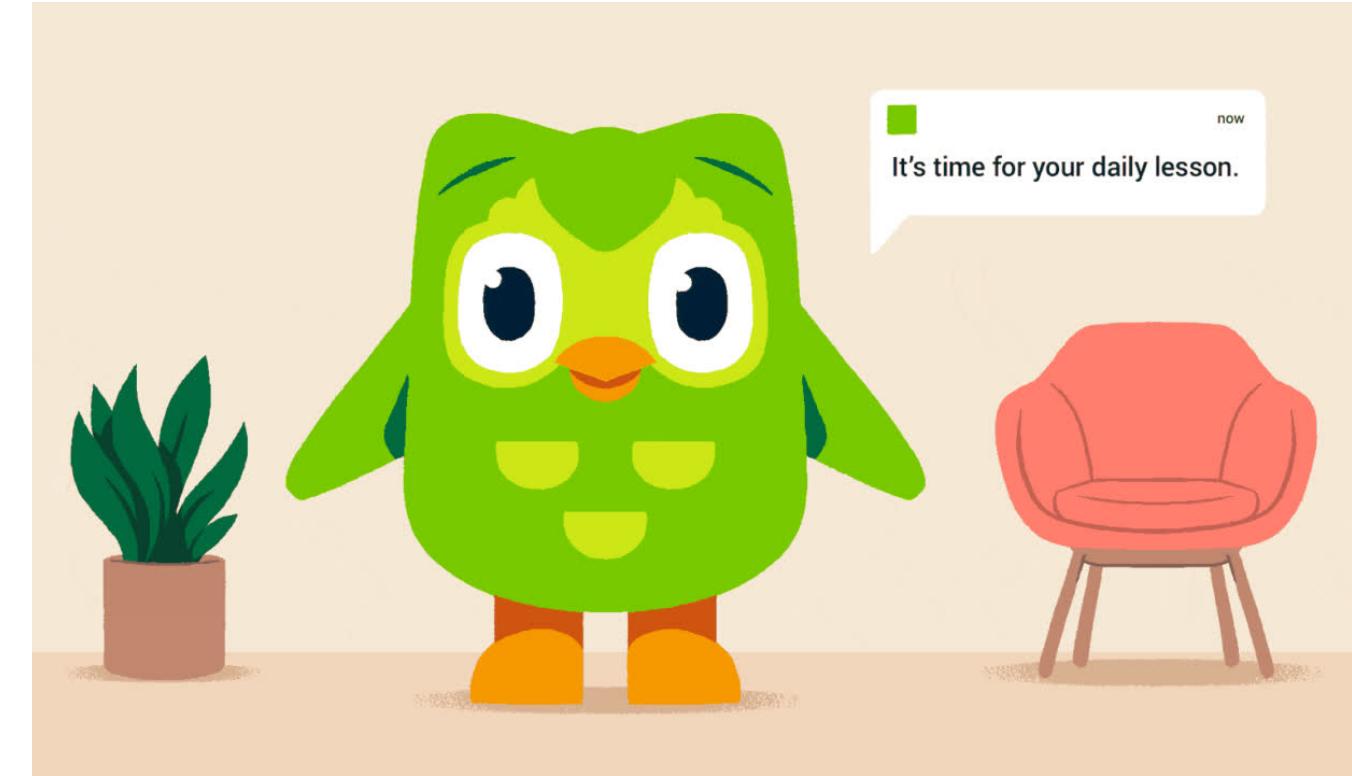
- Google Translate
- Grammarly
- Amazon Alexa
- Chatgpt



amazon alexa

APLIKASI NLP POPULER

Proprietary document of Indonesia AI 2023



Menariknya sekarang kita dapat belajar bahasa alami melalui platform yang memanfaatkan teknologi NLP.

Ayo, siapa yang pernah pakai~

POPULARITAS NLP DI INDONESIA

Proprietary document of Indonesia AI 2023

Saat ini ada 3 use case NLP yang cukup populer di Indonesia:

- Analisis sentiment saat mendekati periode Pemilu.
- Virtual assistant pada aplikasi yang dapat menangani aspek customer service sampai troubleshooting.
- Pengkategorian jenis topik di media berita.

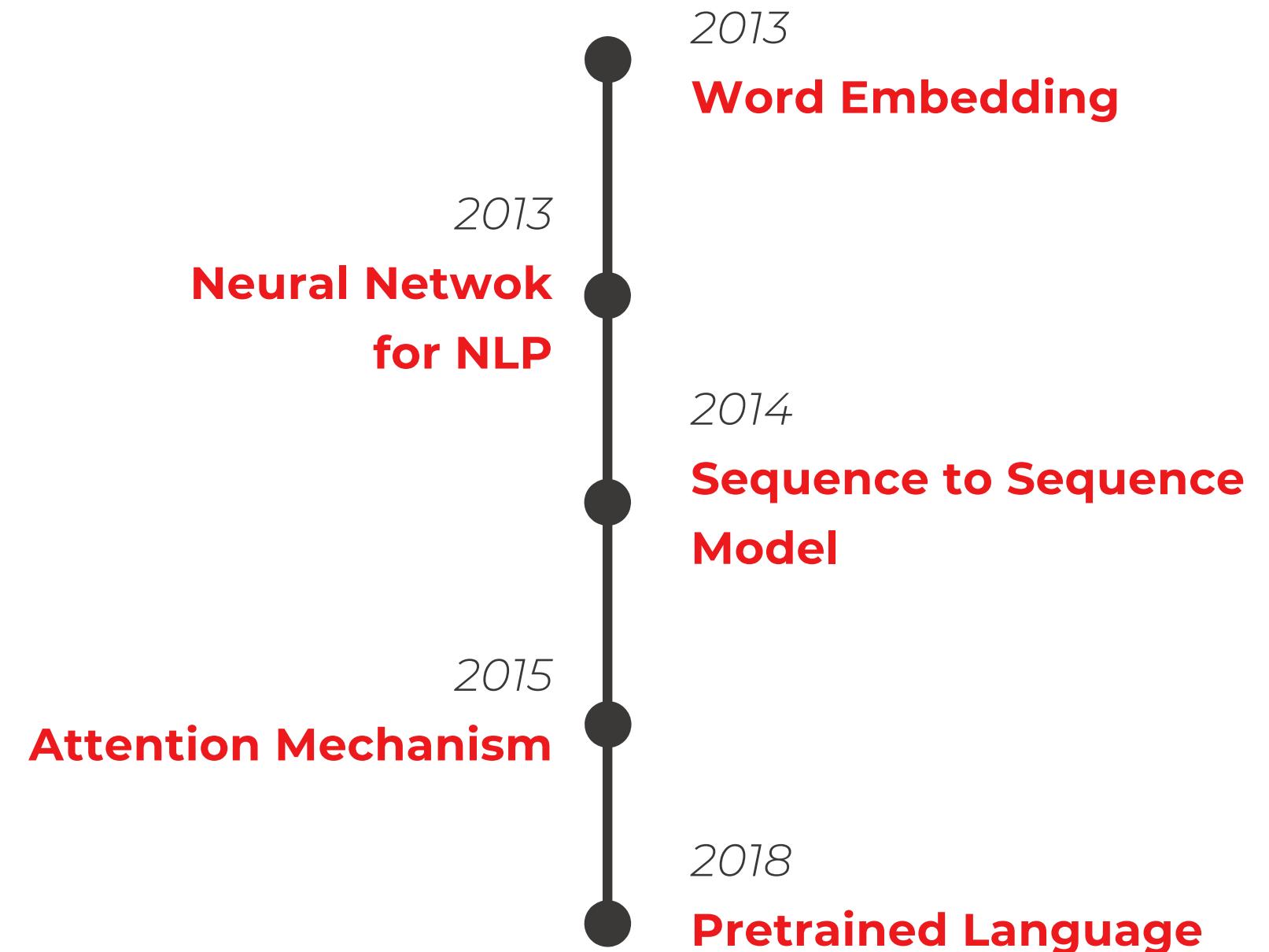


TREND RISET NLP

Proprietary document of Indonesia AI 2023

Dalam pengembangan model-model NLP terbaru, ada dua konsep yang sangat penting dan populer, yaitu **Attention Mechanism** dan **Pretrained Language**.

Kedua konsep ini telah menjadi tren utama dalam menciptakan model-model NLP terkini yang menghasilkan kinerja yang sangat baik.

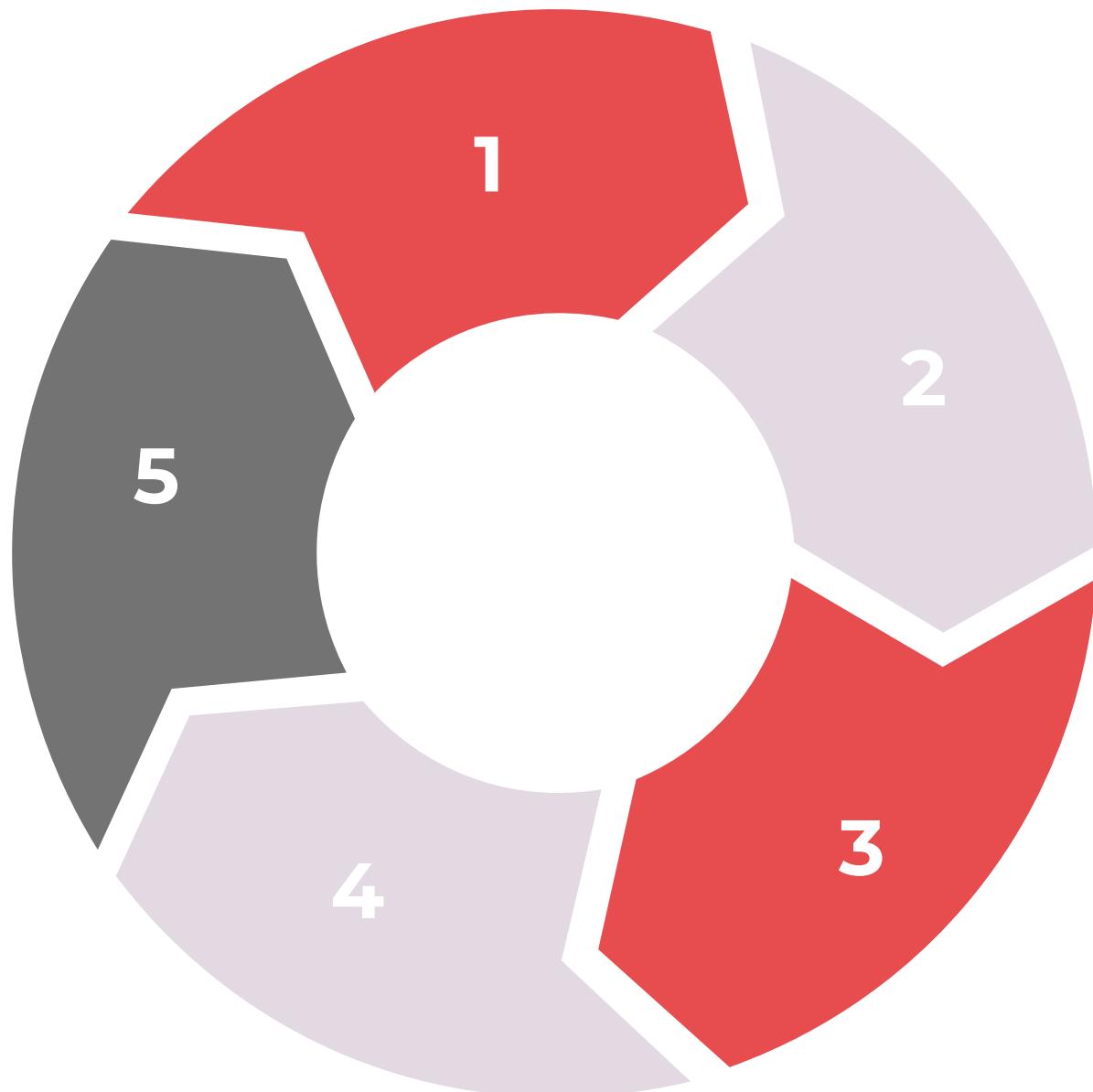


Any question guys ~

— Kaidah Dasar dalam NLP

MASIH INGAT MACHINE LEARNING LIFECYCLE ?

Proprietary document of Indonesia AI 2023

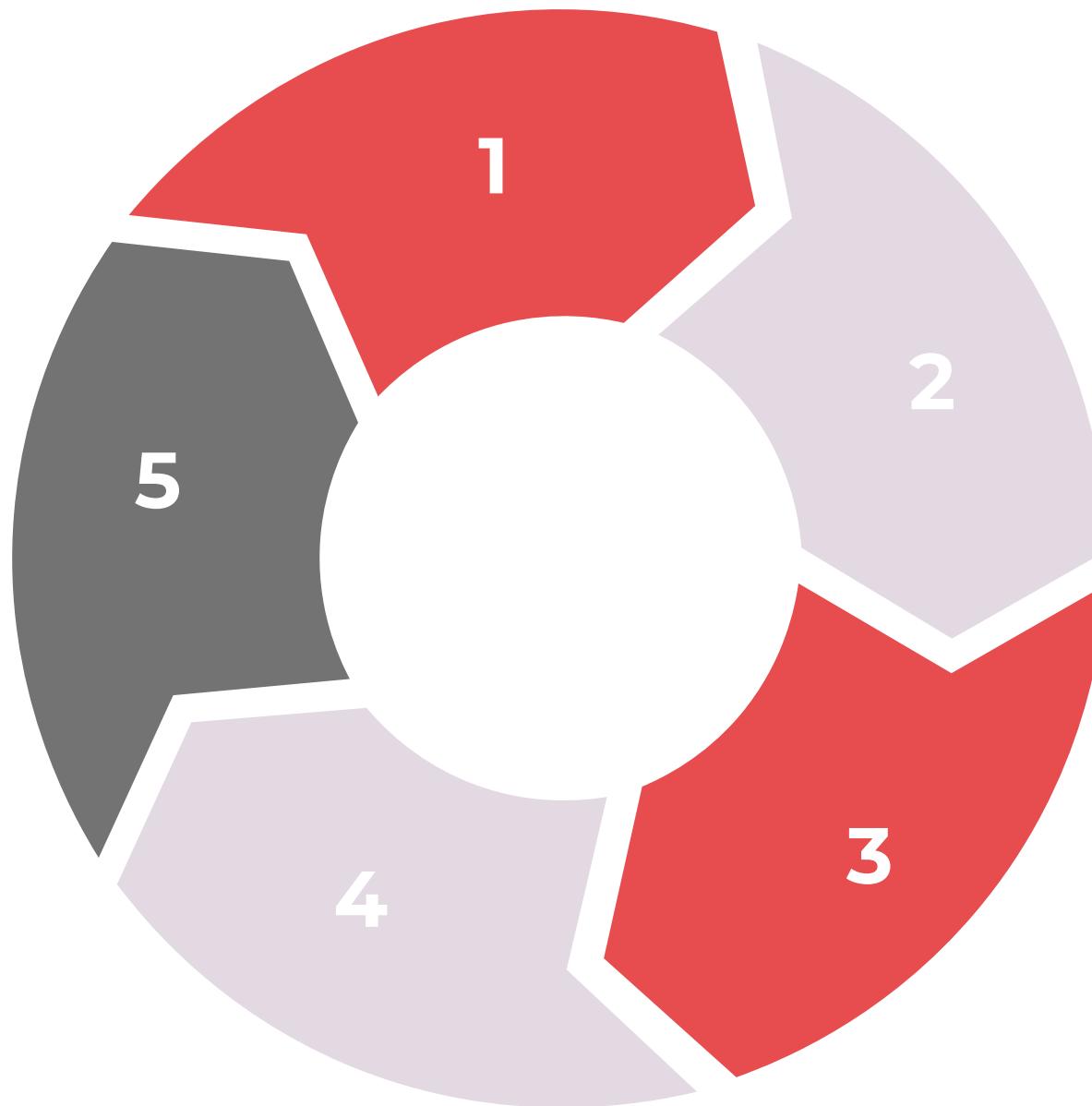


1. Data Collection
2. Data Preprocessing
3. Data Modeling
4. Model Evaluation
5. Model Deployment & Maintenance

Berlaku untuk projek AI/Machine Learning ataupun Data Science...

NLP LIFECYCLE

Proprietary document of Indonesia AI 2023

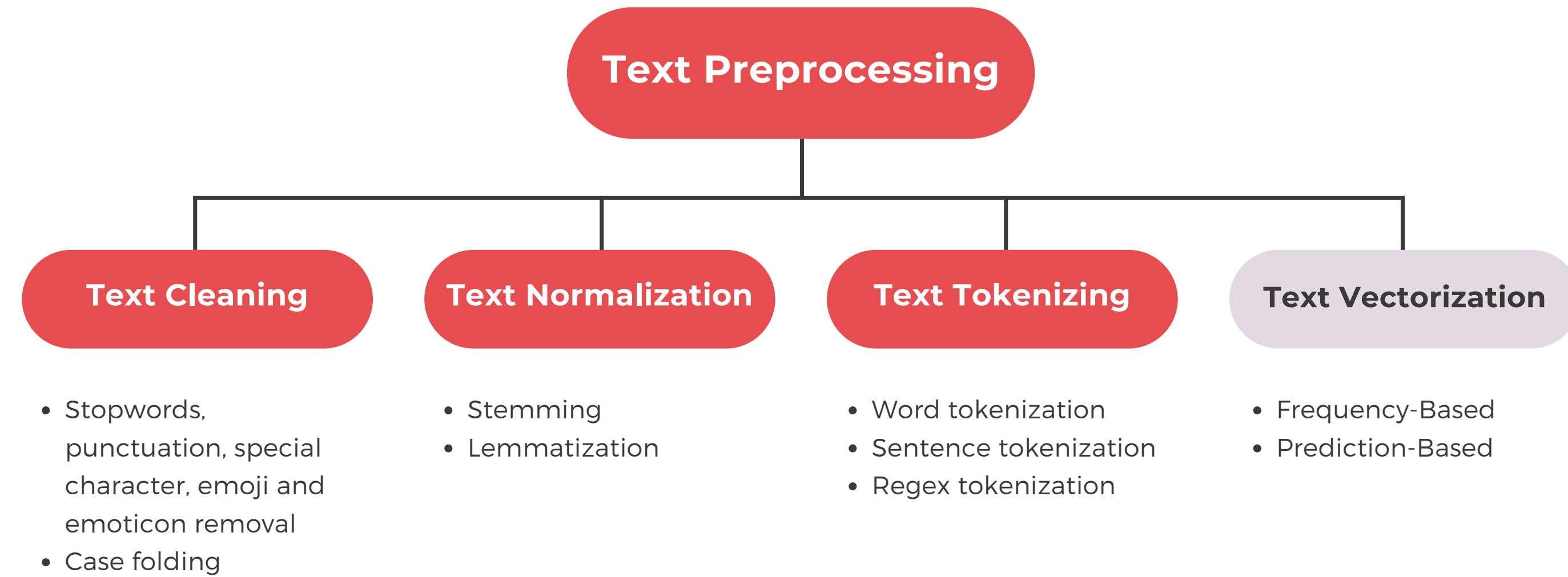


1. Data Collection
2. **Natural language Text Preprocessing**
3. Data Modeling
4. Model Evaluation
5. Model Deployment & Maintenance

Text preprocessing memberikan dampak yang signifikan pada hasil analisis karena dalam bahasa yang digunakan manusia, terdapat konteks dan makna yang berbeda ketika disajikan dalam bentuk yang sedikit berbeda.

TEKNIK PREPROCESSING

Proprietary document of Indonesia AI 2023



Text cleaning, text normalization, dan text tokenizing memanfaatkan teknik regex secara langsung.

DATA MODELING

Proprietary document of Indonesia AI 2023

Rule-based Approach

VS

Machine Learning
Approach

DATA MODELING

Proprietary document of Indonesia AI 2023

Rule-based Approach

VS

Machine Learning Approach

- Menggunakan aturan dan heuristik yang ditentukan secara manual.
- Mudah dipahami dan diinterpretasikan oleh manusia.
- Membutuhkan pengetahuan linguistik dan pemodelan yang mendalam.

- Menggunakan algoritma machine learning untuk mempelajari pola dan aturan bahasa secara otomatis.
- Dapat belajar pola bahasa secara otomatis dari data tanpa memerlukan pengetahuan linguistik yang mendalam.
- Relatif sulit untuk diinterpretasikan oleh manusia.

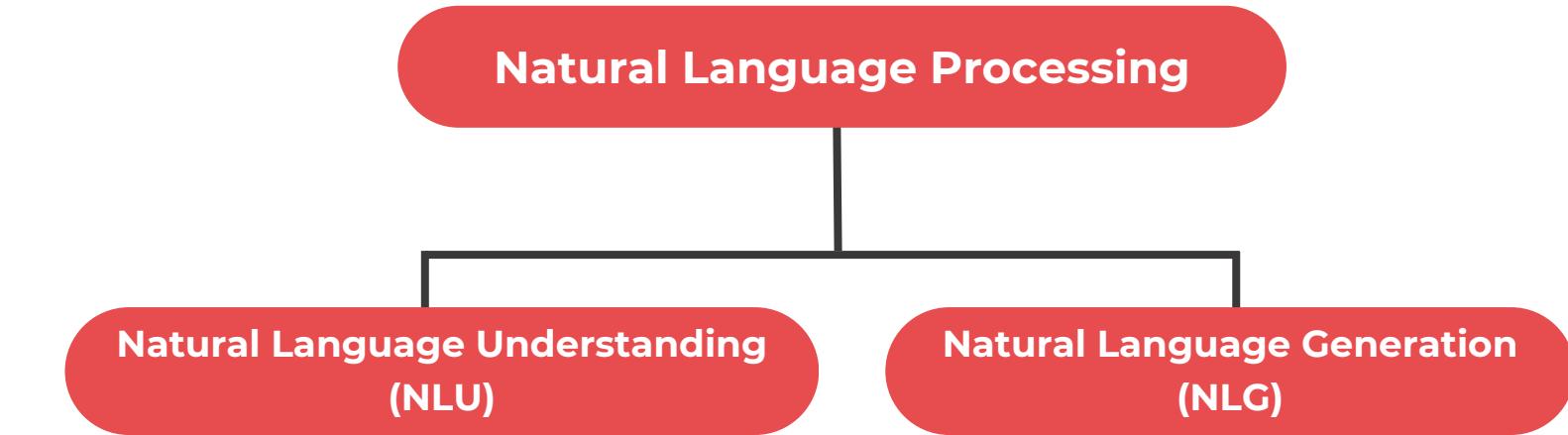
JENIS TASK NLP

Proprietary document of Indonesia AI 2023

Secara umum, task NLP dapat dibagi menjadi Natural Language Understanding (NLU) dan Natural Language Generation (NLG).

NLU berfokus pada **analisis teks** untuk mengekstrak informasi, mengenali entitas, mengidentifikasi tujuan, mengekstrak sentimen, memahami sintaksis, dan memodelkan wacana.

NLG berfokus pada **menghasilkan sebuah teks**. Output dari NLG dapat berupa teks lengkap, ringkasan, dialog, cerita, atau konten bahasa alami lainnya.



Any question guys ~

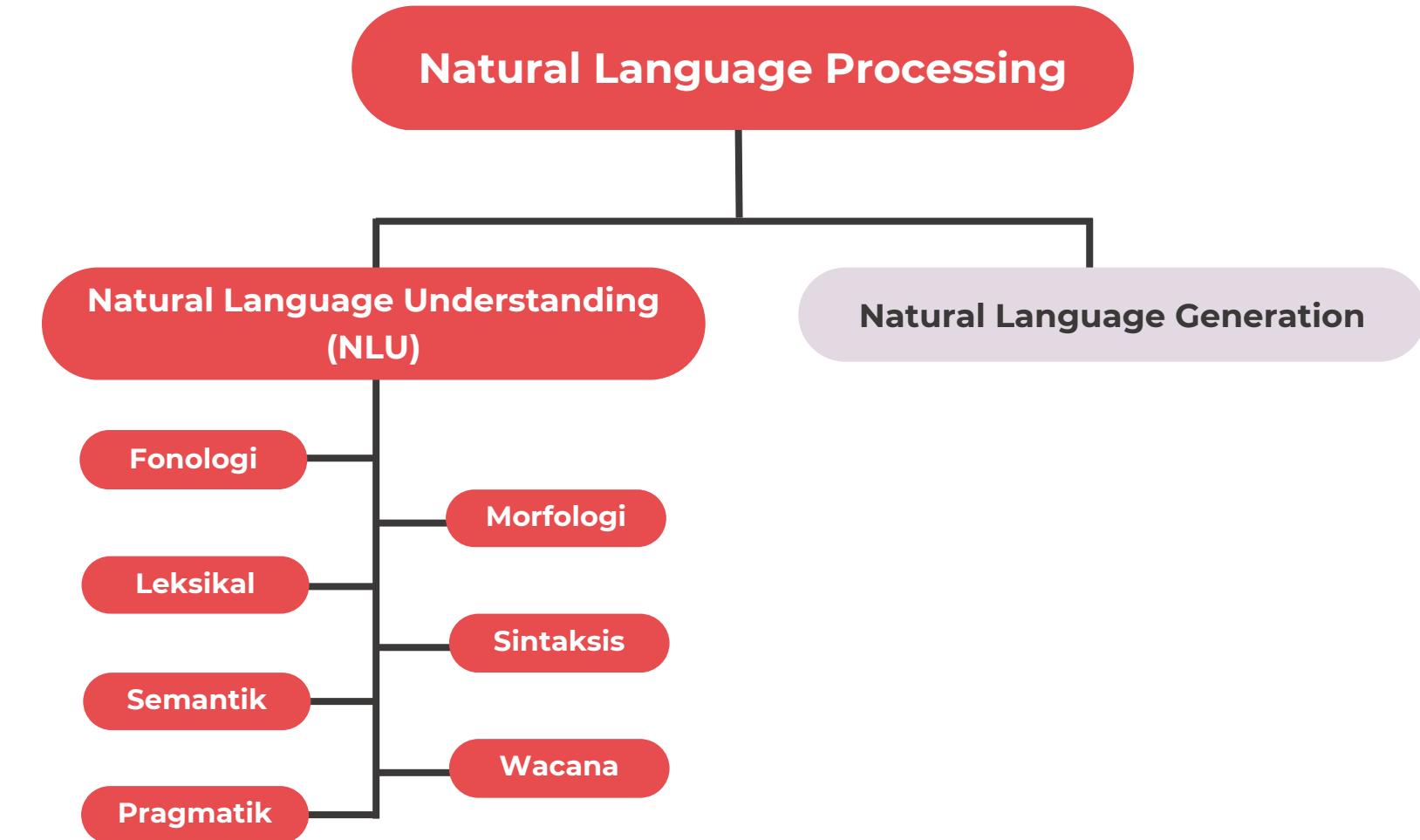
Aspek dasar dalam Ilmu Linguistik

LINGUISTIK

Proprietary document of Indonesia AI 2023

Linguistik adalah ilmu yang melibatkan arti bahasa, konteks bahasa, dan berbagai bentuk bahasa.

Oleh karena itu, penting untuk memahami berbagai terminologi penting pada linguistik dan tingkatannya yang berbeda agar ketika melakukan NLP, hasilnya dapat lebih baik dan akurat.



NATURAL LANGUAGE UNDERSTANDING (NLU)

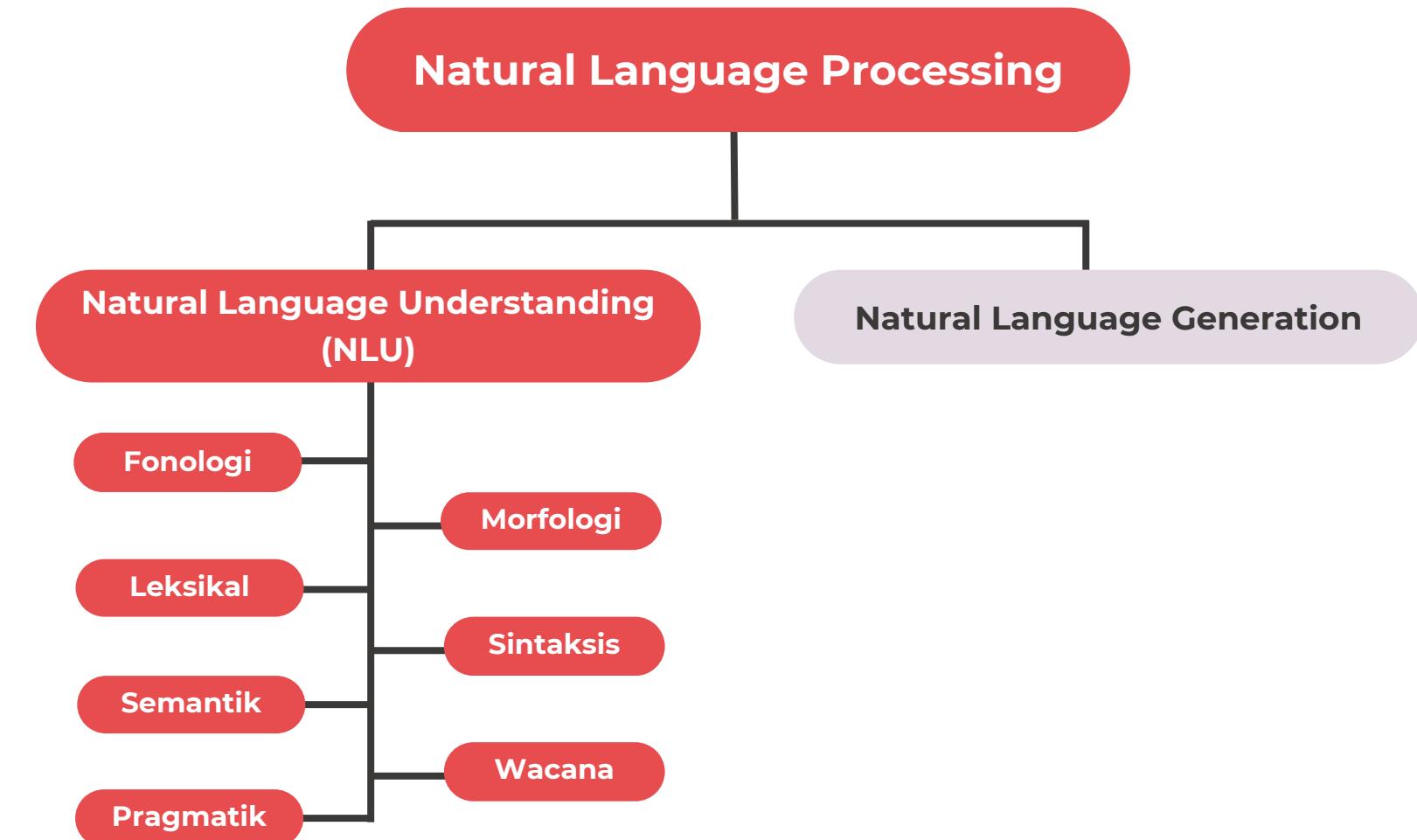
Proprietary document of Indonesia AI 2023

Natural Language Understanding/NLU memungkinkan mesin untuk memahami bahasa manusia dan menganalisisnya dengan mengekstrak konsep, entitas, emosi, kata kunci, dan lainnya.

Istilah dalam ilmu linguistik yang penting pada NLU:

1. Fonologi

Mempelajari suara-suara dalam bahasa dan cara mereka diatur. Contohnya, bagaimana bunyi "b" berbeda dengan bunyi "p".



NATURAL LANGUAGE UNDERSTANDING (NLU)

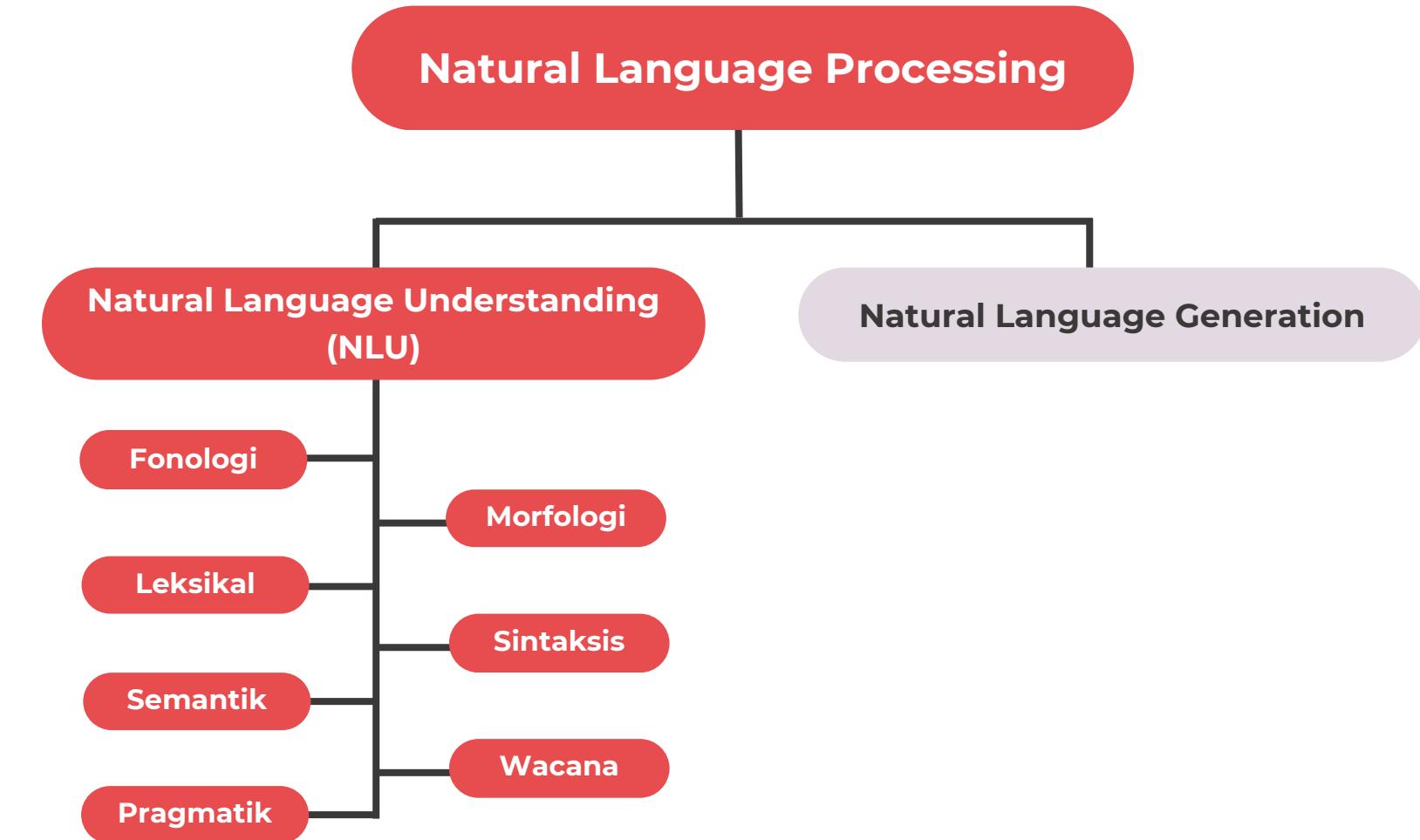
Proprietary document of Indonesia AI 2023

2. Morfologi

Mempelajari struktur kata-kata dalam bahasa. Contohnya, bagaimana kata-kata dibentuk oleh bagian-bagian kecil seperti awalan, akar kata, dan akhiran.

3. Leksikal

Mempelajari arti kata-kata dalam bahasa. Contohnya, memahami makna kata-kata secara individu tanpa mempertimbangkan kalimat keseluruhan.



NATURAL LANGUAGE UNDERSTANDING (NLU)

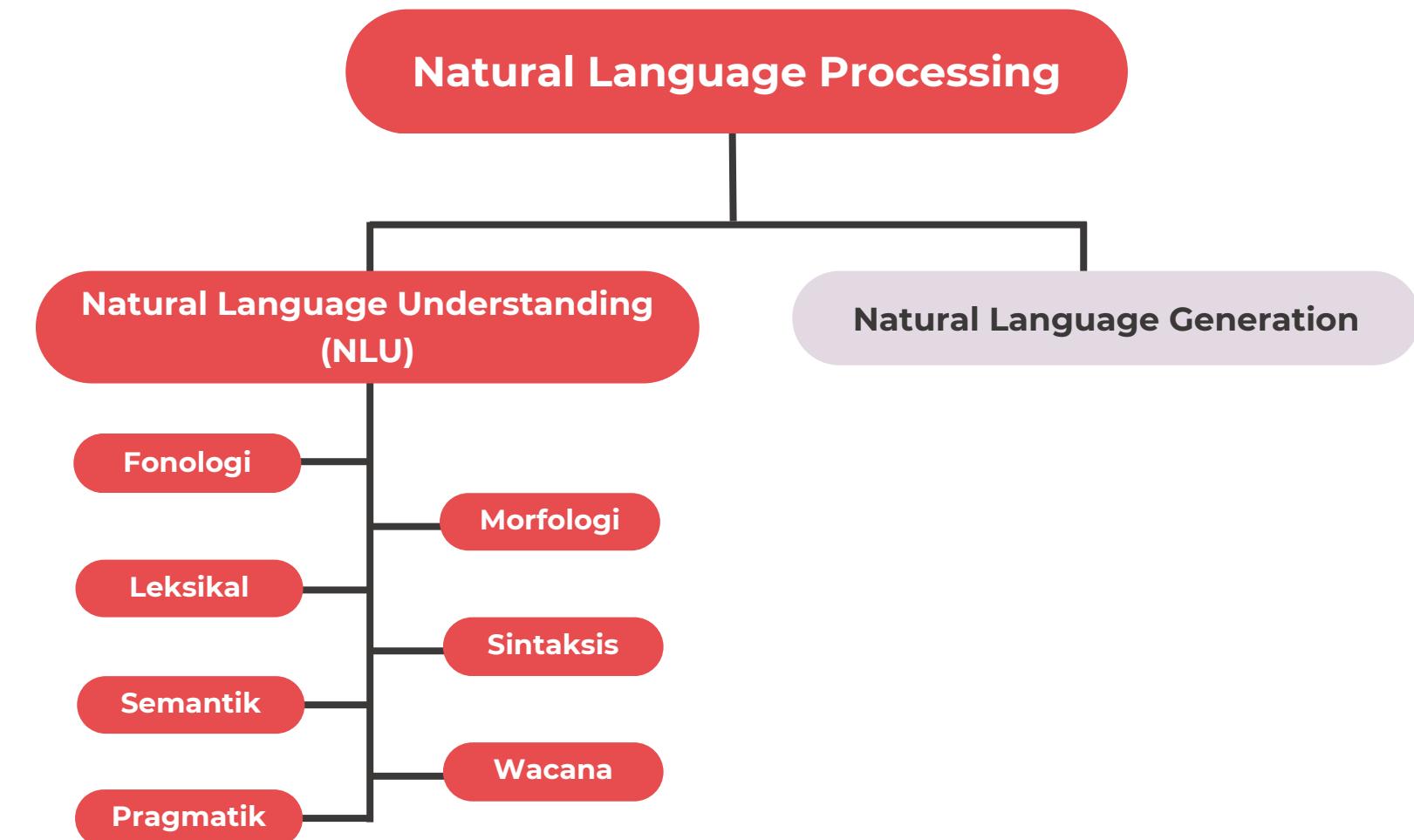
Proprietary document of Indonesia AI 2023

4. Sintaksis

Mempelajari tata bahasa dan bagaimana kata-kata diatur dalam kalimat. Contohnya, bagaimana kata-kata dihubungkan satu sama lain untuk membentuk kalimat yang bermakna.

5. Semantik

Mempelajari makna dalam bahasa. Contohnya, memahami makna kata-kata, frasa, dan kalimat serta hubungan antara mereka.



NATURAL LANGUAGE UNDERSTANDING (NLU)

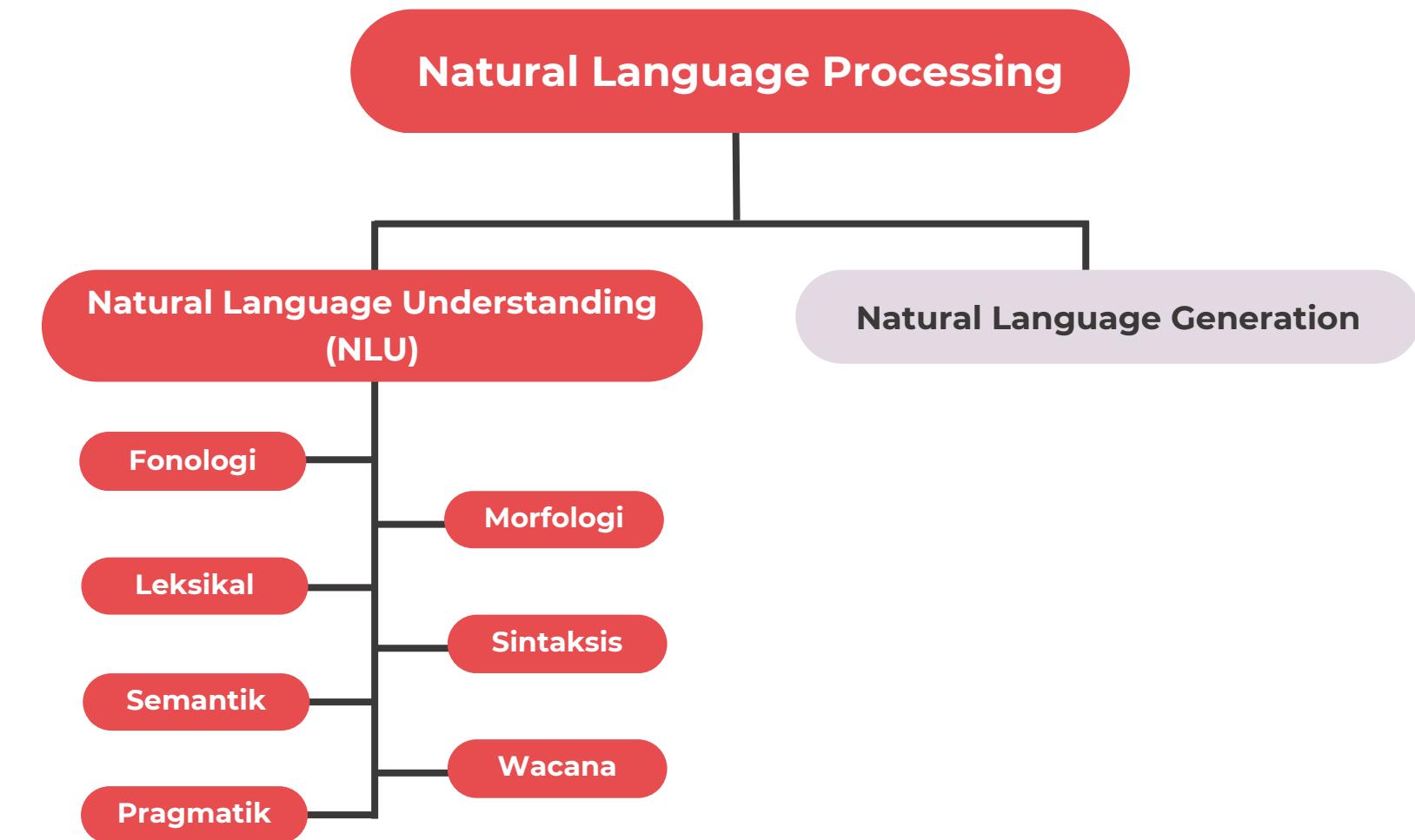
Proprietary document of Indonesia AI 2023

6. Wacana

Mempelajari bagaimana kalimat-kalimat membentuk teks atau percakapan yang lebih besar. Contohnya, bagaimana kalimat-kalimat saling terkait dan membentuk makna yang lebih kompleks.

7. Pragmatik

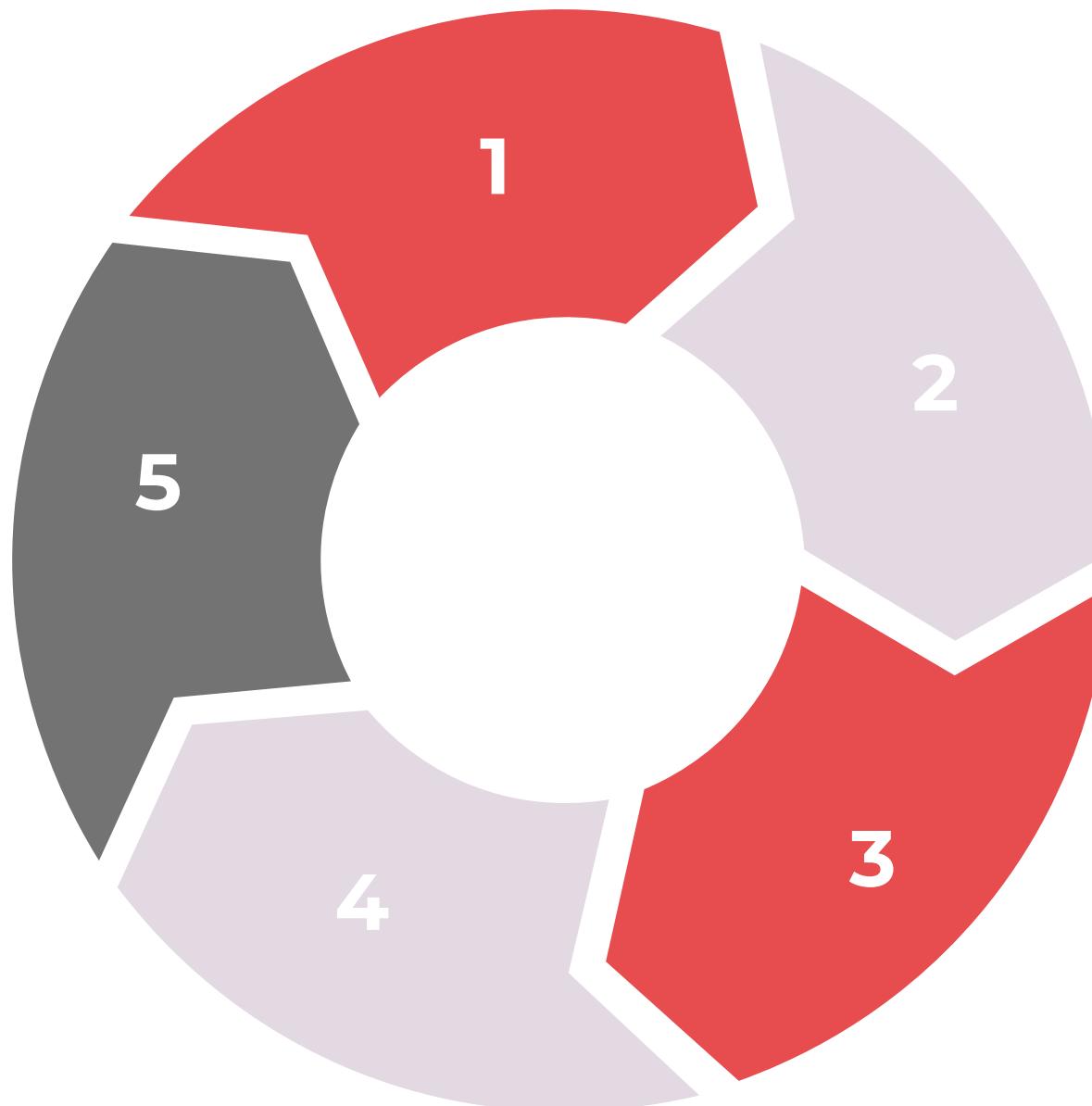
Mempelajari penggunaan bahasa dalam konteks sosial. Contohnya, memahami bagaimana makna dipengaruhi oleh konteks, tujuan komunikasi, dan pengetahuan latar belakang.



Any question guys ~

NLP LIFECYCLE

Proprietary document of Indonesia AI 2023



1. **Data Collection**
2. Natural language Text Preprocessing
3. Data Modeling
4. Model Evaluation
5. Model Deployment & Maintenance

Mari kita coba lakukan tahap pertama dalam setiap projek Machine learning.

Kali ini untuk melakukan tahapan data collection, kita akan mencoba melakukan web scraping sederhana~

Terimakasih!