Introduction to Natural Language Processing (NLP)

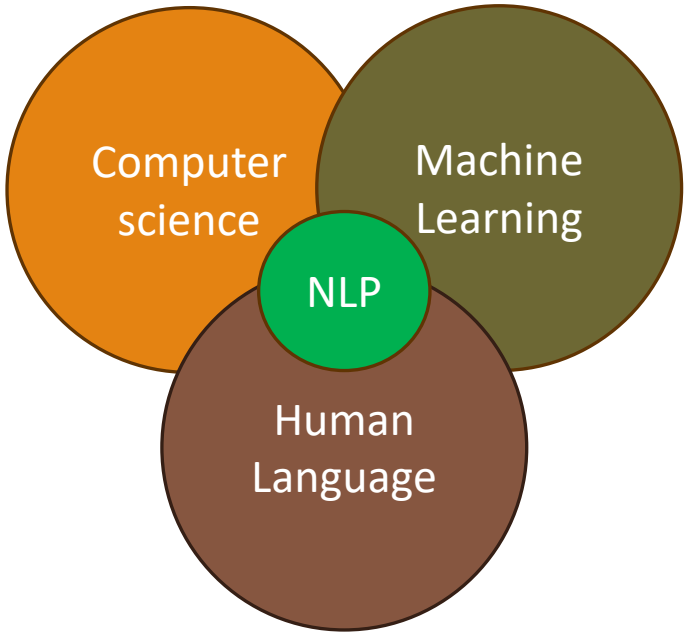
PREPARED BY: AHMAD ALAA ALDINE

PRESENTED BY: AHMAD ALAA ALDINE

What is NLP?

NLP is the intersection of computer science, human language, and machine learning.

Goal: Help computers to **understand** and **generate** human language as well as we do.



Where have you seen NLP used in your daily life?

Illustrating the interplay and overlap

AI

Machines that perform tasks needing intelligence.

Machine learning

Learning patterns from data.

Deep Learning

Multi-layered neural networks to learn patterns

Generative AI

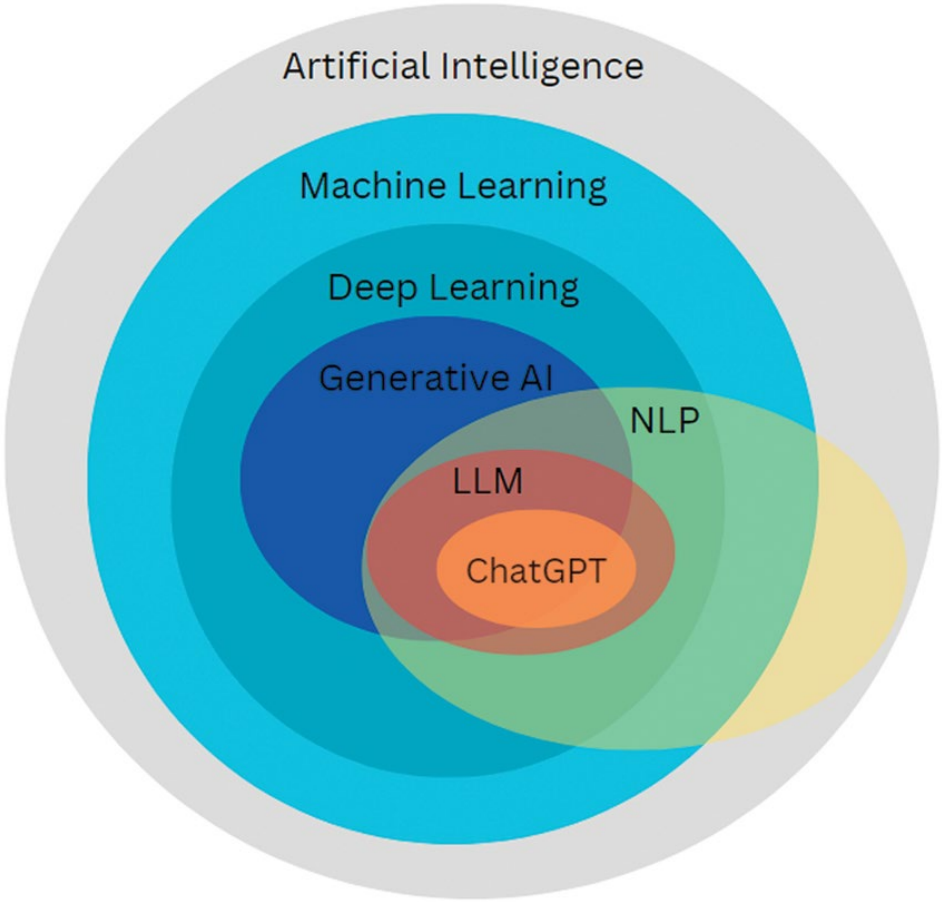
AI that generate contents in response to prompt.

NLP

Language is the data.

LLMs

Very large models trained on text.



NLP Use Cases

Question Answering

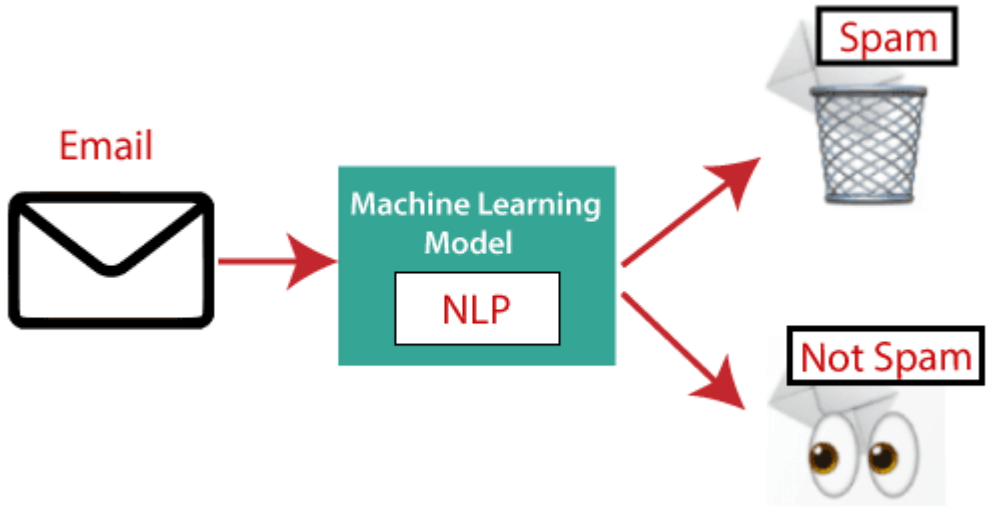
Question Answering focuses on building systems that automatically answer the questions asked by humans in a natural language.



NLP Use Cases

Spam Detection

Spam detection is used to detect unwanted e-mails getting to a user's inbox.



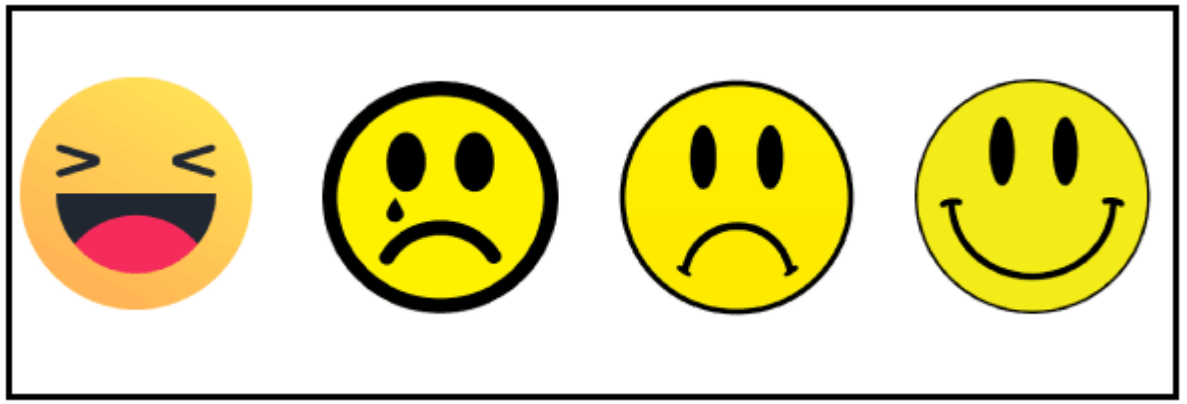
NLP Use Cases

Sentiment Analysis

It is used to analyze the attitude, behavior, and emotional state of a text writer.

This application is implemented through a combination of NLP (Natural Language Processing) and machine learning to:

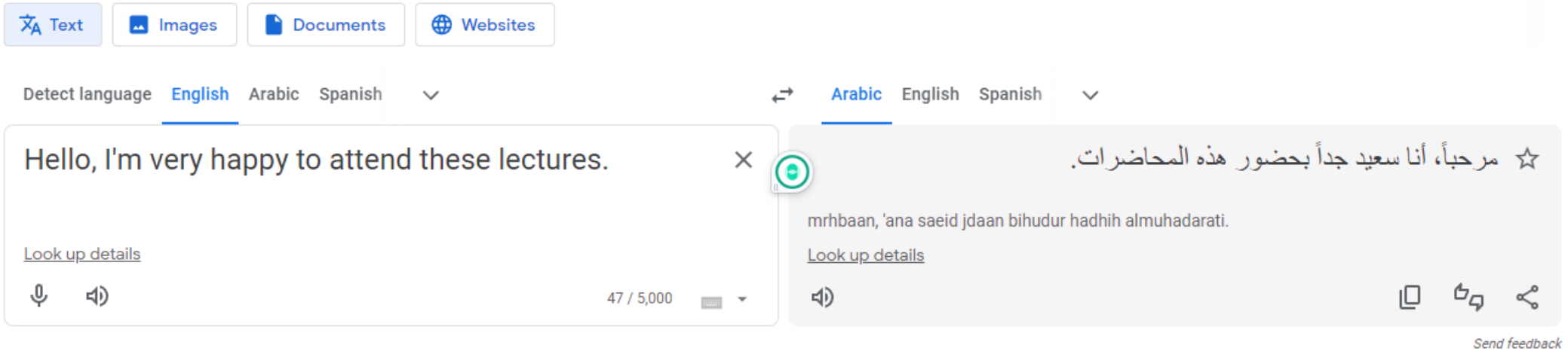
- classify the text as positive, negative, or neutral
- identify the mood of the context as happy, sad, angry, etc



NLP Use Cases

Machine Translation

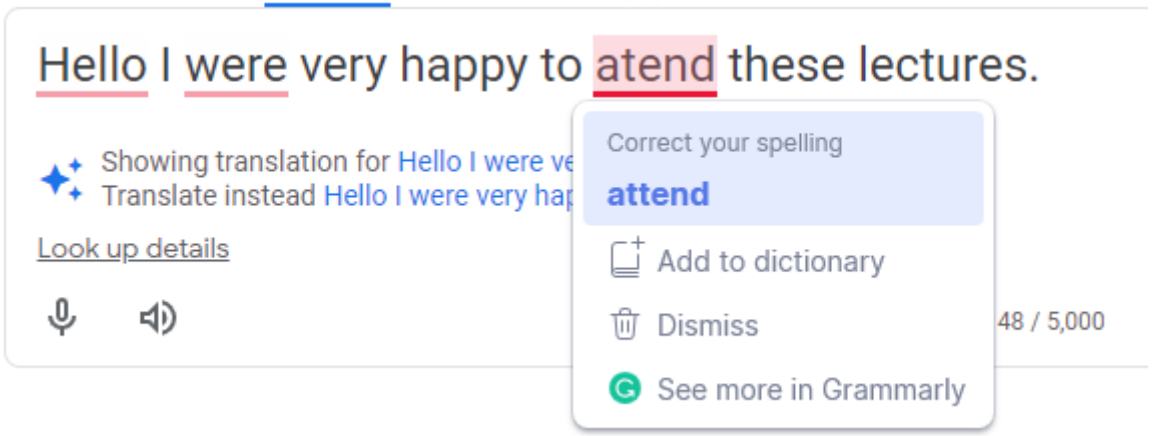
Machine translation is used to translate text or speech from one natural language to another natural language.



NLP Use Cases

Text Correction

Text correction is used to correct text from spelling, grammar, and punctuation mistakes.



NLP Use Cases

Information Extraction

Information extraction is one of the most important applications of NLP. It is used for extracting structured information from unstructured or semi-structured machine-readable documents.

Musk was born in Pretoria, South Africa, and briefly attended the University of Pretoria before immigrating to Canada at age 18, acquiring citizenship through his Canadian-born mother.

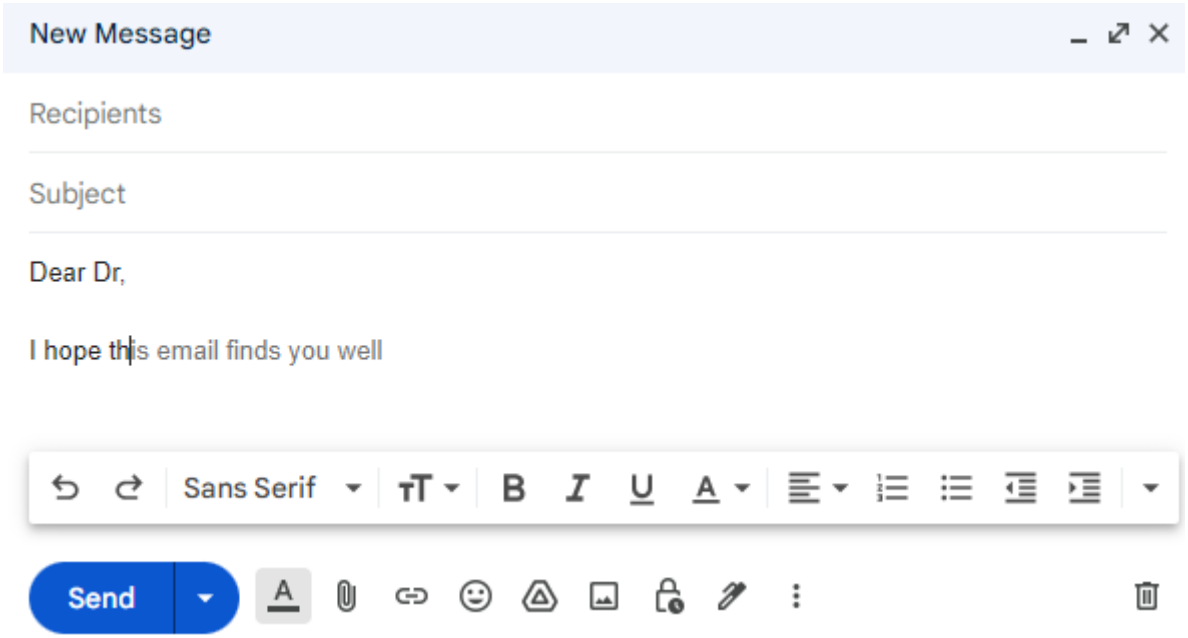
Born(Musk, Pretoria)

Hassan Kamel Al-Sabbah was a Lebanese electrical and electronics research engineer, mathematician and inventor. He was born in Nabatieh in present-day Lebanon.

Born(Hassan Kamel Al-Sabbah, Nabatieh)

NLP Use Cases

Text Generation



The NLP Explosion

From Hidden Technology to Everyday Intelligence

- Billions of users interact with NLP every day.
- Huge increase in publications, startups, and open-source models.

Why Now?

Why Now?

Textual Resources Availability

Amount of online textual data

- 70 billion web-pages online
- 55 million Wikipedia articles

Growing fast

- 9000 tweets/second
- 3 million mail/second

Why Now?

Free Available Pretrained Models

Fast text

- https://fasttext.cc/

Tensorflow hub

- https://www.tensorflow.org/hub

GPT3

- https://platform.openai.com/docs/models/gpt-3

Why Now?

Open ecosystem

















Why Now?

Cheap and High Performance Hardware







Why Now?

Cloud Resources







Levels of linguistic analysis

"The students read the book."

Morphology	students = student + s
Syntax	subject, verb, object
Semantics	who did what to what
Pragmatics	what the speaker intends

Why NLP is Hard?

1. Productivity
   - New words, senses, structure are introduced in languages all the time
   - Examples: social distance was added to the Oxford Dictionary in 2021
2. Diversity
   - About 7000 languages spoken in the world
3. Ambiguity
   - Lexical Ambiguity: It was cool. → Temperature or approval?
   - Syntactic Ambiguity: I saw the girl with the binocular. → Who has them?
   - Referential Ambiguity: Kiran went to Sunita. She said, I am hungry. → Who is “she”?

Arabic: a case study

Diacritics

كَتَبَ(he wrote) and كُتُب(books) share the same letters.

Morphology

One root ك-ت-ب gives write, book, library, writer.

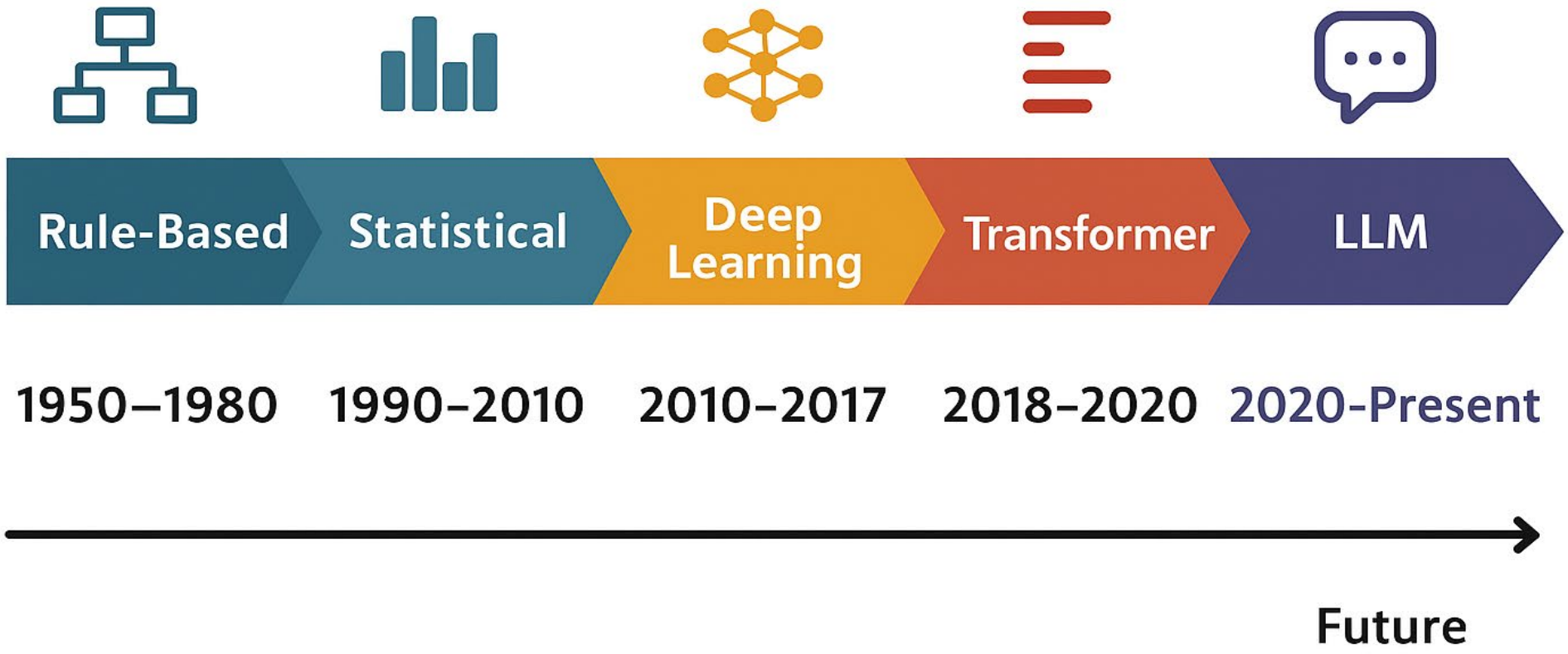
Dialects

Lebanese, Egyptian and Gulf Arabic differ from Modern Standard.

Code-switching

“Hi, kifak? Ça va?” mixes three languages.

EVOLUTION OF NLP



Evolution of NLP

Rules and heuristics	Machine Learning	Deep Learning	LLMs
• Regular expression • Morphological Analysis • Syntactic Analysis • Semantic Analysis	• Count Vectorizer • Tf-idf Vectorizer • SVM • Naïve Bayes • KNN • Decision Tree	• RNN • CNN • LSTM • Word2Vec	• Transformers • BERT • GPT • Prompt Engineering • RAG • Agentic AI

From rules to LLMs

Era	Key idea	Limit
Rules	Hand-written patterns	Fragile, hard to scale
Statistical	Counts from large corpora	Weak on meaning
Neural	Learned word vectors	Long text is hard
Transformers	Attention over the whole text	Costly to train
LLMs	Huge models, simple prompts	Hallucination, cost