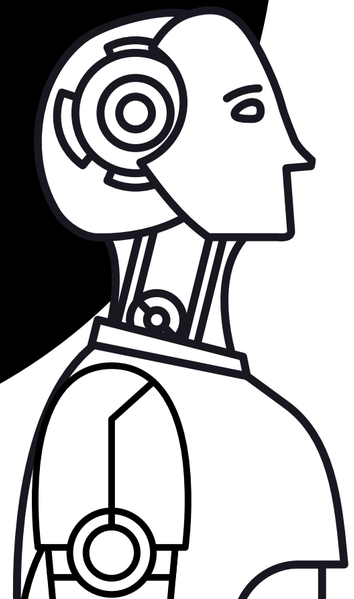


INTRODUCTION TO

NATURAL LANGUAGE PROCESSING (NLP)

WITH

HUGGING FACE TRANSFORMERS

(A GUIDED PROJECT)

PREPARED BY : NUR AISYAH ATHIRAH BINTI HISHAMUDDIN

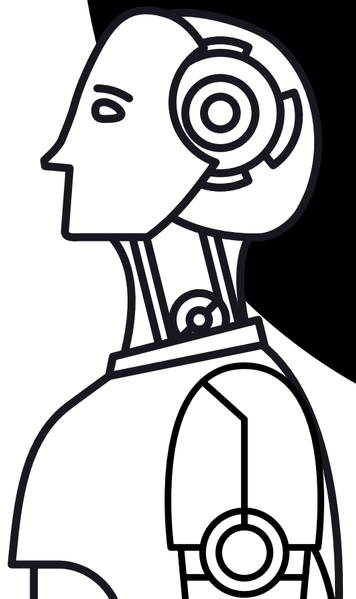


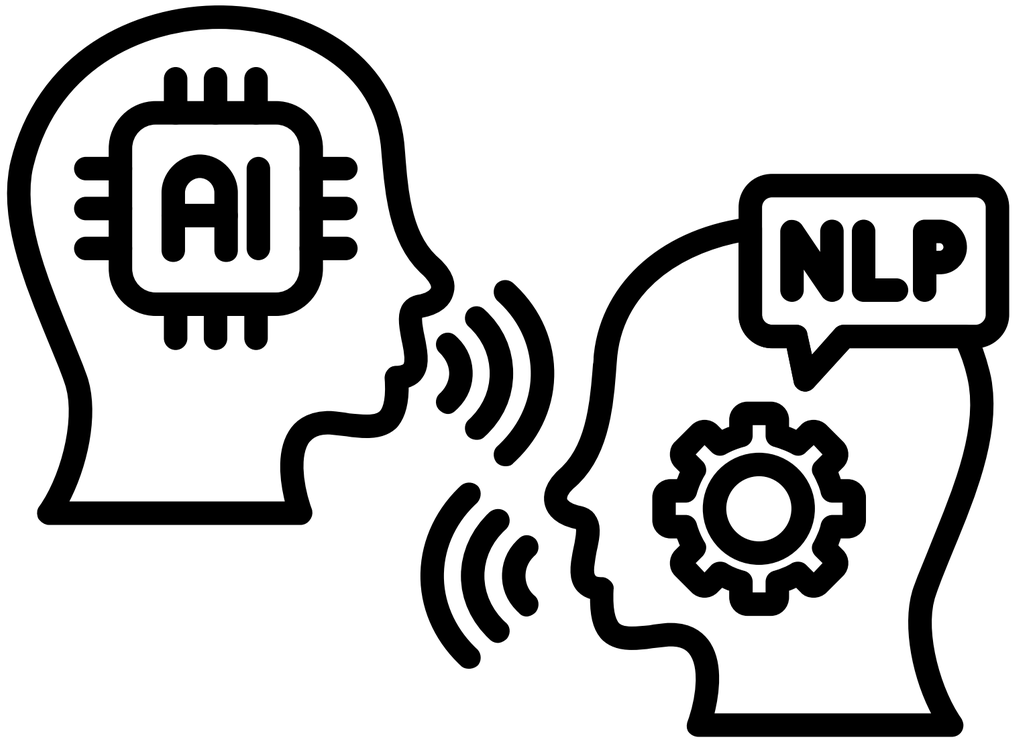
TABLE OF CONTENTS

NATURAL LANGUAGE PROCESSING (NLP)

A subfield of computer science and artificial intelligence (AI) that uses machine learning to enable computers to understand and communicate with human language.

Benefits :

- Improved Communication
- Enhanced Data Analysis
- Increased Efficiency



HUGGING FACE TRANSFORMERS

An open-source framework for deep learning created by Hugging Face. offering APIs and tools for using and fine-tuning state-of-the-art pre-trained models across NLP, computer vision, audio, and multi-modal tasks.

HUGGING FACE

IMPORTED LIBRARIES

- **PIPELINE**

  Allows you to sequentially apply a list of transformers to preprocess the data and, if desired, conclude the sequence with a final predictor for predictive modeling.

- **AUTOTOKENIZER**

  A generic tokenizer class that will be instantiated as one of the tokenizer classes of the library when created with the AutoTokenizer.from_pretrained(pretrained_model_name_or_path) class method.

- **AUTOMODEL**

  A generic model class that will be instantiated as one of the base model classes of the library when created with the AutoModel.from_pretrained(pretrained_model_name_or_path) or the AutoModel.from_config(config) class methods.

SENTIMENT ANALYSIS

WHAT?

NLP technique used to identify and analyze the **sentiment expressed** in text, classifying it as **positive**, **negative**, or **neutral**.

WHY?

To understand customer emotions, improve services, and track brand reputation.

WHO?

- Businesses
- Marketers
- Analyst
- Researchers

WHERE?

- Social Media
- Surveys
- Support Chats
- Emails
- Reviews

WHEN?

During product reviews, social media monitoring and customer feedbacks.

HOW?

1. Load sentiment analysis pipeline from pipeline() using "**sentiment-analysis**" task identifier.
2. Use, "**distilbert-base-uncased-finetuned-sst-2-english**" model for sentiment analysis. [Alternative model : **cardiffnlp/twitter-roberta-base-sentiment**]
3. Input sentence (e.g: random product review) into selected classifier

Output :

The sentence will be classified as POSITIVE / NEGATIVE / NEUTRAL with accuracy score.

Alternative Output :

The sentence will be classified as 0 -> Negative / 1 -> Neutral / 2 -> Positive with accuracy score.

TOPIC CLASSIFICATION

WHAT?

NLP technique that automatically **assigning categories** or labels to a text **based on its content**. For example, a news article might be classified under "Politics", "Health", or "Technology". This helps in grouping and organizing large volumes of text data.

WHY?

To automate content organization and improve searchability.

WHO?

- News Agencies
- Customer Support
- Researchers
- App Developers

WHERE?

- Forums
- News Articles
- Support Systems
- Academic Databases

WHEN?

It is used in systems like news aggregators or when organizing documents, emails, or user queries.

HOW?

1. Load text classification pipeline from pipeline() using "**zero-shot-classification**" task identifier.
2. Use, "**facebook/bart-large-mnli**" model for topic classification.
3. Input sentence (e.g: text paragraph) into selected classifier and a list of candidate labels.

Output :

The model assign corresponding labels to the input. Among the label candidates, the most accurate one for the input will have the highest score.

TEXT GENERATOR

WHAT?

NLP models to automatically create **coherent, human-like text** based on a given prompt or context.

WHY?

To save time, improve productivity or create engaging content

WHO?

- Content Creators
- Chatbot Developers
- Writers
- Educators

WHERE?

- Messaging Apps
- Blogs
- Storytelling Tools
- Code Assistant

WHEN?

It is used for writing stories, answering questions, generating code, or completing sentences.

HOW?

1. Load text generator pipeline from pipeline() using "**text-generation**" task identifier.
2. Use, "**gpt2**" model for text generator. [Alternative model : “**distilgpt2**”]
3. Input sentence (e.g: incomplete sentence) into selected generator with parameters, such as length and number of the sentences needed.

Output:

The model continued the incomplete sentence into full sentence.

NAME ENTITY RECOGNITION (NER)

WHAT?

NLP technique that scans text to identify and **classify named entities** such as people's names, locations, organizations, dates, monetary values, and more.

WHY?

To extract structured information from unstructured text

WHO?

- Researchers
- Data Scientist
- Legal Professionals
- Medical Profesionals

WHERE?

- News Articles
- Legal Documents
- Research Papers

WHEN?

During information extraction or knowledge base creation.

HOW?

1. Load text generator pipeline from pipeline() using "**text-generation**" task identifier.
2. Use, "**dbmdz/bert-large-cased-finetuned-conll03-english**" model for NER.
3. Input sentence (e.g: text paragraph) into NER.

Output:

The model identifies all entities (PER, ORG, LOC) in the sentence with highest confidence score.

QUESTION ANSWERING

WHAT?

NLP model that **responds to user questions** by either **extracting answers** from a given context or **generating answers**. It mimics how humans answer questions based on reading comprehension.

WHY?

To understand the context and infer the correct answer quickly and logically while maintaining accurate information.

WHO?

- Developers of AI assistant
- Search Engines
- Customer Service Bots

WHERE?

- Websites
- Mobile Apps
- Knowledge Bases

WHEN?

It is commonly used in virtual assistants, chatbots, search engines, and automated FAQ systems.

HOW?

1. Load question answering pipeline from pipeline() using "**question-answering**" task identifier.
2. Use, "**distilbert-base-cased-distilled-squad**" model for question answering.
3. Input the question and context of the question.

Output:

The model extract the correct answer with its confidence score.

TEXT SUMMARIZATION

WHAT?

NLP technique that uses the algorithm to **reduce a large text** document into a shorter version while **preserving its main ideas** and key information.

WHY?

To quickly understand key information without reading the entire content

WHO?

- Researchers
- Journalist
- Students
- Busy Professionals

WHERE?

- News Sites
- Academic Tools
- Corporate Reports

WHEN?

It’s particularly useful when in need for speeding up reading and decision-making processes in domains like journalism, academic research, and legal work.

HOW?

1. Load text summarization pipeline from pipeline() using "**summarization**” task identifier.
2. Use, "**sshleifer/distilbart-cnn-12-6**" model for the summarizer.
3. Input sentence (e.g: text paragraph) into selected summarizer.

Output:

The model produce a short summary of the text paragraph.

TRANSLATION

WHAT?

NLP technique that able to automatically **converting text from one language to another**, maintaining the original meaning and context.

WHY?

This is essential for cross-lingual communication, enabling access to content and services globally.

WHO?

- Businesses
- Travelers
- Language Learners
- Global Content Creators

WHERE?

- Websites
- Translator Apps
- Customer Support
- Education Platforms

WHEN?

It used during international communication especially at the embassy or for travelers to understand local language and ease the mobility throughout the travelling journey.

HOW?

1. Load translator pipeline from pipeline() using "**translation**" task identifier. [Alternative identifier : "**translation_en_to_fr**"]
2. Use a specific model that is from one specific language to another (e.g : French-English), "**Helsinki-NLP/opus-mt-fr-en**" for the translator. [Alternative model : "**t5-small**"]
3. Input the text that want to be translated.

Output:

The model translate the sentence from English to French.

THANK YOU FOR YOUR ATTENTION

Thank you for joining this exploration on Natural Language Processing.
I hope this presentation has inspired new ideas and perspectives about how technology shapes our world.

IMPLEMENTATION :

https://github.com/Aisyah-Athirah/Introduction_to_NLP