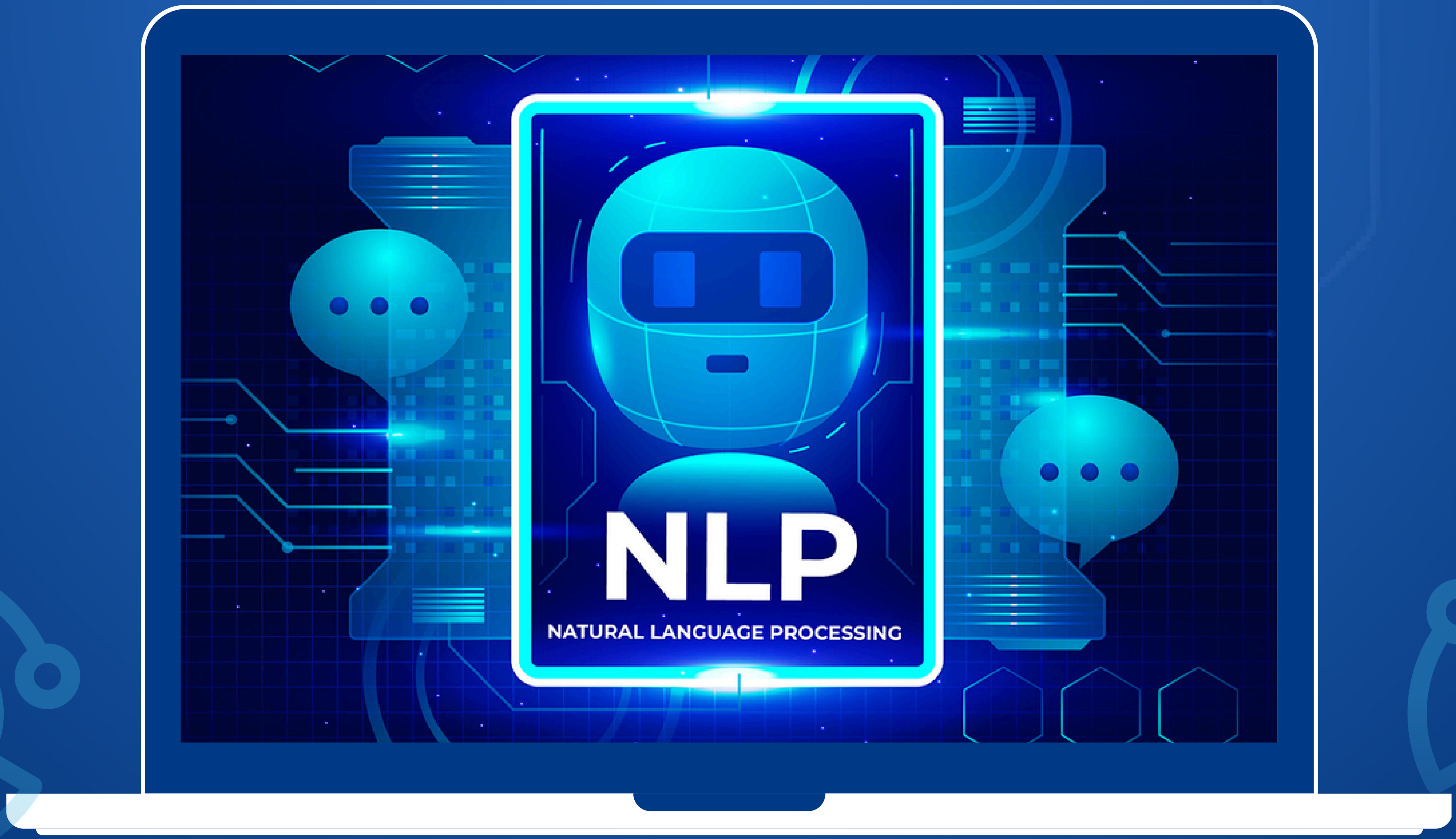
TOKENIZATION IN NLP

Let's explore the different types of tokenization



✉ statfusionai@gmail.com

▶ statfusionai

WHY TOKENIZATION?

Before AI can understand text, it must break it down into meaningful units—words, sentences, or subwords. Tokenization is the first step in **Natural Language Processing (NLP)**, enabling search engines, chatbots, and voice assistants to make sense of human language.

📊 From simple word splitting to advanced subword processing, tokenization shapes how machines read, analyze, and generate text.

1. WORD TOKENIZATION

Definition

Splits text into individual words, considering spaces and some punctuation.

Example

📌"Machine learning is powerful!"

🔹 Word Tokenization:

["Machine", "learning", "is", "powerful", "!"]

Use Cases

✅ Used in search engines, chatbots, text classification.

✅ Simple and fast but struggles with hyphenated words and abbreviations.

2. WHITESPACE TOKENIZATION

Definition

Splits text using spaces only, without handling punctuation.

Example

📌 Machine learning is powerful!

🔹 Whitespace Tokenization:

["Machine", "learning", "is", "powerful!"]

Use Cases

✅ Used in simple text analysis, search engines.

✅ Not ideal for complex languages with no spaces (e.g., Chinese, Japanese).

3.SENTENCE TOKENIZATION

Definition

Splits text into individual sentences.

Example

📌 Dr. Smith is an AI expert. He works in NLP.

🔹 Sentence Tokenization:

["Dr. Smith is an AI expert.", "He works in NLP."]

Use Cases

✅ Used in summarization, text parsing, and translation.

✅ Handles sentence boundaries but struggles with abbreviations like "U.S.A.".

4.RULE-BASED TOKENIZATION

Definition

Uses predefined rules to handle specific cases like dates, abbreviations, and proper nouns.

Example

📌 Prof. Dumbledore lives at No. 4, Privet Drive.

🔹 Rule-Based Tokenization:

["Prof. Dumbledore", "lives at", "No. 4,", "Privet Drive."]

Use Cases

✅ Used in legal, medical, and financial documents.

✅ More accurate but requires extensive rule writing.

5.CHARACTER TOKENIZATION

Definition

Splits text into individual characters.

Example

📌 AI!

🔹 Character Tokenization: **["A", "I", "!"]**

Use Cases

✅ Used in OCR (Optical Character Recognition), spell-checkers, and CAPTCHA solvers.

✅ Works for languages without spaces (e.g., Chinese, Japanese, Korean).

6. BYTE-PAIR ENCODING (BPE) TOKENIZATION

Definition

A subword tokenization method that merges the most common character pairs iteratively.

Example

📌 unhappiness

🔹 BPE tokens: **["un", "happiness"]**

Use Cases

✅ Used in GPT, BERT, and multilingual NLP models.

✅ Helps handle rare words and new vocabulary.

7. MORPHOLOGICAL TOKENIZATION

Definition

Breaks words into their smallest meaningful units (morphemes).

Example

📌 unhappiness

🔹 Morphological Tokens: **["un", "happy", "ness"]**

Use Cases

✅ Used in linguistics, machine translation, and low-resource languages.

✅ Works well for agglutinative languages (e.g., Turkish, Finnish, Korean).

8. PHONETIC TOKENIZATION

Definition

Groups words that sound similar for better speech recognition.

Example

📌 **"AI"** and **"Aye"** may be mapped to the same phonetic token.

Use Cases

✅ Used in speech recognition, voice assistants, and search engines.

✅ Helps in handling homophones (e.g., "night" vs. "knight").

9. PUNCTUATION-BASED TOKENIZATION

Definition

Separates words while preserving punctuation marks as separate tokens.

Example

📌 Hello, how are you?

🔹 Punctuation-Based Tokenization:

["Hello", ",", "how", "are", "you", "?"]

Use Cases

✅ Used in chatbots, text analysis, and parsing dialogues.

✅ Improves sentiment analysis by handling punctuation emphasis.

10. N-GRAM TOKENIZATION

Definition

Creates sequences of N words for better context understanding.

Example

📌 Natural Language Processing

🔹 2-gram (Bigram):

[("Natural", "Language"), ("Language", "Processing")]

🔹 3-gram (Trigram):

[("Natural", "Language", "Processing")]

Use Cases

✅ Used in text prediction (autocomplete), machine translation, and speech recognition.

✅ Helps AI understand word relationships.

11. HYBRID TOKENIZATION

Definition

Combines multiple tokenization techniques (e.g., word + punctuation + BPE).

Example

📌 Don't tokenize me!

🔹 Hybrid Tokenization:

["Do", "n't", "tokenize", "me", "!"]

Use Cases

✅ Used in advanced NLP models like ChatGPT and Google Translate.

✅ Balances speed, accuracy, and efficiency.

SUMMARY TABLE

Tokenization Type	Key Feature	Used In
Word Tokenization	Splits by spaces	Search Engines, NLP
Whitespace Tokenization	Ignores punctuation	Simple Text Processing
Sentence Tokenization	Splits by sentences	Summarization, Chatbots
Rule-Based Tokenization	Uses grammar rules	Legal, Medical NLP
Character Tokenization	Splits into characters	OCR, Spell-checkers
Byte-Pair Encoding (BPE)	Subword merging	GPT, BERT
Morphological Tokenization	Extracts root words	Linguistics, Translation
Phonetic Tokenization	Groups similar sounds	Speech Recognition
Punctuation-Based Tokenization	Preserves punctuation	Sentiment Analysis
N-Gram Tokenization	Sequences of words	Autocomplete, AI Models
Hybrid Tokenization	Combination method	Advanced NLP

STATFUSION AI

FOLLOW US ON

/ STATFUSIONAI

statfusionai@gmail.com

/statfusionai