Introduction to Natural Language Processing

Agenda

- Natural Language Processing Quiz
- NLP and its applications
- Text Cleaning
- Bag of Words (BOW) model
- n-grams model
- Sentiment Analysis

Let's begin the discussion by answering a few questions on natural language processing (NLP) and text cleaning

NLP and Text Cleaning Quiz

What is the primary goal of Natural Language Processing (NLP)?

A. To translate human languages into computer programming languages

B. To enable computers to understand, interpret, and generate human language

C. To convert speech into text documents

D. To create artificial languages for communication between computers

NLP and Text Cleaning Quiz

What is the primary goal of Natural Language Processing (NLP)?

A. To translate human languages into computer programming languages

B. To enable computers to understand, interpret, and generate human language

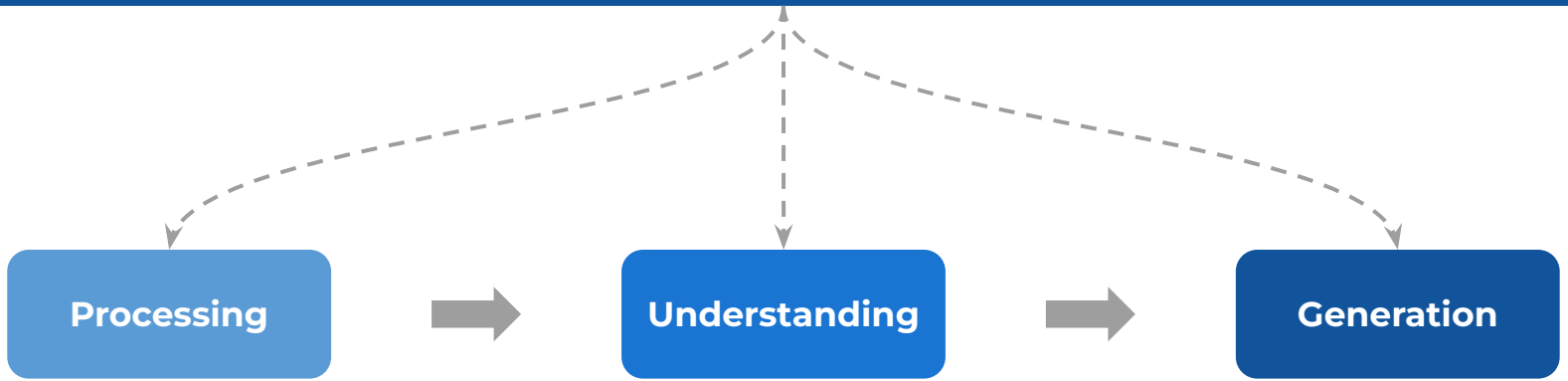
C. To convert speech into text documents

D. To create artificial languages for communication between computers

Natural Language Processing

Branch of artificial intelligence (AI) that deals with the interaction between machines and human languages

Aims to **automate the reading, interpretation, and understanding of human language**, also called natural language



NLP and Text Cleaning Quiz

Which of the following are the applications of Natural Language Processing?

A. Sentiment Analysis

B. Machine Translation

C. Chatbot

D. Document Summarization

NLP and Text Cleaning Quiz

Which of the following are the applications of Natural Language Processing?

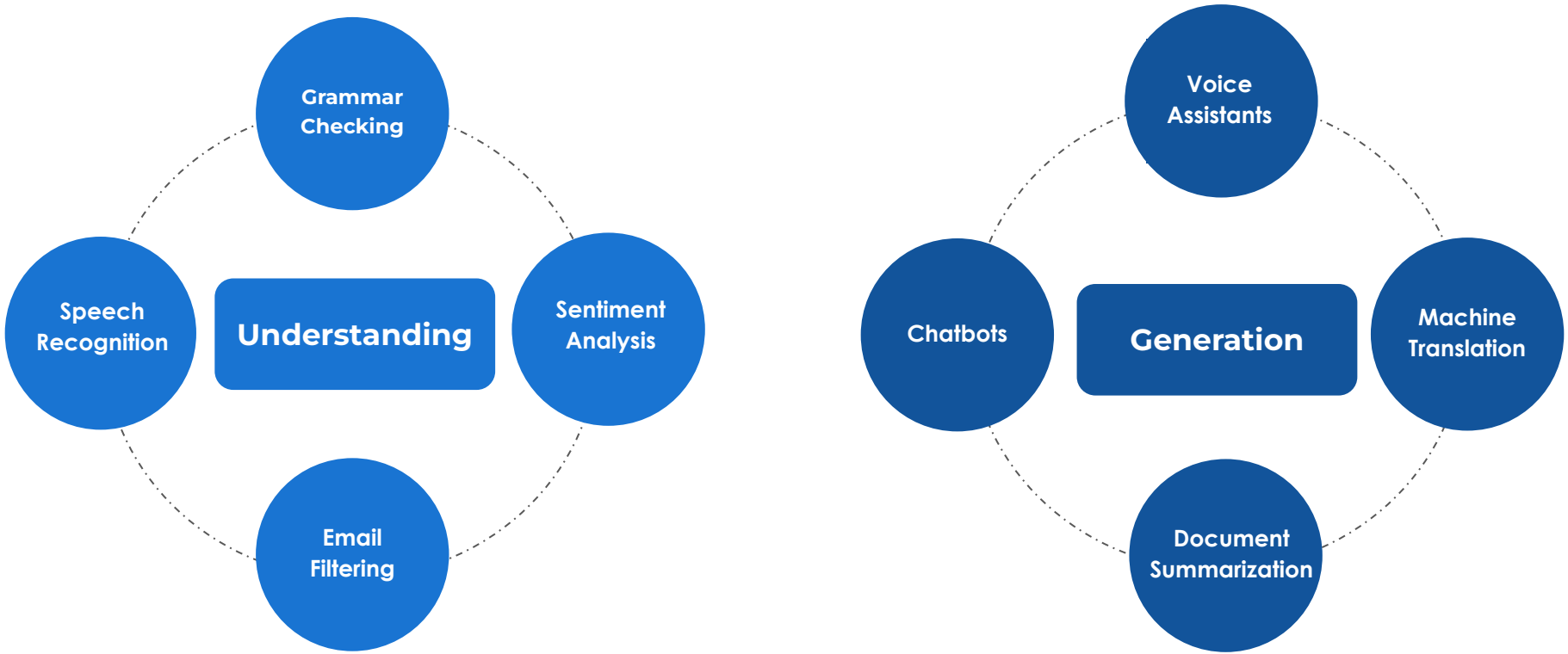
A. Sentiment Analysis

B. Machine Translation

C. Chatbot

D. Document Summarization

Applications of NLP



NLP and Text Cleaning Quiz

Which of the following tasks are performed during text cleaning?

A. Stemming

B. Lowercasing

C. Removal of Special Characters

D. Adding extra white spaces

NLP and Text Cleaning Quiz

Which of the following tasks are performed during text cleaning?

A. Stemming

B. Lowercasing

C. Removal of Special Characters

D. Adding extra white spaces

Text Cleaning

Process of preparing and refining raw text data by removing noise, such as special characters, punctuation, stop words, and irrelevant symbols, to standardize text data

Improves the suitability of the data for analysis and modeling

Stopword Removal	Stemming	Lowercasing	Remove special characters	Strip extra white spaces
Removes common words like “and”, “the”, “is”, etc which often appears frequently and generally do not add ‘contextual value’ to the text	Converts the word into its root form, reducing it to its base or stem, to capture the core meaning	Converts all the words into lower case letters	Removes special characters like “, “”, !, @, etc	Removes extra spaces between the words

Text Cleaning - Example

Text preprocessing is key in natural language processing (NLP). It involves refining raw text for machine understanding and analysis. This process includes eliminating redundant words, reducing words to their root form, standardizing text to lowercase, excluding symbols (!@#$), ensuring clean spaces, and filtering out numerical characters 1234.

Text Cleaning

text preprocess key natur languag process nlp involv refin raw text machin understand analysi process includ elimin redund word reduc word root form standard text lowercas exclud symbol ensur clean space filter numer charact 1234

NLP and Text Cleaning Quiz

Great Learning
POWER AHEAD

Consider the following three sentences:

"The cat jumped"
"The dog barked"
"The cat chased the dog"

What is the number of dimensions in the vector representation after applying the bag-of-words (BoW) model to all these sentences?

NOTE: Stopwords are not to be removed from the sentences.

A 6

B 9

C 12

D 15

NLP and Text Cleaning Quiz

Great Learning
POWER AHEAD

Consider the following three sentences:

"The cat jumped"
"The dog barked"
"The cat chased the dog"

What is the number of dimensions in the vector representation after applying the bag-of-words (BoW) model to all these sentences?

NOTE: Stopwords are not to be removed from the sentences.

A 6

B 9

C 12

D 15

Bag of Words (BoW) model

Represents text by **counting the frequency of unique words** in a document **without considering the order or structure** of the words

Creates a "bag" (or set) of words in a text corpus, ignoring grammar and word order

Example:

Sentence	the	cat	jumped	dog	barked	chased
The cat jumped	1	1	1	0	0	0
The dog barked	1	0	0	1	1	0
The cat chased the dog	1	1	0	1	0	1

NLP and Text Cleaning Quiz

Great Learning
POWER AHEAD

Words with high frequency are always considered to be stop words.

A. True

B. False

NLP and Text Cleaning Quiz

Great Learning
POWER AHEAD

Words with high frequency are always considered to be stop words.

A. True

B. False

Stopwords

Words in any language which do not add much meaning to a sentence.

Can safely be ignored without sacrificing the meaning of the sentence.

Example:

Consider you are analyzing a collection of documents related to cooking recipes.

The word "salt" might appear very frequently across these documents due to its significance in cooking.

However, "salt" will not be considered a stop word in this context because it carries essential meaning within the domain.

NLP and Text Cleaning Quiz

Great Learning
POWER AHEAD

Consider the following three sentences:

"The cat jumped"
"The dog barked"
"The cat chased the dog"

What is the number of dimensions in the vector representation after applying n-gram model with n = 2 to all these sentences?

NOTE: Stopwords are not to be removed from the sentences.

A 6

B 9

C 12

D 15

NLP and Text Cleaning Quiz

Great Learning
POWER AHEAD

Consider the following three sentences:

"The cat jumped"
"The dog barked"
"The cat chased the dog"

What is the number of dimensions in the vector representation after applying n-gram model with n = 2 to all these sentences?

NOTE: Stopwords are not to be removed from the sentences.

A 6

B 9

C 12

D 15

n-gram model

Similar to the BoW model but **considers sequences of 'n' consecutive words at a time** (called n-grams)

Takes into account the sequence, and through that, the context of words

Example:

Sentence	the cat	cat jumped	the dog	dog barked	cat chased	chased the
The cat jumped	1	1	0	0	0	0
The dog barked	0	0	1	1	0	0
The cat chased the dog	1	0	0	0	1	1

NLP and Text Cleaning Quiz

What is the main characteristic of a lexicon-based approach in sentiment analysis?

A. Relies on machine learning algorithms to infer sentiment

B. Utilizes a predefined set of words with assigned sentiment scores

C. Analyzes sentiment through deep neural networks

NLP and Text Cleaning Quiz

What is the main characteristic of a lexicon-based approach in sentiment analysis?

A. Relies on machine learning algorithms to infer sentiment

B. Utilizes a predefined set of words with assigned sentiment scores

C. Analyzes sentiment through deep neural networks

Sentiment Analysis

The process of analyzing a piece of text and categorizing it based on the context and emotions expressed within the text

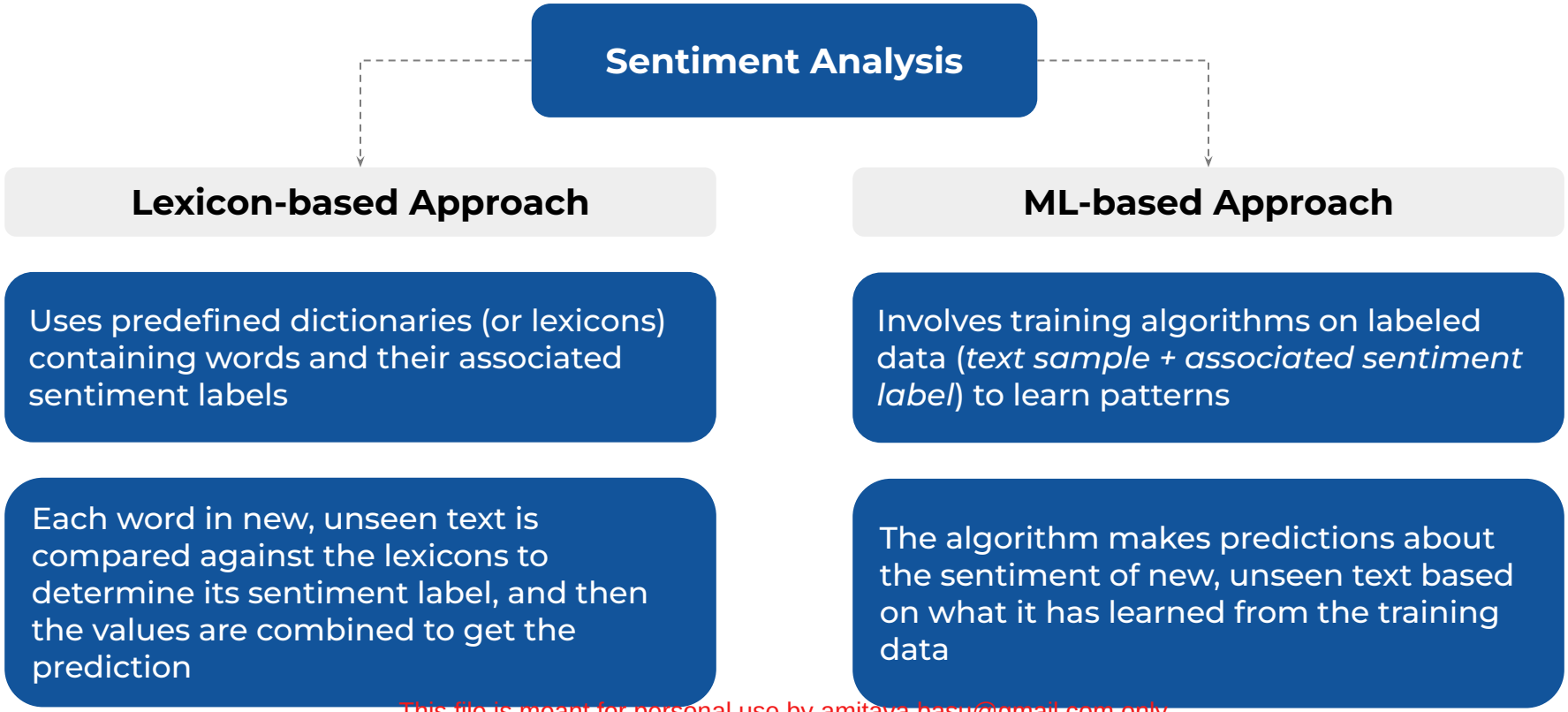
In general, categorization is done as positive, negative, or neutral

Used extensively to analyze end-user feedback, gain insights on the sentiment of the user towards a product/service, and identify areas of improvement

Example:

“The movie was fun, brisk, and imaginative” => **Positive**

Sentiment Analysis Approaches



Great Learning

Happy Learning !