Day: 09 Jan 2023

Natural Language Processing (NLP)

by mentor Dr. Ayan Debnath, IIT Delhi & Harvard Alumni

Computer Science → NLP

Applications

① chatGPT → Write me a work job... → NLP

② Email → next word Prediction

Autocorrect, autoplete

NLP is subset of AI

③ Text classification

④ Sentiment analysis

I am a good boy → +ve Sentiment

The food is bad → -ve Sentiment

⑤ Recommendation system

The food is delicious → +ve review

" - " bad → -ve "

⑥ Hate speech

website → story, anything →

——— → +ve ✓

-ve ✓

⑦ chatbot (Conversational AI)

⑧ QA

⑨ Text summarization

——— 10 sentences

100 paragraphs

⑩ Information Retrieval

google search

⑪ Language Translation

English ⇄ spanish

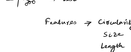
⑫ Sentiment Review

Machine Learning (ML)

Supervised → Data → annotated → Labelled / Class category

Unsupervised → Data → Not Labelled

Robot / Child



Features → Creamy, Size, Length → apple / Banana

Independent Variable → Dependent

f_1	f_2	f_3	y / output
Creamy	Size	Length	
—	✓	✓	class +1 → 0

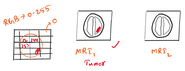
f_1	f_2	f_3	f_4	f_5	y / output
HR	BMR	Pulse	-	-	Diseases / Non Diseases
1 ✓	✓	✓	✓	✓	D (1)
2					ND (0)
3					
4					
5					
⋮					
100					

Features

Task → input → model → prediction → +ve / -ve, 1 / 0

Task Non Data

RGB + 0-255



Training

	mean,	median,	mode,	skewness,	Roundness	
①	50	25	50	100	0.9	Tumor (1)
②	—	—	—	—	—	Tumor (0) (1)
③	10	5	9	10	0	No Tumor

100

AI/ML model → algorithm

100 → 25 25 75 125 0.2 → Tumor

Data → ① text (- - - words) → NLP ✓

② Numerical values ✓ → NLP

Text: Sub/f_1 Body/f_2 Store/f_3 y

Q4 Mail → Food is good ★★★★ +1

Computer → Don't understand Text/Language

↳ understand Number

12 14 50

Food is good

Numerical representation

Vector

NLP Text → Machine → Processing / ML

Nos → " → " / ML

Roadmap

Level 0 Coding / Python

Level 1 Text Preprocessing → Data cleaning, Tokenization

Basic

Email a widely used

Level 2 Text Processing → text to No (BOW, TF-IDF)

Level 3 Core concepts of ML/DL

Intermediate

Level 4 DL → RNN, LSTM, GRU

Level 5 Transformer

Advance

Level 6 chatGPT, BERT

Text Pre-processing

① Data Cleaning → Food is good! [A-Z a-z]

Food is good

12 14 50

Food is nice

Food is nice

" " 9.1 8.5

② Remove case / Lowercasing

Eg ⇄ Spanish

Twice was has 86 to go to... → ✓ No

[a-z, A-z, 0-9]

Tokenization → splitting paragraph / sentences into smaller units

Q5 Food is not good. Book not come to this place again

① ② ③ ④

Paragraph → sentences → Sentence Tokenization

0 1 2 3

Sentence → word → Word Tokenization

Food is good

① ② ③

Q7 Why Tokenization is needed?

Food is good ...

13 145 243

easy for computation

easy - processing

1000

12

100 → 5-6

https://www.linkedin.com/in/ayan-debnath/