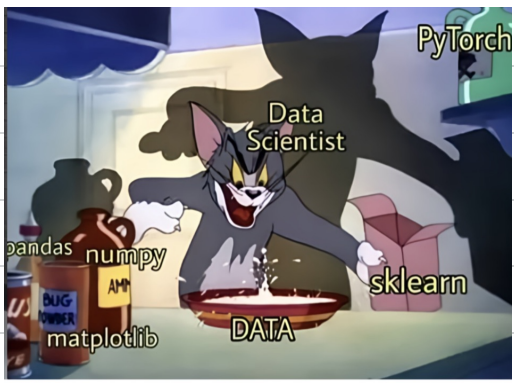
Session -1

Intro To NLP

April 12, 2024



Agenda

1. Overview of NLP module + applications.
2. Basic terms & preprocessing
3. Task: Sentiment Classifications
4. Encode text as Vector
5. Build a Simple model.

NLP → Text
→ Speech → 2 years
→ audio / signal
→ Sequential Time series

→ Projects → End of module

(i) Text Classification (Sentiment)
(ii) LLM (GPT)
(iii) NMT (neural machine translation)
(iv) Chatbots
(v) Text → Coming From LLM
→ Plagiarism → 7%
(vi) Text Summarization → Pointer generator
→ Abstractive
text summar
(vii) CC - Closed captions (Youtube)
(viii) Spelling Checker

(viii) Number plate detection → Yolo-V5

↓

$50,000

↓

A2126

↓

LSTM

↓

deciphering text

CV → Special Kind → preserving Spatial di

RNN → " " " Sequential Structure

are, can, can't, shall, will

NLP (couple of decade)

↳ Rule-based / Heuristic

↳ BOW, TF-IDF

DL → Word2Vec → word → vector

↳ 2012-2013 40-50 / 100

→ Nurse → Woman

↓

Femal

Remove bias

Doctor → male

→ Probabilistic Techniq → LDA, CRF

15-20 years

RNN → LSTM, GRU → Last 10 yea

→ Transformer → Attention

↳ Key idea

→ BERT - GPT

OpenAI

↓

25,000 A100-

2 months

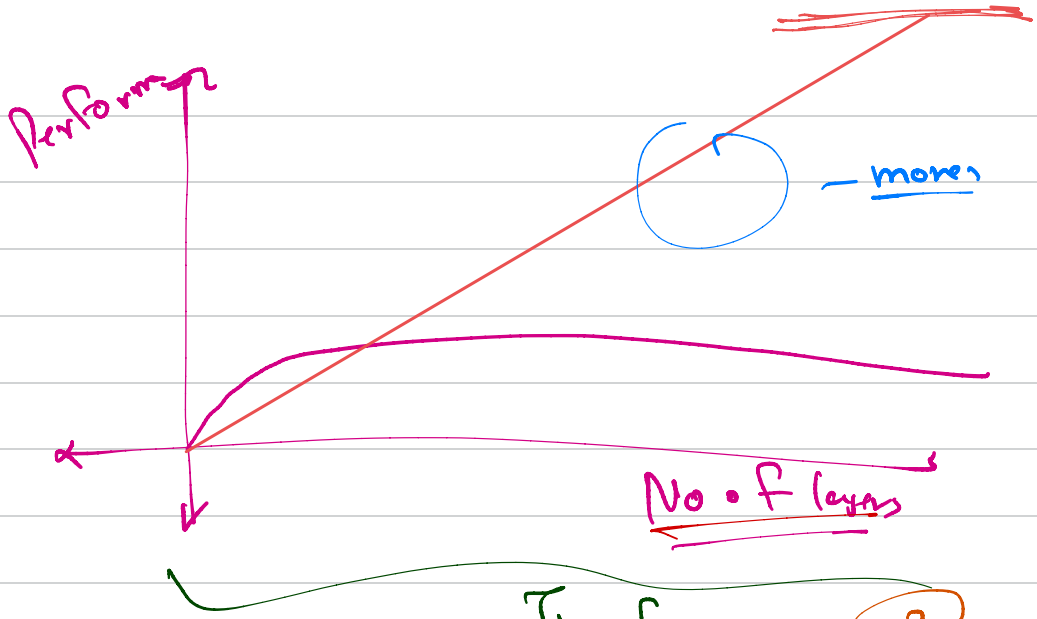
165 million $

URAM

80 GB

RTX-4090 → 24 GB

→ 2.3 lakhs INR



Transformer

Transformer → 1. w_1 w_2 2

GPT-4.5 -- 1.8 Trillion

1t. Brain → 1000 trillion

10% ✗ True appln

min amt of switches

↓

word

Sent - Classification

→ ① micron

Govt — Situation

↓

lots of tweets / Cases

↓

Posith / Negativ

→ Some pre-processing:

"I love WFH, thank covid"

① Tokenization → break down sentences + words

② Stop words

③ Stemming / Lemmatization

Crude algo

↓

v. fast

running → run

Carrying → Car

gone → go

went → go

P's statement:

banned keywords: Product 1

→ allowed / or not

Cannabis indica marijuana → can't allow

8 - cr - > Product → I/O

marijuana

Paste → garlic

1. Spelling mistakes.
2. Hindi + English + Tamil + Telugu
3. Synonym name

Govardhan Parvat stone

→ BOIL

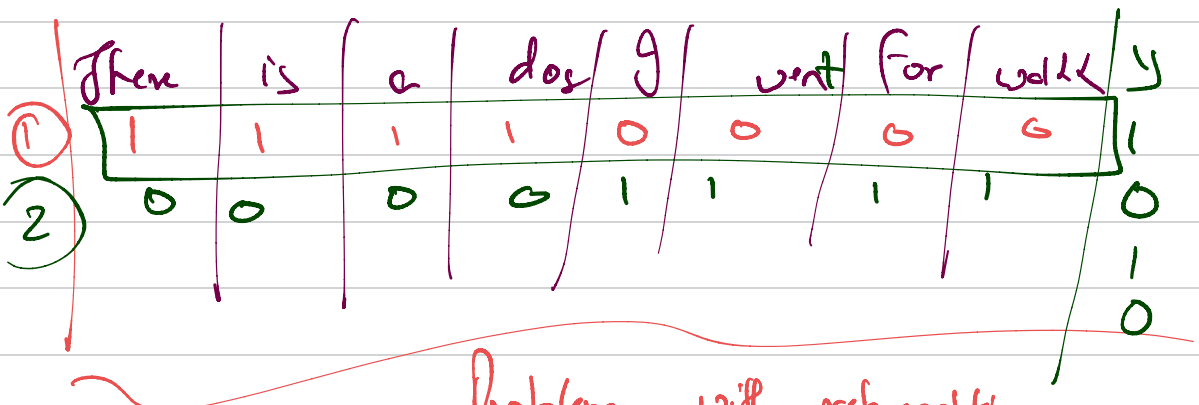
→ Character level - LSTM network — 99.6%

naive bays

BOW →

There is a dog

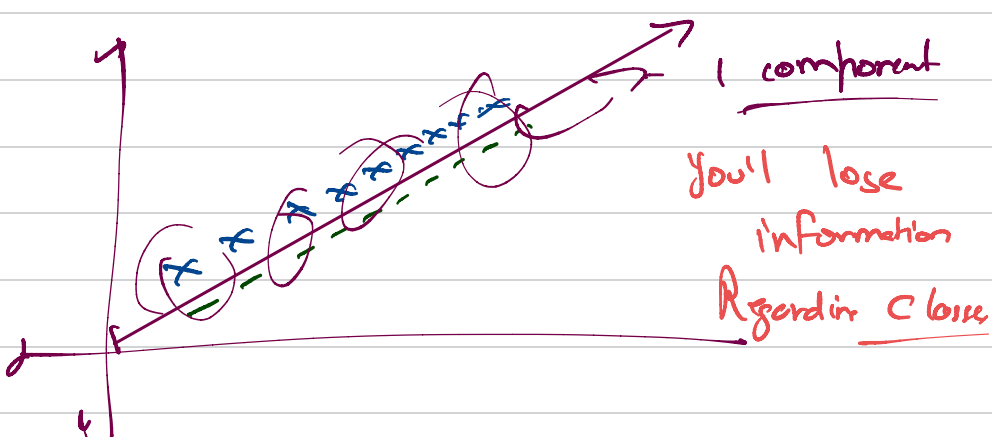
I went for walk.



Problem with representation

U. V. sparse

→ Need to reduce dimensionality.



Orange → across all the tweets

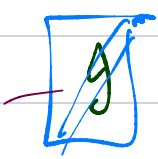
↳ (+) → tweets → Freq of Oran

↳ (−) → tweets → Freq of Orang

Orange , (+) Freq , (−) Freq

↓

∀ the words in a give sent



like

cooking

99 (+)

10 (−)

(80) (+)

20 (−)

80+20 , 30 → d dimension def

+

2 dimensi

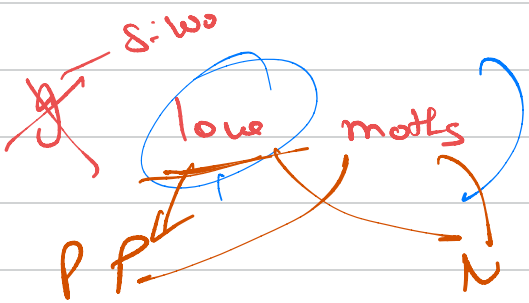
10K

10K, 2

Train. —→

1 million rows
↳ 500,000
(+)
500, 000
(=)

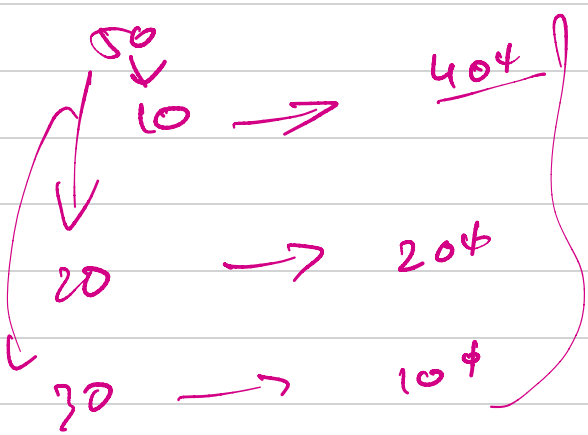
~~This~~ ~~is~~ ~~my~~ day —→ 50000 ↑
↳ 500 (=)



test dat

→ Valid limitation

You rely → on u. by default



Dil relief