Natural Language Processing [NLP]

application

1. spam classifier → importance / not "
2. Resturant Review → Food is good/bad
3. Alexa, Google assistant →
4. chatbot / Q&A
5. Sentiment Analysis → +ve / -ve

→ slang

NLP → **Roadmap**

Text Preprocessing (1) → Tokenization, stop words, stemming

(2) Bag of words, TF IDF, word2vec

ANN, RNN, LSTM, GRU

seq2seq, attention, Encoders, decoders

Transformers (GenAI)

BERT, GPT

Text processing

words, sentences, paragraph, Document

→ model → +ve/ -ve

1. Food is good → +ve
2. " " bad → -ve

- - - - - - - - - → - - - - (relevant/imp info)

Processing → (1) Data cleaning

?, ..., #

Food is $ good

↳ Food is good

(2) Lowering (Lower case)

Food is not upto the mark but the food can be made good
(upper case) (lower case)

Tokenization → splitting sentences/paragraph into smaller units

eg Food isn't good → Food (1) is not (2) good (3/4)

Food is good → Food (1) is (2) good (3)

we are learning NLP. NLP has wide rang of app. NLP is widely used now
(1) (2) (3)

we (1) are (2)

why Tokenization is needed?

Text → computer " only understands no.

NLP, Food is good ✗

123 546 234

[5000] Food

- - - - ? autocompletion

next word prediction

- - - Food is good / bad ✗

model

Stopwords

I am a boy and I love to eat chicken → boy love eat chicken

stemming → Process of converting infected word to stem/base word

history, historical → histori (base word/stem word)

loving, love → love

final, finalization, finally → fina

Lemmatization → converts the words to base word with some meaning to it

eg history, historical → history

final, finalization, finally → final

Stemming	Lemmatization
(1) convert to base word	(1) convert to base word but meaningful word
(2) Faster process; historical → histori	(2) Slower process wrt stemming; historical → history
eg Sentiment analysis	eg chatbot

murali (BOW)

7469, 10,000

Transformer

murali → 169 → [0, 169, 11, 12, 134, 0] Vector

text

stemmer