LECTURE-2

NLP Pipeline

↳ NLP Pipeline is a set of steps followed to build an end-to-end NLP Software.

↳ NLP Software consists of the following steps :-

(1) Data Acquisition

(2) Text preparation.
- ↳ Text Cleanup
- ↳ Basic Preprocessing
- ↳ Advance preprocessing

(3) Feature Engineering.

(4) Modelling
- ↳ Model Building
- ↳ Evaluation.

(5) Deployment
- ↳ Deployment
- ↳ Monitoring
- ↳ Model update.

① Data Acquisition

(a) Data is Available.

↳ In a csv file.

↳ Database - use data engineering to extract data from the database.

↳ Less Data - Data Augmentation is used

Synonym &

Bigram

Flip

↳ Back Translation.

↳ Additional Noise.

Tools used for Data Augmentation

(b) Some other party has ~~a~~ data

↳ Public dataset

↳ API (requests)

↳ Image

↳ web scrapping (beautiful soup)

↳ pdf

↳ Audio.

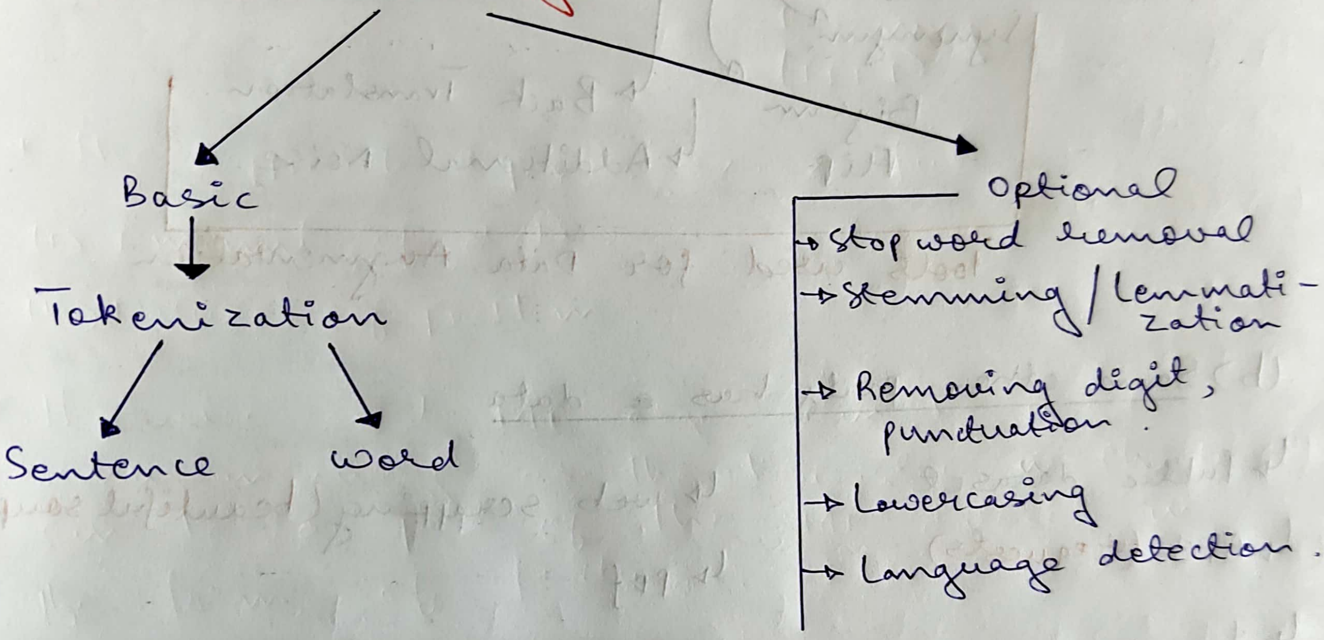
(c) Nobody has the data

② Text Preparation

(a) Cleaning

- ↳ html tag cleaning
- ↳ emojis for sentiment analysis
- ↳ Spelling check

(b) Basic Preprocessing



(c) Advance Preprocessing

- ↳ Part of Speech tagging
- ↳ Parsing
- ↳ Coreference Resolution

③ Feature Engineering

↳ It is the process of transforming raw data into meaningful features to enhance the performance of ML models.

Eg: Bag of words, Tf-idf, One Hot encoding, Word2Vec.

Feature engineering stage

ML Algorithm

Data → Pre-processing → feature
↓
Algorithm

Advantage
↳ Can justify the o/p and the accuracy of the model.

Disadvantage
↳ Takes a lot of time to build feature.
↳ Need domain knowledge.

DL Algorithm

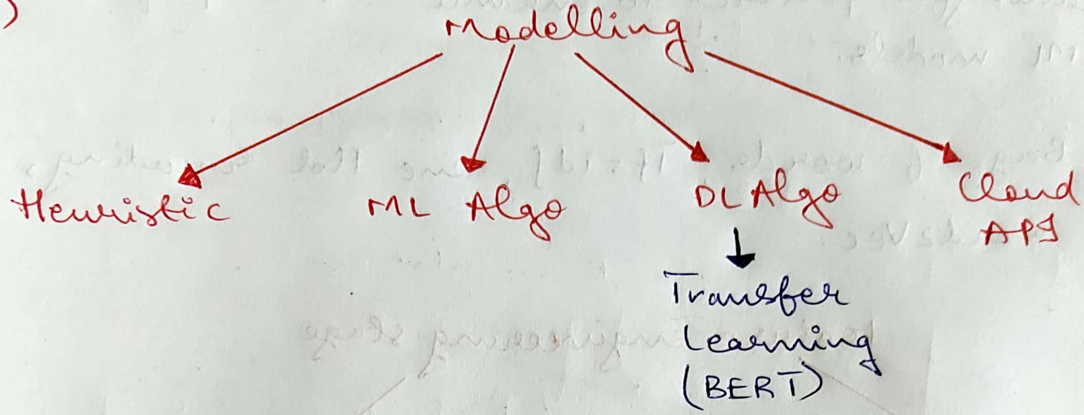
Data → Pre-processing
↓
Algorithm

Advantage
↳ No need to build features as they are formed automatically.
↳ No need of domain knowledge.

Disadvantage
↳ Loss of Interpretability.

④ Modelling

(a)



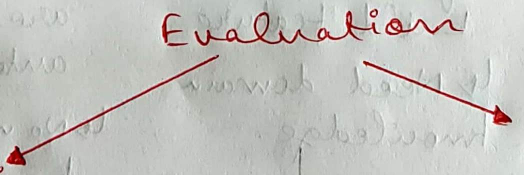
* The approach depends on 2 factors :-

(i) Amount of data.

(ii) Nature of problem.

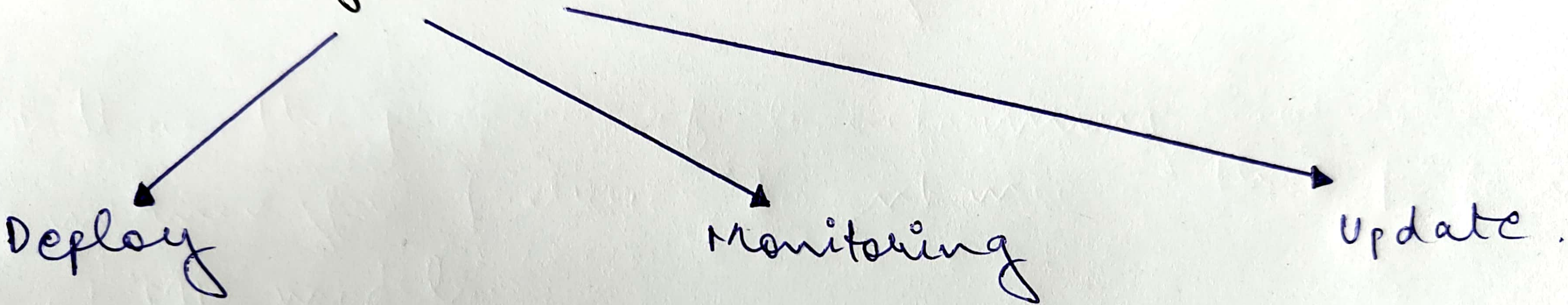
* Heuristic and ML algo. can be used simultaneously as the heuristic methods can be a part of the ML dataset in the form of features.

(b)



Intrinsic eval.	Extrinsic eval.
↳ Assesses a model's performance directly on specific tasks without relying on external context, measuring inherent capabilities or quality.	↳ Measures a model's performance in real-world scenarios, considering its effectiveness within broader applications or systems beyond isolated tasks.

⑤ Deployment



Deploy

Monitoring

Update.

↳ API (Cloudservices)

↳ Chatbot