Natural Language Processing

→ NLP :- Is the ability of a Computer program to understand human language as it Spoken and written -- referred to as natural language. It is a Component of AI. It is used in a wide range of applications, including machine translation, Sentiment analysis, Speech recognition.

- Some techniques used in NLP include :

i) Tokenization : the process of breaking text into individual words or phrase.

ii) Part-of-Speech tagging : the process of labeling each word in a Sentence with its grammatical part of Speech.

iii) Named entity recognition : the process of identifying and categorizing named entities.

iv) Sentiment analysis : the process of determining the sentiment of a piece of text.

v) Machine translation : the process of automatically translating text from one language to another.

vi) Text classification : the process of categorizing text into predefined categories and topics.

* The field is divided into three different parts :

1. Speech Recognition - The translation of Spoken language into text.
2. Natural Language Understanding - The Computer's ability to understand what we say.
3. Natural Language Generation - The generation of natural language by a Computer.

Input → Automatic Speech Recognition → NLU → NLG → Output

→ Steps in NLP :-

i) Tokenization
↓
ii) Stemming
↓
iii) Lemmatization
↓
iv) Part of Speech tag
↓
v) Name Entity Recognition
↓
vi) Chunking

i) Tokenization :-

Ex: Dipesh is abnormal human.

ii) Stemming :-

Ex :- Knows → Knaut ← Knowing ← Known

Date → Dat ← Dating ← Dated

iii) Lemmatization :-

Ex :- Die → Die ← Died ← Dead

iv) Part of Speech tag :-

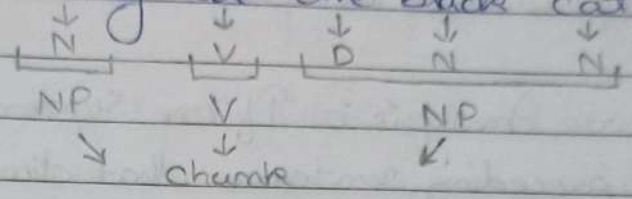
Ex: Maroon Killed a bat and ate it.

noun verb det Noun Con Verb Pron

iv) Name Entity Recognition :-
Ex :- Company, location, name

Google it!

v) Chunking :-
Ex :- Anurag ate the black cat



∴ It helps in getting insight and meaningful info from text.

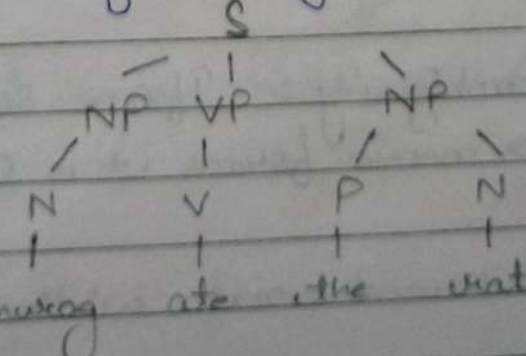
→ Stages / level of Natural Language Processing :-

i) Morphological Analysis :-
- first stage
- Breakwords into tokens
- Morphology i) roots of words
  ii) Study of words

ifeuthfullness
→ truth full ness → tokens

ii) Syntax Analysis :-
eg :- Dog ate my homeworks ∝

eg :-



iii) Sem Semantic Analysis :- meaning Sentences

eg :- Plant : Industry / Organism ∝

iv) Pragmatic Analysis :- One sentence may have two meaning

eg :- Dipesh loves his girlfriend and Manan does not.

v) Discourse Analysis :- If one sentence get effected by their preceding sentence that time Discourse analysis will work.

→ Ambiguity in NLP :- One sentence is having more than one meaning.

* Word Ambiguity :- Anurag ate a bat
I don't even have my baseball bat yet

* Sentence Ambiguity :- Dipesh loves his girlfriend and Manan does too.

i) Lexical Ambiguity :- Here one word has different meaning in different sentences.

eg :- I can play cricket
give me that can

ii) Syntactic Ambiguity :- If one sentence can pass in two or more forms it is called Syntactic amb.

eg :- Abid saw the man with binoculars

- Abid saw the man carrying binoculars.
- Abid saw the man through binoculars

iii) Semantic Ambiguity :- If one sentence may have more than one meaning.

eg :- Monm loves his cat and Dipesh does too.

iv) Anaphoric Ambiguity :- Anaphoric means when noun replace with with pronoun it is called Anaphoric Ambiguity.

eg :- The house is on a long street
It is very dirty.

v) Pragmatic Ambiguity :- when a certain sentence is not specific about something.

eg :- I love you too.

→ Application of NLP :-
i) Voice Assistants
ii) Chat bots
iii) Language translator
iv) Grammar checkers
v) Email Classification and Filtering

→ Morphological Parsing :- From one word we have to find morphins.

Morphine means :- from one smallest word

→ Morphology Parsing :- It means from one word we have to find morphine one.

Morphine one means :- A language smallest meaningfull unit that will no longer divided.

eg :- Mangoes
↓ ↓
Mango es

Morpheme → Stem → Root word
└→ Affix → Prefix
reform
→ Suffix
loved
→ Infix
Passers by

* Design of Morphological parser :-

i) Lexicon :- Stem, Affix

ii) Morphotactics :- It decide which word come first or come last.

eg :- useableness
↓ ↓ ↓
use able ness

iii) Orthographic rules :- lady + s = ~~lady~~ ladys ×

lady + es = Ladies

→ Morphology Analysis :- Morphology means a word which have meaning and Analysis means deep knowledge about that word.

eg :- ShowCase
↓
Show Case.

→ Inflectional and Derivational Morphology :-

Morphenes :- Smallest meaningful unit that will no longer divided.

Morphenes
↙ ↘
Stem Affix
[root word] - Prefix - Unwell
eg- Love - Infix - Passersby
- Suffix - Killer

Free and Bound Morphene :-

Free Morphene means words which itself has a meaning and it independent. eg : pen, board

Types :- i) Lexical - it means pic words.
ii) Grammatical - it consider grammar words it.
eg : it, and etc.

Bound Morphene means it does not have own meaning but after adding to another word it consider as a meaningful sentences.

Types :- i) Inflectional
ii) Derivational

i) Inflectional :- Words which combine to free morphemes then their part of Speech doesn't change.

eg :- Cat + s = Cats
N N

ii) Derivational types :-

a) Class Changing :- when it get combined to free morpheme then their part of Speech and class will change.

eg :- danger + ous = dangerous
N Adj

b) Class Maintaining :- when it get combined to free morpheme then their words gets changed but their class didn't get changed.

eg :- Laws + yer = Lawyer
N N

→ Difference between Inflectional and Derivational :-

i) Inflectional :- • It is a morphological process that adapts existing words so that they function effectively in sentences without changing pos of base Morpheme
• Regularity :- more Regular eg :- cats
• Use :- Can only be Suffix or infix.
• Change in POS :- Never change the grammatical category or P.O.S

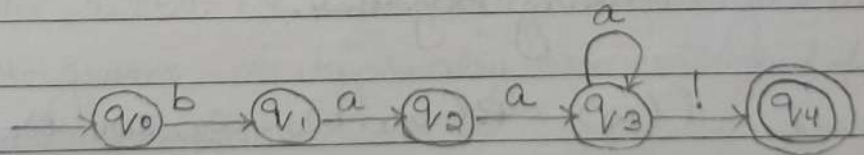
ii) Derivational :- • It is concerned with the way morphemes are connected to existing lexical forms as affixes.
- Regularity :- Less regular.
- Use :- Can be both Prefix or Suffix
- Change in POS :- Can change.
- Eg :- danger + ous = dangerous
  N Adj

→ Regular Expression :- are a sequence of characters that define a search pattern. They are used in NLP to search for specific patterns or structures in text data. Regular Expression are highly expressive and can match many patterns, including numbers, dates etc.

→ Finite State Automata :- An automaton having a finite number of states is called a Finite Automaton or Finite State Automata.

- baa+!
  baa! ✓ ba! × b! ×
  baaa! ✓
  baaaa! ✓

* Diagram :-



* Transition Table :-

States \ I/P	a	b	!
q_0	ϕ	q_1	ϕ
q_1	q_2	ϕ	ϕ
q_2	q_3	ϕ	ϕ
q_3	q_3	ϕ	q_4
q_4	ϕ	ϕ	ϕ

* Q: finite set of states
  q_0, q_1, q_2, q_3, q_4

Σ : Set of input alphabets
  : { a, b, ! }

q_0 : Start State

F : Set of final sets
  : $F \subseteq Q$

$\delta(q, i)$: the transition function or transition matrix between states.

→ Language Model :- A language model in NLP is a probabilistic statistical ~~model~~ model that determines the probability of a given sequence of words occurring in a sentence based on the previous words. It helps to predict which word is more likely to appear next in sentence.

- Joint probability :- The probability that two ~~event~~ events will both occur and is the likelihood of two events occurring together.

$$P(w) = P(w_1, w_2, w_3, w_4, w_5)$$

- Conditional probability :- Is the probability of one event occurring in the presence of a second event.

$$P(w_5 | w_1, w_2, w_3, w_4)$$

$$P(x|y) = \frac{P(x,y)}{P(y)}$$

$$P(x,y) = P(x|y)\,P(y)$$

$$P(x,y,z) = P(x)\,P(y|x)\,P(z|x,y)$$

- Markov Assumption :-

Sentence = I wish I was a Unicorn

$$P(\text{Unicorn} \mid \text{I wish I was a}) \approx P(\text{Unicorn} \mid \text{a})$$

or

$$P(\text{Unicorn} \mid \text{was a})$$

→ N-Grams :- A N-gram language model predicts the probability of a given N-gram within any sequence of words in language. A good N-gram model can predict the next word in the sentence i.e value of $p(w|h)$. It is known as n-Gram language model.

and →

- N-gram can be defined as the contiguous sequence of n items from a given sample of text or speech. The item can be letters, words or base pairs according to the application. The N-grams typically are collected from a text or speech corpus.

Unigram :- $P(w_1, w_2 \ldots w_n) \approx \prod_i P(w_i)$

Bigram :- $P(w_i \mid w_1, w_2 \ldots w_{i-1}) \approx P(w_i \mid w_{i-1})$

→ Substitution :- banguage = Language

→ Deletion :- Penu = Pen

→ Addition :- Roaste = Roasted

→ Transposition :- Hangoever = Hangover

→ Corpus :- A Corpus is a large and structured set of machine-readable texts that have been produced in a natural Communicative setting.

→ Syntax Analysis :- Syntax analysis, also known as parsing, is the process of analyzing a string of symbols, either in natural language or in a computer language, according to the rules of formal grammar. It involves checking whether a given ~~out~~ input is correctly structured according to the Syntax of the language.

- Syntax refers to a arrangement of words in a sentence such that they make grammatical Sense.
- In NLP Syntactic analysis is used to access how NLP aligns with the grammatical rules.
- Syntactic analysis helps us to understand the role played by different words in the body of text.

Eg :- Innocent peacefully Cats sleep little.
- Innocent little cats sleep peacefully.

→ Part of Speech tagging :- It is a process of converting a sentence to forms - list of words, list of tuples. The tag in case of is a part of & Speech tag and Signifies the word is a noun, adjective, verb etc.

i) ~~classes~~ Tag set for English :-
Open class :- Noun, Verb
adjective
adverb
closed class :- Preposition
determiner,
Conjunction
Pronoun, Participles

ii) Tagger :- It means a word is assigned to which part of Speech tag.

iii) Application :- a) Text to Speech
b) Search bar

iv) Problem :- a) Ambiguity

→ Rulebased POS tagging :- Rule based taggers use dictionary or lexicon for getting possible tags for tagging each word. If the word has more than one possible tag, then rule based taggers used hand written rules to identify the correct tag.

Eg :- I play (V) Cricket everyday.

I want to perform a play. (D N)

Note :- An ambiguous word is noun rather than verb if it follows a determiner.

Input

↓

Tag Dictionary
or
Lexicon

↓

Hand written rules

↓

Output
(Word, tag)

Stochastic POS tagging :- The model that includes frequency or probability can be called stochastic. Any ~~number~~ number of different approaches to the problem of part of Speech tagging can be referred to as Stochastic tagger.

Word ~~frequency~~ frequency Approach :-
The Stochastic taggers disambiguate the words based on the probability that a word with a particular tag. We can also say that the tag encountered most frequently with the word in the training set is the one assigned to an ambiguous instance of that word. The main issue with this approach is that it may yield inadmissible sequence of tags.

Camlin Page
Date / /

- Tag Sequence Probabilities :- It is another approach of stochastic tagging, where the tagger calculates the probability of a given sequence of tags occuring. It is also called n-gram approach. It is called so because the best tag for a given word is determined by the probability at which it occurs with the n previous tags.

→ Multiple tags, Words, Unknown words :-

- Multiple tags :- • Penn treebank, • British National Corpus are the two corpus that will help to solve ambiguity.

- Multiple words :- It divide words and give them their POS.

eg :- ~~Cate~~ Couldn't → Could → POS
→ not/nt → POS

- Unknown words :- If in English language day by day new words are forming so to assign their POS Unknown words uses 3 types.

1. Ambiguous Declaration :- trigram

2. Spell - ed → ~~Verb~~ Verb → past particible
→ Capital → Noun

3. Transformation based learning ex lengths, start, ending ~~Content~~ Content etc.

→ Context free grammar :-
Context free grammar is a formal grammar which is used to generate all possible strings in a given formal language.

LHS → Syntactic Category

RHS → has alternative Components parts defined.

Constituency, grammatical relations, subcategorization relations

→ grammatical relations :- means to find the subject & object.

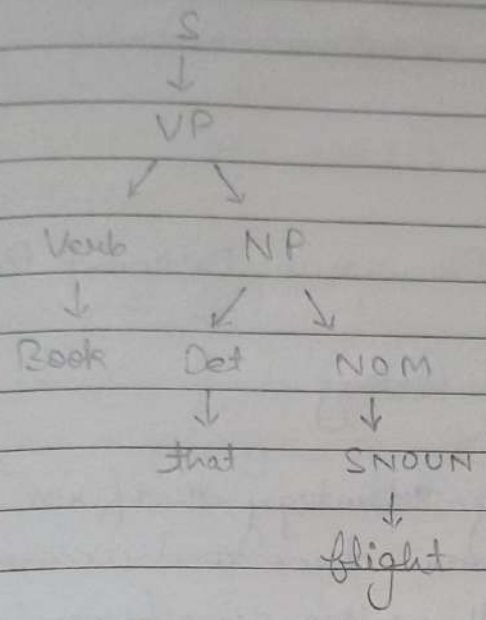
eg :- She ate a monoth breakfast
S O

→ Subcategorization ~~ob~~ relations :- words and phrases ke relation ko refer karta hai.

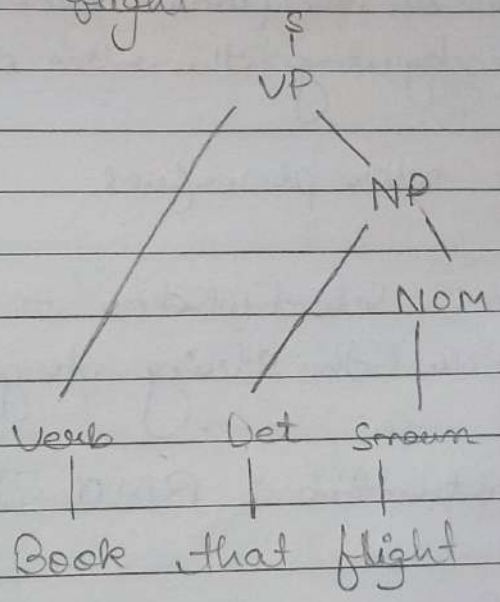
→ Top Down :-

S → VP
VP → Verb NP
NP → Det NOM
ND → Det NOM
Det → that
NOM → Singular Noun
Verb → Book
Singular noun → flight

input → Book that flight.



→ Bottom UP :-



→ Difference Top down and Bottom up parsing :-

i) Parameter : Top down Parsing

ii) Definition : It performs the parsing from the Starting Symbol to the input starting. It starts from the root level of the Parse tree and works down by using the rules of the formal grammar.

iii) Strength : Moderate.

iv) Main Decision : To select what production rule to use in order to construct a string.

v) Method of Construction : LMD

vi) Example : ~~Recursive~~ Recursive Descent Parser.

a) Parameter : Bottom up Parsing

b) Definition : A parsing strategy that first looks at the level of the parse tree and works up the parse tree by using the rules of a formal grammar

c) Strength : more powerful

d) Main Decision : Select when to use a production rule reduce the string to get the starting Symbol

e) Method of Construction : RMD

f) Example : SR parser

→ Semantic Analysis :- Semantic Analysis is a subfield of Natural Language Processing that attempts to understand the meaning of Natural language. Understanding Natural language might seem a straightforward process to us as humans. However due to the vast complexity and subjectivity involved in human language, interpreting is quite a complicated task for machines. Semantic Analysis of Natural Language captures the meaning of the given text while taking into account context, logical structuring and grammar rules.

Camlin Page
Date / /

* Applications:-
i) Information Extraction
ii) Text Summerization
iii) Information Retrieval
iv) Expert Systems

* Why Semantic Semantic Analysis is important?
Ans:- Due to the vast complexity and subjectivity involved in human language, interpreting it is quite a complicated task for machines. Semantic Analysis of Natural Language captures the meaning of the given text while taking into account context, logical structuring of sentences and grammar roles.

* Why Semantic Analysis important?
Ans. Semantic analysis is the process of drawing meaning from text. It allows Computers to understand and interpret ~~sentences~~, sentences, paragraphs by analyzing their grammatical structure, and identifying relationships between individual words in a particular ~~cont~~ Content.

* Elements of Semantic Analysis:-

1) Hyponymy: Hyponymys refers to a term that is an instance of a generic term. They can be understood by taking class-object as an apology analogy.

eg:- Color is a hypernymy while 'grey' 'blue' is hyponymys.

ii) Homonymy :- refers to two or more distinctly lexical terms either with the same Spellings but completely distinct in meaning.

eg :- Rose is a flower
Rose in a bed.

iii) Synonymy :- when two or more lexical terms that might be spelt distinctly have the same or ~~sim~~ Similar meaning, they are called Synonymy

eg :- Large, Big

iv) Antonymy :- Antonymy refers to a pair of lexical ~~ter~~ terms that ~~the~~ have contrasting meaning - they are ~~Q~~ Symmetric to a Semantic axis.

eg :- Large, Small.

v) Polysemy :- Polysemy refers to lexical terms that ~~ha~~ have the same ~~meaning~~ Spelling but multiply closely related meanings. It differs from homonymy because the meanings of the terms need not be closely related in the case of homonymy.

eg :- man, male human adult.

vi) Meronomy :- Meronomy refers to a relationship wherein one lexical term is a Constituent of some larger entity.

eg :- 'wheel' is a meronym of 'automobile'

→ WordNet :- is the lexical database i.e dictionary for the English language, specifically designed for natural language processing.

Synset :- is a special kind of a simple interface that is present in NLTK to look up words in WordNet. Synset instances are the groupings of synonymous words that express the same concept.

→ Word Sense disambiguation in NLP is the problem of identifying which "sense" of a word is activated by the use of the word in a particular context or scenario.

WSD is a subfield of Natural Language Processing that deals with determining the intended meaning of a word in a given content. It is the process of identifying the correct sense of a word from a set of possible senses, based on the content in which the word appears.

It is a challenging task because it requires understanding the content in which the word is used and the different senses in which the word can be used.

Some common approaches to WSD include :-

Supervised learning :- This involves training a machine learning model on a dataset of annotated examples, where each example contains a target word and its sense in a particular content. The model then learns to predict the correct sense of the target word in new contents.

Example :- 'Can you pull the car over?"

- Actual meaning : Are you capable of pulling the car?
- Pragmatic meaning : This means 'Can you stop the car'.

* Aspects of Pragmatics :-

i) ~~Deixis~~ Deinis :- Deinis refers to words and phrases that show time, place or situation when someone is talking. It refers to words or phrases such as "me", "here" etc, which are difficult to understand without additional information.

Ex :- i) Meet me here
ii) I wish you'd been here yesterday.

ii) Implicature :- Implicature means that more information is is communication that being said.

- Conversational Implicature :- It happens when the speaker says something that requires interpretation. It is an indirect way of saying something. It relies upon the the cooperative principle.

Person A : I am out of gas

Person B : The gas station is around the corner.

- Conventional Implicature :- Conventional ~~implicature~~ is directly implicature is directly attached to the literal meaning of the words. It does not rely on the cooperative principle.

Ex :- Gaurav is rich but happy.

iii) Presupposition :- A presupposition is when the speaker assumes something as a case before making an utterance.

Ex :- Jane no longer writes fiction

Here it is assumed that Jane once wrote fiction.

iv) Speech Act :- A Speech Act is when the sentence conveys an action rather than saying something.

Ex :- I smashed a potato.

Here it is the action of smashing a potato is depicted.

v) Conversational Structure :- Conversational structure refers to the underlying framework / ~~structure~~ structure that tells the flow of a Conversation. It follows rules, principles and conventions in a meaningful dialogue.

→ Discourse Analysis :-

Discourse Analysis in NLP is nothing but Coherent groups of sentences. When we are dealing with NLP the provided language consists of structured, collective and consistent groups of sentences, which are termed discourse in NLP. The relationship between words makes the training of the NLP model quite easy and more predictable than the actual results.

Discourse Analysis :- is extracting the meaning out of the corpus or text. Discourse Analysis is very important in NLP and helps train the NLP model better.

Information Retrieval System :- Can be defined as a software program that deals with the organization, storage, Retrieval and evaluation of information from document repositories, particularly textual information. Information Retrieval is the activity of obtaining model material that can usually be ~~docum~~ documented on an unstructured nature i.e usually ~~text~~ text which satisfies an information need from within large collections which is stored on Computers.

Machine Translation :-
Machine translation or robotized interpretation is simply a procedure when a Computer Software translate text from one language to another without human contribution. At its fundamental level, machine translation performs a straightforward replacement of atomic words in a single characteristics language for words in another.

In Simple language, we can say that machine translation works by using Computer Software to translate the text from one Source language to another target language.

Types of Machine translation in NLP :-

Statistical Machine Translation :- It works by alluding to statistical models that depend on the ~~imag~~

investigation of huge volumes of bilingual Content. It accepts expects to decide the correspondence between a word from the source language and a word from the the objective language. A genuine illustration of this is Google Translate.

ii) Rule based Machine Translation :- RBMT basically translates the basics of grammatical rules. It directs a grammatical examination of the source language and the objective language to create the translated Sentence. But, RBMT requires broad editing, and its substantial reliance on dictionaries implies that proficiency is accomplished ~~after~~ after a significant period.

iii) Hybrid Machine Translation :- It is a mix of RBMT and SMT. It uses a translation memory, making it questionably more successful regarding quality. Even HMT has a set of downsides, the biggest of which is the requirement for enormous editing, and human translators will also be needed.

iv) Neural Machine Translation :- NMT is a type of machine translation that relies upon neural networks models to build statistical model models with the end goal of translation. The essential advantage NMT is that it gives a solitary system that can be prepared to unravel the source and target text.

→ Text classification ?

Text classification also known as text tagging or

Text categorization is the process of categorizing text into organized groups. By using NLP, text classifiers can automatically analyze text and then assign a set of pre-defined tags or categories based on its content.

* Text classification Examples :-

- Sentiments analysis :- the process of understanding if a given text is talking positively or negatively about given subject.

- Topic Detection :- the task of identifying the theme or topic of a piece of text.

- Language Detection :- the procedure of detecting the language of a given text.

→ Sentiment Analysis :-

Is the process of classifying whether a block of text is positive, negative or neutral. The goal which sentiment analysis tries to gain is to be analyzed people's opinions in a way that can help businesses expand. It focuses not only on polarity but also on emotions.

* Types of Sentiment Analysis :-

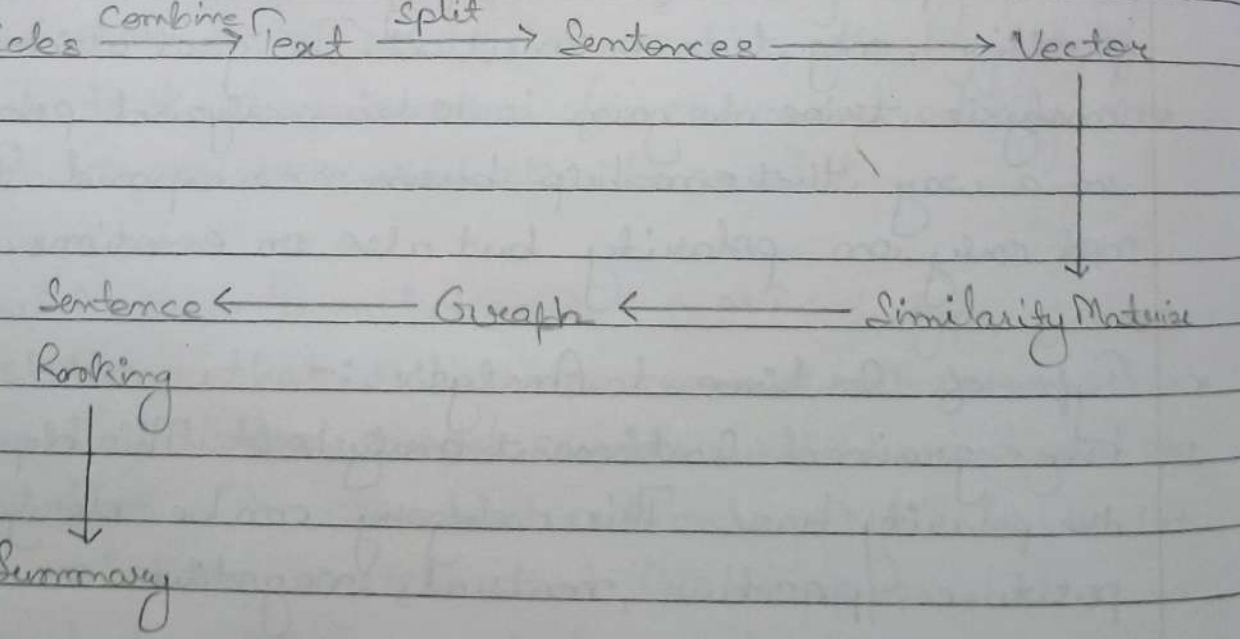
1. Fine-grained Sentiment analysis :- This depends on the polarity base. This category can be designed as very positive, positive, neutral, negative or very negative.

2. Emotion detection :- The sentiments happy, sad, angry and so on come under emotion detection. It is also known as a lexicon method.

3) Aspect Based Sentiment Analysis :- It focuses on a particular aspect for instance if a person wants to check the feature of the cell phone then it checks the aspects such as the battery, screen.

4) Multilingual Sentiment Analysis :- Multilingual consists of different languages where the classification needs to be done as positive, negative and neutral. This is highly challenging and comparatively difficult.

→ Text Summarization :- Suppose we have too many lines of text data in any form, such as from articles or magazines or on a social media, we have time scarcity so we want only a nutshell report of that text. We can Summarize our text in few lines by removing unimportant text and converting the same text into smaller Semantic text form.



Two approaches are :-

Extractive approaches :- In this, we store all the important words and frequency of all the those words in the dictionary. On the basis of high frequency words, we store the sentences containing that word in our final summary. This means the words which are in our summary confirm that they are part of the given text.

Abstractive approaches :- On the basis of their requirements we exchange some sentences for smaller sentences with the same semantic approaches of our text data.