Q) Why we need NLP?

\* We need applications which directly interact with humans like human say something and machine can understand that. That’s why we need NLP. There are many applications came of NLP like Alexa, Siri, etc.

CHALLENGES IN NLP

\* Technology is booming now days and this is because of AI and ML. AI and ML makes things convenient that even uneducated people can use this.

\* Pipeline is depended on the NLP Engineer, company will come and say I need this application and as NLP engineer you have to think that whether to collect data and how to make application, what pipeline should be used. Seniors will help in this, but if you are a senior then you have to do it by your own.

\* These are challenges face in NLP and due to challenges accuracy will don’t come out to good.

1) Ambiguity: One sentence has 2 meanings. E.g.: I saw the boy on beach with my binoculars.

\* First meaning of this sentence is: mene beach pe ladka dekha apni binoculars se.

\* Second meaning of this sentence is: mene beach pe dekha ek ladka jiske pass meri binoculars thi.

2) Same Word with 2 meaning in same sentence: E.g.: I ran to the store because we ran out of the milk.

\* Meaning: Me bhaga kyuki hamara milk khtm ho gya tha. In this sentence we have ran 2 times, meaning of ran is different in both places.

\* As humans we understand this thing, but machine will understand that both words are same.

3) Slangs: E.g.: (i) Pulling your leg. (ii) Bed of roses. (iii) Piece of cake.

\* These phrases have another meaning but machine understands that pulling your leg means literally someone is pulling leg of another person.

4) Sarcasm: One word has multiple meanings at different places and machine will understand same meaning of that words at every place. E.g.: aacha (Hindi word).

\* Uses of aacha words at multiple places and at every place meaning is different but words will same. (i) Acha papa kr dunga ye kaam. (ii) Aacha tu mje samjha raha hai. (iii) Aacha yaad aaya. In these sentences, tones is different but word is same.

5) Spelling error: Auto-correction is helpful but somethings it will confuse that at which place what word is used. E.g.: I live in Idia. Here model gets confuse from which word Idia gets replaced. ‘India’ or ‘Idea’. In WhatsApp we write our text in English by using Hindi like kya haal vhal hai bhai, here spelling error becomes big problem. Here auto-correction do not perform well. E.g.: we want to write ‘Hi, Jaan’ so there may chance that model auto correct it to ‘Hi Japan’.

\* For which language we create NLP model, challenges and complexity are comes in that way.

6) Poems/Dialogues: There is a famous dialogue in Munna Bhai MBBS movie, ‘Jaadu ki jhappi’, as humans we understand that what is it meaning but out model understands that something is related to magic.

7) Synonyms: Different words have nearly same meaning, such as large and big, Happy and Joyful, etc.

8) It will not understand sentence properly: E.g.: I am new in New Delhi. Here it considers New of New Delhi as new.

9) Convert one language into another can be challenging task because loss of information is happening. E.g.: Practically do this: Write a paragraph in Hindi convert it into English using google translate and again translate it into Hindi, you will see the difference. This is information loss.

\* These are some problems and challenges which we faced in NLP and due to these problems accuracy of NLP model comes out to be not so good.

--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

NLP Pipeline / Flowchart / Working Process

\* It provides thinking framework that from where to start where to go and what the steps involved in between. It is like a simple flowchart. For example: In machine learning pipeline is collecting data, perform EDA, feature engineering, model building, evaluating model and monitoring same to make, NLP applications we need pipeline.

Steps involved in NLP Pipeline: -

1) Data Acquisition: It is a very tedious task. If you are having data then you are the luckiest guy otherwise data acquisition is a very head tic task.

\* For acquisition we have multiple ways:

(i) Data is available: There are multiple situations when data is available.

(a) Data is available in any format and you have to work on data from first day.

(b) Data is present in database but you don’t have access to pull data, so you go the data engineering team and ask them to provide you data. First, they give sample of data and if approved it then they will give more data which you want. This whole process 10-15 days because in this way cross team communicates.

(c) Data is available but, in less quantity. So, we will data augmentation to increase data but here data augmentation is performed in different way.

\* How data augmentation is performed in NLP:

\* Use synonyms: Generate more sentences by using synonyms. Synonyms means 2 different words have same meaning. In this way it creates multiple sentences using single sentence by using synonyms. Meaning of sentence will remain same after using synonyms.

\* Back Translate: It means translate text from one language to another then again translate it into original language. By doing this formation of text gets changed. This is called back translate. We can use one or multiple languages in it. E.g.: Text is in English, translate it into Hindi then into German then into French then again translate back it into English.

\* Add or remove noise to text: Means add more words to sentence like add irrelevant words, it means words are relevant but there is no need of that words in the text. E.g.: text is ‘product is good’ add very before good so text became ‘product is very good.’ In this way we add noise. If text contains noises, then we can also remove it. By doing this we can increase or decrease the length of text.

\* These are the ways why which we can perform data augmentation in NLP.

(ii) Someone has data and you get data from him:

\* Data is not available so get data from public datasets like Kaggle. Kaggle is a public dataset from which anyone can download dataset.

\* There are UCI repositories from where we get data.

\* Do web scrapping from competitors: If you want to perform sentiment analysis of a particular product on your website and you don’t have enough data (reviews or comment), go to competitor’s website and fetch data of that particular product.

\* Data is in text, pdf, images or audio form. In reviews people attach pdf or images of product and angry customers attach recording of phone call with customer care. So, this vey difficult task to collect data from multiple sources. If data is in audio, then you apply speech to text, to convert audio into text and pre-process that text pass it into model.

(iii) Nobody has data

\* If you want to sentiment analysis or review of a product and that product is new in market. So, you don’t have enough data anywhere in that case we call customers personally which uses that product and take their review. We have data of customers because they purchase product online and make a separate team for calling customers personally and take their review on that product. In this task more cost will take.

\* These are the ways why which we can collect data in NLP.

2) Text Preparation: In NLP text preparation is different from ML in ML we perform EDA (outliers treatment, drop duplicate rows, etc.) to prepare the data for model building but in NLP we do not perform these steps in NLP we are dealing with languages, such as comments or reviews of something.

Steps Involved in text preparation:

In text preparation 3 steps are involved:

(i) Cleaning / Text clean up

(ii) Basic Preparation

(iii) Advance Preparation

(i) Text clean-up: we are removing emojis, URL’s, punctuation marks, all irrelevant information.

\* There are 10-12 steps involved in text clean-up which we study later in detail.

\* We removing HTML tags: Some websites are not compactable with version of browser so in that case if user uploads an image or text of review and when we fetch data by web scrapping data contains unnecessary HTML tags etc.

\* Remove Emojis

\* Use spell checker for checking the spellings of words.

\* We have short form words in corpus like ASAP, TTYL so we will convert these into full form.

(ii) Basic Pre-processing: In basic pre-processing we remove stop words, such as is, am, are, such, as, the, etc because stop words are useless in model building.

Basic Pre-processing includes 2 steps:

(a) Basic: There are 2 types of basic pre-processing: -

1) Tokenization: It means breakdown text.

(1.1) Sentence Tokenization: It will break down whole text into sentences.

(1.2) Word Tokenization: It will break down whole sentence into words.

(b) Advance:

(1) Convert whole text into lower case: otherwise, will treat ‘He’ and ‘he’ differently. We will perform this in every project.

(2) Removing Stop words: To deal with stop words in multiple language we have libraries in python. If we do not remove stop words train will take more time to train but there is no effect on accuracy. Stop words are irrelevant.

\* If we do not remove stop words unnecessarily it makes model complex and lengthy, means model takes more time train.

\* Stop words do not help model to determine whether it is a positive or negative review.

\* We have direct library available to deal with stop words of multiple languages. Every language has their own stop words.

\* While working on multiple languages in NLP, it is difficult because when we translate one language from another there is loss information.

\* Practically apply on google translator, translate one paragraph from hind to English then again translate that English to Hindi, you see that there is a difference between them. This is called information loss.

(b) Advance Pre-Processing

(3) Stemming: When stemming is performed it is not necessary that convert root words have meaning or not, they exist or not. Convert root word can be anything. E.g.: It will convert many to ‘mani’ and ‘go’ to ‘goe’

(4) Lemmatization: In lemmatization converted root word has actual English dictionary meaning. Where you want that model will reply user i.e., chatbots we are using lemmatization.

Note: This is the only and 1 important difference between stemming and lemmatization.

\* These both are interlinked words, work of them is to convert words into its root form / base form. E.g.: go, going, gone will convert into go. This will help to perform calculations faster because number of columns will decrease by doing this. There is separate column for separate word. More vocabulary equivalent to a greater number of columns. We will see this feature engineering. That’s why we use stemming and lemmatization to reduce number of columns and speed up the training process.

\* When interviews ask that what is the need of stemming and lemmatization then speak about these points.

(5) Language Detection

(6) Removal of Digits: Removal of digits depends on tasks. Like you build model for sentiment analysis of a product where review is like ‘My delivery was supposed to reach here by 29th of May but it reached at 6th of June. So, I am very disappointing of service because delivery was late.’ Here we don’t need digits so we will remove them because ‘delivery was late’ mentioned in review.

(iii) Advance Pre-Processing: (i) Coreference Resolution and (ii) Part of Speech Tagging.

(a) Coreference Resolution

\* E.g.: ‘Donald Trump and his wife are the citizens of USA. He is a great man.’ Model is not able to recognise that ‘his’ and ‘he’ are used for whom, model thinks ‘his’ and he’ are separate word.

\* To deal with this problem we use advance pre-processing called ‘coreference resolution’.

\* It means we referred one word to another like he is for Donald Trump. It makes linkage between words.

(b) Part of Speech Tagging

\* It includes noun, pronoun, verb, adverb etc.

\* We see that 1 word has 2 meaning. E.g.:   
‘I would like to work in google.’ Here: - Google is noun

‘I like to google things.’ Here: - Google is verb.

In both sentences google is used but their meaning is not same in both, means the part of speech is not same, means noun, pronoun, verb, adverb, etc. all these are called ‘Part of Speech’.

\* If we not perform part of speech, model will recognise all same words as same part of speech.

\* That’s why part of speech is important and there is a separate topic of part of speech, which we study later.

**Note**: Whenever discussion is happening in interview about challenges and steps involved in NLP, speak about all these above points.

3) Feature Engineering: In feature engineering we convert text into numbers. Like in ML we convert text into numbers using Label Encoder, Target-Mean Encoding, One Hot Encoding (Dummy Variables) etc. In NLP we have multiple technique to do this.

(i) One Hot Encoding

(ii) Bag of Words

(iii) Ngrams

(iv) TF.IDF

(v) Custom Features

(vi) Word2Vec

These are the different ways by which you can convert text into numbers.

\* This is the Data Preparation part, now after perform these steps data is prepared now.

4) Model Building: We apply ML algorithms to make NLP models.

5) Model Evaluation: Check the performance of model on test data.

6) Deployment

\* In interviews mainly questions are asked from text preparation and feature engineering because these 2 steps are related from NLP remaining other are related to ML.

--------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------

Terminologies

1) Corpus: It means entire universe of the words which is given in dataset. Suppose 20K sentences are given in dataset which consists of 1L words. These 1L words are called ‘corpus’. Corpus have stop words, punctuation marks, duplicates mean one word is repeating more than 1 time in dataset.

2) Dictionary / Vocabulary: All the unique words of corpus are called dictionary / vocabulary. E.g.: Among 1L words 60K are unique words so dictionary contain only these unique words and size of dictionary / vocabulary is 60K.

3) Document: Every single observation is called document or we can every single review or comment of dataset is a document. One user upload 4 lines of review so all 4 lines are considered as 1 document.

4) Word: Every single word of document is known as word.

\* These are terminologies which we are using going forward.