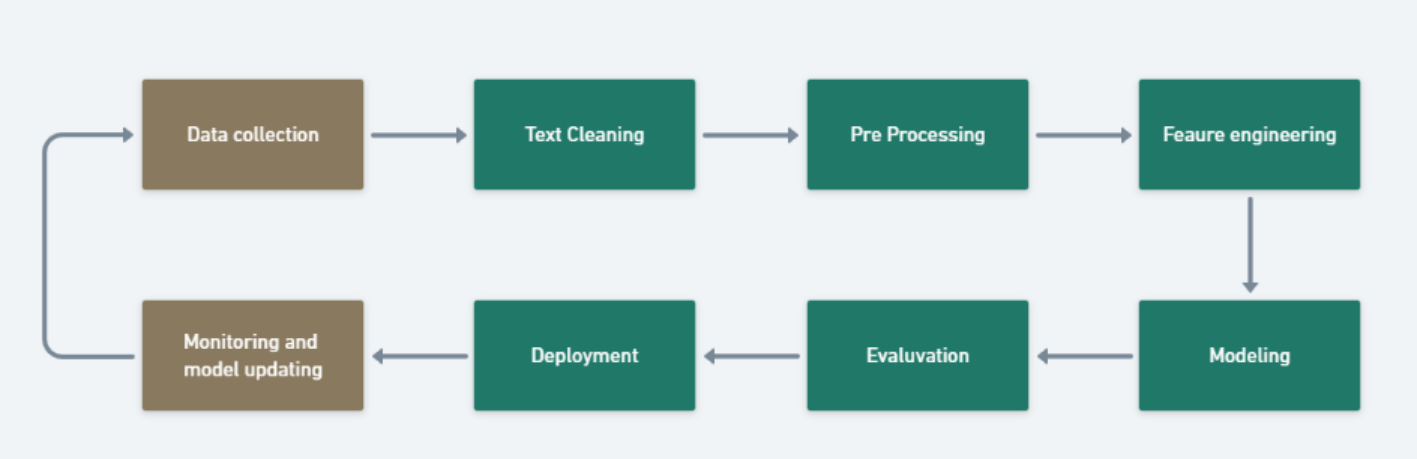
# WEEK- 11

# NATURAL LANGUAGE PROCESSING

## Session 7

## 11.13 NLP Pipeline

The set of ordered stages one should go through from a labeled dataset to creating a classifier that can be applied to new samples is called the NLP pipeline. NLP Pipeline is a set of steps followed to build an end to end NLP software.




### 11.13.1 Steps involved in building NLP Pipeline

**1. Data Acquisition**

In the data acquisition step, we collect the data required for building our NLP software. We can collect the data using any of the following methods:

a) We can conduct to survey to collect data and then manually give a label to the data
b) Public Dataset – If a public dataset is available for our problem statement.
c) Web Scrapping – Scrapping data using beautiful soup or other libraries

**2. Text Preprocessing**

Once the data collection step is done, we cannot use this data as is for model building. We have to do text preprocessing. It helps to remove unhelpful parts of the data, or noise, by converting all characters to lowercase, removing stop words, punctuation marks, and typos in the data. After doing data preprocessing accuracy of the model get increases.

**Steps involved in Text Preprocessing –**

1. Text Cleaning – In-text cleaning we do HTML tag removing, removing punctuations, Spelling checker, etc.

2. Basic Preprocessing — In basic preprocessing we do tokenization (word or sent tokenization), stop word removal, removing digits, lower casing etc.

3. Advance Preprocessing — In this step we do POS tagging, Named entity recognition etc.

### 3. Feature Engineering

After text cleaning and normalization, the processed text is converted to feature vectors so that we can feed it to machine learning applications. Feature Engineering means converting text data to numerical data. But why it is required to convert text data to numerical data? Because many Machine Learning algorithms and almost all Deep Learning Architectures are not capable of processing strings or plain text in their raw form. This step is also called Feature extraction from text.

In this step, we use multiple techniques to convert text to numerical vectors.

1. One Hot Encoder
2. Bag Of Word(BOW)
3. n-grams
4. Tf-Idf
5. Word2vec

### 4. Modelling / Model Building

In the modeling step, we try to create a model based on the cleaned data. Here also we can use multiple approaches to build the model based on the problem statement.

Approaches to building model –

1. Heuristic Approach
2. Machine Learning Approach
3. Deep Learning Approach

### 5. Model Evaluation

In the model evaluation, we can use different metrics for evaluation such as Accuracy, Recall, Confusion Metrics, Perplexity, etc.

### 6. Deployment

In the deployment step, we have to deploy our model on the cloud for the users. Deployment has three stages deployment, monitoring, and retraining or model update.

Three stages of deployment –

1. **Deployment** – model deploying on the cloud for users.
2. **Monitoring** – In the monitoring phase, we have to watch the model continuously. Here we have to create

a dashboard to show evaluation metrics.

3. **Update**- Retrain the model on new data and again deploy.

## References:


1. https://www.analyticsvidhya.com/blog/2022/05/nlp-preprocessing-steps-in-easy-way/
2. https://libguides.library.usyd.edu.au/text_data_mining/cleaning
3. https://www.analyticsvidhya.com/blog/2022/01/text-cleaning-methods-in-nlp/
4. https://www.analyticsvidhya.com/blog/2022/06/an-end-to-end-guide-on-nlp-pipeline
5. Infosys Springboard – Natural Language Processing for developers