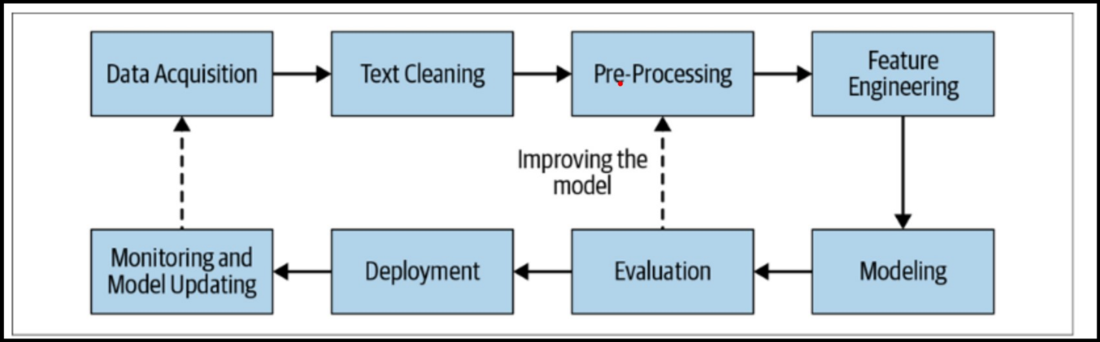
NLP pipeline



1. Data Acquisition
2. Text cleaning
3. pre - processing
4. feature engineering
5. Modeling
6. evaluation
7. Deployment
8. Monitoring and Model updating

Data Acquisition

1. This is the initial stage where raw textual data is collected
   The data may come from various sources Web scraping
   APIs
   data repositories
   user input

2. The goal is to gather a large amount of diverse and relevant text data that can be used for NLP task

Text cleaning

1. once data is acquired It contains noise — HTML tags, punctuation
   spot words

2. Text cleaning → lowering all text
   → Removal special character, extra spaces, number
   → Handling missing or values

pre-processing

After cleaning, the data undergoes pre-processing to prepare it for modeling

1. Tokenization :- text ⌒ words or sub words
2. Stemming or lemmatization :- Reducing words to their base or root form   acting → act
3. Stopword Removal :- eliminating common words (the, and, is) that are not informative
4. pos → noun, verb → structure of sentence

Feature Engineering

In this stage, the cleaned and pre-processed text is transformed into features that the model can understand

1. TF- IDF — text ⌒ vector
2. Word embedding → Word2vec, BERT
3. N- grams

Modeling

ML Model
Deep learning

This is where machine learning or deep learning models are trained using the features

evaluation