**Preprocessing text for sentiment analysis:**

Test preprocessing steps:

1. Tokenization:
   1. ML models - entire text X
   2. Break text into smaller chunks: tokens(words, numbers/ punctuation marks)
2. Lemmatization:
   1. Reducing words to their base or root form
   2. Considers context of words
   3. Stemming: reducing word to root - chopping off ending - more crude that lemmatization
3. Stop word removal:
   1. Stop words: occur frequently and provide little unique info.

Sentiment Analysis - 3 approaches:

1. Rule based: rules by experts - in if and else statements- very naive but can be tweaked
2. Machine learning: build and train a neural network model
3. Hybrid: rule based+Ml based

Ml models process numeric data - then how to process text??

Express words as numbers - for ML to process

**Tokens in numeric form:**

1. One hot encoding: encoding of words in 0/1s
2. Count vector encoding: text or word - array of numbers
   1. Create vocab of tokens(words)
   2. Doc - represented by tensor -length =vocab size
   3. Words are represented : Count of number of times the word is present in a sentence
3. TF-IDF encoding: to represent tokens
   1. Create vocab of token or words
   2. Document - tensor whose length = vocab size
   3. All words in vocab = represented by TF-IDF score - term freq inverse doc frequency
   4. Frequent words = high TF score
   5. Less freq words have high IDF score
   6. The above 2 dont capture semantic relation and meaning of words
   7. The size of vocab large and hence the length of the vectors are large
4. Word embeddings:
   1. Words - vectors
   2. Numbers in vectors capture semantic similarity
   3. Help understand context
   4. Found during model training process/ use pre trained word embeddings

**Types of sentiment analysis:**

1. Graded Sentiment Analysis:
   1. Categorize sentiment - multiple emotion levels
   2. Rating systems - in e-commerce systems - 5 star ratings
   3. Granular analysis of sentiments - ensures precision and polarity of emotions
   4. Understand customer feedback
2. Aspect-based Sentiment Analysis:
   1. Focus on **Features/ Aspects** of service - eg laptop sounds or keyboard quality
   2. Relate customer intent to features - get feature specific sentiment
   3. Gain enhanced product insight
3. Intent based sentiment analysis:
   1. Used for market research
   2. Aims to make customer targeted campaigns
   3. Learn purchase intentions
   4. Strategic customer engagement
4. Emotional detection:
   1. Get writers psychological state
   2. Identify emotions on a broad emotional spectrum - ‘joy, anger, frustration and happiness’’
   3. Use ML to detect emotion of words - complexity of emotions

**Sentiment Analysis approaches:**

1. Rule-based -
   1. Human crafted rules - get subjectivity, polity or subject matter
   2. Set scores for positive and negative words
   3. Simple to set up
   4. Difficult to scale
   5. Not account for position of words in a sentence
2. Machine learning approaches:
   1. Positive negative or neutral
   2. ML or DL classification models - trained on large dataset with diverse eg
   3. Different word arrangements are considered
   4. Prereq: Need label data with right sentiment categories
   5. If you have labeled data - then ML models are Very scalable and accurate
3. Hybrid: ML + rule based:
   1. Uses features from both approaches
   2. May be more accurate
   3. More time and effort

**Challenges in Sentiment analysis:**

1. Difficult to identify Irony or sarcasm
2. Negation
3. Multiplority - sentence with more than one sentiment - aspect based sentiment analysis
4. Comparisons
5. Defining neutral - is difficult and tricky

Stratify - helps to get equal instances of all the sentiments positive, negative and neutral