Q1.

Mainly there are two ways of doing so:

Extractive and Abstractive summarization. An extractive summarization method is concatenating important sentences or paragraphs without understanding the meaning of those sentences. An abstractive summarization method is generating a meaningful summary.

In our case, we have to mark some insightful sentences out of the given text data which is similar to the extractive summarization task.

So we can use algorithms like TextRank, LexRank, and LSA (Latent Semantic Analysis). But in case we had to work on abstractive summarization then we can use T5 Transformers, BART Transformers, GPT-2 Transformers.

To get this task done I used the TextRank algorithm plus some of my ideas (which came after analyzing the required output). TextRank is a graph-based ranking algorithm like Google’s PageRank algorithm.

In short, my approach includes assigning some scores to each sentence (using a normalized word-frequency approach) and then filtering them out using some criteria and using the TextRank algorithm on those sentences. I had observed that most of the required sentences lie within the range of 0 to 3, so I filtered them out accordingly and then used the TextRank algorithm on those filtered sentences.

Q2.

In case we want to mark sentences from raw text data as insightful or not (binary classification):

For some particular domains of interest, we can make a bag of words (or bag of n-grams) and assign some score to each word (or to n-gram) according to their importance for those topics, whenever we receive new text data we can use topic modeling to get the topics out of the text data. After that, we can break that text data into sentences and calculate scores for each sentence. Set some threshold score. So those sentences which satisfy the threshold score criteria, we can mark those sentences as insightful sentences. In this way, we can label the data for training.