a) TF-IDF firstly creates a vocabulary of words. It then creates a term frequency vector per document containing the frequency of each word from the vocabulary in the document. These are normalised by the inverse document frequency that assigns greater weighting to rare words under the assumption they convey more unique information about a document.

TF-IDF is useful because it is easy to compute, and it is easy to compare documents by using cosine similarity between the vectors. However, it is a bag of words model, so we lose information about the order of the tokens which is valuable information. These vectors are also high-dimensional and sparse which is expensive for computation.

b) I extract features using both DistilBERT and DistilRoBERTa from HuggingFace to compare their efficacy later on. Instead of a vocabulary of words, BERT uses sub-word segmentation called WordPiece. This breaks words down into pieces based on their value to the language model. RoBERTa uses… A tokenizer is used to convert the text to token IDs (and extract attention masks) which I pass to the transformer to extract [CLS] tokens for use later on.

Transformers work by applying attention… [see slides]

The advantages of using transformers are… the disadvantages are…