```
print(predict_sentiment(model, tokenizer, "This film is a clever and humorous interpretation"))

print(predict_sentiment(model, tokenizer, "This film is boring and sloppy"))
```

```
0.9814117550849915
0.006266661919653416
```

1. The residual connection to the input and output layer is the multi-head attention and dropout layers. The multi-head attention computer the attention weights for the inputs and produces an output vector with encoded information of how each word should attend to other words in the sequence. When we add the residual connection, we provide the transformer model a means to identify importance of each word in the sequence with regards to other words in the sequence. Without this connection, we would not be able to use the attention values, which are highly useful in identifying important tokens. The residual connection layer allows the gradients to flow through the network.
2. The layer normalization is useful to normalize intermediate activations in order to control/normalize distribution. In this manner the mean and variance can be manipulated according to our preference and prevents skewed distributions. This can significantly lead to greater accuracy and reduced training time. This can be useful in situations where data is biased or distributions are inherently skewed.
3. When we multiply the dot product between Q and K with $1/\sqrt{dk}$, we end up with a matrix which has mean 0 and variance 1. This is because Q and K are dk dimensional independent vectors which have mean 0 and variance 1. Upon taking the dot product, we get the new variance as dk. Each of the single Q.K pairs have variance of 1 and since we sum up the variances of all the dk possible values, we end up with resultant variance of dk. If we remove it, it may lead to decreased accuracy, longer training time and exploding gradients.