Accuracy for Amazon 25k:

The accuracy of the Bert model was 93.3%

Accuracy for the Amazon Book reviews:

The accuracy of the ANN was 68.5%

The accuracy of the Bert model was 73.5%

The accuracy of the RoBERTa model was 74.1%

There was a decision to try the Bert model onto the given, much smaller amazon dataset. The accuracy increased dramatically. A cool 93.3%. Given the difference in sizes of the datasets one can argue that this had a dramatic impact onto the result. Besides, the amazon 25k dataset had only binary classification whereas the larger book rating dataset had six classes.

There are multiple decisions as to why one would pick an ANN model over the more complex transformer models. The main idea of picking a similar model lies in the complexity and granularity of the data. A dataset with fewer datapoints and with fewer classes does not require complex models in the sense that a model would achieve similar accuracy as a simpler model but taking longer to train doing so. This was experienced by the team when training the ANN and Bert models on a simpler dataset (the given amazon dataset).

As to when to use one transformer over another is a more difficult question to answer. In many cases one would have to do an extensive dive into what the fundamental differences between the models are. In our case, RoBERTA performed marginally better. Our main reasoning is that this model is based on the other Bert model, but with more extensive training. An upgrade basically. It is also important to note that both models trained in a similar amount of time.
However, choosing one transformer over another is also relevant to the size and complexity of the data as stated previously.

The idea of classing arbitrary rating scores on, for instance, book reviews is a very interesting idea in the sense that even us humans are relatively bad at determining scores. What exactly does it mean to rate a book as “2 stars”? Every individual might have a different complex reason but it is always clear and it is definitely uniform between people. To summarize, it is our belief that there is incredible noise in the data and it is faulty from the beginning.
Comparatively with the given amazon dataset, which has very distinct binary classification, this second classification problem is much harder and complex. For further improvement of the models, one could consider clumping classes together (maybe four and five stars together) or completely removing other classes.