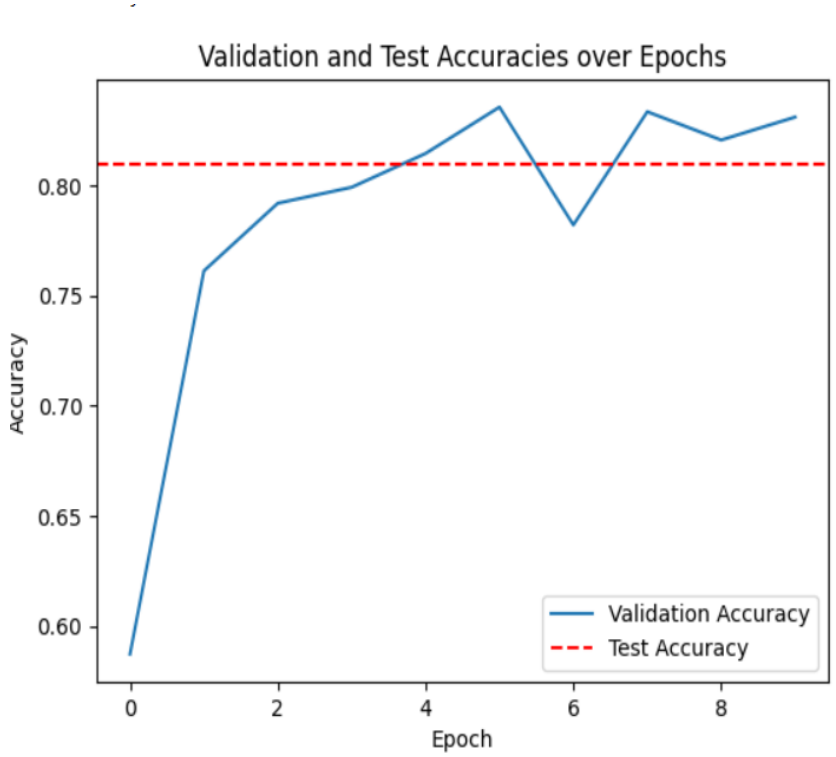
SUMMARY

INTRODUCTION:

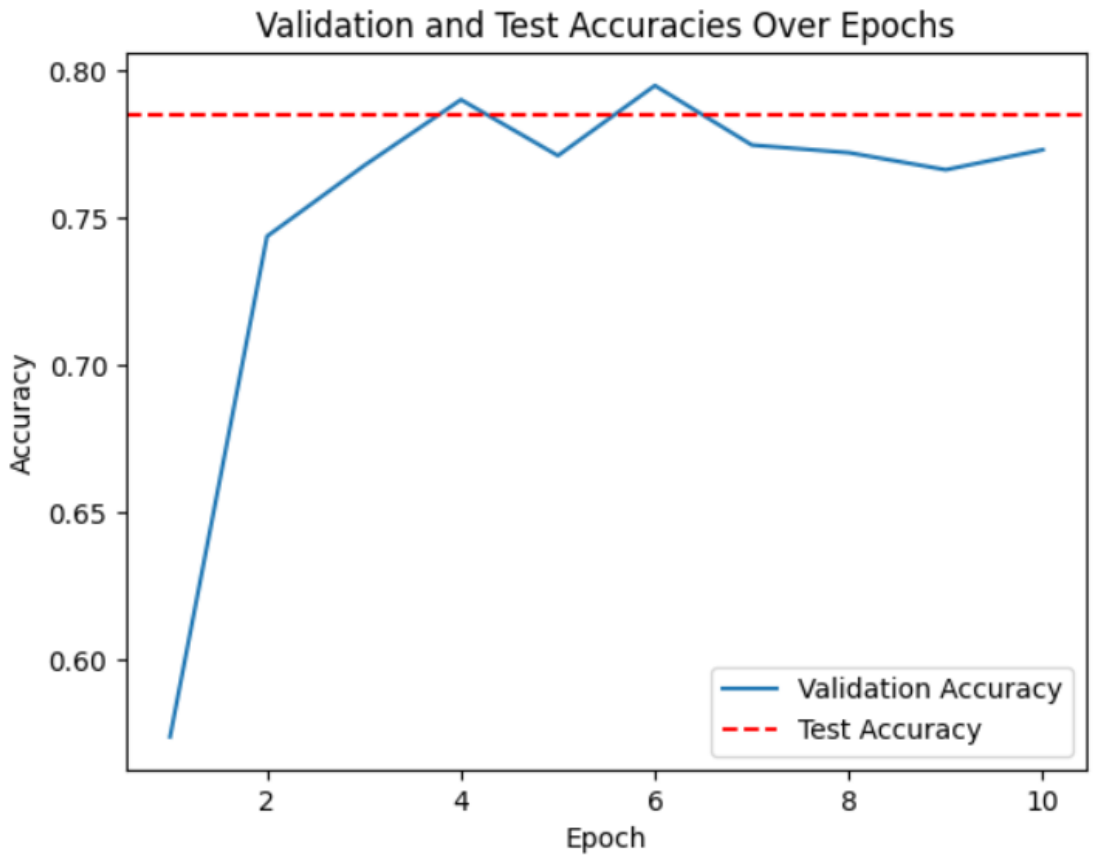
We began by getting a dataset of movie reviews from IMDb. These reviews were already categorized as positive or negative. We then removed any reviews that weren't labeled and focused solely on the categorized ones. Next, we split the data into three groups: training, validation, and testing. To achieve this, we started with a planned division and moved some files from the original training set to the validation set. This allows us to train the model on one group of data, test its performance on another, and use a third group to fine-tune its accuracy.

RESULTS FOR THE LIST OF QUESTIONS:

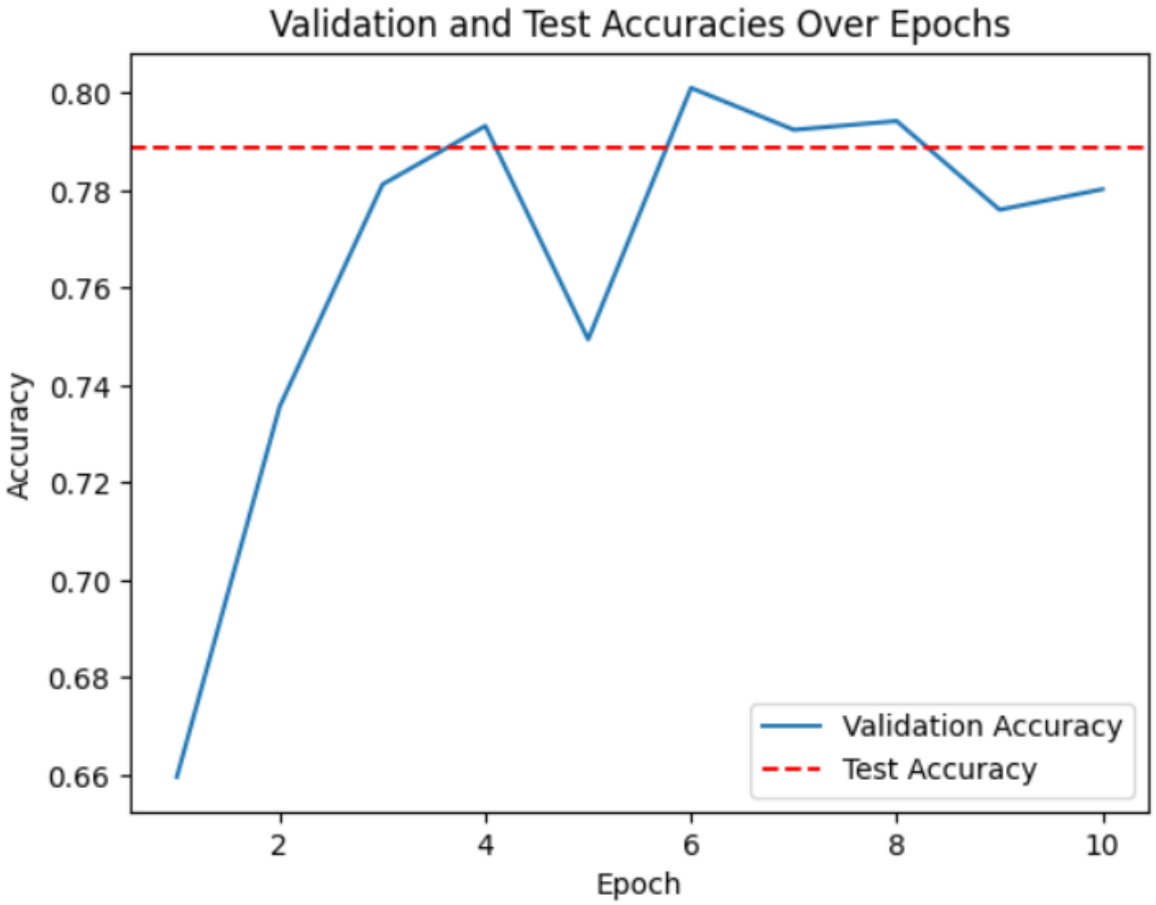
1. using just 150-word reviews and a mere 100 training samples, the model achieved good accuracy (83.1% validation, 81.0% test). This shows sentiment analysis can be useful even with limited data.



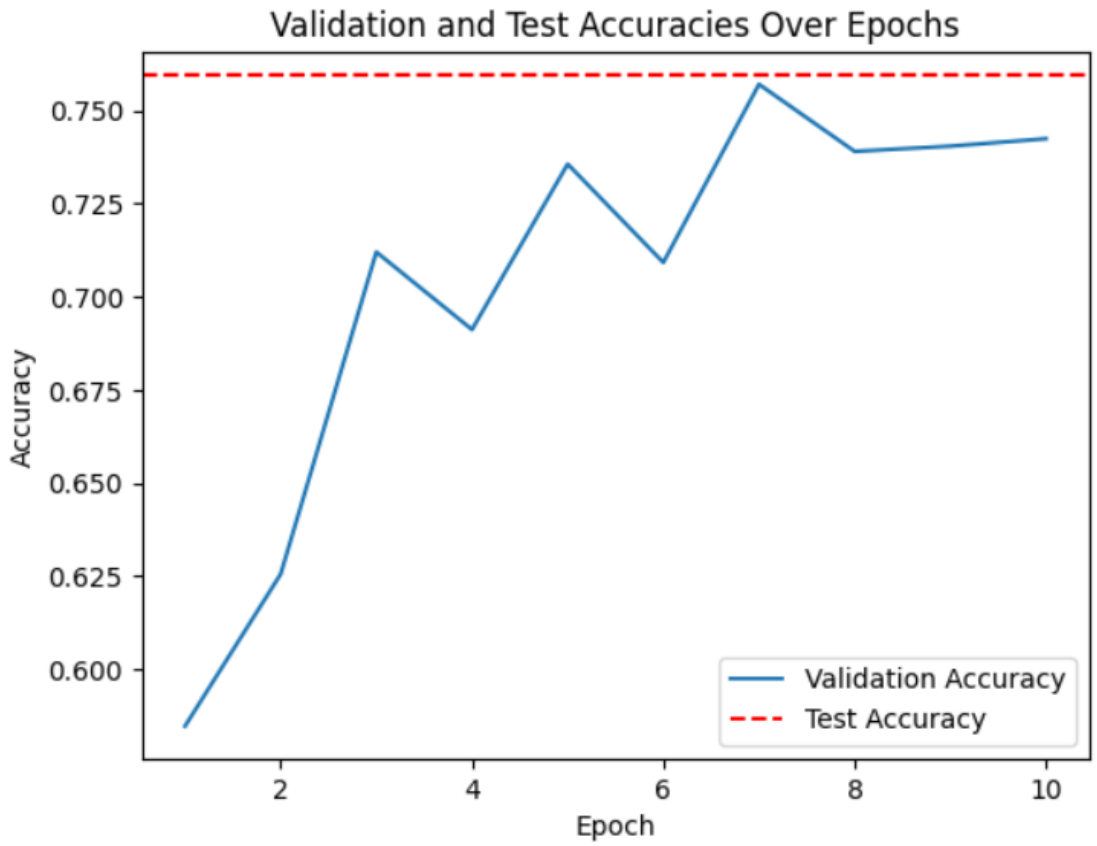
2. Limiting ourselves to only 100 training samples turned out to be a bad idea. The model's performance was terrible, with a validation accuracy of around 50% - basically just random guessing. This lack of training data left the model "underfit," meaning it couldn't learn any meaningful patterns from the reviews. In general, especially for complex tasks like sentiment analysis, using larger datasets for training deep learning models is crucial for better results.



3. To create a validation set with 10,000 samples, we used a technique called validation_split=0.2 and subset="validation". This essentially split the original data into training and validation sets, with the validation set containing 20% (or 10,000 samples) of the data. This gave the model a substantial amount of data for training and a significant validation set for evaluation. The validation process helped assess the model's ability to handle unseen data and provided valuable insights into its overall effectiveness for real-world use.



4. We capped the model's vocabulary at 10,000 words, the all-stars of the dataset. This means rare words get the boot, keeping the focus on the most frequent terms. This could be a triple win: less noise, better generalization, and faster training. Bonus: tokenization becomes a breeze!



5. Adding Power with Embeddings:

The model got a boost with a "bidirectional embedding layer" that turns word sequences into meaningful vectors. This helps understand the context and meaning of words in reviews. We can take this a step further by using pre-trained word embeddings (like GloVe) which capture deeper relationships between words than the model can learn on its own.

Two Strategies, One Goal:

There are two ways to use pre-trained embeddings:

1. Load and Learn: Load pre-trained embeddings and fine-tune them during training. (Easier to set up data splits)

2. Full Control: Load embeddings separately and manage data splits yourself. (More flexible for advanced users)

Both approaches improve sentiment analysis by capturing word meaning more effectively.

More Data, Better Embeddings:

We tested the models with different training sizes. The results are clear: validation accuracy increases with more data. This suggests that the embedding layer performs better when it has more information to learn from.