Answers

Question 1:

We can see from the model accuracy that the Word2Vec log-linear model performs better than the simple log-linear model. This is due to the fact that Word2Vec embedding is much more expressive than the simple One-hot encoding since it can capture more informative features and semantics, that the simple one-hot encoding cannot (For example, two different words with a similar semantics will have two different One-hot vectors but some similarities I their Word2Vec vectors). Another reason that might explain these results is that with One-hot encoding we have much more parameters to learn (about ~16000) in compare to the 300 parameters of Word2Vec encoding. With limited dataset like we have it will be harder task to learn 16,000 params than 300 params.

Question 2:

We can see that the LSTM model outperform the log-linear models (The One-hot encoding or Word2Vec encoding). The LSTM model can better capture the semantics between words since the hidden states can capture relations with all past words in the sentence, at each points in the training process. In addition, the LSTM bidirectional feature enables to capture these relations front to back, and back to front, and producing results considering them both.

Question 3:

Negated polarity

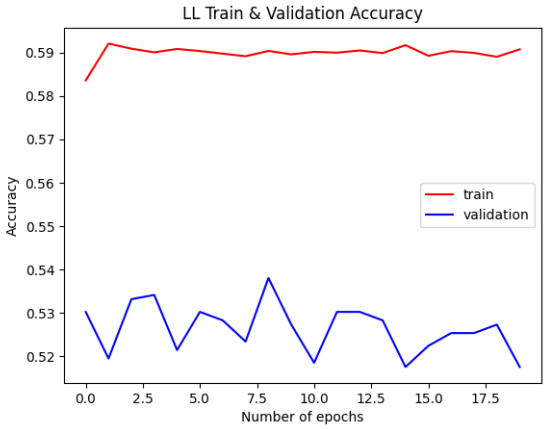
Highest accuracy: 0.6935483813285828 (**LSTM**)
Lowest accuracy: 0.45729293174201757 (**simple one-hot log linear**)

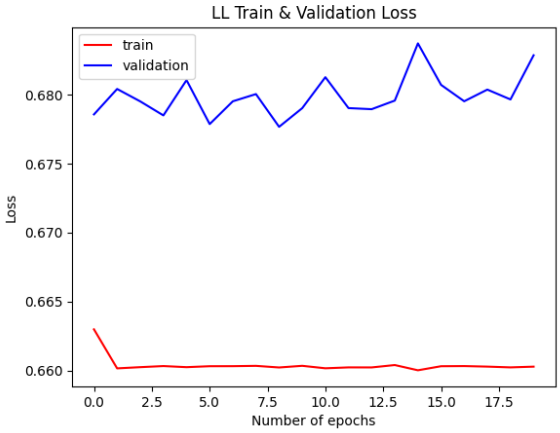
The simple log-linear model has the lowest accuracy because it cannot consider relations between parts of the sentence, and it will predict the most common sentiment it sees in the sentence.
Also, since in Word2Vec encoding (W2V and LSTM) we use less features for predictions, it might prevents overfitting to the polarity learned in compare to the One-hot encoding.
The LSTM works better than W2V since the hidden state and the fact it works both ways, might help it capturing better the negated polarities.
It is important to notice that since the hidden state in LSTM captures relations to previous parts in the sentence (and their polarity) might hurt the prediction of negated polarity sentences, and therefore we can see reduced performance of the LSTM model with those sentences.

Rare words:

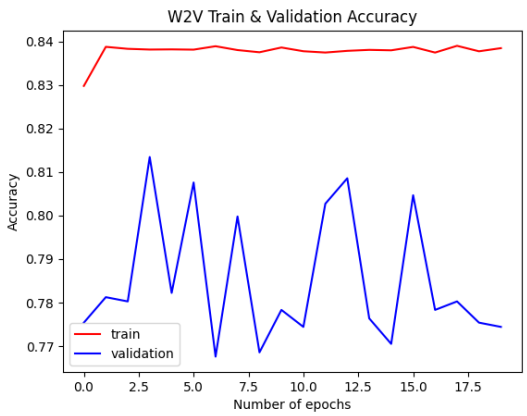
Highest accuracy: 0.7799999713897705 (**LSTM**)
Lowest accuracy: 0.30000001192092896 (**simple one-hot log linear**)
Since the models that uses Word2Vec embedding can capture semantics they can handle better unknown words using known words with similar semantics. The LSTM model works better because it can capture semantics better (Hidden states, bidirectional as mentioned above). Compared to these two models, the simple one-hot log-linear model cannot capture semantics at all and therefore don't know how to handle unknown words, and probably having trouble handling rare words with little data on them.

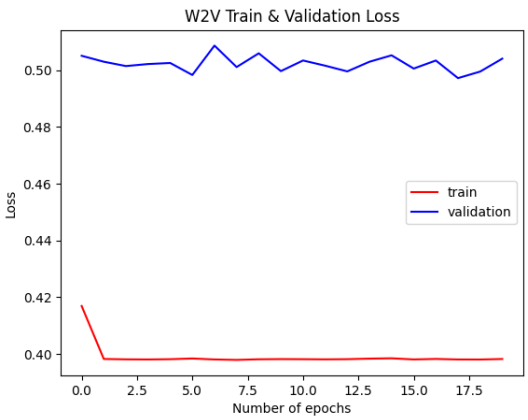
Plots



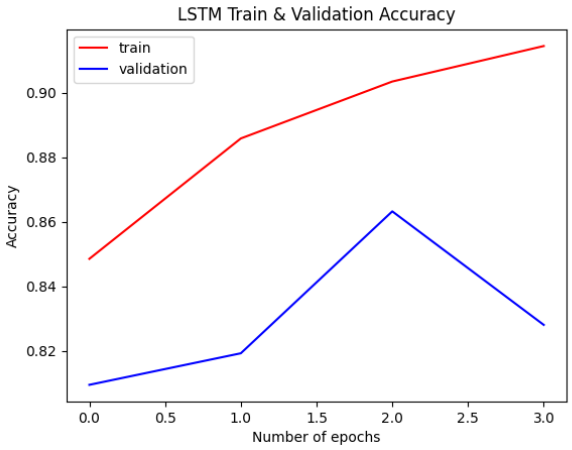


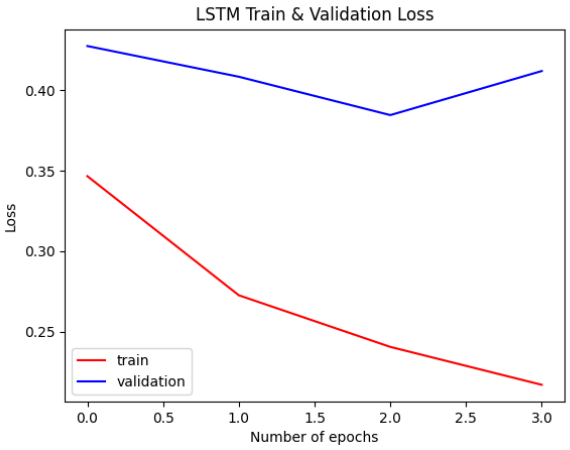

```
LOG LINEAR | test loss: 0.6755067259728094
LOG LINEAR | test accuracy: 0.5517578125
LOG LINEAR | test rare words loss: 0.749246523231268
LOG LINEAR | test rare words accuracy: 0.30000001192092896
LOG LINEAR | test negated polarity loss: 0.7019793864700102
LOG LINEAR | test negated polarity accuracy: 0.4838709533214569
```





```
W2V MODEL | test loss: 0.45729293174201757
W2V MODEL | test accuracy: 0.8291015625
W2V MODEL | test rare words loss: 0.6507316616806201
W2V MODEL | test rare words accuracy: 0.7400000095367432
W2V MODEL | test negated polarity loss: 0.7473153365670794
W2V MODEL | test negated polarity accuracy: 0.5967742204666138
```






```
LSTM MODEL | test loss: 0.3176068885143408
LSTM MODEL | test accuracy: 0.865234375
LSTM MODEL | test rare words loss: 0.5402750285086222
LSTM MODEL | test rare words accuracy: 0.7799999713897705
LSTM MODEL | test negated polarity loss: 0.7935812148667362
LSTM MODEL | test negated polarity accuracy: 0.6935483813285828
```