Assignment-4

Name: Aksa Taniya

Problem Statement: Our task is to apply RNN's to text and Sequence data. Also, we need to validate model using the basic model and pretrained word embedding.

Approach:

As per instructions mentioned in the assignment, using the below conditions varying training sample size starting from 100, 800, 1500, 3000 and no restriction to training sample which means including all the training samples.

Changes as per the instructions,

1) Cutoff reviews after 150 words

2) Restrict training samples to 100

3) Validate on 10,000 samples

4) Consider only the top 10,000 words

5) Before the layers. Bidirectional layer, consider a) an embedding layer, and b) a pretrained word embedding.

```
max_length = 150
max_tokens = 10000
```

Considering 1 & 3 points here.

```
train_ds = train_ds.take(4000)
int_train_ds = train_ds.map(
    lambda x, y: (text_vectorization(x), y),
    num_parallel_calls=4)
val_ds = val_ds.take(10000)
int_val_ds = val_ds.map(
```

Here we changed train, validate sample size, and 5 th point screenshot below,

```
inputs = keras.Input(shape=(None,), dtype="int64")
embedded = layers.Embedding(input_dim=max_tokens, output_dim=256)(inputs)
x = layers.Bidirectional(layers.LSTM(32))(embedded)
x = layers.Dropout(0.5)(x)
outputs = layers.Dense(1, activation="sigmoid")(x)
```

About Glove Pretrained Embedding:

GloVe Embeddings are a type of word embedding that encodes the co-occurrence probability ratio between two words as vector differences. GloVe uses a weighted least squares objective that minimizes the difference between the dot product of the vectors of two words and the logarithm of their number of co-occurrences.

Performance Results:

Conditions	First Basic Sequence Model	Embedding Layer	Using Padding & Masking	Using Pretrained word embedding
1.Cutoff reviews after 100 words. 2.Restrict training samples to 100. 3. Validate 10,000 samples. 4. Consider only the top 10,000 words. 5. Before Embedding layer, add a & b and b) a pretrained word embedding	80.2	77.5	80.2	77.0
1.Cutoff reviews after 300 words. 2.Restrict training samples to 800. 3. Validate 10,000 samples. 4. Consider only the top 10,000 words. 5. Before Embedding layer, add a & b and b) a pretrained word embedding	83.6	81.3	81.6	81.9
1.Cutoff reviews after 800 words. 2.Restrict training samples to 2400. 3. Validate 10,000 samples. 4. Consider only the top 10,000 words. 5. Before Embedding layer, add a & b and b) a	84.9	83.5	83.5	84.0

pretrained word embedding				
1.Cutoff reviews after 1500 words. 2.Restrict training samples to 3 3. Validate 10,000 samples. 4. Consider only the top 10,000 words. 5. Before Embedding layer, add a & b and b) a pretrained word embedding	82.2	82.2	82.5	84.5
1.Cutoff reviews after 150 words. 2.Restrict training samples to 3000. 3. Validate 10,000 samples. 4. Consider only the top 10,000 words. 5. Before Embedding layer, add a & b and b) a pretrained word embedding	79.3	82.6	82.0	83.9
1.Cutoff reviews after 150 words. 2.N0 Restriction. 3. Validate 10,000 samples. 4. Consider only the top 10,000 words. 5. Before Embedding layer, add a & b and b) a pretrained word embedding	80.0	78.5	79.7	77.2

Conclusion:

The performance of the first basic sequence model performance comparison is almost equal to pretrained word embedding as per the results. Also, we need to consider the sample size, as we increase the sample size there is

increment in the accuracy of the model. Initially we started with 100 samples and other conditions, there was performance decay, later as we increased the maximum amount, the overall performance is good,

Later we compare the embedded and masked embedded, Embedded is better than the mask embedded in some of the cases. Overall, the pretrained (glove) technique is better than the first basic sequence.