Name: Nada Taher Elwazane ID: 6876
Name: Mariem Mostafa Mahmoud ID: 6873
Name: Haguar Tarek Dessouky ID: 6878
Name: Manar Amgad Helal ID: 7113

Artificial Intelligence
Project 4

Contents

Name: Nada Taher Elwazane ID: 6876
Name: Mariem Mostafa Mahmoud ID: 6873
Name: Haguar Tarek Dessouky ID: 6878
Name: Manar Amgad Helal ID: 7113

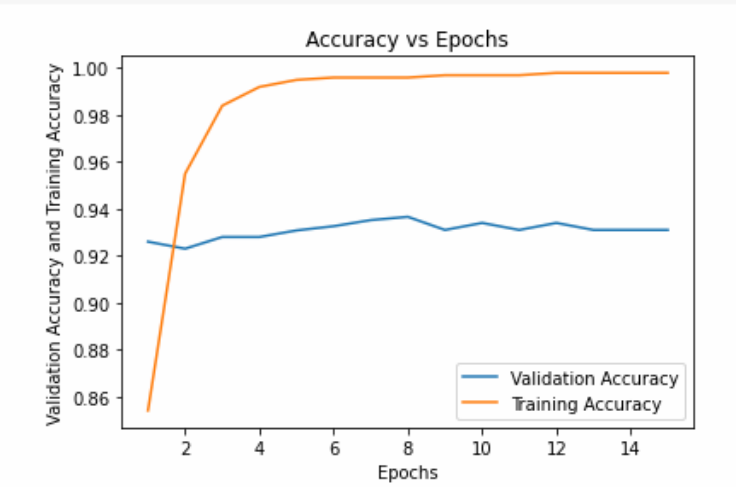
Artificial Intelligence
Project 4

Movies Reviews Classification Using BERT

Graphs representing the change of training and validation accuracies with the number of training epochs:

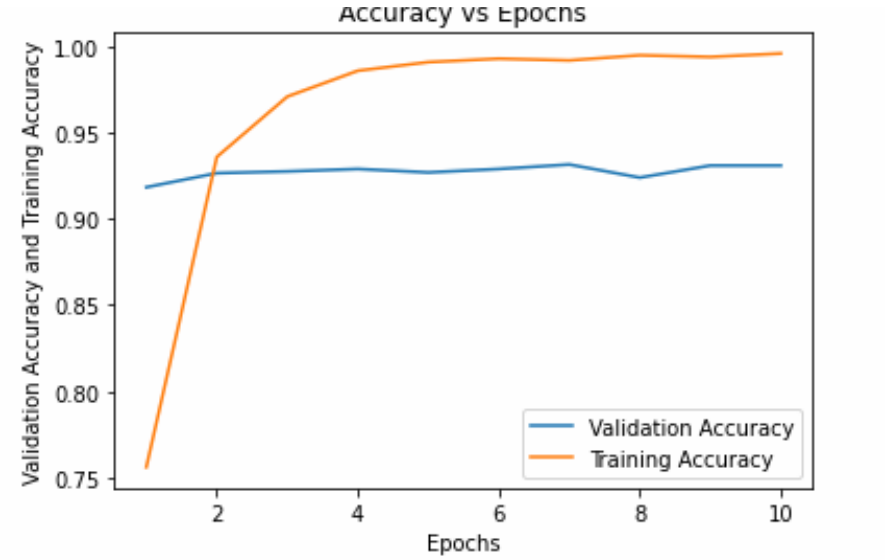
Without Preprocessing:

LR = 5e^-6



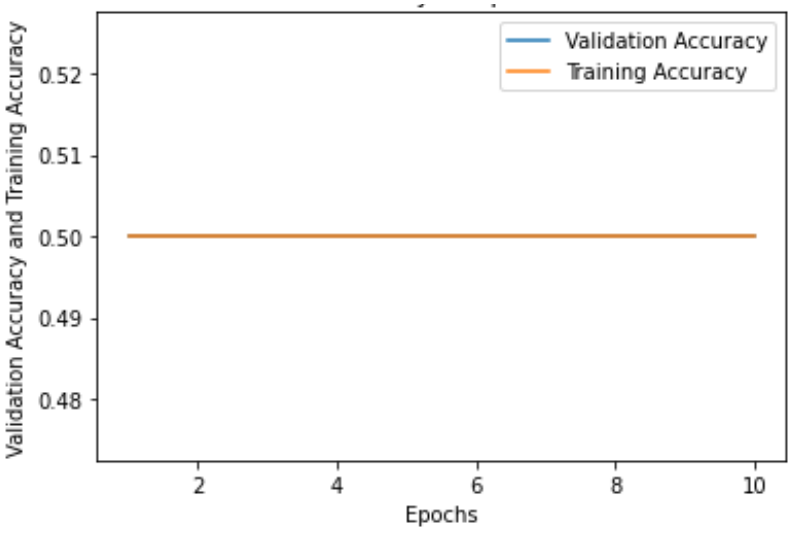
The training accuracy increases until it reaches 0.99 then remains constant while the validation accuracy stabilizes around 0.930.

LR = 5e^-5



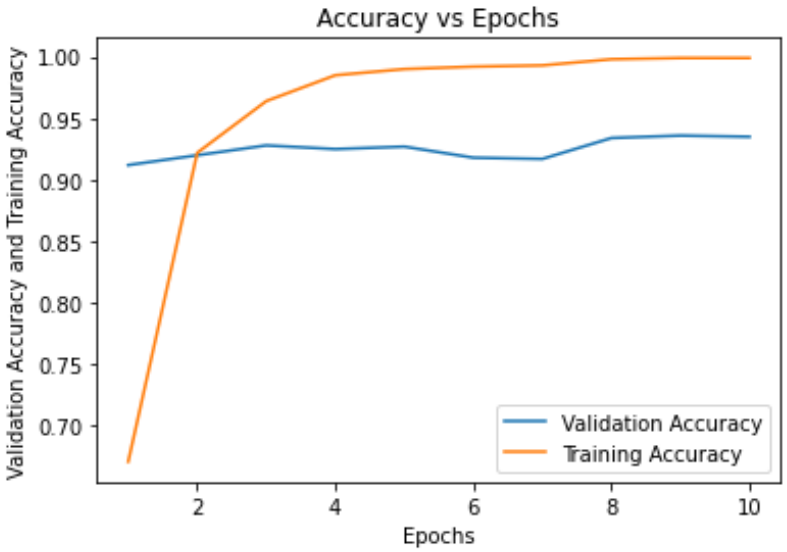
The training accuracy increases until it reaches 0.99 then remains constant while the validation accuracy stabilizes around 0.930.

LR = 5e^-4



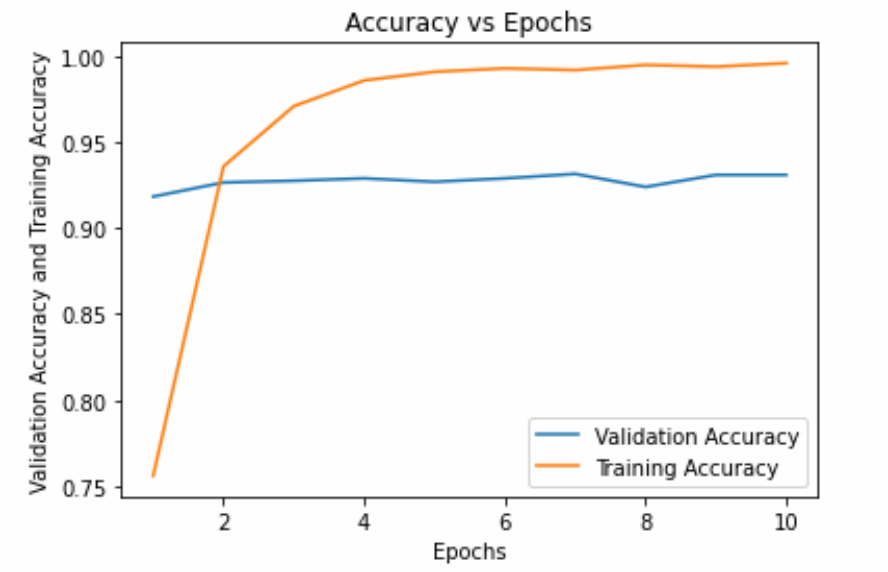
Both validation and training accuracy are constant with a value of 0.500.

LR = 1e^-6



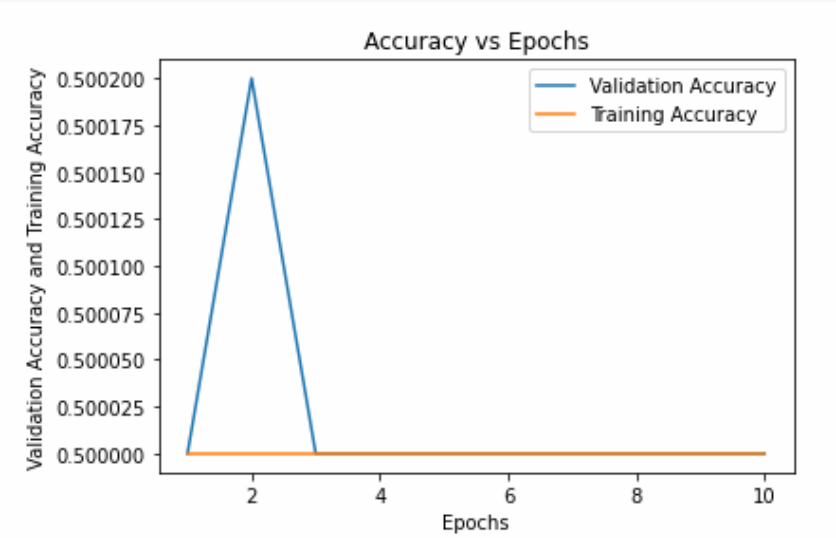
The training accuracy increases until it reaches 0.99 then remains constant while the validation accuracy stabilizes around 0.92.

LR = 1e^-5



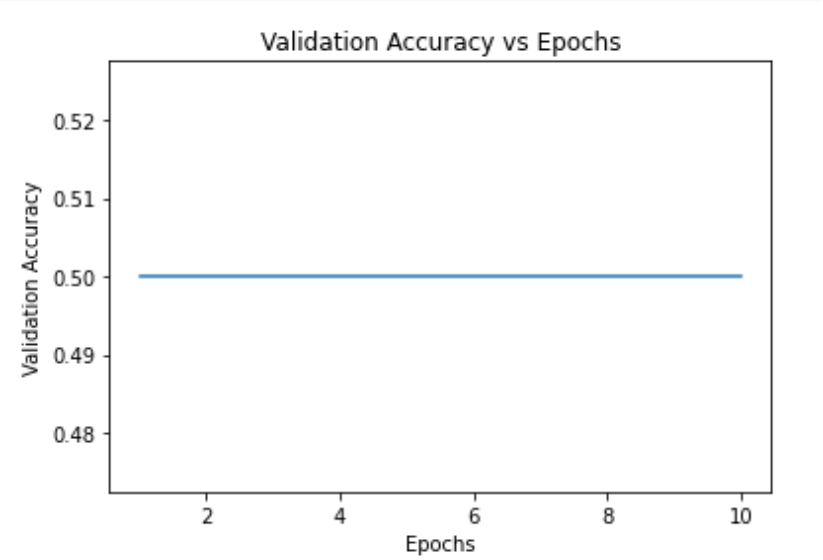
The training accuracy increases until it reaches 0.99 then remains constant while the validation accuracy stabilizes around 0.93.

LR = 1e^-4



Both validation and training accuracy are constant value around 0.500.

LR = 1e^-3



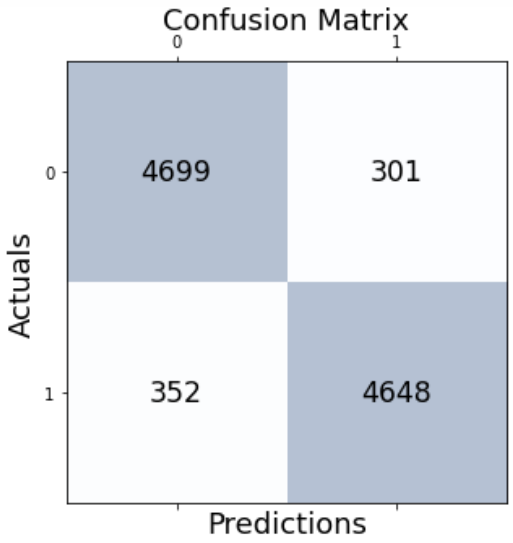
Both validation and training accuracy are constant with a value of 0.500.

Comparison of all the learning rates for that model

```
[ ] LR

    5e-06
```

```
[ ] best_epoch

    8
```

```
    valid_accuracy_max

    0.9366
```




```
    Test Accuracy:  0.935
```

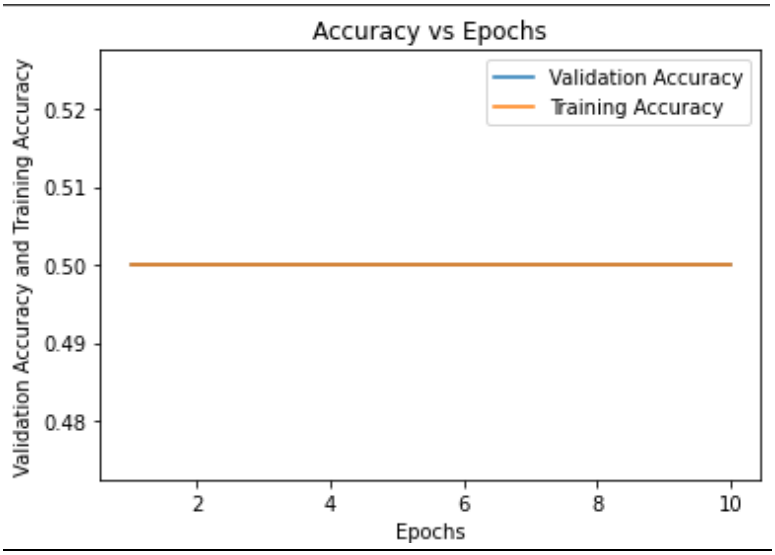
```
[ ] tp=4699
    tn=4648
    fp=352
    fn=301

    evalscore(tp,tn,fp,fn)
```

```
    F1-score:    0.935
    Precision:   0.930
    Recall:      0.940
    Specificity: 0.930
```

5e-6 has the best performance out of all the learning rates with a testing accuracy of 0.935 and the same f1-score, which is expected as the data is balanced. Recall is impressive at a 0.940. The recall measures the model's ability to detect positive samples. The higher the recall, the more positive samples detected.

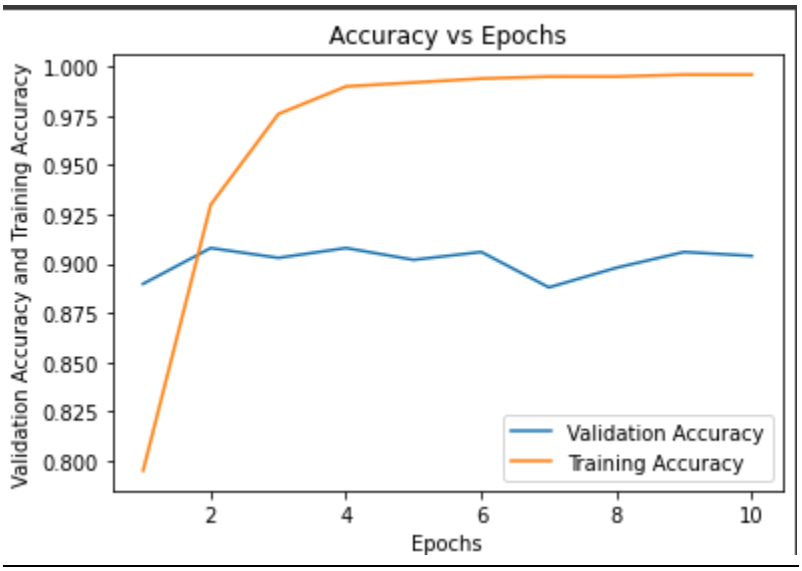
With Preprocessing:

LR = 1e^-3



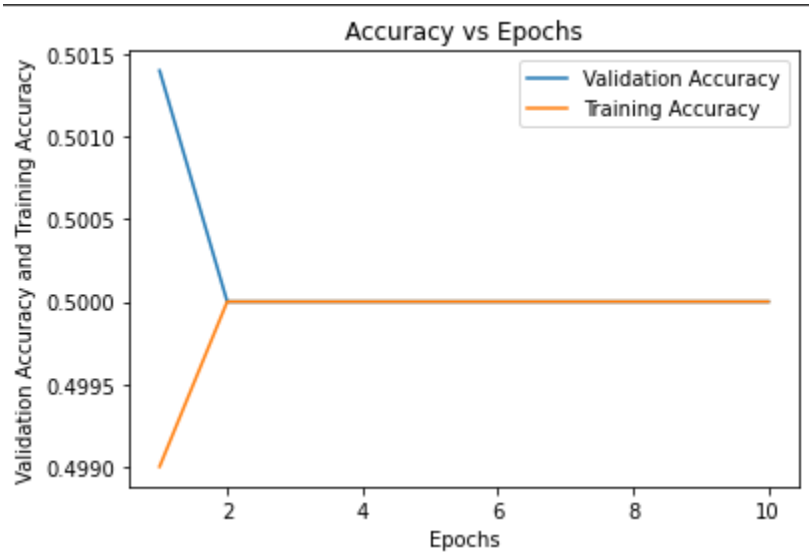
Both validation and training accuracy are constant with a value of 0.500.

LR = 1e^-5



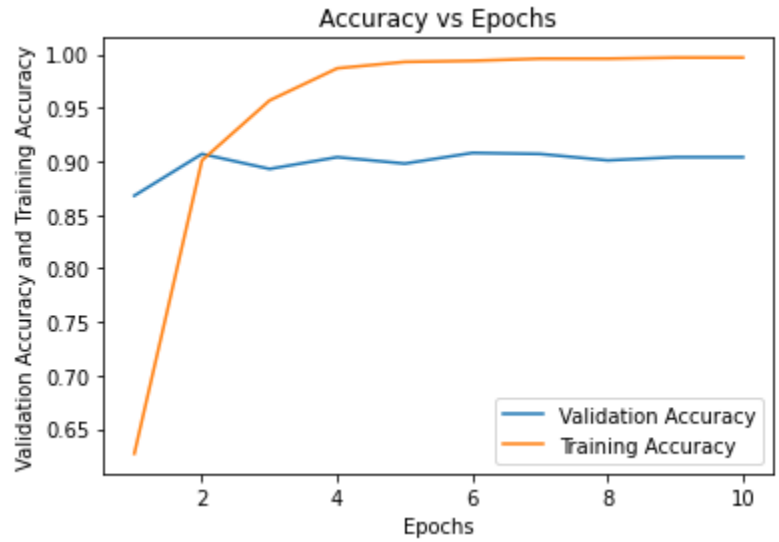
The training accuracy increases until it reaches 0.99 then remains constant while the validation accuracy stabilizes around 0.88.

LR = 5e^-5



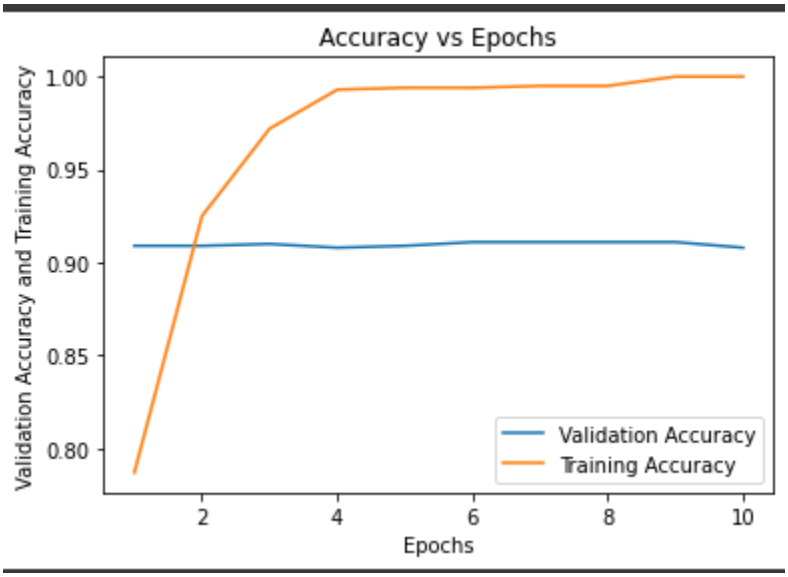
Both validation and training accuracy are constant with a value of 0.500.

LR = 5e^-6



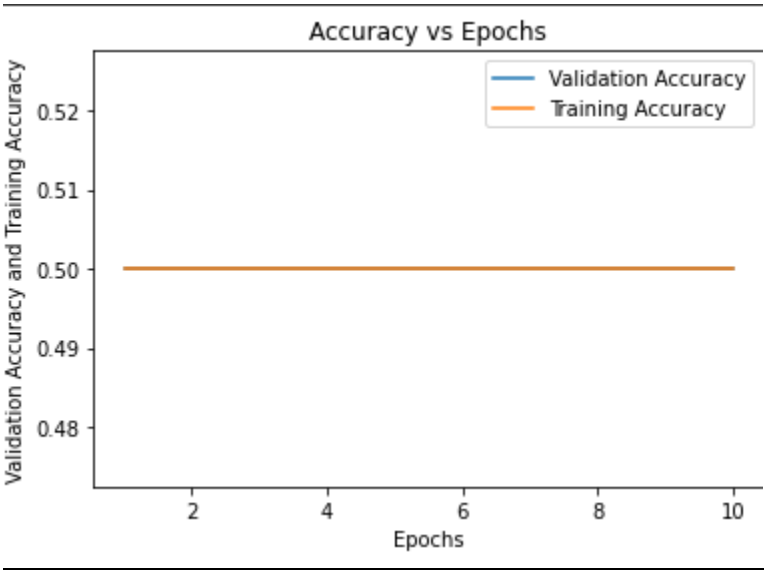
The training accuracy increases and approaches 1 while the validation accuracy stabilizes around 0.908.

LR = 1e^-6



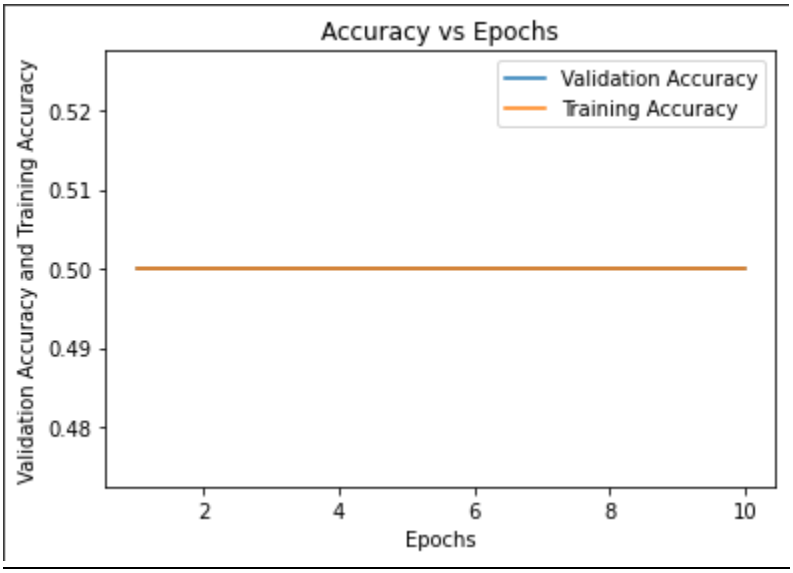
The training accuracy increases until it reaches 0.99 then remains constant while the validation accuracy stabilizes around 0.91.

LR = 5e^-4

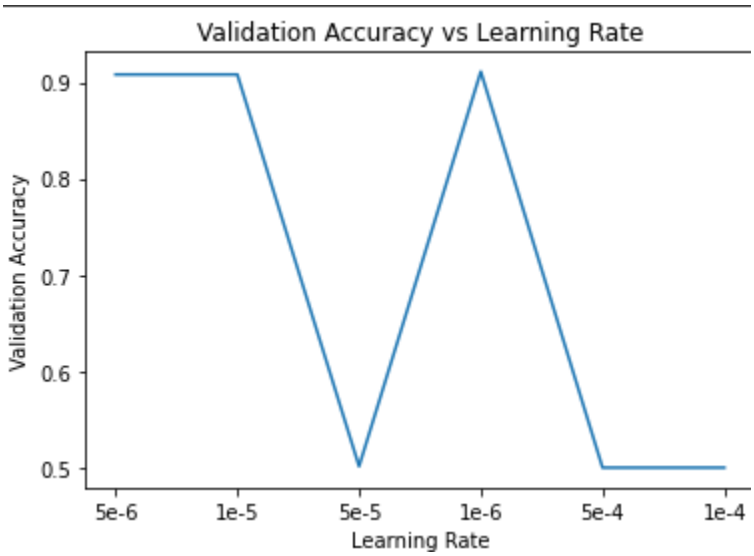


Both validation and training accuracy are constant with a value of 0.500.

LR = 1e^-4



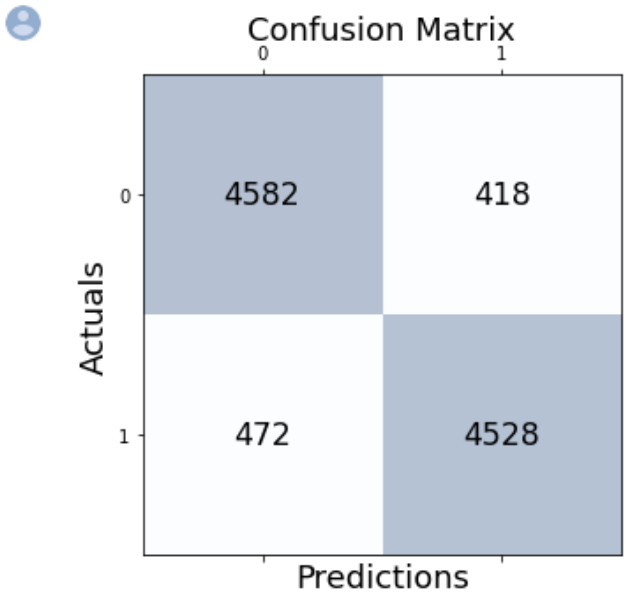
Both validation and training accuracy are constant with a value of 0.500.



Comparison of all the learning rates for that model

```
[ ]  LR

     1e-06

[ ]  valid_accuracy_max

     0.9114

[ ]  best_epoch

     6
```

```
F1-score:    0.911
Precision:   0.907
Recall:      0.916
Specificity: 0.906
Accuracy:    0.911
```



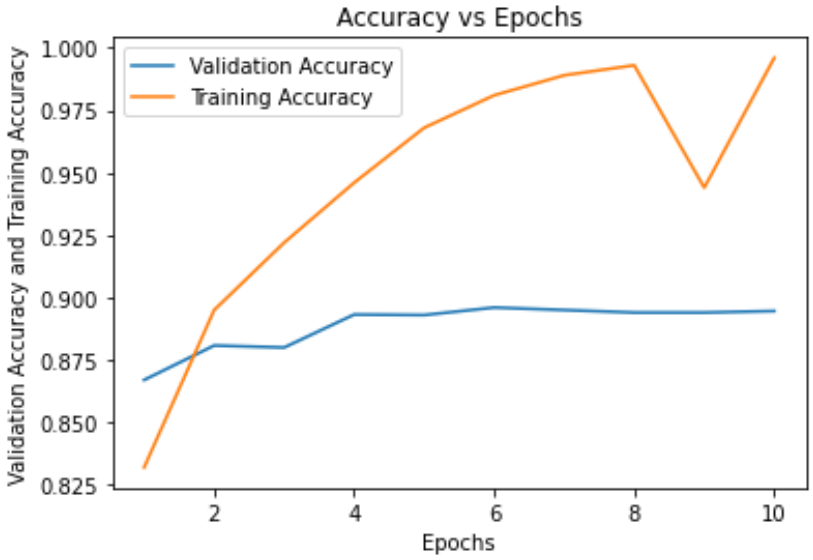
1e-6 has the best performance out of all the learning rates with a testing accuracy of 0.911 and the same f1-score, which is expected as the data is balanced. Recall is impressive at a 0.916. The recall measures the model's ability to detect positive samples. The higher the recall, the more positive samples detected.

Bonus:

Preprocessing:

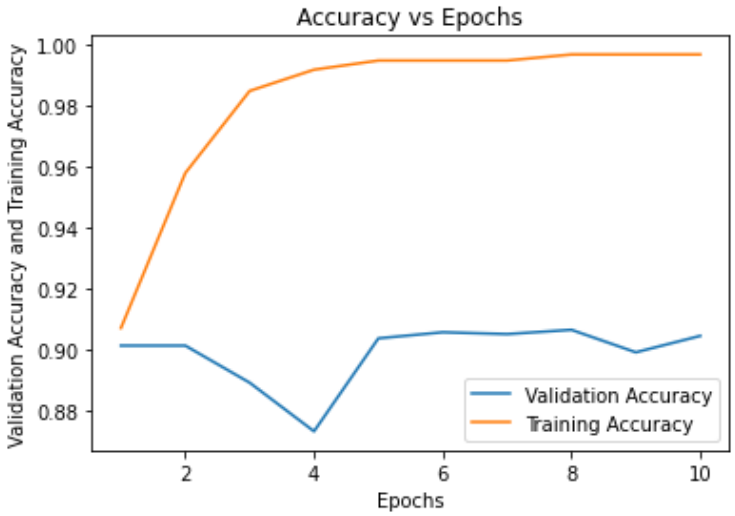
768,64,128,256,128,64,2

1e-6



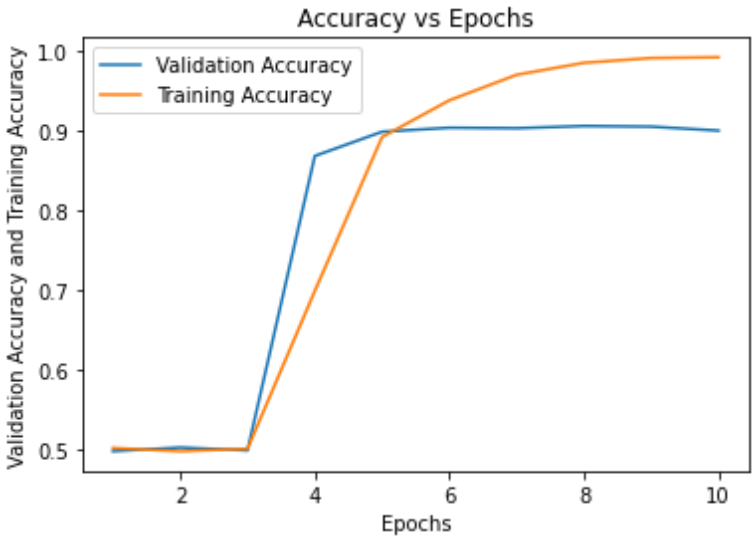
This model has reached 100% training accuracy. And validation accuracy is mostly stable at 90%.

5e-6



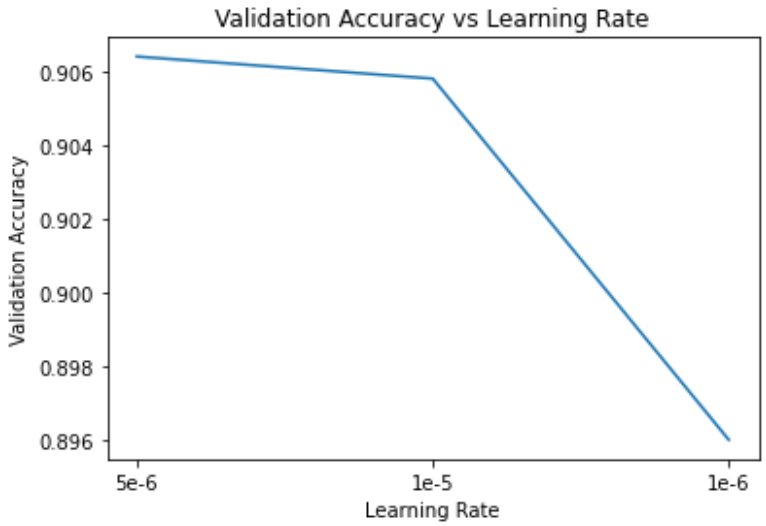
Validation accuracy is capped at 90-91% which training accuracy jumps from 0.907 in the first epoch to 0.997

1e-5



Training accuracy increases to 99.2% which validation accuracy has a max of 0.9058

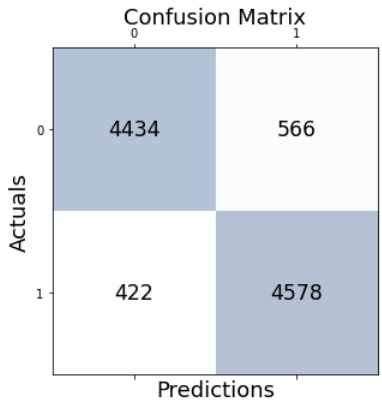
Comparison of all the learning rates for that model



Best Model has a learning rate of 5e-6 and epoch of 8

```
[ ] evaluate(model, test_data)

    Test Accuracy:  0.901
```

```
conf_mat(test_labels,pred_labels)
```



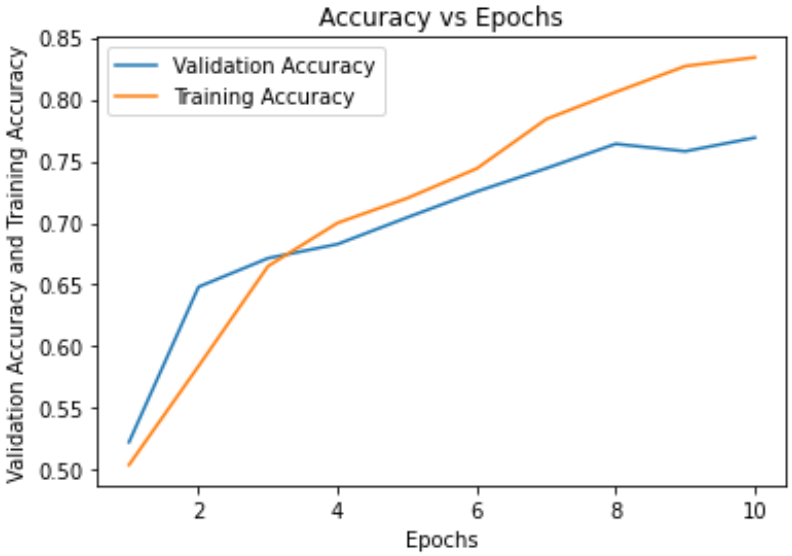

```
evalscore(tp,tn,fp,fn)
```

```
F1-score:    0.900
Precision:   0.913
Recall:      0.887
Specificity: 0.916
```

5e-6 has the best performance out of all the learning rates with a testing accuracy of 0.901 and almost the same f1-score, which is expected as the data is balanced ,specificity is impressive at a 0.916. specificity is the proportion of true negatives that are correctly predicted by the model.

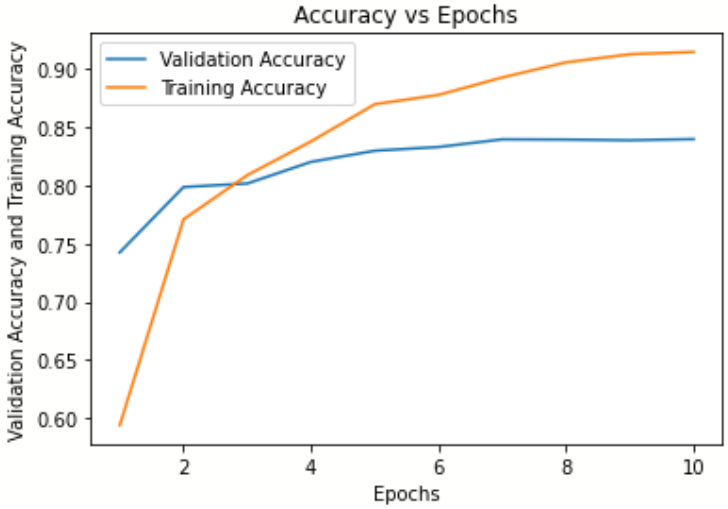
768,1024,256,16,4,2

1e-6



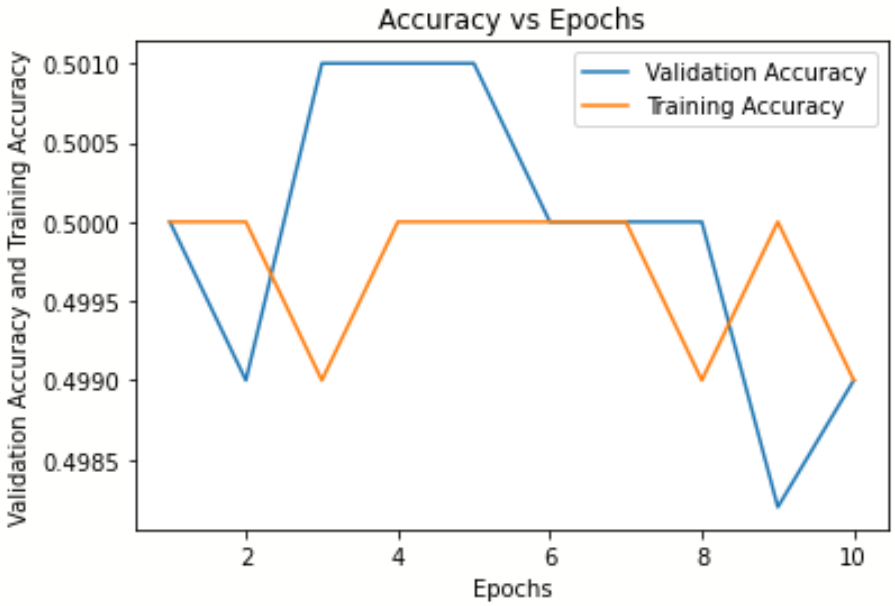
Validation and Training accuracy are steadily increasing, with a maximum validation accuracy of 0.769 and a maximum training accuracy of 0.834.

5e-6



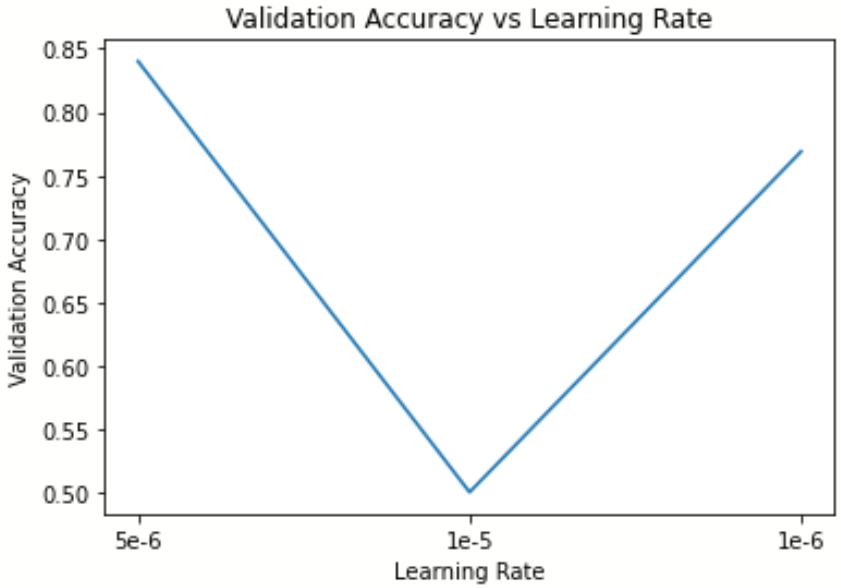
Training Accuracy is steadily increasing with a maximum of 0.915, and validation is mostly steady at 0.80 give or take. And a maximum of 0.840

1e-5



Model does not appear to learn with each epoch further epochs might be needed.

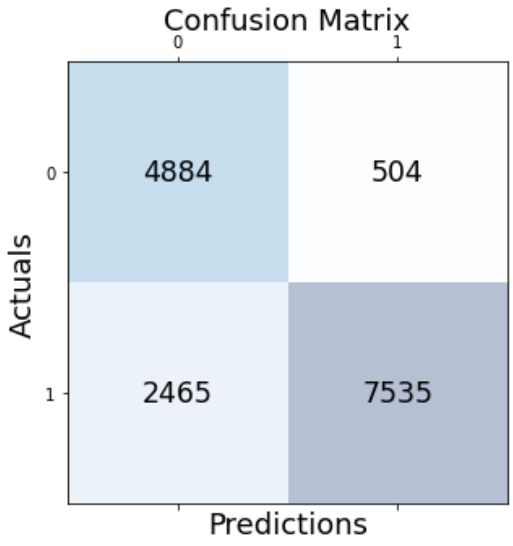
Comparison of the learning rates of that model



5e-6 has the maximum validation accuracy, epoch 10 and a validation accuracy of 0.84

```
[ ]  evaluate(model, test_data)

     Test Accuracy:  0.840
```

```
conf_mat(test_labels,pred_labels)
```



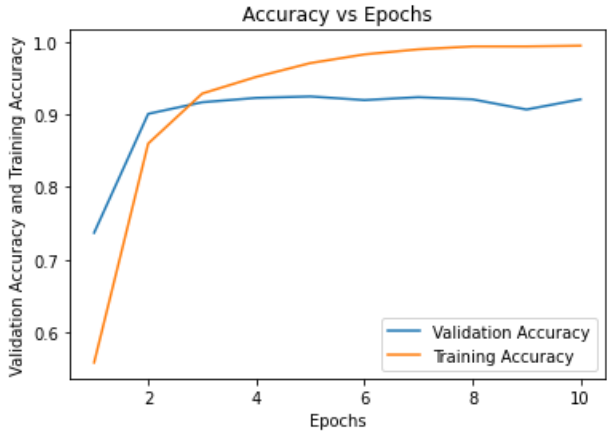

```
evalscore(tp,tn,fp,fn)
```

```
F1-score:    0.767
Precision:   0.665
Recall:      0.906
Specificity: 0.753
```

The best model has poor performance comparing to the rest of the models. Recall is impressive at a 0.906. The recall measures the model's ability to detect positive samples. The higher the recall, the more positive samples detected.

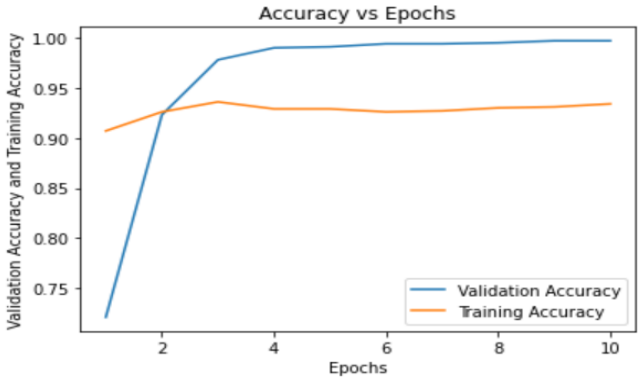
768,32,128,512,128,32,2

1e-6



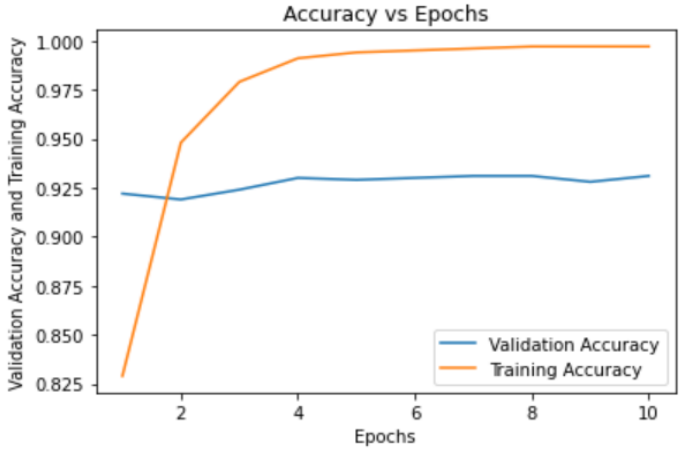
The validation accuracy started from around 0.7 then increased until it reached 0.9 in which it nearly stabilized. The training accuracy increased from below 0.6 until it reached 0.99 and then it stabilized.

1e-5



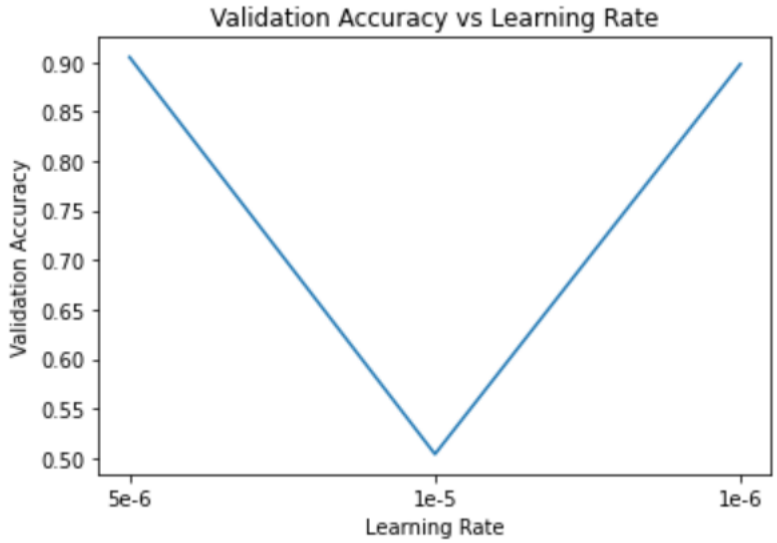
Validation accuracy is quickly approaching 1 (with a maximum of 0.996). While training accuracy is fluctuating around 0.925, with a maximum of 0.9328.

5e-6



Validation accuracy is quickly approaching 1 (with a maximum of 0.996). While training accuracy is fluctuating around 0.925, with a maximum of 0.9328.

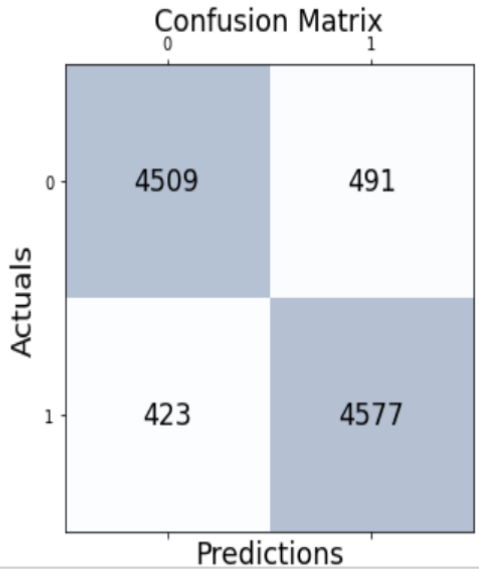
Comparison of the learning rates of that model



Best model is at 5e-6 with a validation accuracy of 0.905, epoch 3.

```
[ ]  evaluate(model, test_data)

     Test Accuracy:  0.909
```

```
[ ]  conf_mat(test_labels,pred_labels)
```




```
evalscore(tp,tn,fp,fn)
```

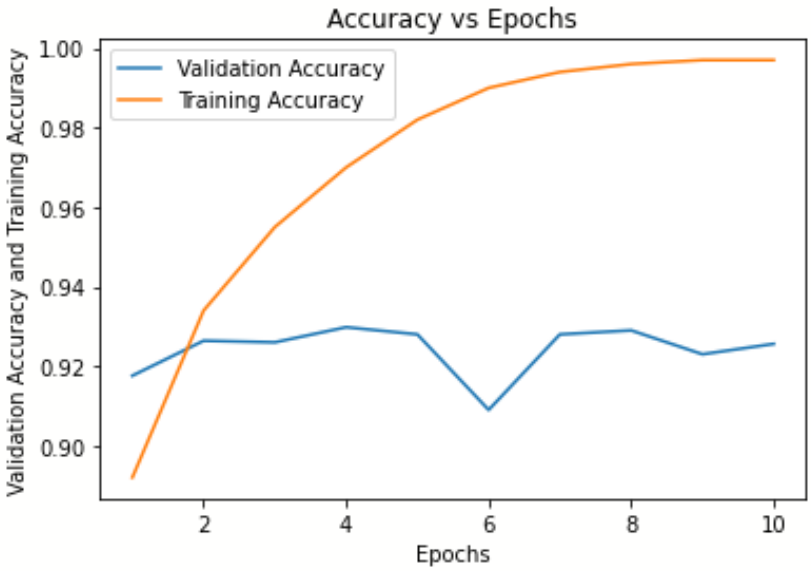
```
F1-score:    0.908
Precision:   0.914
Recall:      0.902
Specificity: 0.915
```

Compared to other models with preprocessing it has good performance .F1-score,precision,recall,specificity are high . specificity is impressive at a 0.915. specificity is the proportion of true negatives that are correctly predicted by the model.

No preprocessing:

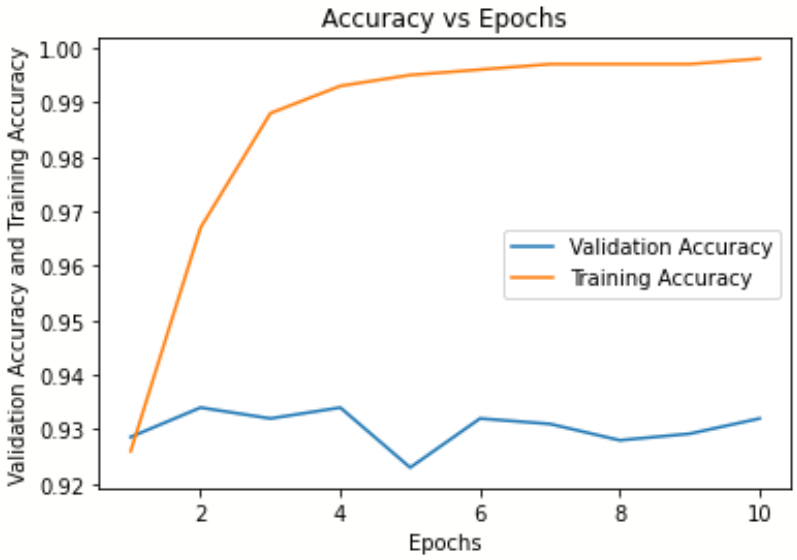
768,64,128,256,128,64,2

1e-6



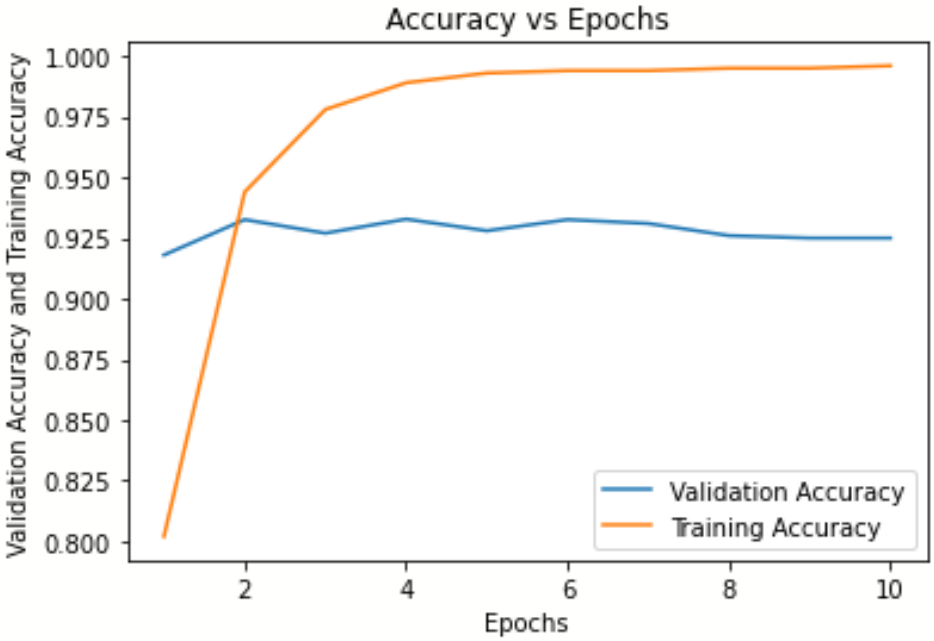
Training accuracy is increasing and quickly approaching 1. While validation accuracy is fluctuating around 0.92, with a maximum of 0.9298.

5e-6



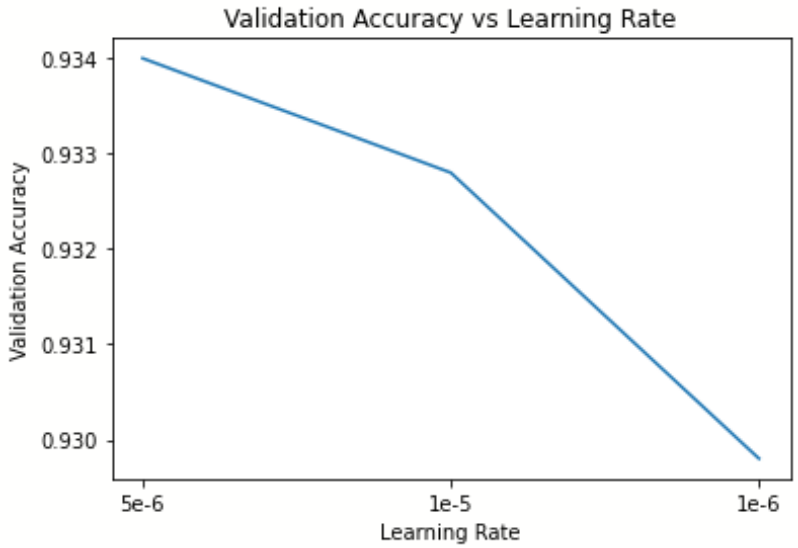
Training accuracy is increasing and quickly approaching 1 (with a maximum of 0.998). While validation accuracy is fluctuating around 0.93, with a maximum of 0.934.

1e-5



Training accuracy is quickly approaching 1 (with a maximum of 0.996). While validation accuracy is fluctuating around 0.925, with a maximum of 0.9328.

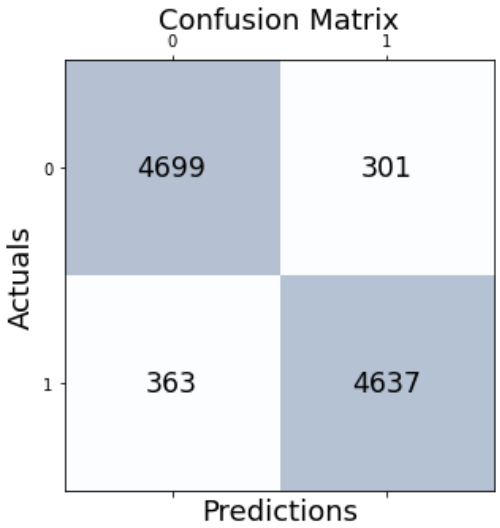
Comparison of the learning rates of that model



The best model has a learning rate of 5e-6, epoch 4 and a validation accuracy of 0.9344

```
evaluate(model, test_data)
Test Accuracy:  0.934
```

```
[ ] conf_mat(test_labels,pred_labels)
```




```
evalscore(tp,tn,fp,fn)
```

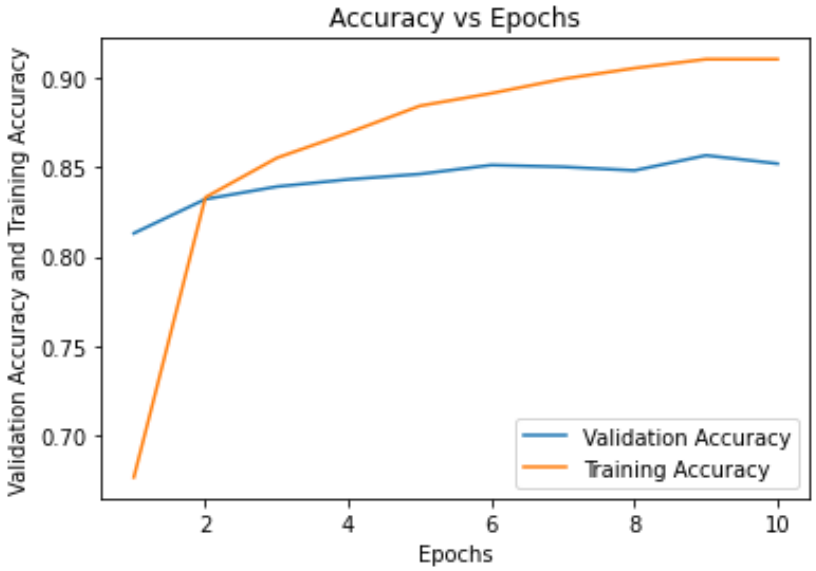
```
F1-score:    0.934
Precision:   0.928
Recall:      0.940
Specificity: 0.927
```

The f1 score and test accuracy are similar because the data is balanced.

Recall is impressive at a 0.940The recall measures the model's ability to detect positive samples. The higher the recall, the more positive samples detected.

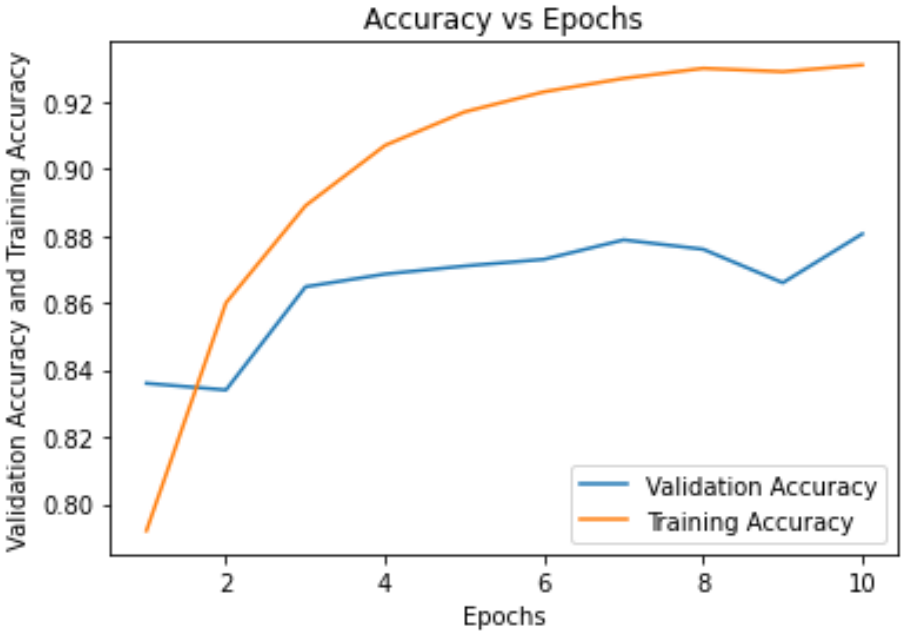
768,1024,256,16,4

1e-6



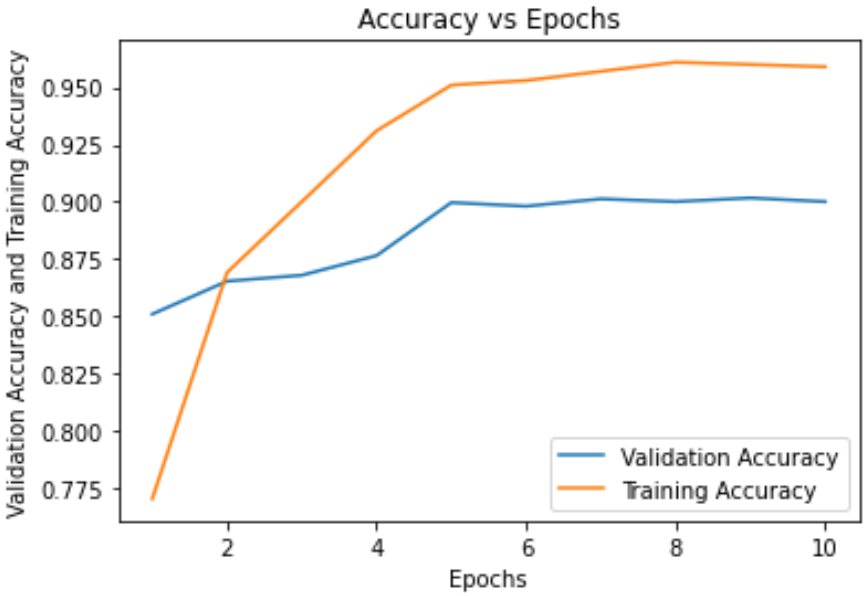
Training accuracy is increasing and stabilizing at around 0.90 (with a maximum of 0.910), and a validation accuracy fluctuating 0.84 (with a maximum of 0.8564)

5e-6



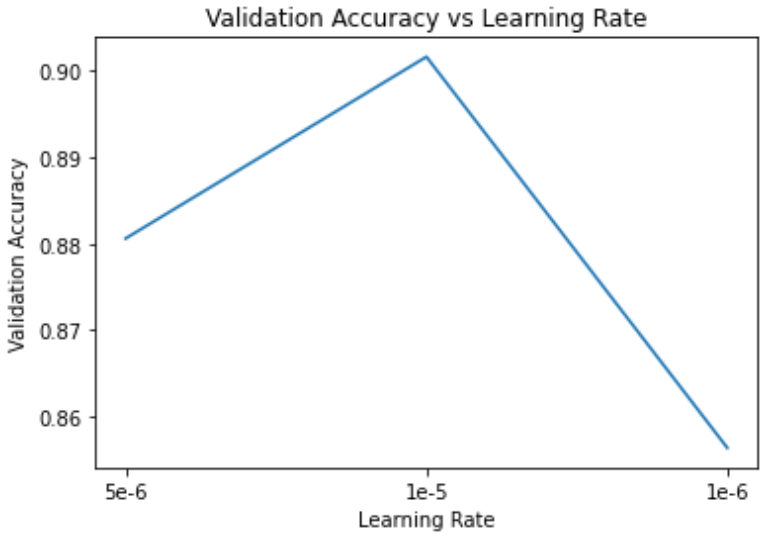
Training accuracy is stabilizing around 0.92 (with a maximum of 0.931) and is a maximum validation accuracy 0.8806

1e-5



Training Accuracy stabilizes around 0.950, while validation accuracy has a maximum of 0.9016.

Comparison of the learning rates of that model

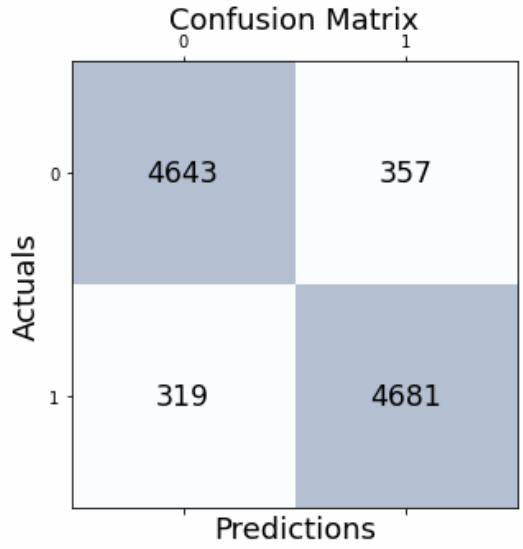


Best model is at 1e-5 with a validation accuracy of 0.9016, epoch 9.

```
evaluate(model, test_data)
```

```
Test Accuracy:  0.932
```

```
conf_mat(test_labels,pred_labels)
```



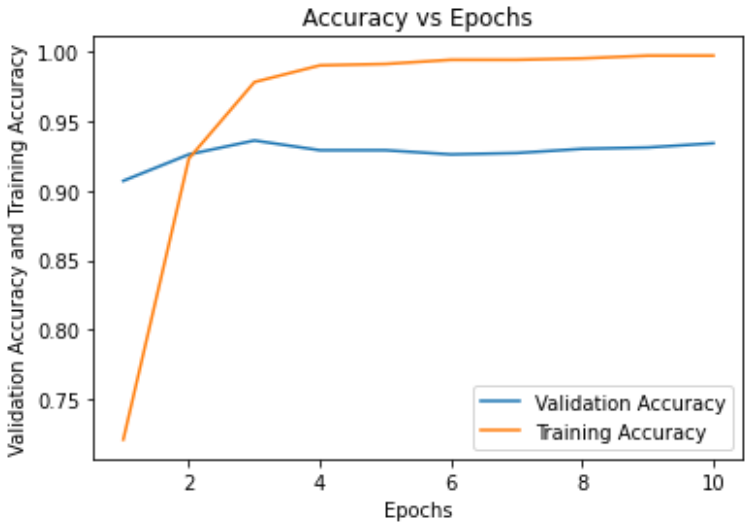

```
evalscore(tp,tn,fp,fn)
```

```
F1-score:    0.932
Precision:   0.936
Recall:      0.929
Specificity: 0.936
```

The model has good performance. Specificity is impressive at a 0.936. Specificity is the proportion of true negatives that are correctly predicted by the model.

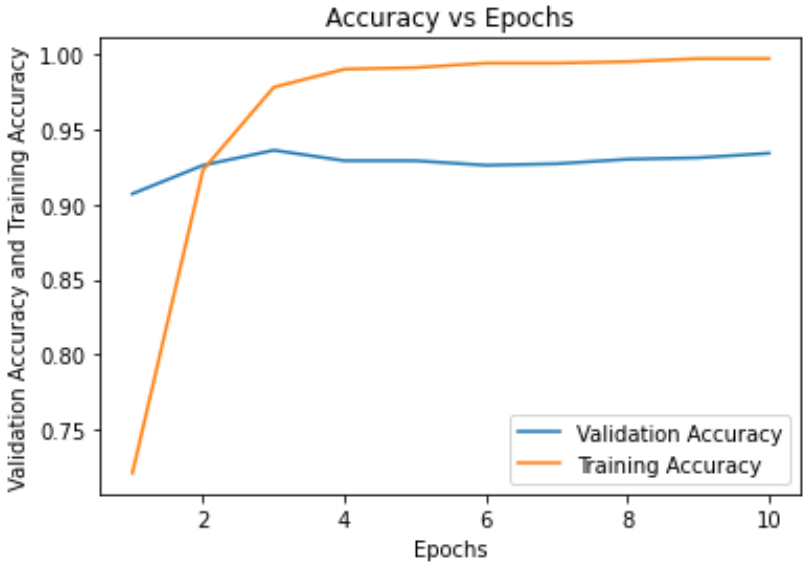
768,32,128,512,128,32,2

1e-6



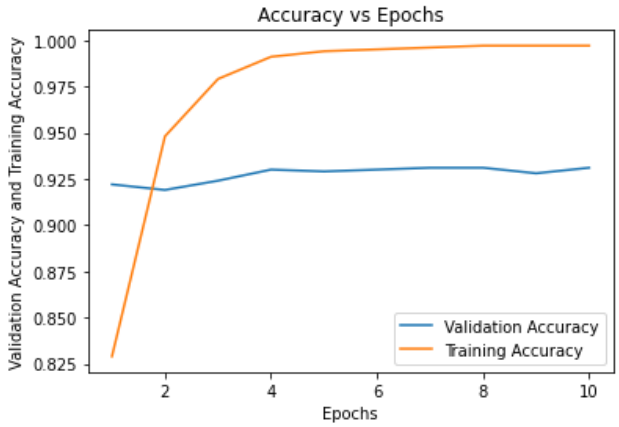
The Training accuracy started from around 0.7 then increased until it reached 0.995. The validation accuracy is stable around 0.92 with a maximum of 0.925.

1e-5



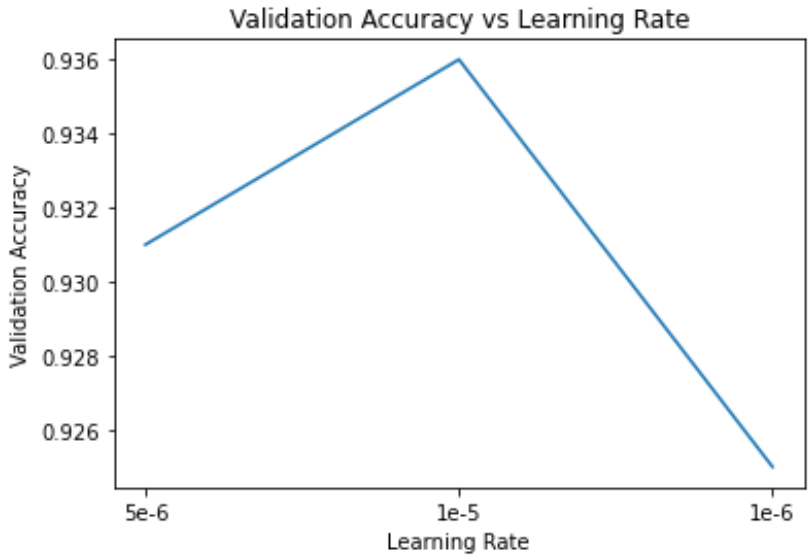
Training accuracy is quickly approaching 1 (with a maximum of 0.996). While validation accuracy is fluctuating around 0.930, with a maximum of 0.936.

5e-6



Training accuracy is quickly approaching 1 (with a maximum of 0.997). While validation accuracy is fluctuating around 0.930, with a maximum of 0.931.

Comparison of the learning rates of that model

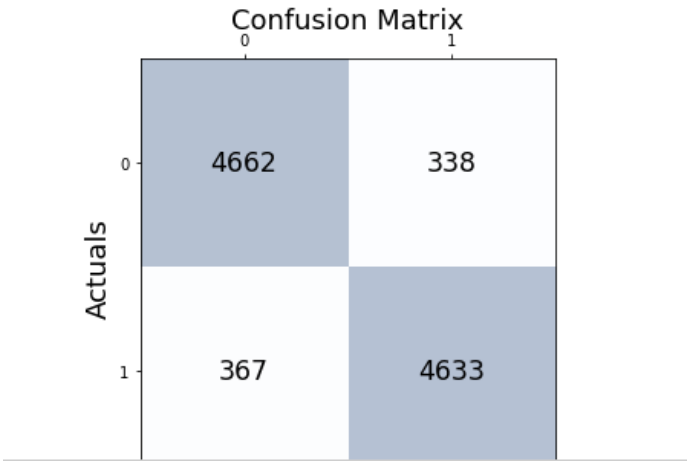


Best model is at 1e-5 with a validation accuracy of 0.936, epoch 3.

```
[ ]  valid_accuracy_max

     0.936
```

```
     evaluate(model, test_data)

     Test Accuracy:  0.929
```

```
[ ]  conf_mat(test_labels,pred_labels)
```




```
[ ]  tp=4662
     tn=4633
     fp=367
     fn=338

     evalscore(tp,tn,fp,fn)
```

```
     F1-score:    0.930
     Precision:   0.927
     Recall:      0.932
     Specificity: 0.927
```

The model has good performance. Recall is impressive at a 0.932.The recall measures the model's ability to detect positive samples. The higher the recall, the more positive samples detected.

Overall Comparison:

A general pattern of these models is that applying preprocessing tends to reduce accuracy of the best model. After preprocessing, BERT's performance of text classification decreased.

Preprocessing is not needed when using pre-trained language representation models like BERT. In particular, it uses all of the information in a sentence, even punctuation and stop-words.

Best model is 768,512,256,128,64,2 for both preprocessing and no preprocessing

For no preprocessing the validation accuracy:0.936

For no preprocessing the Testing accuracy:0.935

For preprocessing the validation accuracy: 0.9114

For preprocessing the Training accuracy:0.911