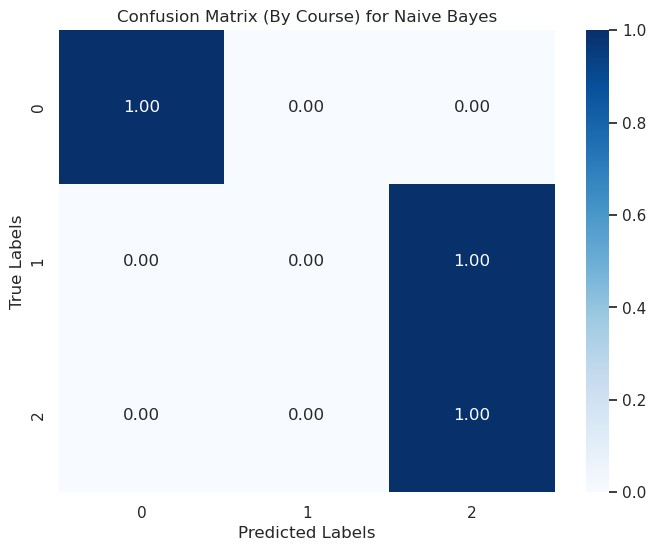
|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Overall Accuracy | “Easy” (Class 0) | | | “Medium” (Class 1) | | | “Hard” (Class 2) | | |
|  | Precision | Recall | F1 | Precision | Recall | F1 | Precision | Recall | F1 |
| Review Classification  (n = 333) | 0.69 | 0.67 | 0.47 | 0.55 | 0.57 | 0.27 | 0.31 | 0.77 | 0.86 | 0.61 |
| Course Classification  (n = 7) | 0.71 | 1.0 | 1.0 | 1.0 | 0.0 | 0.0 | 0.0 | 0.67 | 1.0 | 0.80 |

# **NAIVE BAYES**

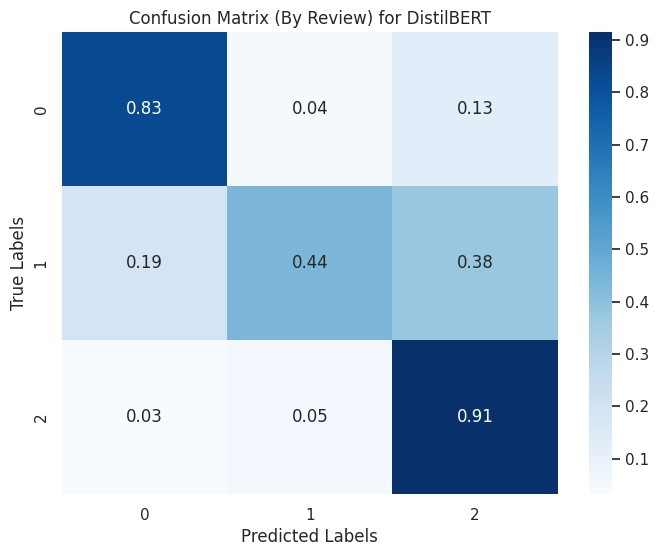


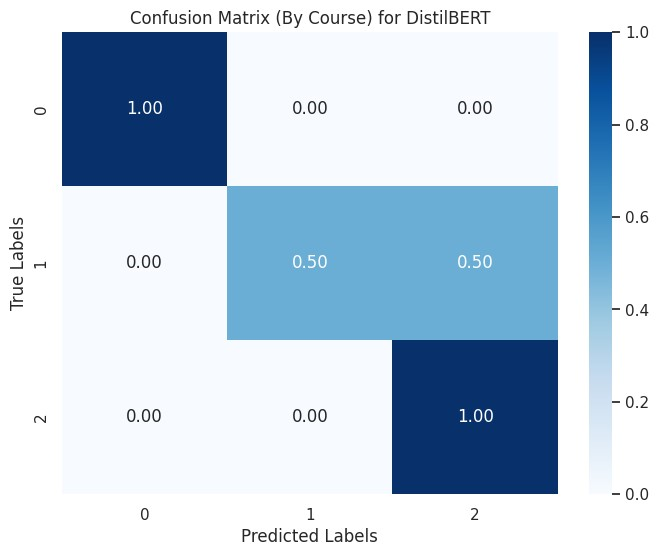
We classify the difficulty of both course reviews and courses because the main objective of our paper is to be able to determine the difficulty of a course, however we have significantly more data points for reviews (test data: n = 133) than courses (test data: n = 7). The difficulty rating of a given course is the desired output, but the difference

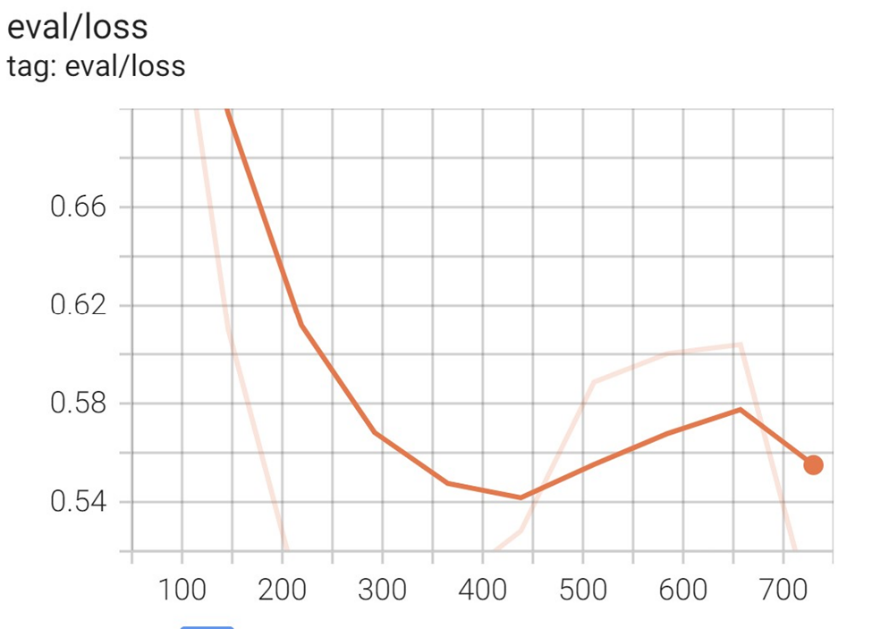
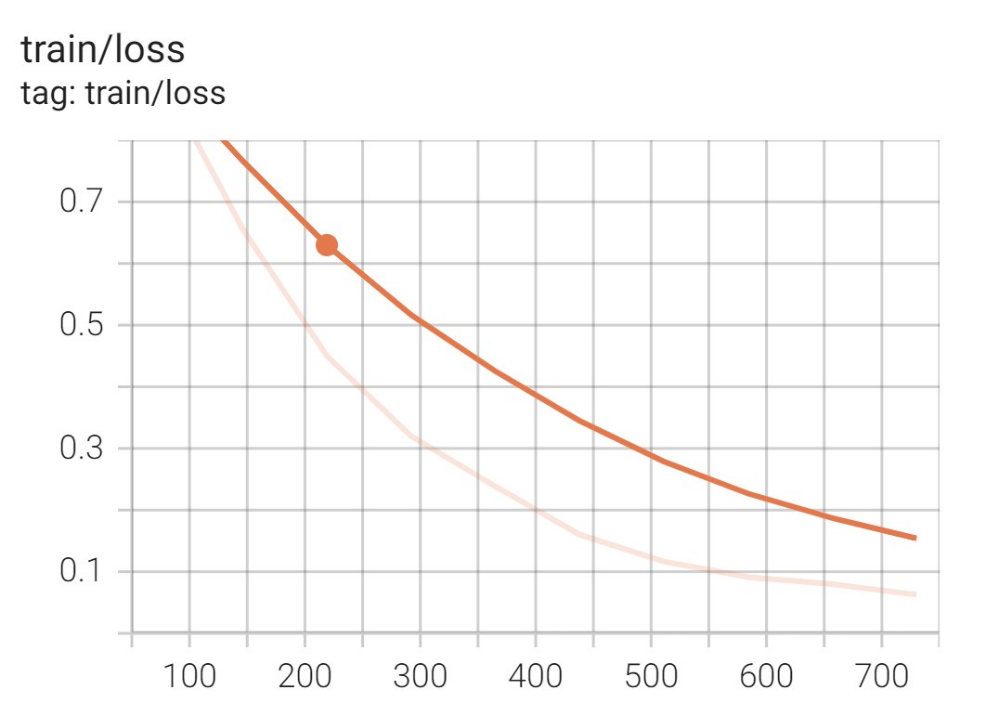
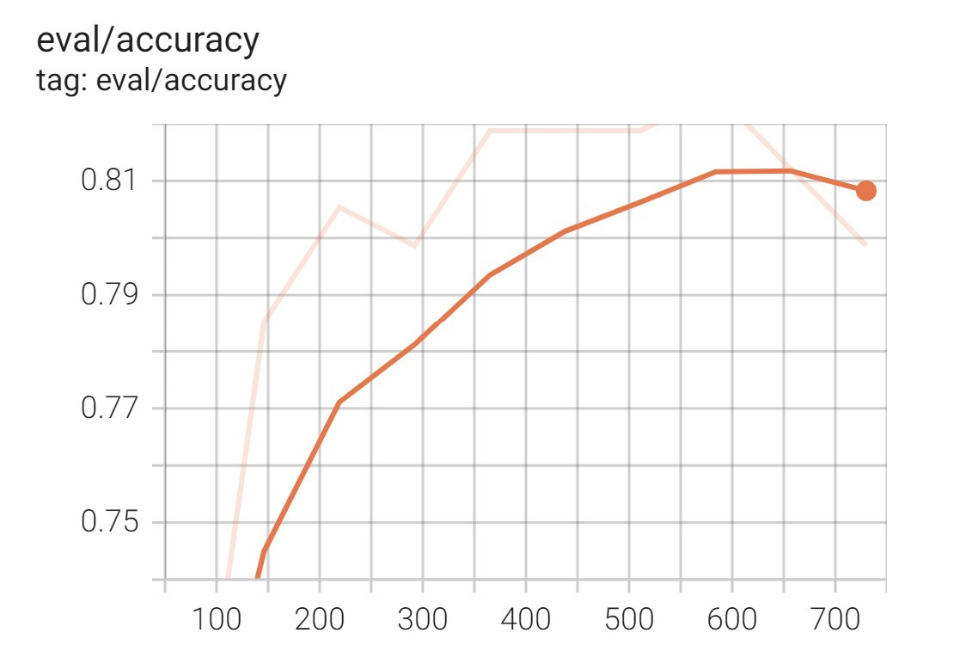
On the test dataset, we see that Naive Bayes achieved an accuracy of 69% on the test for review classification (n = 333) and 71% for course classification (n = 7). We provide

# **BERT**

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | Overall Accuracy | “Easy” (Class 0) | | | “Medium” (Class 1) | | | “Hard” (Class 2) | | |
|  | Precision | Recall | F1 | Precision | Recall | F1 | Precision | Recall | F1 |
| Review Classification  (n = 333) | 0.80 | 0.71 | 0.70 | 0.70 | 0.76 | 0.52 | 0.62 | 0.68 | 0.90 | 0.78 |
| Course Classification  (n = 7) | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |







# Key Insights:

* BERT achieves overall higher accuracy for both review and course classification
* Easy class -> priority is minimizing false positives (precision)
  + Review classification
    - Both BERT and Naïve Bayes achieved 70% precision, which is a little lower than we want it to be
    - However, BERT only misclassifies 13% of easy reviews as hard while Naïve Bayes misclassifies 35%
  + Course classification
    - Both BERT and Naïve Bayes achieved 100% precision which is great and do not misclassify any easy classes
* Medium class -> want to have balance between both precision (minimize false positives) and recall (minimizing false negatives)
  + Review classification
    - From Naïve Bayes to BERT, improved f1 from 0.38 to 0.54
    - 0.54 f1 is still not as high as we want though, for both models, medium class has lowest f1 score compared to other classes (ambiguous, hard to correctly classify)
  + Course classification
    - From Naïve Bayes to BERT, improved f1 from 0.0 to 0.67
    - Naïve Bayes did extremely poor in classifying medium courses, also representative of limited sample size of courses
* Hard class -> priority is minimizing false negatives or recall (misclassifying hard classes as something else)
  + Review classification
    - Both BERT and Naïve Bayes achieved 91% recall which is great, misclassifying only 9% of classes as something else
  + Course classification
    - 100% recall for courses which is also great
* BERT (Training and Test Loss)
  + Darker curve is smoothed out version of other curves
  + Able to achieve training loss of 15% over 700 steps, takes 700 steps, consistently decreases
  + Test accuracy converges around 700 steps at 80% (that is why stopped there)
  + Test loss minimized at around 450 steps and then actually increases and then decreases that (overfitting?)