Assignment 4

Jackson Thissell, Jakub Suran, Joe Rumery

April 9, 2024

Abstract

This paper is a report for Assignment 4 of COS 470 - Text Mining. We focused on the DistilRoBERTa model and its use for genre classification from song lyrics. We also investigated the effect of three hyperparameters (namely batch size, learning rate, and Adam β_1) on the DistilRoBERTa model. Each team member explored one of DistilRoBERTa's hyperparameters as described later in the paper. Lastly, we combined all of our codes together and collaborated on this report.

1 Model Description

RoBERTa was developed by researchers at the University of Washington and Facebook as an improvement to Google's BERT model. RoBERTa is based upon the same transformer architecture of BERT, with improvements to the pretraining methodology and model hyperparameters. It also is pretrained on a larger corpus. The BERT model masks tokens as a data preprocessing step, leaving them fixed during training, while RoBERTa masks tokens dynamically during training to increase the breadth of the token space that the model learns to encode.

DistilRoBERTa, a RoBERTa model with fewer parameters and an increased speed, retains its accuracy and performance using a Student-Teacher methodology known as distillation, where the loss of the student model is informed by the teacher model, and the architectures of the student and teacher are identical outside of layer count and depth. For this assignment, we chose to use DistilRoBERTa for sequence classification since it combines RoBERTa's accuracy and breadth of knowledge with the speed and lightweight form of a distilled model.

2 Hyperparameter Tuning

We picked three hyperparameters to investigate. The selected hyperparameters were Batch Size, Learning Rate, and Adam β_1 picked by Jakub, Jackson, and Joe respectively. Next, we provide a description of each hyperparameter and its effect on the DistilRoBERTa model.

Batch size	F_1	Learning rate	F_1	Adam Beta 1	F_1
4	65.59	1×10^{-5}	65.81	0.85	70.24
8	64.46	5×10^{-5}	70.24	0.9	69.88
16	62.79	1×10^{-4}	63.53	0.99	68.35

Table 1: F_1 scores of the RoBERTa model with different hyperparameter values

2.1 Bacth Size

The Batch Size hyperparameter determines the number of training examples used in one iteration. A smaller batch size means the model updates its weights more frequently, potentially leading to faster learning. However, it can also result in a less stable gradient, as each update is based on less data. On the contrary, a larger batch size provides a more stable estimate of the gradient but requires more memory and can lead to slower learning.

For the DistilRoBERTa model, the F_1 score varies inversely with the batch size: as the batch size increases, the F_1 score decreases. The specific F_1 scores for each batch size can be observed in Table 1. This trend suggests that for this particular model and dataset, smaller batch sizes are more effective, possibly due to more frequent updates allowing the model to learn more effectively from the dataset.

2.2 Learning Rate

The Learning Rate hyperparameter controls how much are the model's weights adjusted with respect to the loss gradient. A too small learning rate can lead to very slow convergence, while a too large learning rate can cause the model to oscillate around or diverge from the minimum of the loss function.

The data from Table 1 suggest that there exists an optimal learning rate of 5×10^{-5} that maximizes the model's performance, while both lower and higher rates than this optimal value lead to poorer performance.

2.3 Adam β_1

Adam β_1 is a hyperparameter of the Adam optimizer that controls the exponential decay rate for the first moment estimates, essentially how quickly it forgets about previous gradients. Lower Adam β_1 values make the optimizer more responsive to recent changes by decaying the moving average of gradients faster, potentially leading to faster adaptation but increased noise sensitivity and risk of overfitting, while higher values result in smoother, more stable updates at the cost of slower convergence and the potential to overlook recent trends in the data.

In the case of the DistilRoBERTa model, adjusting Adam β_1 shows a clear trend in its impact on model performance. Data from Table 1 suggest that a

smaller value of β_1 helps in achieving better model performance.

3 Final Model

After the investigation of the three hyperparameters (batch size, learning rate, Adam β_1) we decided to fine-tune the DistilRoBERTa model with the hyperparameters that reported the best validation scores. The values for the hyperparameters were:

- Batch size: 4
- Learning rate: 5×10^{-5}
- Adam β_1: 0.85

After fine-tuning the model we evaluated the model on provided test set. In the next section, we compare the evaluation results of our fine-tuned model with the pre-trained (not fine-tuned) version of DistilRoBERTa.

4 Results

We generated predictions for both our fine-tuned model and the baseline pre-trained model. Then, we used the predictions to calculate the F_1 scores with respect to the ground truths. Our fine-tuned model achieved F_1 score of 32.71 which is notably higher than the baseline (pre-trained) model's F_1 score of 4.76.

We expected a much higher F_1 score than 32.71, especially because our fine-tuned model achieved a relatively good F_1 score of 70.24 during training. We believe that the model does not achieve great results because the task of genre classification based solely on song lyrics is extremely difficult. Song lyrics lack audio features like tempo, energy, and instrumentalness which carry much more valuable information about the song genre than just text.

We provide examples of song lyrics where one model worked better than the other in Table 2.

Title	Genre	Predictions	
		Fine-tuned	Baseline
Always Free	Blues	Blues	Country
Ace of Spades	Metal	Metal	Blues
Bat Country	Metal	Metal	Country
White Horse	Country	Rock	Country
Last Night	Country	Blues	Country
Deeper Well	Country	Blues	Country

Table 2: Examples where one model succeeded and the other model failed.