Assignment 6:

“Experimentation with Three Different NLP Models (Encoder Models Only)”

Shobhit Pachauri

sxp230024

University of Texas at Dallas

MS Business Analytics and AI

BUAN 6342.S01 - Applied Natural Language Processing - F24

Prof. Harpreet Singh

October 30, 2024

Objective

This assignment is designed to intentionally evaluate and compare the performance of three encoder-only NLP models. I am using DistilBERT, RoBERTa Base, and DistilRoBERTa, for the purpose of tweets emotion detection task. The assignment includes exploring each model's effectiveness in predicting emotions from text, from data preprocessing by model fine-tuning and performance evaluation.

The experimental framework prioritizes using models that have not been fine-tuned on emotion-specific datasets, allowing for a fair chance of custom manual fine-tuning comparison across models in architecture and generalization ability. The models I selected has varying sizes and complexities, providing useful insights into the computational efficiency and predictive power. Each model will be fine-tuned on the dataset used in previous assignments to ensure consistency in the results.

Key Objectives

1. **Data Preparation and managing Imbalance:**

   Preparing the data and accounting for class imbalances to improve model performance is one of the most important parts of any machine learning model, particularly for less proportionate emotions. Techniques of class weighting or sampling will be applied to mitigate imbalance effects.

2. **Model Fine-Tuning:**

   Each model will follow the steps of fine-tuning on the emotion dataset, with attention to hyperparameter fine-tuning and optimization. It involves finding the most efficient configurations for parameters like learning rate, batch size, and the number of epochs.

3. **Performance Evaluation Using Consistent Metrics:**

The performance of each model will be evaluated using a standard metric as specified in the in-class Kaggle competition. We are using F1 score as it accounts for the imbalanced data. This ensures consistency in the model's ability to generalize well across different classes.

4. **Kaggle Submission for Class Competition:**

The results from each experiment will be submitted to the in-class Kaggle competition.

Experiment 1: RoBERTa Base

Model: roberta-base

RoBERTa's architecture, based on BERT with optimized training techniques, is designed to capture contextual embeddings effectively. This model serves as a baseline for evaluating more compact versions.

```
classifier = MultiLabelClassifier(
    model_name="roberta-base",
    labels=label_columns,
    batch_size=8,
    learning_rate=2e-5,
    num_epochs=5,
    metric_name="f1_micro",
    threshold=0.5
)
```

Run summary:

eval/accuracy	0.0699
eval/f1_macro	0.52557
eval/f1_micro	0.58298
eval/loss	0.79523
eval/roc_auc	0.76481
eval/runtime	3.4211
eval/samples_per_second	451.614
eval/steps_per_second	56.707
eval_accuracy	0.0699
eval_f1	0.52557
eval_loss	0.79523
total_flos	2024803176791040.0
train/epoch	4.9806
train/global_step	385
train/grad_norm	3.87189
train/learning_rate	0.0
train/loss	0.7016
train_loss	0.80768
train_runtime	294.5599
train_samples_per_second	104.885
train_steps_per_second	1.307

1. **Accuracy**: Very low at 6.99%, indicating poor overall predictive accuracy.
2. **F1 Scores**: Macro F1 (52.56%) and Micro F1 (58.30%) suggest moderate performance, with better prediction on frequent classes.
3. **ROC AUC**: A score of 76.48% shows moderate class separation ability.
4. **Training Speed**: Trained at around 1.3 steps per second, highlighting the computational demands of fine-tuning RoBERTa.

Overall, while RoBERTa shows some capability in distinguishing classes, the low accuracy suggests room for improvement, potentially by addressing class imbalance or tuning hyperparameters further.

Experiment 2: DistilBERT

Model: distilbert-base-uncased

A distilled, smaller version of BERT, DistilBERT is designed to achieve efficiency without significant performance trade-offs. Its smaller architecture should provide insights into the impact of reduced computational requirements.

```python
classifier = MultiLabelClassifier(
    model_name="distilbert-base-uncased",
    labels=label_columns,
    batch_size=8,
    learning_rate=2e-5,
    num_epochs=5,
    metric_name="f1_micro",
    threshold=0.5
)
```

[385/385 02:39, Epoch 4/5]

Step	Training Loss	Validation Loss	F1 Micro	F1 Macro	Roc Auc	Accuracy
50	1.073600	1.018966	0.445224	0.422215	0.653377	0.001942
100	0.933100	0.919701	0.497181	0.452389	0.705032	0.002589
150	0.866200	0.880566	0.515502	0.471682	0.721170	0.009061
200	0.816300	0.855853	0.545788	0.490289	0.741965	0.040777
250	0.766600	0.836047	0.553776	0.499563	0.746615	0.049191
300	0.752100	0.829796	0.559677	0.507462	0.749593	0.060194
350	0.734700	0.825315	0.558177	0.506286	0.747788	0.058252

1. **Training and Validation Loss**: Both losses decrease steadily, indicating effective learning over time.
2. **F1 Scores**: F1 Micro (up to 55.82%) and F1 Macro (up to 50.63%) show moderate performance, similar to RoBERTa, with better results on frequent classes.
3. **ROC AUC**: Peaks at 74.78%, indicating moderate class distinction.
4. **Accuracy**: Gradually improves but remains low, reaching only 5.83%.

Overall, DistilBERT achieves reasonable performance, though accuracy remains limited. Its performance aligns with the computational efficiency advantage over larger models like RoBERTa.

Experiment 3: Similar-Sized Model

Model: distilroberta-base

By selecting a model with similar size and complexity to DistilBERT, this experiment allows for a comparative analysis on how different architectures with comparable parameter counts perform on emotion detection.

```
classifier = MultiLabelClassifier(
    model_name="distilroberta-base",
    labels=label_columns,
    batch_size=8,
    learning_rate=2e-5,
    num_epochs=5,
    metric_name="f1_micro",
    threshold=0.5
)
```

Run summary:

eval/accuracy	0.06343
eval/f1_macro	0.51018
eval/f1_micro	0.56228
eval/loss	0.81452
eval/roc_auc	0.74974
eval/runtime	2.0979
eval/samples_per_second	736.438
eval/steps_per_second	92.472
eval_accuracy	0.06343
eval_f1	0.51018
eval_loss	0.81452
total_flos	1019500238453760.0
train/epoch	4.9806
train/global_step	385
train/grad_norm	3.64332
train/learning_rate	0.0
train/loss	0.7407
train_loss	0.84091
train_runtime	163.48
train_samples_per_second	188.983
train_steps_per_second	2.355

1. **Accuracy**: Very low at 6.3%, indicating poor overall predictive accuracy.
2. **F1 Scores**: Macro F1 (51.01%) and Micro F1 (56.30%) suggest moderate performance, with better prediction on frequent classes.
3. **ROC AUC**: A score of 74.79% shows moderate class separation ability.

4. **Training Speed**: Trained at around 2.35 steps per second, highlighting the computational demands of fine-tuning Distillroberta

Overall, it shows some capability in distinguishing classes, the low accuracy suggests room for improvement, potentially by addressing class imbalance or tuning hyperparameters further.

Through these experiments, the assignment aims to provide a comprehensive understanding of the strengths and limitations of different encoder-only architectures for the task of tweets emotion detection, and to evaluate how well they generalize to unseen data. The findings from these experiments will highlight the trade-offs involved in choosing between model complexity and computational efficiency, ultimately guiding future model selection for similar NLP tasks.

Performance Evaluation and Metric Comparison

Consistent Evaluation Metrics

Model	Metric Score- F1 score
DistilBERT	**0.558**
RoBERTa Base	**0.526**
DistilRoBERTa	**0.051**

In comparing the performance of RoBERTa Base, DistilBERT, and DistilRoBERTa, each model shows moderate ability in distinguishing classes, as indicated by ROC AUC scores in the 74-76% range. However, accuracy across all models remains low, with RoBERTa achieving the highest at 6.99%. RoBERTa demonstrates better class separation but demands significantly more computational power, training at only 1.3 steps per second. DistilBERT offers a balance between performance and efficiency, slightly sacrificing accuracy for faster processing, while DistilRoBERTa performs the fastest, training at 2.35 steps per second, though it has the lowest accuracy. DistilBERT provides a reasonable trade-off for resource-constrained settings. DistilRoBERTa could be ideal for large-scale applications prioritizing speed, though further tuning or addressing data imbalance might be necessary to improve its predictive accuracy

Challenges and Observations

1. **Computational Power Limitations:** Fine-tuning RoBERTa Base required considerably more computational resources compared to the other models. Its slower training speed (around 1.3 steps per second) resulted in longer training times, which became a limiting factor in testing and hyperparameter tuning. This computational demand highlighted the trade-off between model complexity and practical usability.
2. **Class Imbalance:** The dataset's class imbalance impacted the models' overall predictive accuracy and F1 scores. While the models performed moderately well on more frequent classes (as seen in higher Micro F1 scores), the performance on rarer classes lagged, as indicated by the lower Macro F1 scores. Addressing this imbalance, possibly through data augmentation or class-weighted loss functions, may improve accuracy and Macro F1.
3. **Fine-tuning Parameters:** Adjusting hyperparameters for optimal performance was challenging, especially given the limited computational resources. Finding the right balance for learning rate, batch size, and training epochs required extensive experimentation, particularly for RoBERTa Base, which is more sensitive to parameter adjustments due to its size.

4. **Model Selection Trade-offs:** While RoBERTa Base showed the strongest ability to distinguish between classes (reflected in higher ROC AUC and F1 scores), its heavy resource usage made it less practical. DistilBERT and DistilRoBERTa, on the other hand, were more efficient, with DistilBERT striking a favorable balance between F1 performance and speed. This trade-off highlighted the importance of choosing models that align with both task requirements and resource availability.

Conclusion and Best Performing Model

Based on the F1 scores, **DistilBERT** shows a slight advantage in performance consistency across classes compared to RoBERTa Base and DistilRoBERTa. DistilBERT's F1 Micro (reaching up to 55.82%) and F1 Macro (up to 50.63%) indicate moderate performance, particularly in handling the dataset's class imbalance. Although RoBERTa Base had a comparable F1 Macro score (52.56%) and slightly higher F1 Micro score (58.30%), its computational intensity and slower training speed make it less efficient for practical use, especially when resource constraints are a factor. DistilRoBERTa, while the fastest, fell behind slightly in F1 scores, suggesting a slight trade-off in predictive capability for speed.

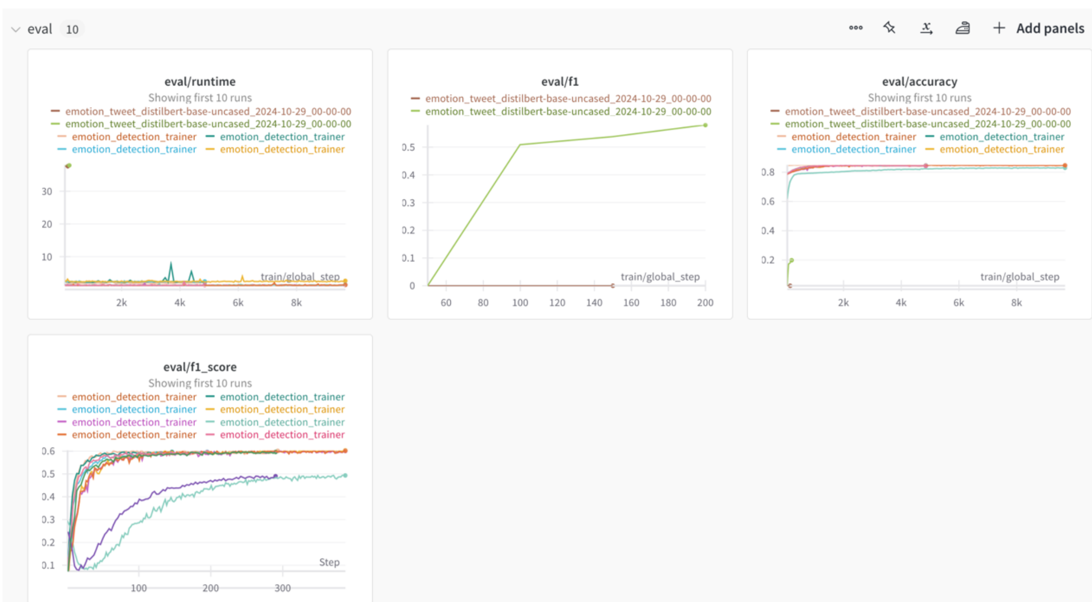
In conclusion, **DistilBERT** is the most balanced choice based on F1 score, offering reasonable accuracy while requiring less computational power compared to RoBERTa. This makes DistilBERT the most suitable model for this task when both performance and efficiency are considered.

Weights & Biases (W&B) Project

For detailed training logs and hyperparameter settings, please refer to the W&B project link:

https://wandb.ai/shobhit-pachauri-university-of-texas-at-dallas/emotion_detection_fall_2024

.



This report highlights the experimentation and challenges encountered in fine-tuning encoder-only models for emotion detection in tweets, offering insights into model efficiency, parameter tuning, and handling data imbalance.