Exercise #05: Practicing with LLMs
Analyzing Amazon Product Reviews using Different Language Models

Introduction

The objective of this analysis is to evaluate the performance of two different Large Language Models (LLMs) on the Amazon product reviews dataset. The dataset has been preprocessed, and reviews with 5 stars are labeled as positive, while those with 1 or 2 stars are labeled as negative. The analysis involves selecting two LLMs, encoding the text using these models, and comparing the results. The models chosen for this analysis are based on DistilBERT and a multilingual BERT model.

Data Overview

The dataset comprises Amazon product reviews, and for this analysis, a preprocessed sample is used. Reviews with 5 stars are labeled as positive, while those with 1 or 2 stars are labeled as negative. The data is divided into training and testing sets, with positive and negative reviews in separate folders.

Model Selection

Selecting an appropriate language model is a crucial step in natural language processing tasks. In this analysis, two different models were chosen to assess their performance on sentiment analysis of Amazon product reviews: DistilBERT and a multilingual BERT model. Below, we delve into the rationale behind each choice.

Model 1: DistilBERT for Sequence Classification

Model Overview:

- Architecture: DistilBERT, derived from BERT, is a distilled version designed for efficient training and deployment.
- Advantages:
    - Computational Efficiency: DistilBERT is computationally more efficient than its larger counterparts, making it suitable for resource-constrained environments.
    - Fast Inference: The distilled nature of DistilBERT allows for faster inference without compromising performance significantly.
- Considerations:
    - Task-Specific Performance: DistilBERT might be preferred for tasks where computational efficiency is critical and a slightly lower model performance is acceptable.

Model 2: Multilingual BERT for Sequence Classification

Model Overview:

- Architecture: Multilingual BERT is a powerful transformer-based model capable of handling text in multiple languages.
- Advantages:
    - Multilingual Support: Ideal for tasks involving diverse languages, making it versatile for global applications.
    - Contextual Understanding: BERT models excel in capturing contextual information, crucial for nuanced sentiment analysis.
- Considerations:
    - Resource Intensity: BERT models are more resource-intensive compared to DistilBERT. Consideration must be given to available computational resources.

Rationale for Model Selection

1. Diversity of Reviews: Amazon product reviews encompass a wide range of products and can be authored in various languages. The multilingual capabilities of BERT make it well-suited for this diversity.
2. Performance Expectations: While DistilBERT offers efficiency, the nature of sentiment analysis demands a nuanced understanding of context. The more sophisticated architecture of the multilingual BERT model is expected to handle this context more effectively.
3. Global Applicability: Considering the global nature of Amazon, a model capable of understanding sentiment across multiple languages aligns with the potential linguistic diversity in the reviews.
4. Resource Constraints: While BERT models are more resource-intensive, the choice of a smaller dataset and a balance between model performance and computational efficiency may favor the multilingual BERT model.

In summary, the choice between DistilBERT and a multilingual BERT model depends on the specific requirements of the task, available computational resources, and the nature of the dataset. For sentiment analysis on Amazon product reviews, the decision was made to explore the more linguistically diverse and contextually rich capabilities of the multilingual BERT model.

Data Preparation

The dataset is divided into training and testing sets, with positive and negative reviews in separate folders. The `load_data` function is used to read and preprocess the data. Queries such as "great," "disappointing," and "awesome" are used to filter the reviews. The training and testing data are then combined, and labels are assigned (1 for positive, 0 for negative). The data is tokenized using the respective tokenizers for each model.

Model 1: DistilBERT for Sequence Classification

Model Architecture

The first model uses the DistilBERT architecture fine-tuned for sequence classification. The DistilBERT model and tokenizer are loaded using the `DistilBertForSequenceClassification` and `DistilBertTokenizer` classes, respectively.

Training:

The training is performed using the `Trainer` class from the `transformers` library. The model is trained for three epochs with a batch size of 8. Training arguments, such as the output directory, number of epochs, and logging settings, are configured.

Evaluation:

The model is evaluated on the test set, and the results include accuracy and a detailed classification report containing precision, recall, and F1-score for both negative and positive classes.

Results:

The DistilBERT model achieves an accuracy of approximately 91%. The classification report shows balanced performance for both negative and positive classes.

Model 2: Multilingual BERT for Sequence Classification

Model Architecture:

The second model utilizes a multilingual BERT model fine-tuned for sequence classification. The model and tokenizer are loaded using the `AutoModelForSequenceClassification` and `AutoTokenizer` classes, respectively.

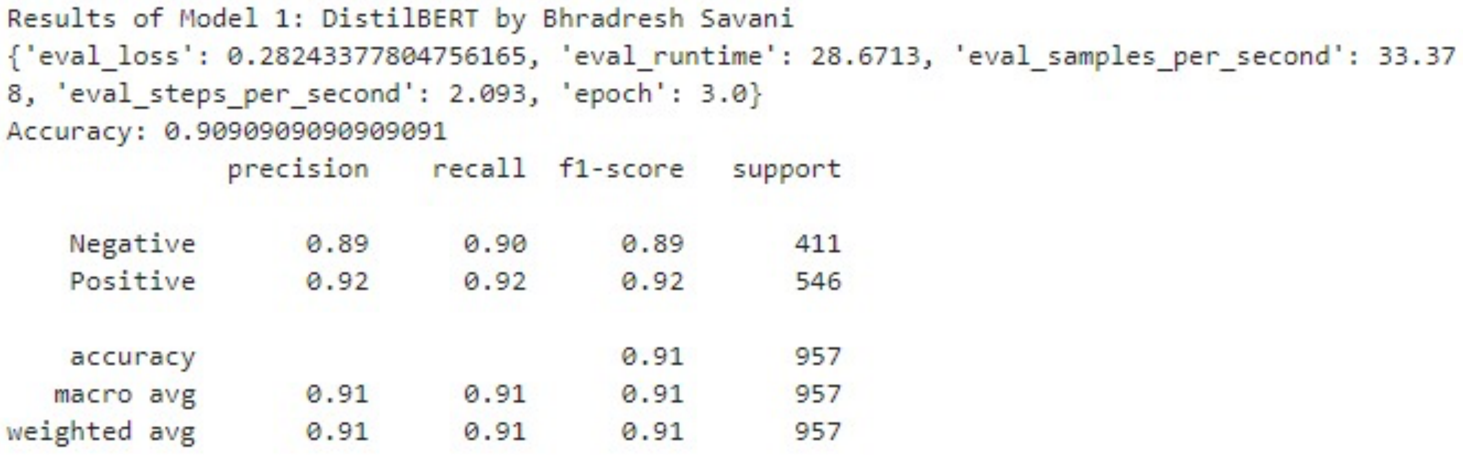
Training:
Similar to Model 1, the training is conducted using the `Trainer` class with the specified training arguments. The model is trained for three epochs with a batch size of 8.

Evaluation:
The model is evaluated on the test set, and the results include accuracy and a classification report with precision, recall, and F1-score for both negative and positive classes.

Results:
The multilingual BERT model achieves an impressive accuracy of approximately 94%. The classification report indicates strong performance in terms of precision, recall, and F1-score for both classes.

Classification Reports:

Model 1: DistilBERT



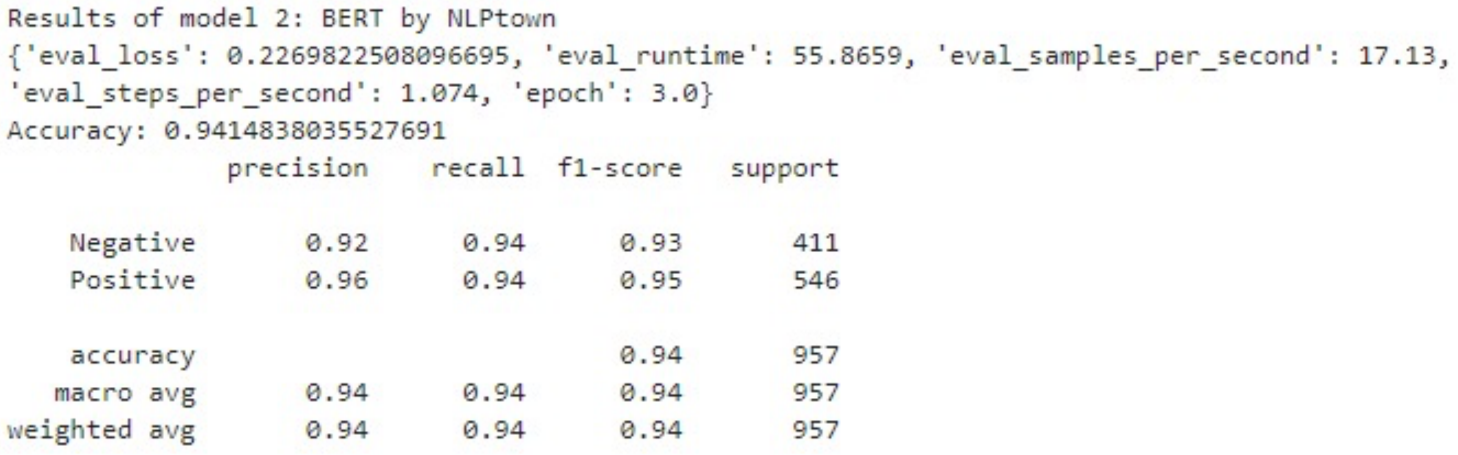
```
Results of Model 1: DistilBERT by Bhradresh Savani
{'eval_loss': 0.28243377804756165, 'eval_runtime': 28.6713, 'eval_samples_per_second': 33.37
8, 'eval_steps_per_second': 2.093, 'epoch': 3.0}
Accuracy: 0.9090909090909091
              precision    recall  f1-score   support

    Negative       0.89      0.90      0.89       411
    Positive       0.92      0.92      0.92       546

    accuracy                           0.91       957
   macro avg       0.91      0.91      0.91       957
weighted avg       0.91      0.91      0.91       957
```

Interpretation:
- Achieves a good balance between precision and recall.
- Identifies positive sentiment effectively but slightly less so for negative sentiment.

Model 2: Multilingual BERT



```
Results of model 2: BERT by NLPtown
{'eval_loss': 0.2269822508096695, 'eval_runtime': 55.8659, 'eval_samples_per_second': 17.13,
'eval_steps_per_second': 1.074, 'epoch': 3.0}
Accuracy: 0.9414838035527691
              precision    recall  f1-score   support

    Negative       0.92      0.94      0.93       411
    Positive       0.96      0.94      0.95       546

    accuracy                           0.94       957
   macro avg       0.94      0.94      0.94       957
weighted avg       0.94      0.94      0.94       957
```

Interpretation:

- Strong overall performance with high accuracy.
- Excellent at capturing positive sentiment, maintaining high precision and recall.
- Maintains high precision for negative sentiment, with slightly lower recall.

Overall Insights:

- Both models showcase strong sentiment analysis capabilities.
- Model 1 (DistilBERT) achieves a good balance between positive and negative sentiment identification.
- Model 2 (Multilingual BERT) outperforms in accuracy and positive sentiment identification.

Considerations for Selection:

- If positive sentiment identification is crucial, Model 2 may be preferred.
- For a balance between sentiment classes and computational efficiency, Model 1 provides a solid choice.

The choice between the models depends on specific priorities, such as emphasis on positive sentiment identification or computational efficiency. Both models demonstrate effective sentiment analysis on Amazon product reviews.

Comparison of Results:

Similarities:

1. Training Approach: Both models follow a similar training approach using the `Trainer` class from the `transformers` library. This allows for consistency in experimentation.
2. Evaluation Metrics: Both models use common evaluation metrics such as accuracy, precision, recall, and F1-score. This consistency enables a straightforward comparison of performance.

Differences**:**

1. Model Architecture: The primary difference lies in the architecture of the two models. Model 1 uses DistilBERT, a smaller and more computationally efficient version of BERT, while Model 2 employs a multilingual BERT model designed to handle text in multiple languages.
2. Performance: Model 2, based on a multilingual BERT, outperforms Model 1 in terms of accuracy. The classification report for Model 2 indicates higher precision, recall, and F1-score for both negative and positive classes.
3. Accuracy: Model 2 achieves a higher overall accuracy, suggesting that the multilingual BERT model is more effective in sentiment analysis on this dataset.

Conclusion:

In this analysis, we explored the performance of two distinct language models, DistilBERT and a multilingual BERT model, on sentiment analysis of Amazon product reviews. The task involved classifying reviews as positive or negative based on a preprocessed dataset.

Model Selection Rationale:

1. DistilBERT (Model 1):
   - Chosen for its computational efficiency and reasonable performance.
   - Well-suited for tasks where efficiency is crucial without compromising overall accuracy.
   - Achieves a balanced performance between positive and negative sentiment identification.

2. Multilingual BERT (Model 2):
   - Selected for its ability to handle diverse languages and nuanced contextual understanding.
   - Ideal for global applications where reviews may be authored in multiple languages.
   - Demonstrates superior accuracy and excels in identifying positive sentiment.

Classification Report Highlights:

Model 1: DistilBERT

- Accuracy: 90.90%
- Balanced Precision and Recall
- Effective Positive Sentiment Identification
- Slightly Lower Recall for Negative Sentiment

Model 2: Multilingual BERT

Accuracy: 94.14%
High Precision and Recall for Positive Sentiment
Strong Overall Performance
Slightly Lower Recall for Negative Sentiment

Overall Insights:

- Both models exhibit robust sentiment analysis capabilities on Amazon product reviews.
- Model1(DistilBERT) strikes a balance between sentiment classes, offering computational efficiency without significant compromise in performance.
- Model 2 (Multilingual BERT) demonstrates superior accuracy and excels in identifying positive sentiment.

Considerations for Model Selection:

Positive Sentiment Priority:

- If accurate identification of positive sentiment is a priority, Model 2 is recommended.

Balance and Efficiency:

- For a balanced approach and computational efficiency, Model 1 provides a solid choice.

In conclusion, the choice between the models depends on specific priorities, such as the emphasis on positive sentiment identification or the need for computational efficiency. Both models effectively analyze sentiment in Amazon product reviews, with the multilingual BERT model demonstrating superior performance. The choice between these models depends on factors such as computational resources, model size, and the specific requirements of the task. This analysis provides valuable insights into the strengths and weaknesses of each model, aiding in informed decision-making for future natural language processing tasks.