Multi-model Sentiment Analysis Documentation

Homework 6

Applied NLP

G, Likith Vinayaka
LXG210034

Introduction:

This report presents a detailed analysis of sentiment analysis experiments conducted using three different models: DistilBERT, RoBERTa, and Google/Flan-T5-Base. The primary objective was to classify tweets into multiple emotional categories, such as joy, sadness, anger, and others. Each model was evaluated based on various performance metrics, challenges encountered, and overall effectiveness in handling the sentiment analysis task.

Methodologies:

Experiment 1: "bert-base-uncased"
Weights & Biases Project: Bert-base-uncased Model

Data Preprocessing:
- Features and Labels: The dataset was split into features (tweets) and labels (emotional categories).
- Data Split: The data was divided into training, validation, and test sets.

Model Configuration:
- Architecture: DistilBERT was chosen due to its lightweight design suitable for quick experimentation.
- Training: Utilized the `Trainer` class from the Transformers library for model training

Evaluation:
- Metrics: Metrics such as accuracy, F1-score were computed on the validation set.

Experiment 2: "distilroberta-base"
Weights & Biases Project: **distilroberta-base**

Data Preprocessing:
- Similar to Experiment 1: The preprocessing steps remained consistent across experiments.

Model Configuration:
- **Architecture:** RoBERTa, an advanced version of BERT, was employed to evaluate if a more complex model could improve performance.
- **Training:** Fine-tuned the RoBERTa model on the tokenized dataset using the `Trainer` class.

Evaluation:
- **Similar Metrics:** The same metrics as Experiment 1 were used for evaluation.

Experiment 3: "google/flan-t5-base"
Weights & Biases Project: Google/flan-t5-base Model

Data Preprocessing:
- **Tokenization:** Utilized the T5 tokenizer for tokenizing the tweets.
- **Custom Preprocessing:** A custom preprocessing function was developed to handle multi-label classification.

Model Configuration:
- **Architecture:** T5 was selected to explore the potential of a transformer-based model specifically designed for sequence classification.
- **Training:** Due to computational constraints, training was conducted on a subset of the dataset.

Evaluation:
- **Similar Metrics:** The same set of metrics used in the previous experiments was employed for evaluation.

Results and Observations:

Experiment 1: Bert-base-uncased

Training Metrics:	**Evaluation Metrics:**
Validation Loss: 0.3261	Validation Loss: 0.3261
Accuracy: 87.02%	Accuracy: 87.02%
F1 Score: 0.6521	F1 Score: 0.6521

Experiment 2: Distilroberta-base

Training Metrics:	**Evaluation Metrics:**
Validation Loss: 0.3334	Validation Loss: 0.3334
Accuracy: 86.75%	Accuracy: 86.75%
F1 Score: 0.6493	F1 Score: 0.6493

Experiment 3: Google/flan-T5-base

Training Metrics:	**Evaluation Metrics:**
Accuracy: 97.22%	Accuracy: 97.22%
F1 Score: 0.0616	F1 Score: 0.0616
Precision: 99.83%	Precision: 99.83%
Recall: 6.25%	Recall: 6.25%
Hamming Loss: 0.0278	Hamming Loss: 0.0278

Challenges Encountered:

- **Limited Computing Power:** The computational demands of the models, especially T5, made it challenging to train on the entire dataset.
- **Data Size:** Due to computational limitations, only a subset of the data could be used for training, affecting the model's performance.
- **Model Complexity:** The complex architectures of RoBERTa and T5 required extensive tuning and longer training times, making them computationally expensive.
- **Evaluation Metrics:** T5's NaN validation loss indicates potential issues in the loss calculation, which needs further investigation.

Comparative Analysis:

Performance Levels:

Bert-base-uncased

- Advantages:
  - Quick training, lower computational requirements.
  - Moderate accuracy at 87.02%.
  - Balanced F1 score of 0.6521.
- Limitations:
  - Lower accuracy and F1-score compared to more complex models.
  - Potential overfitting, as the training and evaluation metrics are closely matched.

Distilroberta-base

- Advantages:
  - Enhanced accuracy and F1-score compared to DistilBERT.
  - Comparable accuracy to DistilBERT at 86.75%.
  - Balanced F1 score of 0.6493.
- Limitations:
  - Increased computational demands, but a balanced performance compared to the other models.
  - Slightly higher validation loss compared to DistilBERT.

Google/flan-T5-base

- Advantages:
  - Demonstrated the highest accuracy but at the cost of poor precision, recall, and F1-score.
- Limitations:
  - Extremely high computational requirements and longer training times.
  - Extremely low F1 score of 0.0616 and very low recall at 6.25%, indicating potential issues with model generalization or dataset size.

Observations:

- **Accuracy:** T5 shows superior accuracy, but the extremely high precision indicates potential class imbalance or overfitting.
- **F1 Score:** DistilBERT and RoBERTa have balanced F1 scores, suggesting a better balance between precision and recall than T5.
- **Recall:** T5's low recall indicates that it might be missing out on predicting some classes correctly.

Conclusions:

- **Bert-base-uncased** is suitable for quick prototyping and applications where computational resources are limited. However, its performance is compromised compared to more complex models.
- **Distilroberta**, base offers a balanced performance with improved accuracy and F1-score. It is a suitable choice for sentiment analysis tasks with a moderate dataset size and computational resources.
- **Google/flan-T5-base**, despite achieving the highest accuracy, requires extensive computational resources and longer training times. It might be suitable for scenarios where extremely high accuracy is crucial, but the associated costs need to be considered.

In summary, the choice of the model depends on the specific requirements of the sentiment analysis task, including the available computational resources, dataset size, and desired performance metrics.

For a comprehensive view of each experiment, please visit our Weights & Biases Project