Classifying Sentiments on the TweetEval Dataset

Ellen Beatriz Shen
Insper
São Paulo, Brasil
ellenbs@al.insper.edu.br

I. Problem Statement

Sentiment analysis involves classifying text into predefined categories—Negative, Neutral, and Positive. This project uses the TweetEval dataset, a benchmark designed for noisy, informal tweet data, characterized by abbreviations, emojis, and diverse contexts. The dataset's imbalance, with the Neutral class dominating, poses challenges for minority class performance. The objective is to fine-tune the `RoBERTa` model on the dataset and evaluate its performance using Macro-Averaged Recall, comparing results to benchmarks from the `TweetEval` paper.

II. Solution Pipeline

A. Model:

- RoBERTa-Base, a transformer-based architecture pre-trained on large-scale corpora, was chosen for its strong performance in text classification tasks.

B. Dataset Preparation:

- The training set was balanced to include 500 examples per class to address the imbalance issue.
- The test set was kept as provided (12,284 examples).

C. Tokenization:

- Tweets were tokenized using the RobertaTokenizer, truncating/padding to a maximum length of 128 tokens.

D. Training Configuration:

- Learning Rate: 2e-5
- Batch Size: 16
- Epochs: 3
- Evaluation Strategy: Metrics calculated at the end of each epoch.

III. Evaluation

A. Metrics

The following metrics were used to assess performance:

- **Precision:** Proportion of true positive predictions out of all predicted positives. High precision indicates fewer false positives.
- **Recall:** Proportion of true positive predictions out of all actual positives. High recall indicates fewer false negatives.
- **F1-Score:** The harmonic mean of precision and recall, providing a balanced measure of classification performance.
- **Macro-Averaged Recall:** Arithmetic mean of recall values across all classes, giving equal importance to each class regardless of support (class size).

B. Class-Level Performance

Table I provides detailed metrics for each sentiment class.

TABLE I
Performance Metrics by Class

Class	Precision	Recall	F1-Score	Support
Negative	0.61	0.86	0.72	3,972
Neutral	0.78	0.47	0.59	5,937
Positive	0.60	0.78	0.68	2,375

C. Macro-Averaged Recall

The model achieved a Macro-Averaged Recall of 0.71, demonstrating its ability to generalize across all classes. This is competitive with the TweetEval paper's reported benchmark of 72.9 for RoBERTa-Twitter.

D. Comparison with Literature

The evaluation revealed several insights:

- The model performed well on the *Negative* and *Positive* classes, with high recall values of 0.86 and 0.78, respectively.
- Performance on the *Neutral* class was lower, with a recall of 0.47, likely due to the ambiguous nature of neutral sentiments and overlaps with other categories.
- The Macro-Averaged Recall of 0.71 is close to the 72.9 benchmark reported for RoBERTa-Twitter in the TweetEval paper.

The evaluation highlights both the strengths and limitations of the current approach, providing a foundation for future improvements.

REFERENCES

1) Barbieri, F., et al. ”TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification.” arXiv preprint arXiv:2010.12421, 2020.
2) Liu, Y., et al. ”RoBERTa: A Robustly Optimized BERT Pretraining Approach.” arXiv preprint arXiv:1907.11692, 2019.