Natural Language Processing - Assignment 2

December 26, 2025

1 Question 5: Sentiment Analysis

1.1 Part (a) - T5 Paper Research

1.1.1 GitHub Repository

The GitHub repository for the T5 project is:

`https://github.com/google-research/text-to-text-transfer-transformer`

Source: T5 Paper, page 1, footnote 1.

1.1.2 Models Made Publicly Available

The authors released five pre-trained T5 model variants with different sizes:

- **T5-Small**: 60 million parameters
- **T5-Base**: 220 million parameters
- **T5-Large**: 770 million parameters
- **T5-3B**: 3 billion parameters
- **T5-11B**: 11 billion parameters

Source: GitHub README.md, section "Released Model Checkpoints".

1.1.3 Dataset for Sentiment Analysis Benchmark

The dataset used to benchmark sentiment analysis in the T5 paper is **SST-2 (Stanford Sentiment Treebank-2)**, which is part of the GLUE benchmark suite. SST-2 is a binary sentiment classification task.
Source: T5 Paper, Section 2.3 "Downstream Tasks", GLUE benchmark description.

1.1.4 Evaluation Metric

The evaluation metric used for the sentiment analysis task (SST-2) is **Accuracy**. The model's predictions are compared against the ground truth labels, and the percentage of correct predictions is reported.
Source: T5 Paper, Table 1 (results table), column header "SST-2: Acc".

1.2 Part (b) - T5-Small Model Fine-tuned on SST2

1.2.1 Selected Model

The selected model for this assignment is:

`lightsout19/t5-sst2`

This model is a T5 variant that has been fine-tuned on the SST-2 dataset for sentiment classification.
Source: Hugging Face Model Hub - `https://huggingface.co/lightsout19/t5-sst2`

1.2.2 Model Details

- **Base Architecture**: T5 (Text-to-Text Transfer Transformer)
- **Fine-tuning Dataset**: SST-2 (Stanford Sentiment Treebank-2)
- **Task**: Binary sentiment classification
- **Framework**: Transformers library (PyTorch/TensorFlow compatible)
- **Usage**: Can be loaded using `AutoTokenizer` and `AutoModelForSequenceClassification`

1.2.3 Alternative Models Considered

During the model selection process, the following alternatives were also identified:

1. `hyyoka/t5-small-finetuned-sst2`: A T5-Small model (60.5M parameters) specifically fine-tuned on SST-2. This model exactly matches the T5-Small size specification.
2. `google/flan-t5-small`: An improved version of T5-Small (77M parameters) that has been instruction fine-tuned on 1000+ tasks. While not specifically fine-tuned on SST-2, FLAN-T5 shows improved performance across many NLP tasks including sentiment analysis.

Source: Hugging Face Model Hub searches for "t5-small sst2" and "flan-t5-small".

1.3 Part (e) - Dataset Balance Analysis

1.3.1 Is the SST-2 Dataset Balanced?

Yes, the SST-2 validation dataset is well-balanced. Our analysis revealed:

- **Negative examples**: 428 (49.1%)
- **Positive examples**: 444 (50.9%)
- **Balance ratio**: 0.964 (close to perfect balance of 1.0)

The difference between the two classes is less than 2%, which indicates excellent balance.

1.3.2 Why is Dataset Balance Important?

Dataset balance is a critical consideration when evaluating model performance with accuracy as the primary metric. With imbalanced datasets, accuracy can be highly misleading. For instance, if a dataset contained 90% positive examples and only 10% negative examples, a naive model that always predicts "positive" would achieve 90% accuracy without learning any meaningful patterns about sentiment classification.

A balanced dataset ensures that the reported accuracy truly reflects the model's ability to distinguish between both classes, rather than merely exploiting the majority class distribution. In the case of SST-2, the near-perfect balance (49.1% vs 50.9%) validates that accuracy is a reliable and meaningful metric for evaluating model performance on this task. If the dataset were imbalanced, we would need to consider additional metrics such as precision, recall, F1-score, or per-class accuracy to properly assess model quality.

1.4 Part (f) - Limitations of Accuracy as an Evaluation Metric

While accuracy measures whether sentiment predictions are correct, human evaluators would notice several important aspects that this metric overlooks:

1.4.1 Confidence Calibration

Accuracy treats all predictions equally regardless of model confidence. Looking at our model's predictions, it assigns 99.92% confidence to "This movie is awesome" (clearly positive) but 95.77% to "Did you like the movie?" (a neutral question). Both count as "correct" if labeled positive, but a human would recognize the second prediction is problematic—the question itself expresses no sentiment.

1.4.2 Ambiguity and Context

The sentence "I'm not sure what I think about this movie" was classified as negative with 97.66% confidence. While this might align with the dataset label, a human evaluator would recognize this as expressing uncertainty rather than negative sentiment. The model's high confidence masks this nuanced interpretation.

1.4.3 Error Severity

Not all misclassifications have equal consequences. Accuracy doesn't distinguish between narrowly missing on borderline cases versus confidently misclassifying obviously opposite sentiments. Human evaluators would weight these errors differently based on their severity and the model's confidence level.