Natural Language Processing - Assignment 2

December 27, 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`

The repository was cloned locally.

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. It shows all the different benchmarks used.

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, Various tables showing SST-2 being evaluated based on accuracy, 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 small 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`

1.3 Part (c) - Sentiment Predictions on Four Sentences

The model was used to predict sentiment on the following four sentences. The results demonstrate the model's ability to classify sentiment with high confidence:

Sentence	Prediction	Confidence
This movie is awesome	POSITIVE	0.9992
I didn't like the movie so much	NEGATIVE	0.9897
I'm not sure what I think about this movie.	NEGATIVE	0.9766
Did you like the movie?	POSITIVE	0.9577

Observations:

- The model correctly identified clear positive sentiment ("This movie is awesome") with very high confidence (99.92%).
- The model correctly identified negative sentiment even with negation ("I didn't like the movie so much") with 98.97% confidence.
- For ambiguous cases like uncertainty expressions ("I'm not sure what I think"), the model classified it as negative, which may reflect how the training data labeled such uncertain statements.
- Interestingly, the model classified the neutral question "Did you like the movie?" as positive, possibly because questions in the training data tended to be associated with positive contexts.

1.4 Part (d) - Model Evaluation on SST-2 Dataset

The model was evaluated on the full SST-2 validation dataset consisting of 872 examples. The evaluation results are as follows:

- **Total examples evaluated**: 872
- **Correct predictions**: [To be filled after running evaluation]
- **Accuracy**: [To be filled after running evaluation]%

Note: The actual accuracy should be reported here after running the evaluation code in the Jupyter notebook. Based on the model card on Hugging Face, T5-Small fine-tuned models on SST-2 typically achieve accuracy in the range of 91-93%.

1.5 Part (e) - Dataset Balance Analysis

1.5.1 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.

1.5.2 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.

1.6 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:

1.6.1 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.

1.6.2 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.

1.6.3 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.

1.7 Part (g) - Consulting Advice for Healthcare Sentiment Analysis

Your challenge is applying sentiment analysis to medical visit summaries, which is quite different from the movie reviews we've been working with. Here are three practical approaches:

Start Simple—Test Existing Models. Try the SST-2 model we used as a quick baseline. It probably won't work great since medical language differs from casual text, but it'll show you the gap and costs nothing to try.

Fine-tune on Your Data. Take a model like T5-Small or BioBERT and train it on labeled examples from your visit summaries. You'll need to manually label maybe 200–500 examples first. This gives the best performance but requires time and ML expertise.

Use LLMs with Few Examples. Tools like GPT-4 can classify sentiment with just a few example visit summaries in the prompt. Quick to set up, but raises privacy concerns with patient data and has ongoing API costs.

Key factors to consider: Make sure your dataset is balanced between both suppliers and positive/negative visits—otherwise accuracy metrics are meaningless (as we saw in Part e). Think about privacy: can you use cloud APIs with medical data, or do you need on-premises solutions? Also, who labels the "ground truth"—different doctors might judge satisfaction differently.

My recommendation: Start with approach 1 to see how bad the domain gap is. Then label 100–200 examples and try approach 2 or 3. Don't just look at accuracy—check if the model catches negative experiences equally well for both suppliers, since missing drug problems is worse than false alarms.