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 results are in the Jupyter notebook, repeating for convince:

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

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

We evaluated the model on the SST-2 validation dataset of 872 examples. The model achieved 90.14% accuracy, correctly classifying 786 out of 872 examples.

1.5 Part (e) - Dataset Balance Analysis

1.5.1 Is the SST-2 Dataset Balanced?

The SST-2 GLUE set is balanced, with 29780 negative examples (44.2%) and 37569 positive examples (55.8%) a balance ratio of 0.793. Which seems fairly well balanced. It appears the researchers have built the dataset to be balanced, as it is likely real distribution skews to a certain side more significantly.

1.5.2 Why is Dataset Balance Important?

Let us consider a hypothetical imbalanced dataset with 90% positive and 10% negative examples. A trivial classifier that always predicts "positive" would achieve 90% accuracy despite not being 'intelligent' what so ever.
With SST-2's balanced distribution, our 90.14% accuracy genuinely reflects the model's classification ability rather than dataset bias. Therefore, the dataset balance is very important in this case as without it bad classifiers can achieve good results, rendering the evaluation objectively poor.

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. **Confidence Calibration:** Accuracy treats all predictions equally regardless of model confidence. Our model 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, yet humans recognize the question expresses no inherent sentiment.

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 match the label, humans recognize this expresses uncertainty rather than actual negative sentiment—a distinction the model's high confidence masks.

3. **Error Severity:** Not all misclassifications have equal consequences. Accuracy doesn't distinguish between narrowly missing borderline cases versus confidently misclassifying clear-cut examples. Human evaluators naturally weight these errors differently based on severity and confidence level.

1.7 Part (g) - Consulting for Healthcare Sentiment Analysis

Three Approaches:
1. Use an Existing Sentiment Model: Deploy the current SST-2 model directly or a similar model. Our 90.14% accuracy shows the model works well on movie reviews, but medical summaries differ significantly in vocabulary and structure. Our results reveal concerning patterns: "I'm not sure what I think about this movie" was classified as negative with 97.66% confidence, showing the model treats uncertainty as negativity which may be problematic for medical contexts where uncertainty is common.
2. Fine-tune on Labeled Medical Data: Collect 500–1000 annotated patient visit summaries and fine-tune T5 or a medical domain model (For example BioBERT). The T5 paper (Section 3.7.2) showed domain-specific pre-training improved performance on in-domain tasks. However, annotation and development requires expertise and will likely be more costly and have a higher development time.
3. LLM Prompting: Use GPT-4 or a similar LLM model with 5-10 example summaries in the prompt. No training required, but raises HIPAA compliance issues for patient data sent to external APIs. Consider local deployment for sensitive data. Running LLM models can also be costly and requires significant infrastructure if ran locally. Another downside is the unpredictably of LLM models, before deployment the LLM suitability and performance needs to be rigorously tested.
Recommendation: Test approach 1 on 100 samples first to quantify domain mismatch. If performance is inadequate, invest in approach 2 with emphasis on per-supplier error analysis, not just overall accuracy. Alternatively Consider testing LLM models performance if they are tested adequate consistently and there is sufficient local infrastructure consider using them.