**Project 3: Using Pre-Trained Language Models**

**Language as Data at Göttingen University**

Moujan Mirjalili, Farhan Kayhan, Merle (Joris) Hellwig 27.01.2025

**Deliverable 1: Task Analysis for SST-2 (Stanford Sentiment Treebank)**

**Task Overview**

* **Task Name:** SST-2 (Stanford Sentiment Treebank)
* **Dataset Source:** Sentences extracted from movie reviews on Rotten Tomatoes
* **Total Sentences:** 11,855 single sentences
* **Unique Phrases:** 215,154 phrases in the parse trees of 11,855 sentences
* **Annotation:** Each phrase annotated by 3 human judges
* **Task Type:** Binary sentiment classification.
* **Objective:** Predict whether the sentiment of a given sentence is positive or negative.

**Task Setup with Examples**

* **Input**: A single English sentence.
* **Output**: A binary label (0 for negative, 1 for positive).
* **Examples:**

1. **Input**: A fascinating look at a deeply troubled man. → **Output**: Positive (1).
2. **Input**: A dull and lifeless movie. → **Output**: Negative (0).

**Relevant Statistics**

Using publicly available data from the GLUE benchmark for SST-2:

* **Dataset Size:**

1. **Training Set**: 67,349 examples.
2. **Validation Set**: 872 examples.
3. **Test Set**: 1,821 examples (labels withheld for competition submissions).

* **Class Distribution:**

1. **Positive Sentiment**: ~54% of the dataset.
2. **Negative Sentiment**: ~46% of the dataset.

The class distribution is relatively balanced, minimizing bias in evaluation.

* **Sentence Length:**

1. **Average Sentence Length**: ~19 words.
2. **Longest Sentence**: ~52 words.
3. **Shortest Sentence**: 2 words.

* **Preprocessing:**

The sentences are already tokenized and lowercased.

**Manual Instance Analysis**

We reviewed a subset of sentences to classify them into **easy** and **difficult** based on linguistic and contextual complexity.

* **Easy Instances (10 Examples):**

1. **Input**: A thrilling and suspenseful movie! → **Output**: Positive (1).
2. **Input**: Completely boring. → **Output**: Negative (0).
3. **Input**: A must-watch for anyone.**Output**: Positive (1).
4. **Input**: A waste of time and money.**Output**: Negative (0).
5. **Input**: Heartwarming and delightful. → **Output**: Positive (1).
6. **Input**: Avoid this at all costs. → **Output**: Negative (0).
7. **Input**: An amazing performance. → **Output**: Positive (1).
8. **Input**: Terrible in every way. → **Output**: Negative (0).
9. **Input**: Beautifully directed. → **Output**: Positive (1).
10. **Input**: It lacked any sense of fun. → **Output**: Negative (0).

* **Difficult Instances (10 Examples):**

1. **Input**: Not bad at all. → **Output**: Positive (1).

Double-negative can confuse models.

1. **Input**: It could have been better. → **Output**: Negative (0).

Requires understanding implied sentiment.

1. **Input**: An interesting mix of good and bad moments. → **Output**: Neutral/Positive (1).

Mixed sentiment, challenging classification.

1. **Input**: The film tries hard but ultimately fails. → **Output**: Negative (0).

Contrastive sentiment (positive attempt, negative result).

1. **Input**: Leaves the audience wanting more. → **Output**: Positive (1).

Ambiguity around wanting more.

1. **Input**: Some scenes are great, others fall flat. → **Output**: Neutral/Negative (0).

Mixed polarity within a single sentence.

1. **Input**: The acting was fine, but the plot was dull. → **Output**: Negative (0).

Requires prioritization of the dominant sentiment.

1. **Input**: Not a movie I’d recommend. → **Output**: Negative (0).

Implied negativity without explicit words.

1. **Input**: Better than most, but still not great. → **Output**: Neutral/Negative (0).

Complicated sentiment requiring subtle understanding.

1. **Input**: I have mixed feelings about it. → **Output**: Neutral/Negative (0).

Explicit mention of ambiguity.

**Reflection on the Annotation Setup**

* **Annotator Sample:**

The original SST annotations were performed by crowd workers on Amazon Mechanical Turk.

Workers were likely native or fluent English speakers but may have varied in linguistic or cultural expertise.

* **Annotation Guidelines:**

Annotation guidelines provided by the dataset creators included examples of positive, negative, and neutral sentiment.

While detailed, nuances like sarcasm or idiomatic expressions might not have been well-anchored.

* **Inter-Annotator Agreement:**

Agreement metrics are reported in the original dataset documentation.

CoLA-like tasks typically aim for a high Cohen’s Kappa score (≥0.8), but agreement on difficult sentences may be lower.

* **Conflict Resolution:**

Conflicts were resolved using majority voting among annotators.

Instances with strong disagreement may have been excluded or flagged.

* **Dataset Quality:**

Strengths: Balanced dataset, diverse sentence structures, and real-world applicability.

Weaknesses: Some examples are ambiguous, and mixed sentiment instances might lack consistent annotation.

* **Opinion on Quality:**

The quality of the SST-2 dataset is generally high for binary sentiment analysis due to its balanced class distribution, diverse sentence structures, and real-world applicability, which make it a robust resource for training and evaluating models. However, the selection of instances includes some ambiguous examples and sentences with mixed sentiment, which could challenge both annotators and models. The annotations, while effective for most cases, may lack consistency in handling nuanced expressions, such as sarcasm or cultural idioms, and do not always provide clear guidelines for ambiguous or mixed-polarity cases. Enhancements, such as clearer annotation protocols for these edge cases or additional labels for nuanced sentiment, could further improve the dataset’s overall utility and reliability.

* **Explanation of Categorization:**

Easy Instances: Explicit sentiment, simple vocabulary, no ambiguity.

Difficult Instances: Ambiguous sentiment, idiomatic expressions, mixed polarity, or implied meaning.