Advanced NLP Assignment Report

Patil Anirudha Vivek (22234)

September 17, 2025

1 For RNN

1.1 Classification Report

Table 1: Classification Report for the Simple RNN Model

Class	Precision	Recall	F1-Score	Support
Neutral	0.23	0.12	0.16	592
Positive	0.29	0.17	0.21	1041
Extremely Negative	0.12	0.06	0.08	619
Negative	0.26	0.62	0.36	947
Extremely Positive	0.23	0.12	0.15	599
Accuracy			**0.25**	**3798**
Macro Avg	0.23	0.22	0.19	3798
Weighted Avg	0.23	0.25	0.21	3798

Final Test Accuracy: 0.2462

1.2 Inference

The RNN model has failed to learn any meaningful patterns, a behavior known as model collapse. It has adopted a simplistic strategy of predicting the ”Negative” class for nearly every input.

This is clearly evidenced by three key metrics in the report:

- The recall for the ”Negative” class is exceptionally high, showing it correctly identifies all true ”Negative” samples by guessing ”Negative” every time.
- The recall for all other classes is extremely low, indicating the model completely fails to identify any other sentiment.
- The final accuracy of ˜25% is not a measure of success, but simply reflects the baseline frequency of the ”Negative” class in the test data.

This outcome is a classic symptom of the vanishing gradient problem in Simple RNNs, which prevents the model from learning complex relationships within the text.

2 For LSTM

2.1 Classification Report

Table 2: Classification Report for the LSTM Model

Class	Precision	Recall	F1-Score	Support
Neutral	0.73	0.66	0.69	592
Positive	0.66	0.67	0.67	1041
Extremely Negative	0.82	0.77	0.79	619
Negative	0.66	0.72	0.69	947
Extremely Positive	0.76	0.75	0.76	599
Accuracy			**0.71**	**3798**
Macro Avg	0.73	0.71	0.72	3798
Weighted Avg	0.71	0.71	0.71	3798

Final Test Accuracy: 0.7098

2.2 Inference

The report indicates that the model has successfully learned to differentiate between all five sentiment classes, showing a deep understanding of the text data with an accuracy of 70%.

- **Balanced Performance:** The model provides consistently strong and balanced F1-scores across all categories, ranging from 0.67 to an impressive 0.79. This proves it is not ignoring any single class and is making nuanced predictions.
- **Architectural Success:** This successful outcome is due to the LSTM's architecture. Its internal gating mechanisms allowed it to capture longer-range dependencies and contextual information within the tweet data, a task where simpler recurrent models often fail.
- **Signal Detection:** The model shows a particular strength in identifying high-intensity sentiment, achieving its best F1-score of 0.79 on the ”Extremely Negative” category.

3 For Transformer

3.1 Classification Report

Final Test Accuracy: 0.7133

3.2 Inference

The Transformer model has proven to be a highly effective and robust classifier for this sentiment analysis task, achieving a final test accuracy of 71.3%. The classification report reveals two key strengths:

Table 3: Classification Report for the Transformer Model

Class	**Precision**	**Recall**	**F1-Score**	**Support**
Neutral	0.73	0.70	0.71	592
Positive	0.65	0.69	0.67	1041
Extremely Negative	0.80	0.80	0.80	619
Negative	0.68	0.65	0.67	947
Extremely Positive	0.77	0.76	0.77	599
Accuracy			**0.71**	**3798**
Macro Avg	0.73	0.72	0.72	3798
Weighted Avg	0.71	0.71	0.71	3798

- **Balanced Multi-Class Performance:** The model successfully learned to distinguish between all five distinct sentiment categories. This is evidenced by the strong and consistent F1-scores across all classes, ranging from 0.67 to a high of 0.80.

- **Strong Signal Detection:** The model demonstrates a particular aptitude for identifying clear sentiment, achieving its best performance on the "Extremely Negative" (0.80 F1-score) and "Extremely Positive" (0.77 F1-score) categories.

Overall, the results indicate that the Transformer architecture successfully captured the complex contextual nuances within the text data to make accurate and reliable predictions.