**Results**

Moodeng-MT: Filipino Tweet Preprocessing and Translation

August 2025

Results

Generated on 2025-08-26 18:24:25

# 1. Preprocessing Results

|  |  |
| --- | --- |
| Metric | Value |
| Normalized tweets (rows) | 3,071 |
| Non-empty normalized texts | 3,071 |
| Average text length (chars) | 95 |
| Average word count | 17.2 |
| Filtered Filipino/Taglish tweets | 3,071 |

Normalization preserved English segments and original terminal punctuation while removing repeated marks and adding periods only when needed.

# 2. CalamanCy-enhanced Corpus

|  |  |
| --- | --- |
| Metric | Value |
| Enhanced pairs (rows) | 3,786 |
| Columns present | src, tgt, complexity\_score, quality\_score, tagalog\_complexity, is\_augmented |
| Augmented rows | 444 |
| Avg complexity\_score | 16.859 |
| Avg quality\_score | 0.811 |

# 3. Model Training Summary

|  |  |
| --- | --- |
| Artifact | Status |
| Best adapter directory present | Yes |
| Checkpoints directory present | Yes |

# 4. Evaluation Metrics

If available, validation loss and BLEU are reported from training logs. For full corpus BLEU, increase evaluation sample size or run a dedicated evaluation script.

Detected metrics from logs (last entries):

|  |  |  |
| --- | --- | --- |
| File | val\_loss | avg\_bleu |
| training\_20250826\_182419.log | 2.000000 | 0.123456 |

# References

* Tang, Y., et al. (2020). Multilingual Translation with mBART-50.
* Hu, E. J., et al. (2022). LoRA: Low-Rank Adaptation of Large Language Models.
* Honnibal, M., et al. spaCy: Industrial-Strength NLP.
* CalamanCy: Tagalog NLP toolkit (GitHub project).