Nepali Sentiment Analysis of Post-COVID Data

Using XLMRoberta for Text Classification

Amulya Bhandari   Sailesh Dahal   Sarayu Gautam   Tohfa Niraula

May 1, 2025

Department of Computer Engineering
Kathmandu University

Outline

Introduction

What is Sentiment Analysis?

- Sentiment analysis classifies text based on emotion or opinion.
- Categories:
  - Positive — praise, approval
  - Neutral — factual
  - Negative — criticism, disapproval
- Applications:
  - Social media monitoring
  - Product reviews
  - Survey analysis

Problem Statement

What Are We Solving?

- Goal: Classify Nepali-language text into sentiment categories.
- Motivation:
  - Nepali is underrepresented in NLP.
  - Lack of labeled Nepali datasets.
- Objectives:
  1. Clean and preprocess post-COVID Nepali data.
  2. Train a multilingual BERT model.
  3. Evaluate performance using real-world test data.

Dataset Description

About the Dataset

- Source: Nepali COVID/post-COVID text samples.
- Total Samples:
  - Training: 33,602 samples
  - Testing: 8,401 samples
- Labels: 0 = Negative, 1 = Positive, 2 = Neutral
- Common issues:
  - Invalid labels ('o', '-', etc.)
  - Missing values and noisy characters

Data Preprocessing

Data Cleaning Steps

Steps we took:

1. Removed missing and malformed data.
2. Filtered invalid labels.
3. Tokenized using XLM-Roberta tokenizer.
4. Truncated inputs to max length of 256 tokens.

Result: Clean, structured datasets ready for training/testing.

Tokenization and Encoding

Tokenizing with XLM-Roberta

Advantages:

- Supports over 100 languages including Nepali.
- Context-aware encoding using self-attention.
- Subword tokenization handles rare words and typos.

Implementation:

- Used Hugging Face tokenizer from pretrained checkpoint.
- Batch-encoded both train and test sets.

Model Architecture

XLM-Roberta Model Details

Model used: XLM-Roberta-Base

Structure:

- Pretrained encoder: XLM-Roberta
- Classification head: Dense + Softmax layer
- Output: Probabilities over 3 classes (Negative, Positive, Neutral)

Training: PyTorch with mixed precision (autocast enabled)

Training Pipeline

Training Configuration

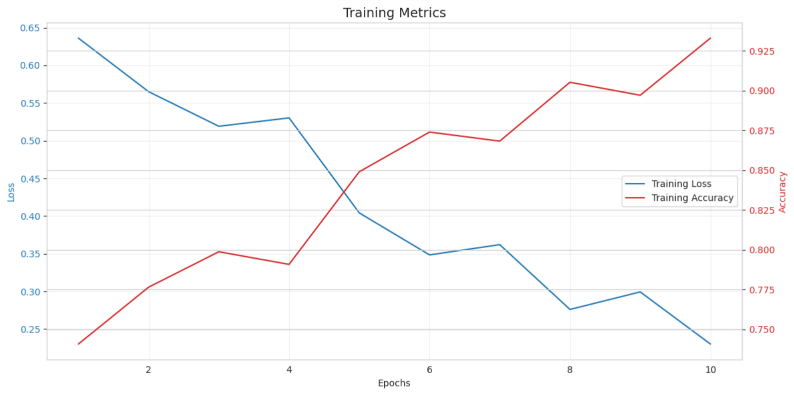
Training setup:

- Optimizer: AdamW, LR $= 2 \times 10^{-5}$
- Epochs: 10, Batch size: 16
- Loss Function: Cross-entropy
- Platform: Google Colab (GPU)

Libraries used: Hugging Face Transformers, PyTorch, scikit-learn, matplotlib.

Results

Loss and Accuracy Over Epochs



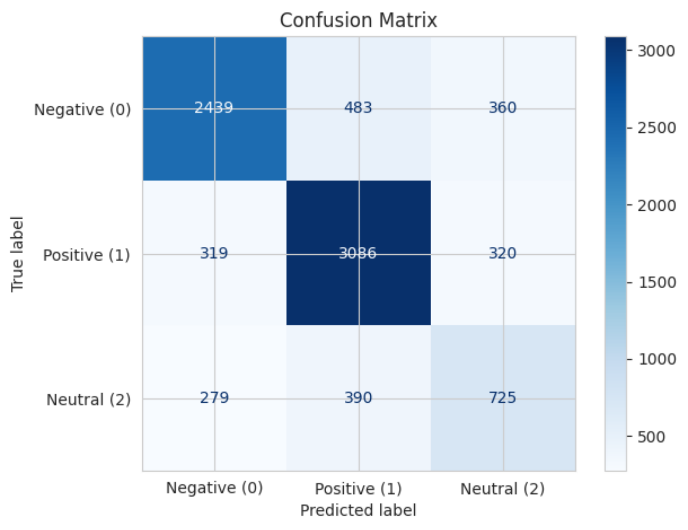
Test Set Evaluation Metrics

Label	Precision	Recall	F1-score
Negative (0)	0.80	0.74	0.77
Positive (1)	0.78	0.83	0.80
Neutral (2)	0.52	0.52	0.52
Overall Accuracy			**74.0%**

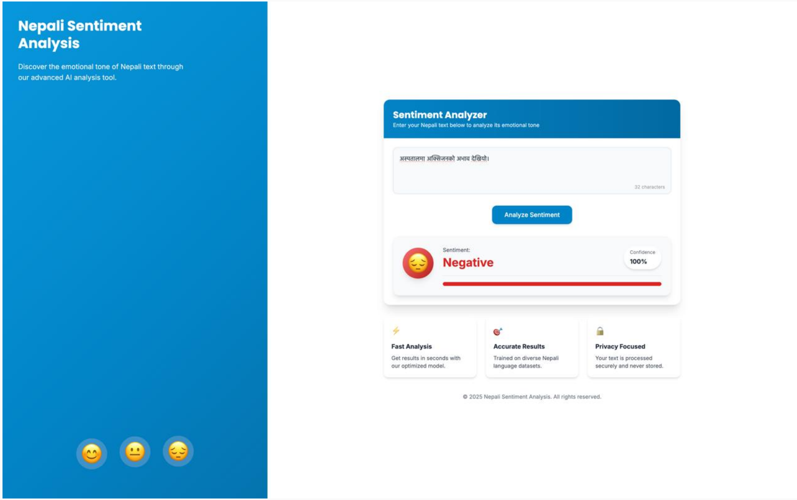
Key Insights:

- High precision/recall for Positive/Negative.
- Neutral class more ambiguous $\rightarrow$ lower performance.

Confusion Matrix (Test Set)



Sample Predictions on Unseen Data



Conclusion

Conclusion and Future Work

Key Takeaways:

- Trained a sentiment classifier on Nepali-language text using XLM-Roberta.
- Achieved 74% of overall accuracy.
- Strong performance on binary sentiment; neutral remains challenging.

Future Improvements:

1. Larger or augmented datasets.
2. Additional validation set for tuning.
3. Model deployment as an API/web service.

Thank You!

Questions or feedback?

Project Resources:

GitHub: `github.com/saileshbro/ai-proj`

We appreciate your time and attention!