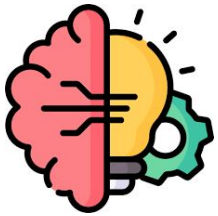


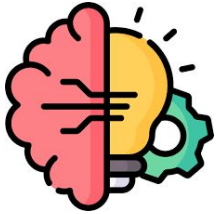
ANLI Round 2: Multiclass Natural Language Inference

Modeling, Evaluation & Deployment

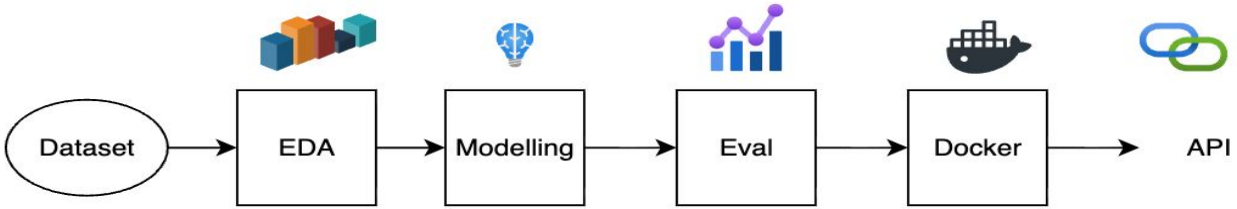
Sanshrit Bakshi

Northeastern University

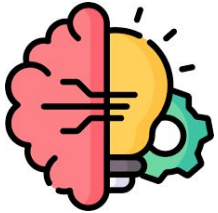
Project Objectives



- Build a complete ML pipeline for ANLI Round 2
- Perform data cleaning & EDA
- Train baselines + a transformer model
- Evaluate on official dev/test splits
- Deploy the final model using Docker + FastAPI
- Host final model on Hugging Face Hub



Dataset Overview



Dataset: ANLI (Adversarial NLI), Round 2

Task: 3-way classification

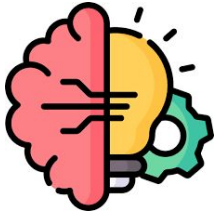
- entailment
- neutral
- contradiction

Splits Used:

- Train: 45,460
- Dev: 1,000
- Test: 1,000

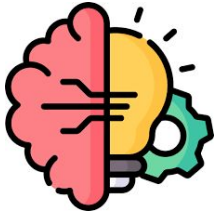
Fields: premise, hypothesis, label, reason

Data Cleaning



- 31 duplicate (premise, hypothesis) pairs found
- All duplicates had identical labels
- Safe to remove; no label conflicts
- Final training size: 45,429 samples
- No missing values in premise/hypothesis

EDA Results



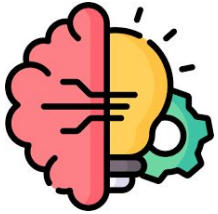
Label Distribution:

- Neutral: 46.1%
- Entailment: 31.8%
- Contradiction: 22.1%

Text Lengths:

- Premise: mean ~54 tokens
- Hypothesis: mean ~10 tokens
- 95th percentile < 256 tokens

Baseline Model: TF-IDF + LR



Why this baseline?

- Fast
- Strong classical method
- Provides lower bound on performance

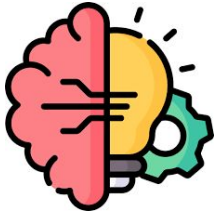
Configuration:

- Unigrams + bigrams
- max_features=100k
- Logistic Regression classifier

Performance (Dev):

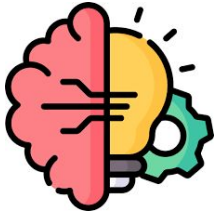
- Accuracy: **39.7**%
- Macro F1: **38.5**%

Why DeBERTa-v3?



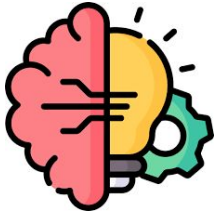
- State-of-the-art on NLI family tasks
- Disentangled attention + enhanced decoding
- Outperforms BERT and RoBERTa
- Better handling of syntax, word-order, semantics

Fine-Tuning Setup



Parameter	Value
Model	microsoft/deberta-v3-base
Max Length	256
Batch Size	16
Epochs	4
Learning Rate	2e-5
Optimizer	AdamW

Training Curve & Overfitting

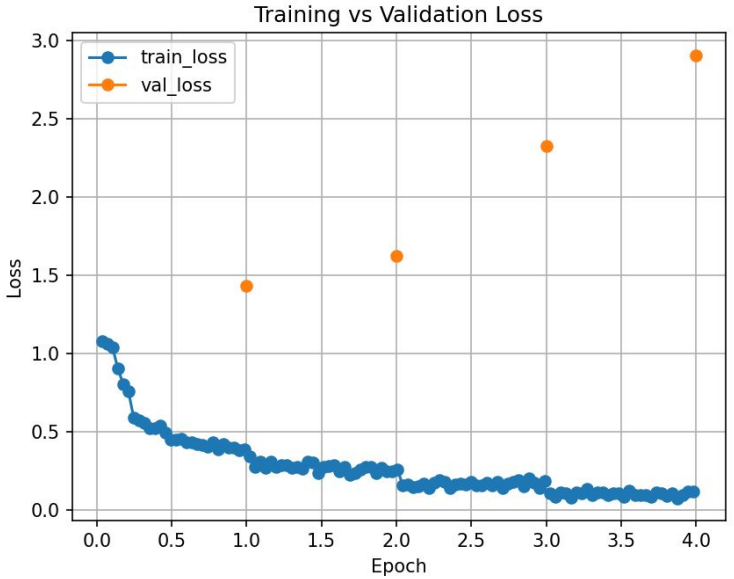


Training Loss: decreases smoothly

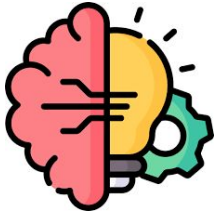
Validation Loss: increases after epoch 1

Interpretation:

- ANLI-R2 is adversarial → generalization is difficult
- Overfitting is expected
- Early stopping helps stabilize performance



Evaluation Results



Validation Set

- Accuracy: **48.4%**
- Macro F1: **0.487**

Test Set

- Accuracy: **49.4%**
- Macro F1: **0.493**