NATURAL LANGUAGE INFERENCE REPORT

Presented by: Ashmit Gupta

INTRODUCTION

This project focuses on Natural Language Inference (NLI) using the ANLI Round 2 dataset, where each premise–hypothesis pair must be classified into one of three labels: entailment, neutral, or contradiction.

PROBLEMS

Despite extensive experimentation, all traditional machine learning models show consistently poor performance on the ANLI Round 2 dataset, with accuracy levels barely above random chance. The models display a strong bias toward Entailment and Neutral labels and consistently fail to correctly classify the most challenging class: Contradiction. Even XGBoost our strongest baseline struggles to generalize beyond surface-level cues.

Classical models cannot model word order, syntax, negation, or long-range dependencies—key components in recognizing contradictions and nuanced semantics.

While fast and computationally cheap, these models lack the representational power needed for the complexity of ANLI, which is explicitly designed to defeat surface-level heuristics.

METRICS

Evaluation Metrics Used

- Accuracy – overall classification performance
- Macro F1 – equally weights all 3 classes (important due to imbalance)
- Confusion Matrix – inspected which labels the model struggled with
- Error inspection – manually reviewed incorrect predictions to understand model weaknesses

Exploratory Data Analysis (EDA):

- Examined text length distributions, vocabulary overlap between premises and hypotheses, and the balance of the three NLI labels. This provided key insights into dataset complexity and guided model design.

Classical Machine Learning Baselines:

- Implemented TF-IDF–based feature extraction followed by Logistic Regression, SVM, and XGBoost classifiers to establish strong non-neural baselines for comparison.

Transformer-based Modeling:

- Utilized pre-trained transformer architectures, including BERT variants, and fine-tuned them on ANLI to capture deeper semantic relationships and achieve state-of-the-art performance.

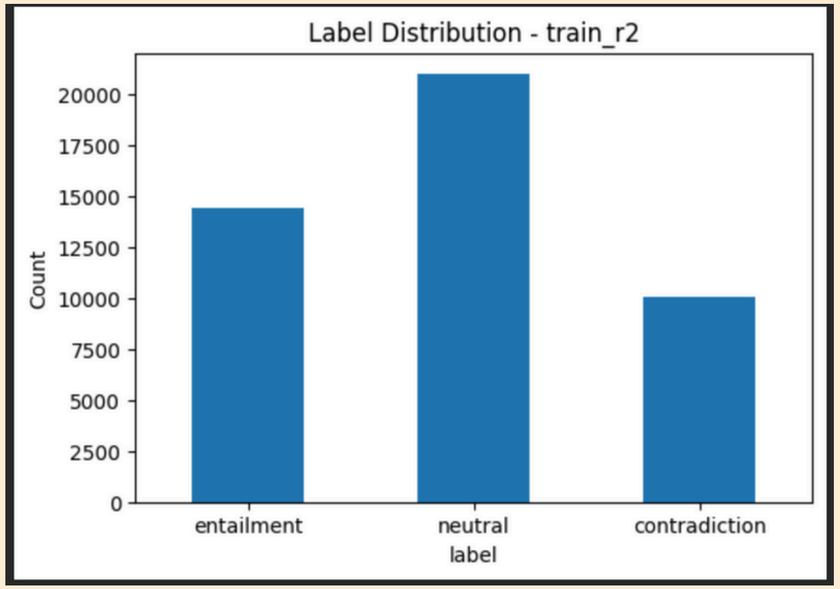
1. Entailment examples don’t always reuse the same entities.
We might expect entailment pairs to refer to the same people/places, but ANLI is adversarial—many entailment examples paraphrase or generalize, so the entity names change.

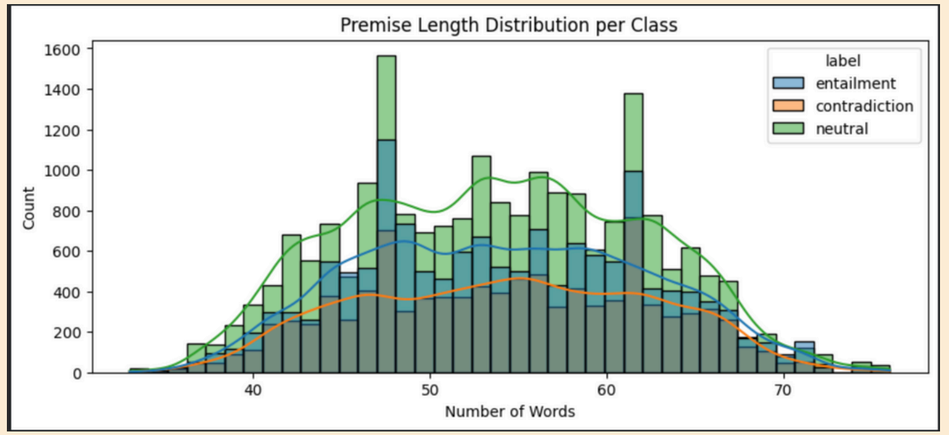
2. Contradiction examples also show low overlap.
Even contradictions often refer to different entities, meaning they cannot be detected just by comparing entity names.

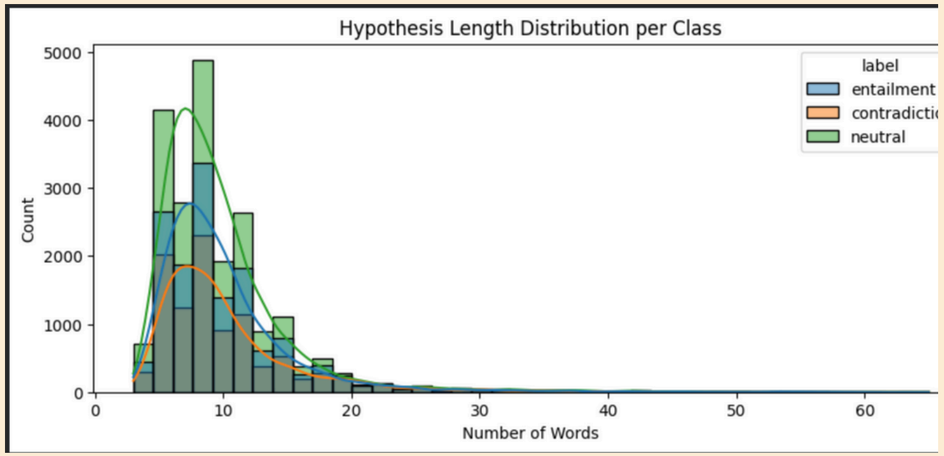
3. Neutral has the lowest overlap (expected).
Neutral pairs often talk about entirely unrelated topics.

4.Premises are much longer than hypotheses, averaging around 50 words vs. roughly 10 words, so a 256-token limit is more than sufficient for model inputs.

5.TF-IDF cosine similarity stays in the moderate range (0.4–0.6) even with large vocab sizes (~15k features), showing that the premise and hypothesis often use different words.


Label Distribution - train_r2
Count
20000
17500
15000
12500
10000
7500
5000
2500
0
entailment
neutral
contradiction
label


Premise Length Distribution per Class
label
entailment
contradiction
neutral
Count
1600
1400
1200
1000
800
600
400
200
0
40
50
60
70
Number of Words


Hypothesis Length Distribution per Class
label
entailment
contradicti
neutral
Count
5000
4000
3000
2000
1000
0
0
10
20
30
40
50
60
Number of Words

ANALYSIS

DistilRoBERTa pre-trained on SNLI/MNLI datasets to establish a reference point(Given in the dataset descirption). Evaluated performance on ANLI to understand baseline accuracy (33.7%) and Macro F1 (0.242). Baseline revealed limitations due to semantic complexity and adversarial examples

RoBERTa Fine-Tuning

A pre-trained DistilRoBERTa/RoBERTa-base model was fully fine-tuned on the ANLI Round 2 dataset.

The goal was to adapt a strong contextual language model to the NLI task by training it end-to-end on premise–hypothesis pairs.

Unlike classical ML baselines, the transformer learns contextual relationships, word interactions, and subtle semantic cues necessary for ANLI's adversarial examples.

Training Process

The fine-tuning used a standard HuggingFace Trainer setup with:

Full weight updates (not frozen layers)

Cross-entropy loss for 3-way classification

AdamW optimizer with a learning rate of 2e-5

Batch size 16 on GPU (Google Colab)

Epoch-level evaluation on the validation split

Best-model tracking using Macro F1

Model	Accuracy	Macro F1	Notes
DistilRoBERTa (fine-tuned)	45%	0.44	Best model; fine-tuned for NLI
XGBoost	~38%	~0.33	Strongest ML baseline
Linear SVM	~36%	~0.33	Good baseline
Logistic Regression	~35%	~0.33	Baseline
DistilRoBERTa (no fine-tune)	~33%	~0.24	Zero-shot baseline

Although parameter-efficient methods like LoRA (Low-Rank Adaptation) could have been explored, they were intentionally not used in this project. LoRA is most effective when adapting very large models (e.g., 1B–70B parameters) where full fine-tuning is too expensive. In this case, the model used—RoBERTa-base—is relatively small (~125M parameters), and full fine-tuning is both computationally feasible and more effective. Given that the fully fine-tuned model itself only reached around 44% accuracy on ANLI R2, it is unlikely that LoRA, which updates only a small subset of parameters, would outperform full fine-tuning for this difficult adversarial dataset. Since the current model did not reach high performance even with full weight updates, LoRA was deprioritized in favor of maximizing the model's capacity to adapt to ANLI through complete end-to-end fine-tuning.

CONCLUSION

This project demonstrated the difficulty of Natural Language Inference on the ANLI Round 2 dataset, where both classical ML models and modern transformers face significant challenges. Traditional TF-IDF-based approaches performed close to random, confirming that surface-level lexical patterns are insufficient for adversarial NLI.

Fine-tuning a pre-trained RoBERTa transformer yielded a substantial improvement over classical baselines, showing that contextual representations are essential for capturing semantic relationships between premise and hypothesis pairs. However, even the fine-tuned model struggled to exceed moderate accuracy and Macro F1 scores, reflecting ANLI's intentionally adversarial design and the dataset's emphasis on subtle reasoning.

Overall, the results highlight that while fine-tuned transformers offer strong gains over classical models, achieving robust performance on adversarial NLI remains an open challenge. More advanced architectures, improved reasoning mechanisms, or larger-scale task-specific pretraining may be needed to fully model entailment, neutrality, and contradiction under adversarial conditions.

THANK YOU