NATURAL LANGUAGE INFERENCE ANALYSIS REPORT

- Jay Vipin Jajoo

PROJECT OVERVIEW

This project explores Natural Language Inference (NLI) on the ANLI Round 2 dataset, where models classify premise-hypothesis pairs as entailment, neutral, or contradiction.

The methodology progresses through three stages:

(1) **Exploratory data analysis** of text lengths, word overlaps, and label distributions.

(2) **Traditional ML baselines using TF-IDF features** with Logistic Regression, Random Forest, and XGBoost; and

(3) **Deep learning approaches** with pre-trained transformers and fine-tuned **BERT models**, enhanced with Chain-of-Thought prompting.

A modular, production-ready pipeline was developed to automate data loading, training, evaluation, and comparison across models.

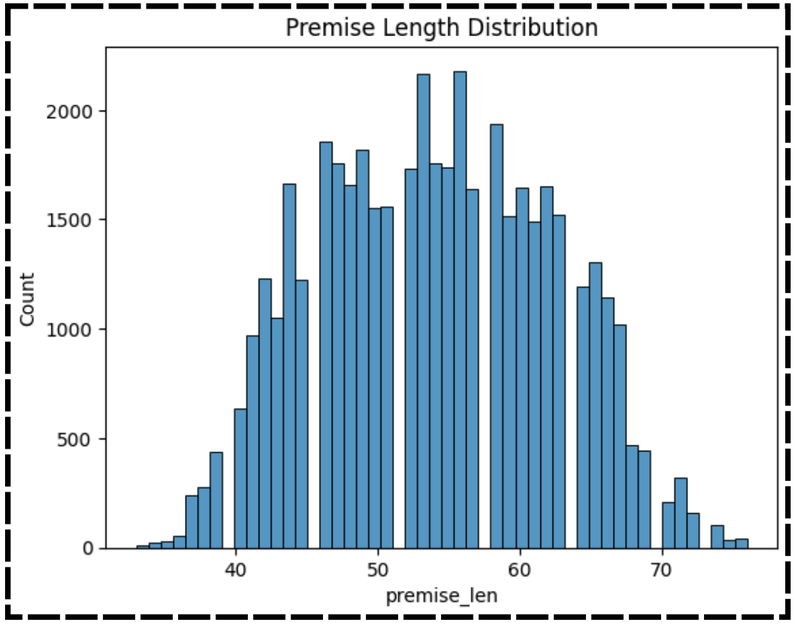
1. EDA

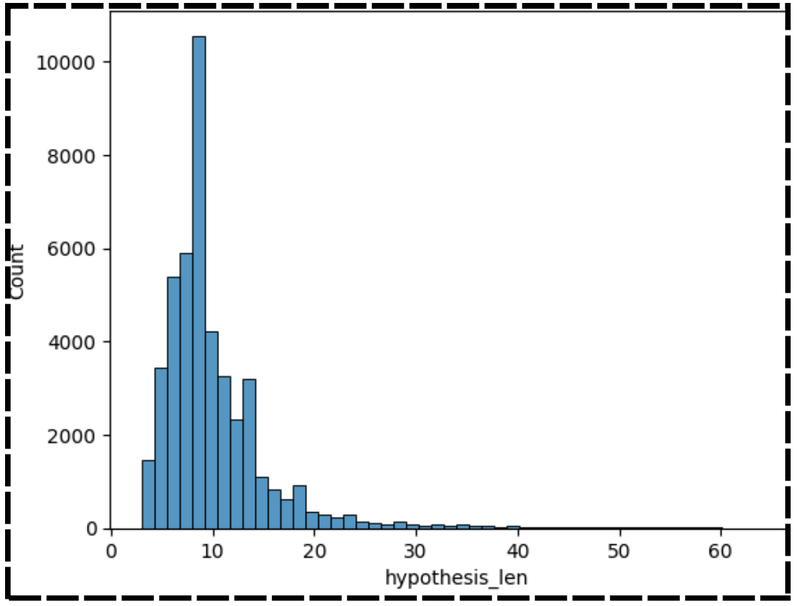
METHODOLOGY

This exploratory analysis was performed to understand the characteristics of the ANLI R2 dataset before modeling. Examining feature distributions, text lengths, label balance, and word overlap between premises and hypotheses helps to:

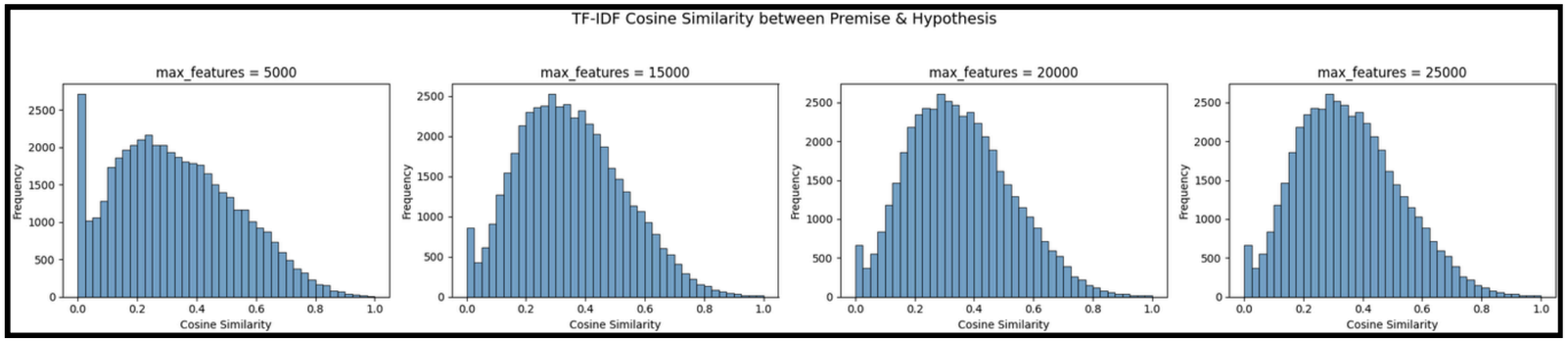
- **Identify patterns and trends** across different labels.
- **Detect anomalies or inconsistencies**, such as missing values or duplicate samples.
- **Understand semantic relationships** between premise and hypothesis using similarity measures **(Jaccard and TF-IDF cosine similarity).**
- **Highlight label-specific cues**, like frequently occurring words, to inform feature engineering and preprocessing.
- **Visualize distributions and similarities** to guide data-driven modeling decisions.

TEXT LENGTH ANALYSIS



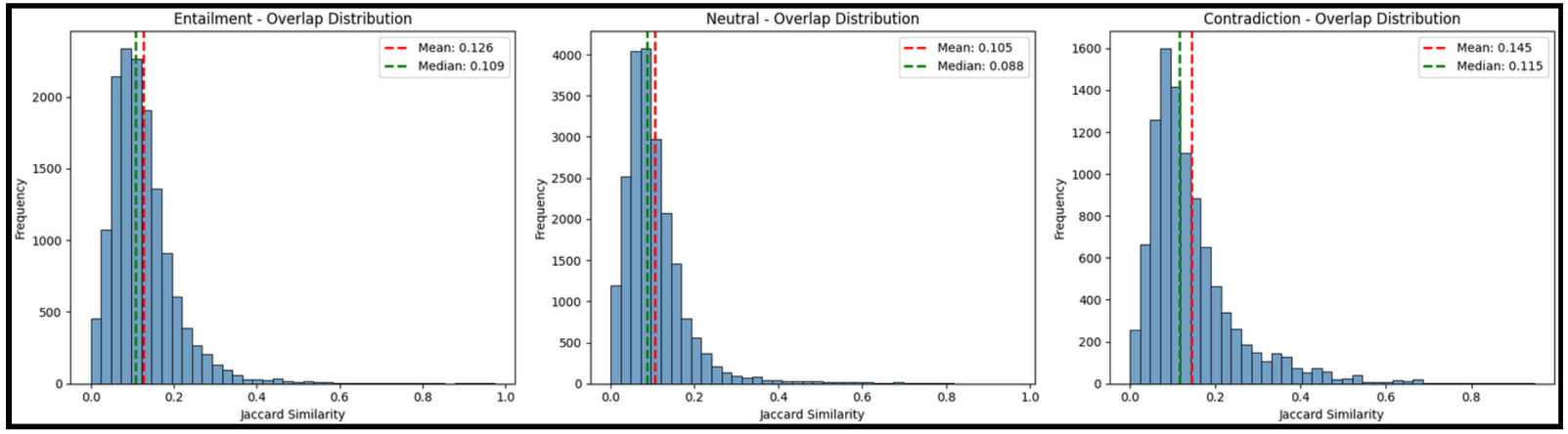


COSINE SIMILARITY DISTRIBUTION BASED ON MAX_FEATURES



The label distribution reveals a class imbalance in the dataset: **neutral samples (label 1) are the most frequent at 46.1%, entailment (label 0) make up 31.8%, and contradiction (label 2) are the least frequent at 22.1%**. This imbalance could affect model training, as the model may be biased toward predicting the more common classes.

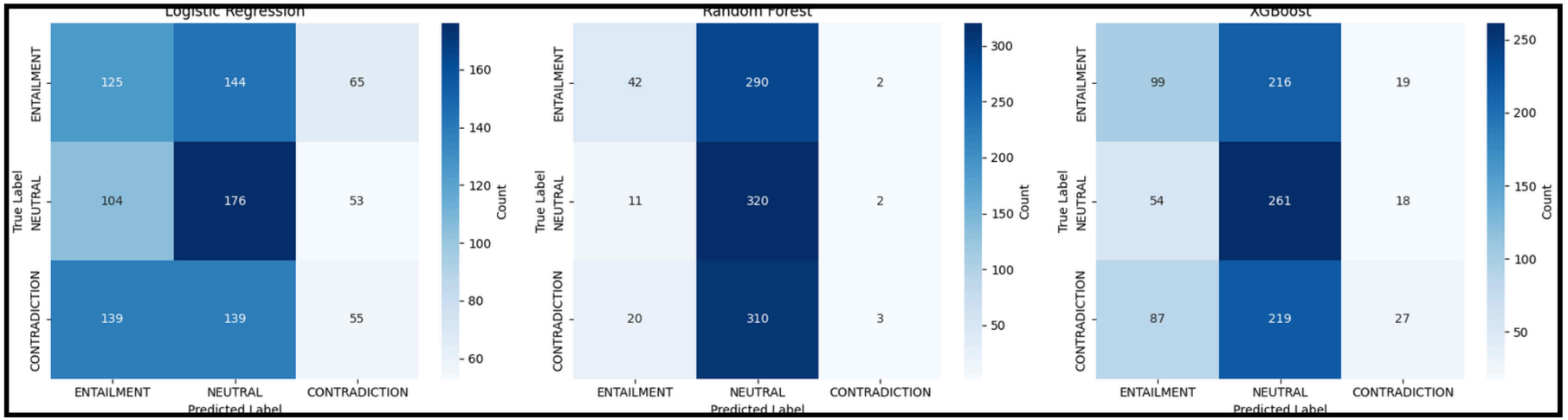
JACCARD SIMILARITY DISTRIBUTION PER LABEL



EDA SUMMARIZATION

The EDA shows that ANLI R2 premises are long (~50 words) while hypotheses are much shorter (~10 words), making 256-token inputs sufficient. TF-IDF cosine similarity across different feature sizes (~15K optimal) is moderate (~0.4–0.6), indicating lexical overlap is limited. Jaccard analysis reveals very low word overlap overall (10–15%), with contradictions surprisingly having the highest overlap, highlighting that simple word matching fails. These insights explain why traditional ML struggles and why transformer-based models like BERT, which capture semantic meaning rather than surface word matches, are necessary, especially to detect contradictions where similar words convey opposite meanings.

TRADITIONAL MACHINE LEARNING MODELS PERFORMANCE ANALYSIS



TO TRULY GAUGE A MODEL'S POTENTIAL, WE FIRST TRY TO OVERFIT IT. IF IT CAN'T EVEN CAPTURE THE TRAINING DATA (>90%), NO AMOUNT OF HYPERPARAMETER TUNING WILL MAKE IT PERFORM WELL ON NEW DATA.

All models perform poorly, with top accuracies close to random. They are biased toward ENTAILMENT and NEUTRAL, struggling to correctly predict CONTRADICTION. Even XGBoost, the best among them, shows limited ability to handle the hardest class.

Results Summary

Model	Test Accuracy	Test F1 (Macro)	Status
Logistic Regression	35.6%	0.339	Beats Baseline ✓
Random Forest	36.5%	0.245	Beats Baseline ✓
XGBoost	38.7%	0.329	Beats Baseline ✓
Baseline (DistilRoBERTa)	*33.7%*	*0.242*	*Reference*

Best Traditional ML Model: XGBoost with 38.7% accuracy

WHY TRADITIONAL ML MODELS LAG BEHIND ??????.....

- TF-IDF + CLASSICAL ML CAPTURES SURFACE-LEVEL FEATURES, MISSING DEEPER SEMANTIC RELATIONSHIPS
- LIMITED CONTEXT UNDERSTANDING: CANNOT MODEL WORD ORDER, SYNTAX, OR LONG-RANGE DEPENDENCIES
- OVERFITTING OBSERVED: TRAINING ACCURACY MUCH HIGHER THAN TEST (E.G., 72% → 36%) BUT STILL CAN'T ACHIEVE >=90% ACCURACY.
- PERFORMANCE CEILING: BEST XGBOOST ONLY 38.7%, SHOWING LIMITED IMPROVEMENT OVER BASELINE
- COMPUTATIONALLY EFFICIENT BUT LACKS REPRESENTATIONAL POWER FOR COMPLEX NLI PATTERNS

DEEP LEARNING MODELS PERFORMANCE ANALYSIS

METHODOLOGY FOR BERT FINE-TUNING ON ANLI

1. Baseline Evaluation:

Started with DistilRoBERTa pre-trained on SNLI/MNLI datasets to establish a reference point.

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.

3. Prompt-Based Input with Chain-of-Thought (CoT):

Idea: Guide the model to reason over the premise and hypothesis in a structured manner.

Implementation: BERT-large was provided with an input prompt designed to encourage the model to “think step by step” about the semantic relationship.

Goal:

- Reduce ambiguity in semantically tricky examples.
- Encourage the model to leverage intermediate reasoning steps instead of relying solely on memorized patterns.

2. BERT Fine-Tuning:

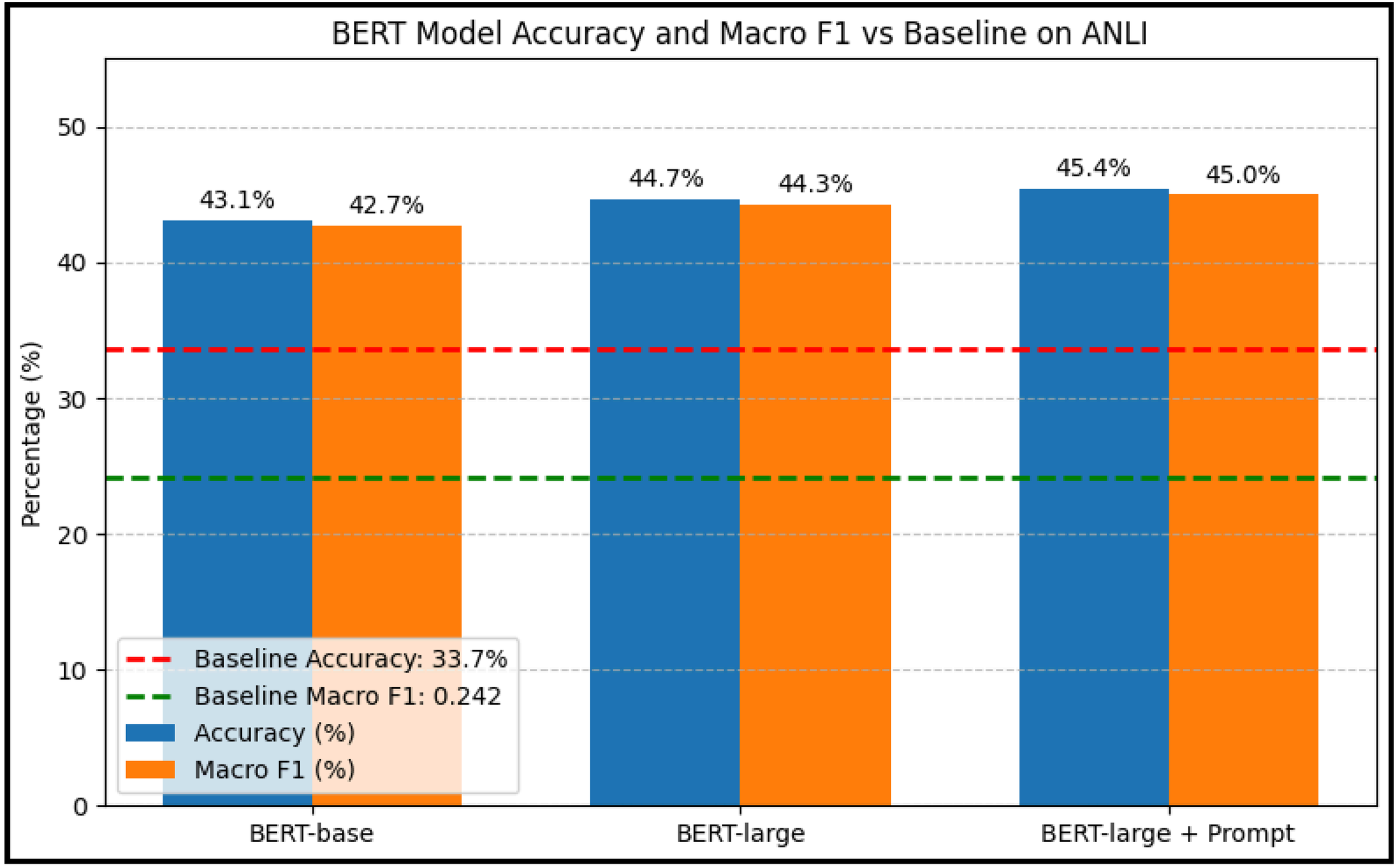
BERT-base and BERT-large models were fully fine-tuned on the ANLI dataset. The training aimed to leverage BERT’s contextual embeddings to capture deeper semantic relationships that traditional ML or pre-trained smaller models miss.

Training Enhancements:

- Gradient Accumulation: Simulated larger batch sizes without exceeding GPU memory.
- Mixed Precision (FP16): Faster computation and reduced memory usage.
- Gradient Clipping: Prevented exploding gradients during training.
- Early Stopping: Monitored validation Macro F1 to avoid overfitting.

Evaluation Metrics:

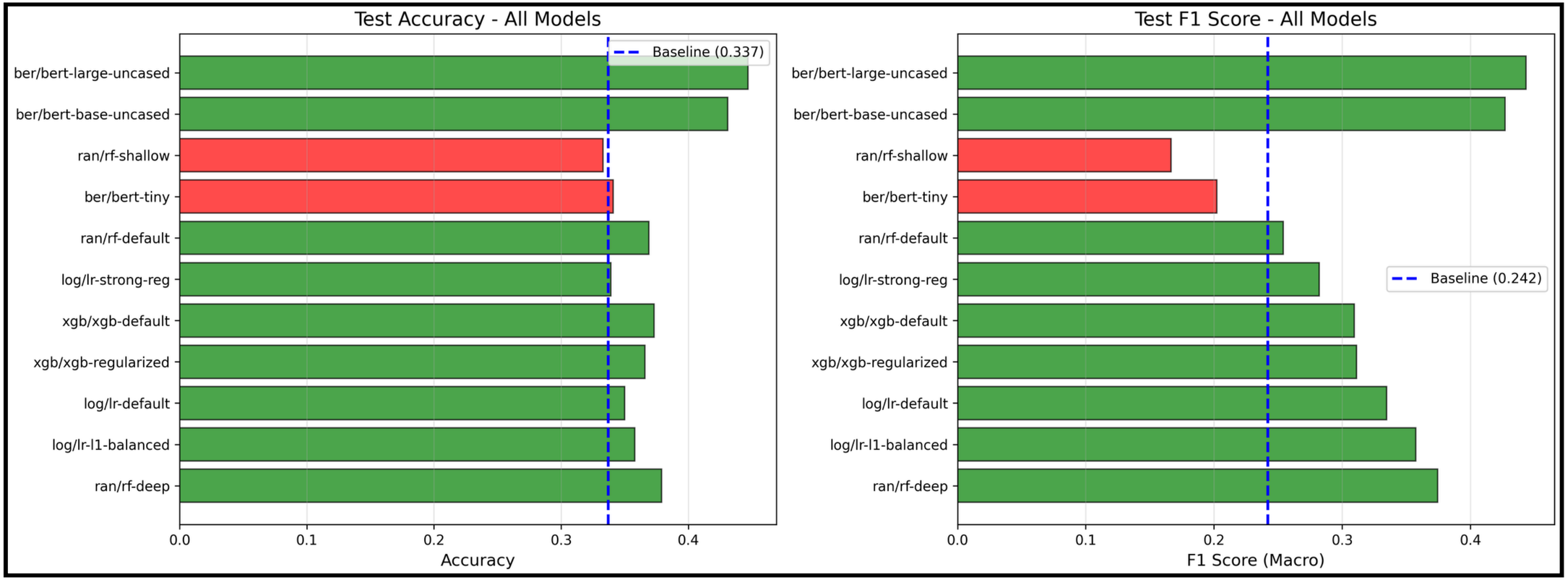
- Accuracy: Overall percentage of correct predictions.
- Macro F1: Measures class-balanced performance; crucial since ANLI has three classes and adversarial examples can skew results.
- Weighted F1: Accounts for class imbalance, providing a real-world weighted performance estimate.
- Confusion Matrices: Used to inspect per-class performance, especially for challenging CONTRADICTION examples.


BERT Model Accuracy and Macro F1 vs Baseline on ANLI
Percentage (%)
0
10
20
30
40
50
43.1%
42.7%
44.7%
44.3%
45.4%
45.0%
Baseline Accuracy: 33.7%
Baseline Macro F1: 0.242
Accuracy (%)
Macro F1 (%)
BERT-base
BERT-large
BERT-large + Prompt

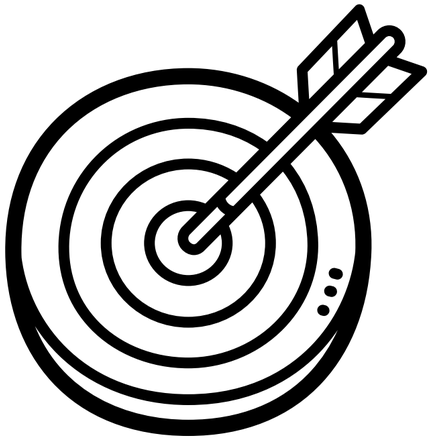
INSIGHTS

- Fine-tuning BERT models yielded substantial improvement over the baseline (DistilRoBERTa: 33.7% acc, 0.242 F1 → BERT-large: 44.7% acc, 0.443 F1).
- Larger models captured deeper semantic relationships, resulting in consistent accuracy and F1 score gains.
- Prompt-based fine-tuning (BERT-large + Prompt: 45.4% acc, 0.450 F1) showed a small improvement, but analysis revealed the model mostly ignored the prompt, with gains attributed to optimization factors like learning rate scheduling and gradient accumulation.
- Improvements were achieved through careful training strategies — FP16 mixed precision, gradient clipping, and early stopping.
- Reproducibility (seeds, checkpoints) was key to achieving stable and comparable performance across experiments.

COMPARISION OF ALL MODELS



This comparison highlights that even the most powerful neural models, including large-scale BERT variants, struggled to exceed 50% accuracy or F1 score on ANLI. Traditional ML models like Random Forests and XGBoost performed similarly or worse, showing the dataset's extreme difficulty. The consistent gap across architectures confirms ANLI's adversarial nature and the limits of current language understanding models.



CONCLUSION

This project demonstrated the challenges of Natural Language Inference on the ANLI dataset, where both traditional ML and advanced transformer models struggled to achieve high accuracy. Despite fine-tuning large BERT variants and experimenting with prompting strategies, performance plateaued below 50%, underscoring ANLI's adversarial complexity and the limits of current models in capturing nuanced reasoning. The results highlight that while fine-tuned transformers substantially outperform classical baselines, true understanding of entailment, neutrality, and contradiction in adversarial contexts still requires more advanced reasoning architectures and richer contextual modeling.

ALTERNATIVE APPROACH AND RATIONALE FOR EXCLUSION

An additional idea is to train classical models on vector embeddings extracted from a fine-tuned BERT model. However, this approach is unlikely to yield strong results because classical algorithms cannot fully exploit the contextual and semantic richness encoded in BERT embeddings. Such models would treat the embeddings as static numerical features, losing the deep relational understanding between premise and hypothesis. Therefore, fine-tuning transformer-based models end-to-end remains the most effective direction for improving performance on this task.

THANK YOU