Università degli Studi di Milano

Data Science for Economics

Text Mining and Sentiment Analysis

Project 8: Safe Clinical NLI

Ketrin Hristova
Matriculation Number: 13209A

1 Introduction

Natural Language Inference (NLI) is a pivotal task in natural language understanding, aiming to ascertain if a statement logically follows from a given premise. This work studies NLI in the clinical domain using the NLI4CT benchmark from SemEval-2024, which targets inference over Clinical Trial Reports (CTRs).
The task requires biomedical and numerical reasoning to decide whether the premise entails or contradicts the statement.
To model the task, transformer encoders are used as binary classifiers, comparing general-purpose and biomedical variants: BERT-base, PubMedBERT, Bio_ClinicalBERT, BioMed-RoBERTa, and DeBERTa-v3.

2 Dataset Description

This study uses two JSON files derived from NLI4CT: the first file (`train.json`) is partitioned into an 80% training split and a 20% validation split, yielding 1,410 training instances and 352 validation instances (1,762 total). The second file (`gold_test.json`) contains 5,500 held–out instances and is consulted only once for the final evaluation.
Each example pairs an expert-written *statement* with a section-specific CTR *premise* and a binary label in {Entailment, Contradiction}.
Premises come from four CTR sections that are relevant for factual claims: Eligibility, Intervention, Results, and Adverse Events. There are two instance types: *Single* (one CTR) and *Comparison* (two CTRs from the same section).

On the training portion, the labels are essentially balanced: contradiction and entailment each account for about half of the examples.
By instance type, *Single* trials are more frequent than *Comparison* trials (about 60% vs. 40%).
Coverage across CTR sections is broad: Adverse Events contributes the largest share, followed by Eligibility, Intervention, and Results.

The length of the statements differs between training and testing sets.
In the training data, statements are shorter, with an average of about 120 tokens.
In the gold test, statements are longer, averaging about 175 tokens.
Test statements are longer by roughly fifty to sixty tokens on average and contain more numerical details and comparisons. This shift toward longer statements is important for interpreting results, as it can make inference harder and increase

the chance that models must reason over counts, rates, or dosages stated in more elaborate ways.

The test set is also moderately imbalanced toward contradiction (about two thirds contradiction versus one third entailment).
Taken longer statements and a skew toward contradiction, the gold test provides a harder and more realistic evaluation of generalization than the balanced, shorter statements seen during training.

For all experiments, only the fields required by the task are used: the example identifier, the instance `Type` (Single or Comparison), the `Section_id`, CTR identifiers (`Primary_id` and, when present, `Secondary_id`), the expert `Statement`, and the gold `Label`.

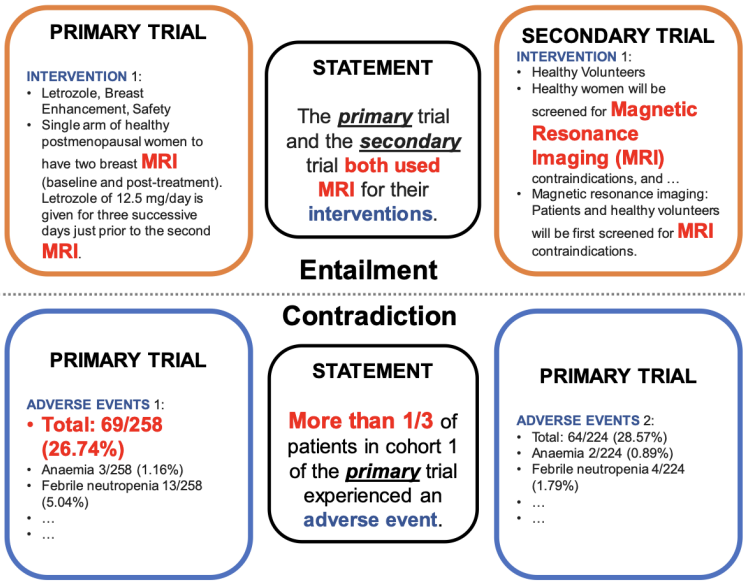

Figure 1: A demonstration of textual entailment and contradiction between the medical statements and clinical trial records. The statement may claim one or two CTRs on a specific section.

Figure 1

3 Data Preprocessing

The raw ClinicalTrials.gov text is turned into inputs that are *consistent for modeling* while *preserving the biomedical cues* (numbers, units, thresholds, acronyms).

All JSON files are read as UTF-8 and passed through a light normalizer (`ftfy` with a fallback to NFKC). This removes mojibake and standardizes symbols so they tokenize deterministically. For example, "It's 75 mg/m^2" renders correctly, and comparators appear as $\leq$ / $\geq$ rather than broken byte sequences.

Whitespace is collapsed and stray newlines removed , but semantics are not changed. Only non-informative punctuation is stripped; symbols needed for quantitative reasoning are *retained*: `% / + - . , < > = ^`, and the comparators $\leq/\geq$. As a result, phrases like "75 mg/m^2", "ORR 32%", or "platelets $\geq$ 100,000/mm^3" remain intact. No lowercasing is applied in cleaning, so medically meaningful case (e.g., `HER2`, `ER`, `MRI`) is preserved for cased encoders; uncased encoders lowercase internally, so the same cleaned text is safe for both.

In NLI4CT the same clinical trial can appear in multiple instances (as a *Single* premise, or paired in a *Comparison*). A random row-wise split would place the same trial text in both train and validation, yielding optimistic scores due to memorization (data leakage).
We target a validation set of roughly 20% while preserving group integrity (no trial appears in both splits) and class balance.
We use `StratifiedGroupKFold(n_splits=5, shuffle=True, random_state=42)`. This keeps label proportions (stratification) and never breaks a connected component (group). If `StratifiedGroupKFold` is unavailable or fails we use `GroupShuffleSplit(test_size=0.20, random_state=42)`. This still honors groups but does not strictly stratify by label.
In both cases we verify leak-freedom by asserting that the sets of trial IDs in train and validation are disjoint.
The held-out `gold_test.json` is never used for splitting or training; it is evaluated once at the end, providing an unbiased measure of out-of-sample performance.

4 Modeling Implementations

For `Single` cases the premise is the requested CTR section. For `Comparison` cases the premise concatenates the two trials' sections with the tokenizer's native separator token `[SEP]`.

We tokenize with `truncation="longest_first"` under a 512–token limit. When the pair is too long, the *longer side* (almost always the premise) is trimmed first while the statement is kept intact. This matters because gold–test statements are often long and number–heavy; preserving the whole hypothesis avoids clipping the claim itself and reduces label noise.

Padding is deferred to the batch via `DataCollatorWithPadding`, so each batch pads only up to its own longest sequence. This saves compute when section lengths vary (e.g., *Eligibility* vs. *Results*), and keeps tensors consistent across models and hyperparameters.

We fix the label indices once (`0` $\rightarrow$ Contradiction, `1` $\rightarrow$ Entailment) and reuse the same mapping in training, validation and testing.

The main metrics reported are accuracy, precision, recall, F1 and macro–F1 (the unweighted average of per–class F1). Macro–F1 is important because the held–out test set is skewed toward Contradiction; macro–averaging prevents the majority class from dominating the score.

5 Model Descriptions

The models used are pretrained Transformer encoders used as cross-encoders for binary NLI. Each model sees the *premise* and *statement* jointly, and a small classification head predicts {Contradiction, Entailment}. The models differ only in their backbone: same BERT-style interface, but with different pretraining corpora/vocabularies or architectural tweaks.

BERT-base-uncased is a bidirectional Transformer pretrained on general English with MLM (Masked Language Modeling) and NSP (Next–Sentence Prediction). It uses WordPiece tokenization, absolute positional embeddings, and segment (A/B) embeddings that naturally separate the two parts of the pair. In this task it preserves clinical symbols (e.g., %, /, $\leq$, $\geq$) and tokenizes numeric strings

such as "75 mg/kg," providing a strong and transparent general baseline for CTR NLI.

PubMedBERT is a BERT-architecture encoder pretrained *from scratch* on PubMed abstracts and full text with a biomedical WordPiece vocabulary. The domain-specific vocabulary markedly reduces subword fragmentation for drug names, biomarkers, units, and outcome acronyms, improving lexical alignment between premises and statements written in clinical-trial language.

BioClinicalBERT is a BERT model continued-pretrained on clinical notes (e.g., MIMIC) with MLM (Masked Language Modeling). This captures clinical shorthand, frequent negation, and eligibility-style phrasing common in trial documents; despite stylistic differences between notes and templated CTR prose, the additional exposure helps with medical abbreviations and rules-like text.

DeBERTa-v3-base is a Transformer encoder with disentangled attention (separate content and position vectors) and relative position biases, plus improved MLM (Masked Language Modeling) training. The relative positions help align evidence that may lie far from the claim within long premises, supporting steadier generalization on CTR NLI where statements are matched against lengthy, structured sections.

6 Model Selection and Final Evaluation

Model	Train Macro-F1	Val Macro-F1	Train Precision	Val Precision	Train Recall	Val Recall
Bio_ClinicalBERT	0.585	**0.525**	**0.614**	**0.557**	0.482	0.372
BERT-base-uncased	**0.597**	0.514	0.612	0.542	0.539	0.359
BiomedNLP-PubMedBERT-base-uncased-abstract-fulltext	0.563	0.506	0.605	0.548	0.421	0.317
DeBERTa-v3-base	0.494	0.498	0.503	0.503	**0.658**	**0.607**

Table 1: Validation performance. Macro-F1 is the primary model-selection metric.

Bio_ClinicalBERT yields the best validation **macro-F1** (0.525) with the highest **precision** (0.557), making it the strongest model overall. **BERT-base-uncased** is close behind (macro-F1 0.514; precision 0.542) and serves as a robust general baseline. **DeBERTa-v3-base** shows very high **recall** (0.607) but a near-chance

macro-F1 (0.498), indicating poor balance .**PubMedBERT** attains solid precision (0.548) but low recall (0.317), behaving conservatively on this split. Based on these trends, we carry forward **Bio_ClinicalBERT** (best macro-F1) and **BERT-base-uncased** to the held-out `gold_test`.

BERT-base-uncased

	Pred: C	**Pred: E**
True: C	1980 (54.1%)	1679 (45.9%)
True: E	853 (46.3%)	988 (53.7%)

	Precision	**Recall**
Contradiction	0.70	0.54
Entailment	0.37	0.54
Accuracy	0.54	
Macro-F1	0.52	

Bio_ClinicalBERT

	Pred: C	**Pred: E**
True: C	1876 (51.3%)	1783 (48.7%)
True: E	874 (47.5%)	967 (52.5%)

	Precision	**Recall**
Contradiction	0.68	0.51
Entailment	0.35	0.53
Accuracy	0.52	
Macro-F1	0.50	

BERT-base-uncased edges out Bio_ClinicalBERT on the test set (accuracy 0.54 vs. 0.52; macro-F1 0.52 vs. 0.50). It also slightly improves *Entailment recall* (0.54 vs. 0.53) and reduces the major error mode—*True Contradiction misclassified as Entailment*—from 48.7% (Bio_ClinicalBERT) to 45.9%. **Bio_ClinicalBERT** retains good sensitivity to *Entailment*, but its precision remains lower.

Bio_ClinicalBERT domain knowledge helps on the in-distribution validation split, but the test set's *length*, *class skew*, and *style* favor BERT-base-uncased's robustness (case-invariant vocabulary, NSP-style pair modeling, steadier calibration), so it generalizes slightly better when the distribution moves.

7 Evidence-Based Retrieval (`bert-base-uncased`)

Finally, a human-readable justification for the trained prediction is provided. We reuse the same tokenizer, input packing, and truncation policy as in training and score each premise sentence against the statement. Using the **BERT-base-uncased**, we return the top-k sentences that the classifier itself scores highest *for its own predicted label*. This procedure does not change the label; it only surfaces the text that best justifies the model's decision.

Below are representative outputs of the sentence re-scoring step.

- *Statement:* "The primary trial dose is lower than the secondary trial."
  Gold: ENTAILMENT *Pred:* ENTAILMENT
  Top evidence: Sentences with explicit doses, e.g., "0.5 mg/kg", "0.2 mg/kg", "0.040 mg/kg/dose".
  Here the evidence is informative: it contains the numbers required to validate the comparative claim, so the justification reads sensibly.

- *Statement:* "There were no urinary tract infections in the primary trial."
  Gold: CONTRADICTION *Pred:* ENTAILMENT
  Top evidence: "11/50 (22.00%) ... 1/50 (2.00%) urinary tract infection ..."
  The retrieved sentence explicitly lists a UTI, which contradicts the statement. The explainer is helpful, but it exposes a model error: the classifier did not align the numeric mention with the negation in the statement.

Sentences that mention the same entities (e.g., "emesis," "UTI," drug names) are regularly ranked highest, even when they do not logically support the claim. This aligns with the observed *false positives for* ENTAILMENT and helps explain the best overall *macro-F1* hovering around 0.52–0.54 : the model localizes the right region but often mistakes "mentions" for "support." Many errors also occur when the claim hinges on negation ("no UTI") or a relation ("twice as likely,"). The top evidence is frequently numeric but not the *right* comparison, matching the confusion-matrix pattern where CONTRADICTION rows are mislabeled as ENTAILMENT.

In short, the evidence view confirms that the model is good at *finding the right topics* but not reliably *verifying the right relations* (negation, ratios, "twice," thresholds), which caps performance.

References

Resources

- NLI4CT: Multi-Evidence Natural Language Inference for Clinical Trial Reports — https://pure.manchester.ac.uk/ws/portalfiles/portal/358459907/2305.02993v2.pdf
- SEME at SemEval-2024 Task 2: Comparing Masked and Generative LMs on NLI for Clinical Trials — https://hal.science/hal-04536273v1/document
- KnowComp at SemEval-2023 Task 7: Fine-tuning Pre-trained Language Models for Clinical Trial Entailment Identification — (https://aclanthology.org/2023.semeval-1.1.pdf)
- NLI4CT — Data formatting (SemEval-2023) — https://sites.google.com/view/nli4ct/semeval-2023/data-formatting
- NLI4CT — CodaLab competition page — https://codalab.lisn.upsaclay.fr/competitions/8937