CSCE 638: Natural Language Processing Fall 2022 Project Report

Team Members: Md Messal Monem Miah, Abdullah Aman Tutul, Adrita Anika, Fardeen Hasib Mozumder, and Md Maklachur Rahman

Multi-evidence Natural Language Inference for Clinical Trial Data

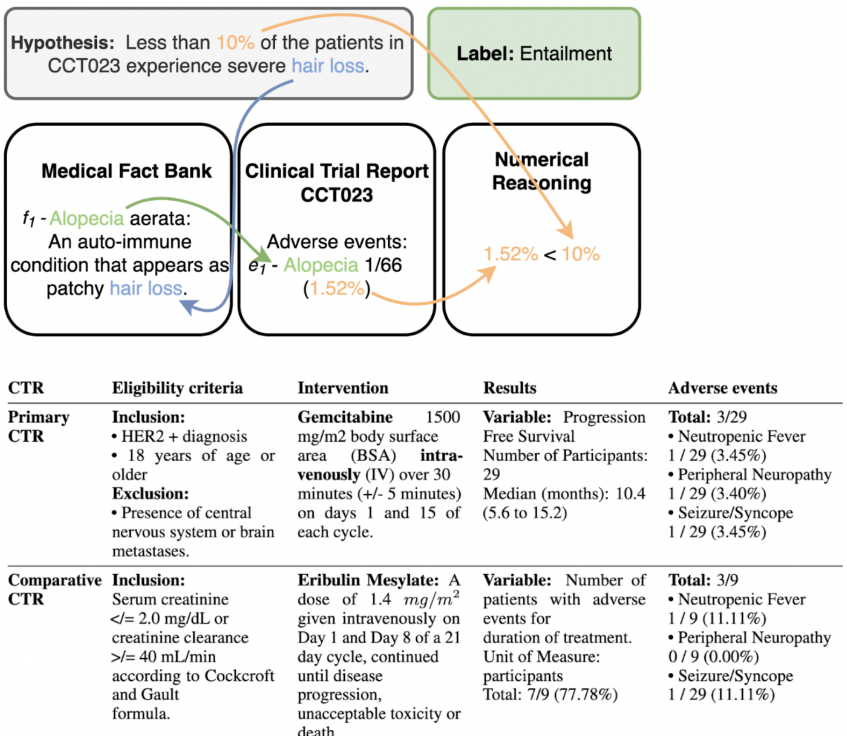
1. Introduction [contributor Adrita]

A clinical trial (CTR) is a medical research study that involves human subjects. It is typically conducted to assess a medicinal, behavioral, or surgical intervention. Clinical trials are the main tool used by researchers to determine whether a new medication, diet, or medical gadget is a safe and effective treatment for patients. It is frequently conducted to see if a novel treatment is superior to the current standard of care in terms of efficacy and adverse effects.[1] As healthcare practitioners depend on clinical trials to treat a medical condition, the ideal approach to gaining information from these CTRs might be a database of CTRs, however, no such database exists. [1] Clinical trial results are typically lengthy natural language articles, and for healthcare practitioners, reading all of them would be prohibitively time-consuming. Additionally, there has been a significant increase in the publication of CTRs in recent years. For example, there have been a total of 435,275 studies recruiting studies registered on ClinicalTrials.gov.[2] Also, there are over 10,000 CTRs for Breast Cancer alone. [1] With such an increase in CTRs, it is becoming more challenging for health practitioners to glean information from all those CTRs.

Natural language processing (NLP) may speed this process up as CTRs are typically lengthy natural language articles. It offers the possibility of assisting with the large-scale interpretation and retrieval of medical evidence. Successful development could significantly improve how we link the most recent evidence to support personalized care. [2] The specific task of NLP that deals with this type of problem is referred to as Natural Language Inference (NLI). The goal of NLI is to determine whether a natural language hypothesis (h) can be justifiedly inferred from a natural language premise (p). A premise is a claim that serves as the foundation for an argument or the basis for a conclusion. A hypothesis is an assumption or an idea proposed for the sake of an argument. The difficulties faced by NLI are very different from those faced by formal deduction since informal reasoning, lexical semantic knowledge, and linguistic expression variety are prioritized. [3] For NLI in the medical domain, the clinical trials are essentially the premises where all the information of the trial is stored. A hypothesis is an assumption determined from the clinical trial (i.e. premise). If the hypothesis is entailed by the premise it is considered an "entailment", and if not it is considered a "contradiction". The term "multi-evidence" refers to the reasoning by propagating information between multiple sentences.

[1] See https://www.nia.nih.gov/health/what-are-clinical-trials-and-studies
[2] See https://clinicaltrials.gov/ct2/resources/trends

To summarize, the task is to determine entailment or contradiction given a premise (CTRs) and hypothesis (assumption) through reasoning between multiple sentences as can be seen in Fig 1.


Hypothesis: Less than 10% of the patients in CCT023 experience severe hair loss.
Label: Entailment
Medical Fact Bank
f_1 - Alopecia aerata: An auto-immune condition that appears as patchy hair loss.
Clinical Trial Report CCT023
Adverse events: e_1 - Alopecia 1/66 (1.52%)
Numerical Reasoning
1.52% < 10%

CTR	Eligibility criteria	Intervention	Results	Adverse events
Primary CTR	**Inclusion:** • HER2 + diagnosis • 18 years of age or older **Exclusion:** • Presence of central nervous system or brain metastases.	**Gemcitabine** 1500 mg/m2 body surface area (BSA) **intra-venously** (IV) over 30 minutes (+/- 5 minutes) on days 1 and 15 of each cycle.	**Variable:** Progression Free Survival Number of Participants: 29 Median (months): 10.4 (5.6 to 15.2)	**Total:** 3/29 • Neutropenic Fever 1 / 29 (3.45%) • Peripheral Neuropathy 1 / 29 (3.40%) • Seizure/Syncope 1 / 29 (3.45%)
Comparative CTR	**Inclusion:** Serum creatinine </= 2.0 mg/dL or creatinine clearance >/= 40 mL/min according to Cockcroft and Gault formula.	**Eribulin Mesylate:** A dose of 1.4 mg/m^2 given intravenously on Day 1 and Day 8 of a 21 day cycle, continued until disease progression, unacceptable toxicity or death	**Variable:** Number of patients with adverse events for duration of treatment. Unit of Measure: participants Total: 7/9 (77.78%)	**Total:** 3/9 • Neutropenic Fever 1 / 9 (11.11%) • Peripheral Neuropathy 0 / 9 (0.00%) • Seizure/Syncope 1 / 29 (11.11%)

Fig 1: An example of the dataset of Breast cancer CTRs

In our work, we have used a collection of breast cancer CTRs[3], hypotheses, explanations, and labels annotated by domain expert annotators. The hypotheses and the evidence are annotated by clinical domain experts, clinical trial organizers, and research oncologists from the Cancer Research UK Manchester Institute. In Fig 1, we can see the proposed tasks utilize the Eligibility criteria, Intervention, Results, and Adverse events sections of CTRs.[4]

- ***The Eligibility Criteria*** *is a set of conditions patients need to satisfy in order to participate in a clinical trial, presented in an unstructured text format.*
- ***The Intervention*** *is a detailed description of the dosage/frequency of the intervention, the route of administration, the duration of the study, and the types of interventions being studied. Also in unstructured text form.*
- ***The Results*** *are split into two parts; an unstructured text describing a clinical arm, and a structured table detailing the variable being evaluated, the units being used, and the results.*
- ***The Adverse Events*** *are structured tables containing the signs and symptoms observed in patients during the clinical trial.*

[3] Extracted from https://clinicaltrials.gov/ct2/home
[4] For details of the problem see https://sites.google.com/view/nli4ct/

2. Challenges [contributor Messal, Adrita]

The task is challenging due to several facts as described below.

3.1. The problem requires substantial amounts of quantitative and numerical reasoning. Hence, it is notoriously difficult for state-of-the-art (SOTA) NLI models. For example,

Premise: *NCT02953860 INTERVENTION 1: Fulvestrant With Enzalutamide 500mg of Fulvestrant will be given IM on days 1, 15, 28, then every 4 weeks as per standard of care (SOC) and 160mg of Enzalutamide will be given, in conjunction with Fulvestrant, PO daily. Fulvestrant with Enzalutamide: 500mg of Fulvestrant will be given IM on days 1, 15, 28, then every 4 weeks as per standard of care (SOC) and 160mg of Enzalutamide will be given PO daily. Patients will receive a tumor biopsy at the start of treatment and 4 weeks after the start of treatment, with an optional 3rd biopsy at the end of treatment.*

Hypothesis: *Patients in NCT02953860 receive more mg of Enzalutamide than Fulvestrant over the course of the study.*

Label: *Entailment*

In the example, we can see for the word "more" in the hypothesis the model needs to determine the relationship between two medicines *Enzalutamide* and *Fulvestrant*. And to that, the model needs to extract information from multiple sentences of the premise which requires strong numerical and numeric reasoning abilities.

3.2. Many SOTA NLI models fail to effectively surmount the word distribution shift from general domain corpora to biomedical corpora. For example,

Premise: *Palbociclib+Letrozole Australia Cohort Participants received Palbociclib orally once a day at 125 mg for 21 days followed by 7 days off treatment for each 28-day cycle. Participants received Letrozole orally at 2.5 mg once daily as continuous daily dosing schedule according to product labeling and in compliance with its local prescribing information*

Premise: *Single injection of SiennaXP in addition to comparator single dose of radioisotope (Technetium Tc99m Sulfur Colloid) and single dose of isosulfan blue dye.*

As seen in the above example, the sentences have words of medical corpora (underlined words) which is very different from general language corpora on which most models are trained. Hence, it is difficult to interpret the standard model.

3.3. The length of the premise of our dataset is much longer compared to the popular NLI datasets.

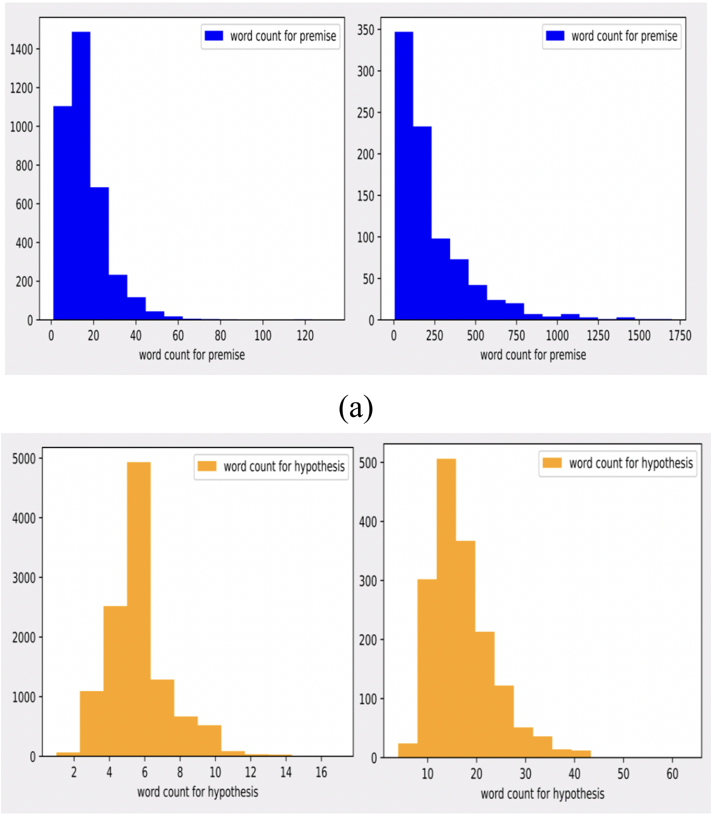
For the NLI task, the common datasets on which the task is evaluated are the Stanford Natural Language Inference (SNLI) Dataset, the Multi-Genre Natural Language Inference (MultiNLI) dataset, etc. Some examples of those datasets (premise) are given below. It is evident that the premise length is much larger than the standard NLI datasets.

Our Dataset: *Palbociclib+Letrozole Australia Cohort Participants received Palbociclib orally once a day at 125 mg for 21 days followed by 7 days off treatment for each 28-day cycle. Participants received Letrozole orally at 2.5 mg once daily as a continuous daily dosing schedule according to product labeling and in compliance with its local prescribing information.*

SNLI Dataset: *A woman with a green headscarf, blue shirt and a very big grin.*

Multi-NLI Dataset: *We sought to identify practices that were commonly implemented by the agencies within the past 5 years.*

3.4. The word count for the premise and the hypothesis are approximately 10 times larger than popular medical natural language inference datasets.



(a)

(b)

Fig 2. Comparison of sentence length between the MedNLI dataset and our dataset. The left graph represents the MedNLI dataset and the right graph represents our dataset. (a) The premise length in terms of word counts (b) The hypothesis length in terms of word count.

In Fig 2, we can see the word count is 10x times higher than MedNLI for both hypothesis and premise in our dataset.

3. Related Work [contributor Maklachur]

Natural language inference, commonly known as NLI, is one of the fundamental tasks associated with natural language comprehension. Its focus is on determining the existence of an inferential relationship, such as entailment or contradiction, between a given premise and a hypothesis. Due to the availability of large annotated datasets such as SNLI [4] and MultiNLI [14] Recently, researchers have developed a number of neural network-based models that are able to produce state-of-the-art performances and could be trained using these large annotated datasets.

In spite of these efforts, natural language understanding has made significant headway in fields such as literature and travel, but it has not yet been thoroughly investigated in the field of medicine. Researchers have been looking at the issue of clinical NLI ever since the release of MedNLI [5], which is an expert-annotated dataset for NLI in the clinical domain. One of the fundamental difficulties involved in obtaining natural language understanding is modeling informal inference, which is widely regarded as a very difficult endeavor. A sentence or word embedding method's potential for downstream applications in the medical domain can be evaluated with the assistance of the MedNLI dataset, which provides this assistance.

Recently, with the advent of powerful contextual word embedding approaches like ELMo [15] and BERT [6], the performances of many natural language processing tasks have improved, setting state-of-the-art results for many of these tasks. Following in the footsteps of this body of research, Lee et al. [7] provide BioBERT, which is a BERT model that has been pretrained on the English Wikipedia and BooksCorpus and then fine-tuned using the PubMed (7.8 billion tokens in total) corpus and PMC full-text articles. Jin et al. [8] offer BioELMo, which is a domain-specific version of ELMo trained on 10 million PubMed abstracts. They then attempt to address the medical NLI problem with these domain-specific embeddings, resulting in state-of-the-art performance. These two efforts point in the direction of a solution to the medical NLI challenge, in which the pre-trained embeddings are fine-tuned on the medical corpus and are used in the most advanced architecture for NLI. Chen et al. [9] proposed the utilization of external knowledge as a way to assist enrich neural-network-based NLI models. This was accomplished by applying components known as Knowledge-enriched co-attention, Local inference collection with External Knowledge, and Knowledge-enhanced inference composition.

Bringing in more subject information from outside sources, such as the Unified Medical Language System (UMLS), is an alternative method of solving the problem [16]. The knowledge-directed attention-based methodologies described in [9] were utilized for the Medical NLI research carried out in [8]. Another attempt of this kind was carried out by Lu et al., who combined domain knowledge in the form of the definitions of medical concepts taken from UMLS with the most recent version of the natural language understanding model known as ESIM [9], as well as the vanilla word embeddings of Glove [18], and fastText [17]. Even if the authors make a significant improvement by merely integrating concept definitions from UMLS, the properties of this clinical knowledge have not yet been completely leveraged to their potential.

4. Methodology [contributor all members]

4.1. Dataset Preprocessing [contributors Adrita, Aman]

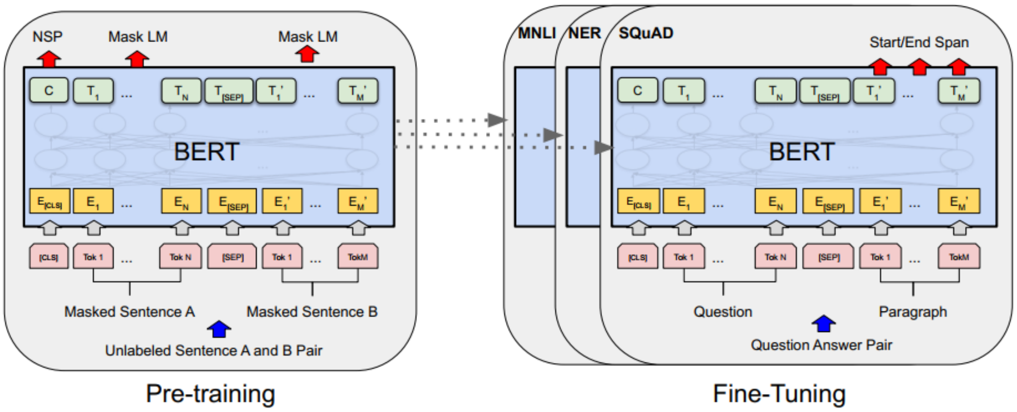
Each instance for the task contains 1-2 CTRs, a statement, a section marker, and an entailment/ contradiction label. Each CTR may have one to two cohorts or arms of patients. These groups might experience different therapies or have different baseline traits. Each CTR has four sections as shown in table 1.

The raw dataset contains 1350 separate CTRs and one file providing the necessary information on 1650 hypotheses and how they are connected to the CTRs with different attributes like primary_id, secondary_id, section_id, statement, etc. We retrieved premise-hypothesis premises by processing all the files. Our input data consists of sentence pairs, we concatenated those forms into a single sentence. Trial numbers have been added to the premise. We used [CLS], and [SEP] tokens to distinguish the hypothesis from the premise. Our experiments show the model performs better when input data is provided in this way.

4.2. BERT Based and RoBERTa-based models [contributors Aman, Adrita]

The pre-trained BERT Based language model [6] was employed. The deep bidirectional representations that BERT pre-trained on allow each token to pay attention to the context to its left and right. A significant improvement over the GPT transformer is the usage of bidirectional self-attention in the BERT transformer. The purpose of the masked language model used by BERT, which masks off randomly selected input ids, is to anticipate the input id of the masked word based on its context. This increases the robustness of the BERT language model. On seven NLP tasks, BERT has already outperformed the most recent results. As fine-tuning the pre-trained BERT models demonstrate greater performance for different NLP tasks in the literature, we used the same approach for our work. A sample pre-training and fine-tuning approach for BERT for QA tasks are shown in Fig 3 below.

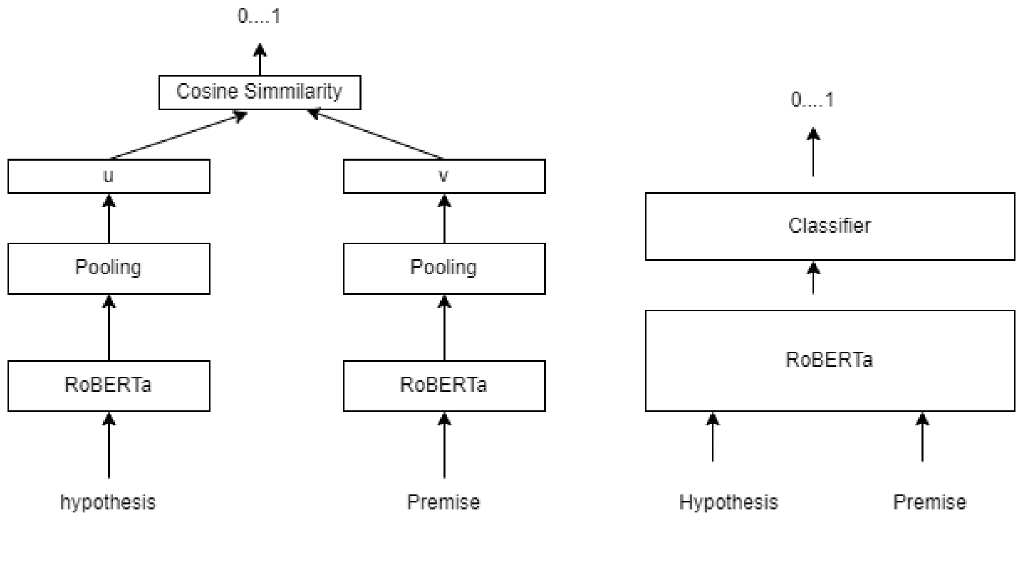
We initially tokenized the hypothesis and the premise pairs for our task. The [SEP] token was then placed in between the premise and the hypothesis. The language model can tell which is the hypothesis and which is the premise by using the separator token [SEP]. The [SEP] token was used for the same purpose in earlier studies that included question-and-answer tasks. For the same reason, we also changed the token type to 1 for all hypothesis tokens and the token type ids to 0 for all premise tokens. Finally, we put these combined tokenized representations into the pre-trained BERT model and added a classification layer at the end of the pre-trained BERT model. Finally, all parameters were jointly fine-tuned for our downstream classification task. We split our dataset into 80-10-10 (Train-Validation-Test). We achieved 49.09% accuracy and 49.2 F1 scores on our task using this approach.



(a) (b)

Fig 3: Pre-training task and Fine-tuning Task for BERT based model where the downstream task is QA task. The [SEP] token is used between pair-wise question-answers to let the model understand which tokens correspond to questions and which tokens correspond to answers (3(b)).

The RoBERTa language model [19] was pre-trained based on the replication study of the Bert based model but it provides the same or better performance compared to the BERT based models. In BERT-based language models, it replaces static masking with dynamic masking, more efficiently tweaks hyper-parameters (such as epsilon in the ADAM optimizer), and removes the next sentence prediction objective. We utilized the RoBERTa language model since it outperforms BERT in the SQuaD (The Stanford Question Answering Dataset), a challenge that is similar to ours in terms of pairwise semantic textual similarity. For pair-wise sentence regression problems (e.g., QA, similarity matching), the Cross-encoder (Nils & Iryna et al. [10]) obtains greater accuracy than the Bi-encoder. When using a typical bi-encoder model, we feed the language model pair-wise sentences separately to obtain sentence embeddings for both pair-wise sentences. We then use cosine similarity to determine how similar these pair-wise sentence embeddings are. However, in the cross-encoder, we pass the language model with both pair-wise sentences together. The individual sentence embedding cannot be obtained independently and is not necessary for our goal. Fig. 4 depicts the cross-encoder and bi-encoder architectures. So, we pass the pairwise hypothesis and premise together in the cross encoder based pre-trained RoBERTa model and we add a classification layer at the end of the language model. We finetune all the parameters of the model jointly. We used an 80-10-10 split (Train-Validation-Test) for our task and we got an accuracy of 52.4% and an F1 score of 52.5.



(a) (b)

Fig 4: (a) Bi-encoder based RoBERTa architecture vs (b) Cross-encoder based RoBERTa architecture

4.3. InferSent Model [contributors Adrita, Messal]

Romanov et al. experimented with different architectures and word embeddings for natural language inference in the clinical domain. [13]. They have found the best performance with glove_bio_asq_mimic embeddings and the Infersent model.

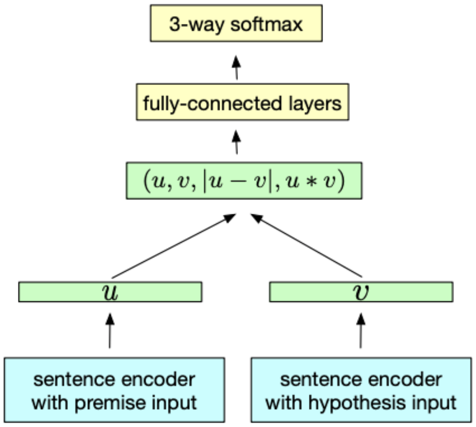

3-way softmax
fully-connected layers
$(u, v, |u - v|, u * v)$
u
v
sentence encoder with premise input
sentence encoder with hypothesis input

Fig 5: Training Scheme of Infersent Model [13]

This is a model for sentence representation where a bidirectional LSTM encoder of input sentences and a max-pooling operation over time steps are used to get a vector for the premise (p) and for the hypothesis (h). It basically used a complex scheme of interaction between the vectors premise (p) and hypothesis (h) to get a single vector z that contains all the information required to produce a decision about the relationship between the input sentences: z = [p, h, |p − h|, p ∗ h] as shown in Fig 5. [11]

The model has been trained on the MedNLI dataset and hence we expect it to have knowledge about the medical domain already. So we used our dataset as the test set to evaluate it. We achieved 48.5% accuracy on our dataset using this model.

4.4. PubMedBert [contributors Fardeen, Messal, Maklachur, Adrita]

The results from this work [22] showed that the domain-specific pretraining from scratch performed significantly better than the continuous pretraining of generic language models. They improved the model's ability to generalize to a greater variety of biomedical natural language tasks by generating this model by extending the pretraining of a BERT Base model over a collection of PubMed abstracts. PubMedBERT is pre-trained from the beginning in two distinct ways: (i) by using only abstracts from PubMed, and (ii) by utilizing both abstracts from PubMed and full-text articles retrieved from PubMedCentral. According to the results of the Biomedical Language Understanding and Reasoning Benchmark, this model is capable of performing at the highest possible level when it comes to a variety of biomedical NLP tasks.

We have tried to leverage this model which is already pre-trained on PubMed abstracts as it performs very well for some related tasks. The authors report 92.7 Pearson similarity score on a sentence similarity task on BIOSSES [20] dataset and 82.32 Micro F1 score on a document classification task on HoC [21] dataset. Although the authors do not report any results on natural language inference tasks, there already exists some work that finetunes the model on the MedNLI dataset for medical natural language inference. An 83.33% accuracy is reported for this work on MedNLI data. We have taken this version of the model from the hugging face and fine-tuned it for our dataset. We have used 80% of the available dataset and the rest of the data have been used for validation and testing. We obtain a 58.5% accuracy and 56.85 F1 score on our testing dataset. This is almost an improvement of 4 F1 points from the previously discussed BERT and RoBERTa based methods.

4.5. BioElectra [contributors Maklachur, Fardeen, Adrita, Messal]

In this research [23], the author proposed a language encoder model that is specialized to the biomedical sector and is adapted from ELECTRA for natural language processing tasks relevant to the biomedical field. BioELECTRA surpasses the earlier models and achieves state of the art (SOTA) on all of the 13 datasets in the BLURB benchmark and on all of the 4 Clinical datasets from the BLUE Benchmark for a total of 7 natural language processing tasks.

BioELECTRA that has been pre trained on and PMC complete text articles achieves excellent results when applied to Clinical datasets. On the MedNLI dataset, BioELECTRA obtains new

SOTA of 86.34% (a 1.39% accuracy improvement), while on the PubMedQA dataset, it achieves new SOTA of 64% (a 2.98% accuracy increase in performance).

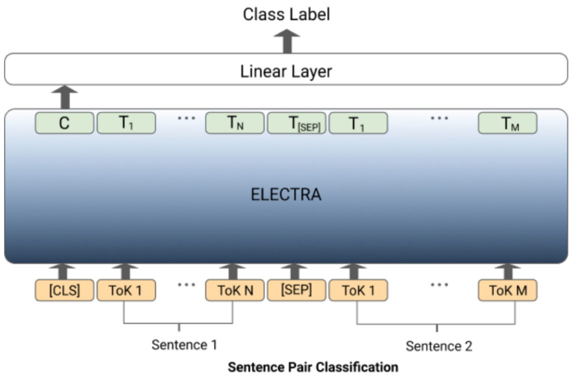


Fig 6: Fine Tuning BioELECTRA for natural language inference

We have used this model and fine tuned it on our dataset for predicting predictions or entailments of the hypotheses with the corresponding premises. Even though the original paper reports a 1.39% accuracy improvement over the BERT based model for MedNLI data, for our case both BioELECTRA and PubMedBert model results in similar performance. However, after infusing breast cancer knowledge into the model, PubMedBert (57.49 F1) performs better than BioELECTRA (59.25 F1).

4.6. DeBERTa [contributors Messal, Maklachur, Fardeen]

DeBERTa, which stands for "Decoding-enhanced BERT with disentangled attention," is a model that enhances both the BERT and RoBERTa models by utilizing two new methods [24]. The first mechanism is known as the disentangled attention mechanism. In this mechanism, each word is represented by two vectors, one of which encodes the word's content and the other its position. The attention weights between words are calculated by using disentangled matrices based on the contents and relative positions of the words. In the second step of the process, an improved mask decoder takes the place of the output softmax layer in order to make predictions about the masked tokens used in the model's pretraining. They demonstrate that utilizing these two strategies results in a significant improvement in both the effectiveness of model pre-training and the execution of downstream tasks. In addition to this, they apply a method called virtual adversarial training for fine-tuning, which helps to increase the generalization of the models. DeBERTaV3 which is an updated version of DeBERTa, improves the original model by replacing the pretraining objective, mask language modeling (MLM) with the replaced token detection (RTD) objective.

One of our main challenges, as explained previously, is the unusually lengthy premises and hypotheses in our dataset compared to a typical NLI dataset such as SNLI and MNLI dataset. To address this challenge we have used a pre-trained DeBERTaV3 model which is specifically fine tuned for NLI tasks on 8 different NLI corpora. One of the corpus used for this fine tuning was DocNLI [25]. This corpus also has long hypotheses (4-279 words) and long premises (20-842 words) like our dataset. We fine tuned this model on our dataset. This model results in a slightly better F1 compared to previous PubMedBert and BioELECTRA models.

4.7. PairSupCon [contributor Messal, Aman, Maklachur]

Previous works on semantic textual similarity (STS) tasks have been done by fine-tuning a language model with triplet loss or Siamese loss. However, it inherently assumes that sentences in contradiction pairs are from different semantic categories. This is not always true, especially for our particular task of project. PairSupCon [25] model overcomes this issue by encoding high level categorical concepts into the representations which improves the performance of the low-level semantic entailment and contradiction reasoning. It jointly optimizes the pairwise semantic reasoning objective with an instance discrimination loss. It selects the hard negative samples as those samples which are not entailment to the anchor but lie very close to the anchor in the representation space. They combine two losses: instance discrimination loss , entailment and contradiction reasoning loss The instance discrimination loss function induced by the model separates each entailment pairs from all other sentence pairs and also, it separates the hard negatives from the anchors. The entailment and contradiction reasoning loss discriminates between contradiction and entailment. By jointly optimizing these two losses, their model is both capable for entailment and contradiction reasoning and at the same time, they can learn high level categorical semantic representations.

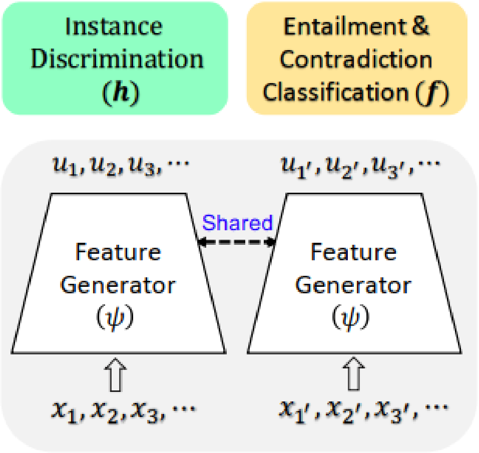

Instance Discrimination (h)
Entailment & Contradiction Classification (f)
$u_1, u_2, u_3, \cdots$
$u_{1'}, u_{2'}, u_{3'}, \cdots$
Shared
Feature Generator (ψ)
Feature Generator (ψ)
$x_1, x_2, x_3, \cdots$
$x_{1'}, x_{2'}, x_{3'}, \cdots$

Fig 7. Joint Optimization Framework for PairSupCon

For our particular task, contradiction and entailment pairs do not differ only in semantic representations, so the high level semantic concept induced by this model helps to improve the performance of our model. The architecture of this model is shown in Fig 5. We have fine tuned the PairSupCon model on our dataset for 80-10-10 split (Train-Validation-Test). This method results in a 63.86% accuracy and 61.54 F1 score, which is the highest.

4.8. Disease Knowledge Infusion [contributors Fardeen, Adrita, Aman, Messal]

Understanding a disease entails learning about its numerous facets, such as its signs and symptoms, diagnosis, and therapy. Many health-related and biological tasks, such as addressing consumer health questions, inferring from medical terminology, and recognizing diseases by name, depend on this illness's knowledge. While pre-trained language models like BERT have demonstrated success in extracting syntactic, semantic, and general knowledge from text, Yun He et al. [12] have demonstrated that these models can also be enhanced by specific knowledge such as understanding of symptoms, diagnoses, treatments, and other disease aspects. As a result, they combined BERT with knowledge of diseases to enhance these crucial duties. They outlined a novel method for disease knowledge infusion training and tested it using a variety of BERT models, including BERT, BioBERT, SciBERT, Clinical BERT, BlueBERT, and ALBERT. Experiments on three different tasks reveal that these models can almost always be improved, proving the efficacy of disease knowledge infusion. Inspired by these findings, we took a similar approach by infusing disease-related knowledge on PubMedBERT, BioElectra, DeBERTA and PairSupCon models. Following their approach, we have considered diseases associated with Breast Cancer listed in Medical Subject Headings(MeSH) (https://meshb.nlm.nih.gov/treeView) (under section C17.800.090) and extracted information on these diseases from Wikipedia as the authors have done. These data have been used to infuse knowledge in our models.

5. Results [contributor Aman]

From Table 1, we can see that InferSent and BERT-Based models perform worse than a random model. However, the RoBERTa-Based model provides better accuracy than the BERT-Based model which is expected as the RoBERTa-Based model was optimized with dynamic masking and better hyper-parameters to get better accuracy than the BERT-Based Model. The PubMedBERT and BioElectra models provide the same performance without knowledge infusion.

Table 1: Results on Test Set for Different Methods

	Without Infusion		**With Infusion**	
Model name	**Accuracy(%)**	**F1-score**	**Accuracy(%)**	**F1-score**
InferSent	48.50	48.05	-	-
BERT-Based Model	49.09	49.2	-	-
RoBERTa-Based Model	52.4	52.5	-	-
PubMedBERT Model	58.5	56.85	59.5 ↑	59.25 ↑
BioElectra Model	58.5	56.85	57.5 ↓	57.49 ↑
DeBERTa Model	58.2	57.5	57.5 ↓	56.50 ↓
PairSupCon	63.86	61.54	**64.24** ↑	**62.04** ↑

The arrow marks indicate the change of performance after knowledge infusion.

Knowledge infusion does not always improve the performance as we can see from Table 1. For PubMedBERT Model, including the knowledge infusion in the model, improves both Accuracy and F1- score but for DeBERTa model, including the knowledge infusion degrades the accuracy and F1-score. We get the best performance for the PairSupCon model with knowledge infusion (accuracy 64.24%, F1-score 62.04%) which is much better compared to the baseline performance reported for this dataset (32%)

6. Qualitative Analysis [contributor Messal, Aman]

The following section demonstrates two sets of examples and the predictions from the model for these examples:

Example 1:

Premise: Outcome Measurement: Progression-Free Survival (PFS). PFS was defined as the time from randomization until the first documented sign of disease progression or death due to any cause. Time frame: Baseline to disease progression or death due to any cause or 30 days after last dose (up to 216 weeks) Results 1: Arm/Group Title: Trastuzumab + Lapatinib. Arm/Group Description: Participants received Lapatinib 1000 milligram (mg) tablets orally daily 1 hour before or after breakfast along with Trastuzumab infusion at a loading dose of 4 milligrams/kilogram (mg/kg) body weight intravenously (IV) over 90 minutes on Day 1, followed by 2 mg/kg IV over 30 minutes weekly, in a 4 week cycle. Overall Number of Participants Analyzed: 146 Median (95% Confidence Interval) **Unit of Measure: weeks 12.0 (8.1 to 16.0)** Results 2: Arm/Group Title: Lapatinib Arm/Group Description: Participants received Lapatinib 1500 mg tablets orally daily 1 hour before or after breakfast. Overall Number of Participants Analyzed: 145 Median (95% Confidence Interval) **Unit of Measure: weeks 8.1 (7.6 to 9.0)** Outcome Measurement: Time to Progression as Evaluated by the Investigator Time to progression (TTP) is defined as the interval between the date of randomization and the earliest date of progression of disease (PD) or death due to breast cancer. The investigator assessed PD based on radiological PD (imaging data) and clinical symptomatic progress (Response Evaluation Criteria in Solid Tumors [RECIST] Criteria: target lesion (TL), at least a 20% increase in the sum of largest diameter (LD) of TLs or the appearance of one or more new lesions; non-TL (NTL), the appearance of one or more new lesions and/or unequivocal progression of existing NTLs). TTP was assessed in participants who died due to breast cancer or progressed, as assessed by the investigator, as well as in those who were censored and completed follow-up and those who were censored but are still being followed. For censored participants (those without a documented date of disease progression/death due to breast cancer), the date of the last radiographic assessment was used. Time frame: Randomization until the date of disease progression or death (average of 26 weeks) Results 1: Arm/Group Title: Lapatinib With Paclitaxel Arm/Group Description: Participants received lapatinib 1500 milligrams (mg) orally once daily (OD) with paclitaxel 175 mg/meters squared (m^2) intravenously (IV) over the course of 3 hours, every 3 weeks. The treatment group was stratified by sites of metastatic disease and stage of disease. Participants were treated until disease progression, unacceptable toxicity, or consent withdrawal. Overall Number of Participants Analyzed: 291 Median (Inter-Quartile Range) **Unit of Measure: weeks 29.0 (13.9 to 46.9)** Results 2: Arm/Group Title: Placebo With Paclitaxel Arm/Group Description: Participants received matching placebo orally OD with paclitaxel (175 mg/m^2 IV) over the course of 3 hours, every 3 weeks. The treatment group was stratified by sites of metastatic disease and stage of disease. Participants were treated until disease progression, unacceptable toxicity, or consent withdrawal. Overall Number of Participants Analyzed: 288 Median (Inter-Quartile Range) **Unit of Measure: weeks 22.9 (12.0 to 38.3)**

Actual Label: ***Entailment*** *| Predicted Label:* ***Entailment***

The median time from randomization until the first documented sign of disease progression or death due to any cause for all participants in NCT00320385, was lower than the median time for patients in NCT00075270.

Actual Label: ***Contradiction*** *| Predicted Label:* ***Contradiction***

Both cohorts in NCT00320385 outperformed cohort 1 of NCT00075270 in median PFS.

Here the premise is a combination of excerpts from 2 studies with IDs NCT00320385 and NCT00075270. The first hypothesis states that the median time for all participants in NCT00320385 (first half) was lower than that of NCT00075270 (second half). This is an entailment because for the first study, the median time will be somewhere close to 10 but for the second study the median would be somewhere close to 25. On the other hand the second hypothesis is a contradiction which is rightly predicted by the model. So we can infer from this example that the model can somewhat do mathematical and logical reasoning to predict entailments and contradictions.

Example 2:

Premise: Outcome Measurement: Mean Total MRI Functional Tumor Volume (FTV) Change From Baseline to Month 3 (V3) Mean total MRI FTV change from baseline to month 3 (V3): For patients with more than one measurable lesion on the MRI, the sum over all measurable lesions on the MRI was calculated at each time point. V3 was calculated by subtracting the total MRI FTV measured (i.e. the sum over all lesions present with MRI FTV measurements) at 3 months from the total MRI FTV measured at baseline. For V3 the raw change in the volume will be calculated for each patient and a mean and 95% confidence interval will be constructed using two-sided t-tests. Time frame: up to 3 months from start of treatment Results 1: Arm/Group Title: Letrozole + MRI Arm/Group Description: Protocol Therapy will consist of 6 months of letrozole, administered orally at a dose of 2.5 mg/day. Patients will have a bilateral MRI for disease evaluation at months 3 and 6. Overall Number of Participants Analyzed: 68 Mean (95% Confidence Interval) Unit of Measure: cubic centimeters -1.93 (-2.87 to -0.98)

Actual Label: ***Entailment*** **|** *Predicted Label:* ***Entailment***

One patient in NCT01439711 had a 2.87 cm3 decrease in Total MRI Functional Tumor Volume (FTV) over 3 months.

Actual Label: ***Contradiction*** *| Predicted Label:* ***Entailment***

One patient in NCT01439711 had a 0.98 cm3 increase in Total MRI Functional Tumor Volume (FTV) over 3 months, almost 1 cm3 less than average.

In this example, the model predicts entailment for both the hypotheses. In the second example, the total MRI FTV decreased in reality but the hypothesis states that it increased. So, it is apparent that the model could not accurately attend to the word increase in the hypothesis and failed to predict the correct label.

7. Future Work [contributor Adrita, Messal, Fardeen]

7.1. Training the tokenizer and model pre-training with that.

We have trained tokenizers on our corpus. Available tokenizers split the words into multiple tokens, and training the tokenizer tends to do better.

Sentence: NCT02953860 INTERVENTION 1: Fulvestrant With Enzalutamide ……
BERT Tokenizer Tokens: ['nc', '##t', '##0', '##29', '##53', '##86', '##0', 'intervention', '1', ':', 'fu', '##lves', '##tra','##nt', 'with', 'en', '##zal', '##uta', '##mide', …]
PubMedBERT Tokenizer Tokens: ['nct0', '##295', '##38', '##60', 'intervention', '1', ':', 'ful', '##ves', '##tran', '##t', 'with', 'enz', '##alu', '##ta', '##mid', '##e',..]
Trained Tokens: ['n', '##ct', '##02', '##95', '##3860', 'intervention', '1', ':', 'fulvestrant', 'with', 'enzalutamide', …]

As a future step, we plan to train the model with this tokenizer.

7.2. We have faced serious limitations of computational resources.

The training of our models has been done on TAMU HPRC GPUs and Google Colab. Often the waiting time was too long, especially for large models to be trained on better GPUs (DeBERTa in A100)

7.3. We want to merge DeBERTa and PairSupCon, this might enhance our best-obtained performance since it is supposed to process longer texts (premises of our dataset)

8. Conclusion [contributor Maklachur]

In this work, we have explored several BERT-based models to infuse medical knowledge, particularly for clinical trial data, and compare our performance in terms of accuracy and F1-score with the state-of-the-art (SOTA) models. The proposed approach outperformed the SOTA performance by a large margin, which achieved 64.24 % and 62.04 % accuracy and F1-score, respectively. We believe that our proposed approach to a specific task for NLI is to determine whether a natural language hypothesis can be justifiably inferred from a natural language premise and may have a broader application on medical knowledge infusion techniques in the future. We used a dataset of breast cancer CTRs3 that specialists had annotated in the relevant field to provide hypotheses, explanations, and labels. Clinical subject matter experts, clinical trial organizers, and research oncologists from the Cancer Research UK Manchester Institute annotated both the hypotheses and the data.

However, we have faced limitations of computational resources (we used TAMU HPRC GPUs and Google Colab, sometimes the waiting time was too long, especially for large models) since we infused knowledge on several BERT-based models and performed important with inevitable preprocessing for preparing our training datasets. In our future work, we plan to focus on resolving such issues and merging DeBERTa and PairSupCon; it might enhance the overall performance since it is supposed to process longer texts.

9. References

1. Jay DeYoung, Eric P. Lehman, Benjamin E. Nye, Iain James Marshall, and Byron C. Wallace. 2020. Evidence inference 2.0: More data, better models. ArXiv, abs/2005.04177.
2. Reed T Sutton, David Pincock, Daniel C Baumgart, Daniel C Sadowski, Richard N Fedorak, and Karen I Kroeker. 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. NPJ digital medicine, 3(1):1–10.
3. MacCartney, Bill. Natural language inference. Stanford University, 2009.
4. Bowman, Samuel R., et al. "A large annotated corpus for learning natural language inference." arXiv preprint arXiv:1508.05326 (2015).
5. Romanov, Alexey, and Chaitanya Shivade. "Lessons from natural language inference in the clinical domain." arXiv preprint arXiv:1808.06752 (2018),
6. Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).
7. Lee, Jinhyuk, et al. "BioBERT: a pre-trained biomedical language representation model for biomedical text mining." Bioinformatics 36.4 (2020): 1234-1240.
8. Jin, Qiao, et al. "Probing biomedical embeddings from language models." arXiv preprint arXiv:1904.02181 (2019).
9. Chen, Qian, et al. "Enhanced LSTM for natural language inference." arXiv preprint arXiv:1609.06038 (2016).
10. Nils, Iryna, et al. “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.” arXiv preprint arXiv:1908.10084 (2019)
11. Romanov, A., & Shivade, C. (2018). Lessons from Natural Language Inference in the Clinical Domain. arXiv preprint arXiv:1808.06752.
12. He, Yun, et al. "Infusing disease knowledge into BERT for health question answering, medical inference and disease name recognition." arXiv preprint arXiv:2010.03746 (2020).
13. Alexis Conneau, Douwe Kiela, Holger Schwenk, Loïc Barrault, and Antoine Bordes. 2017. Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 670–680, Copenhagen, Denmark. Association for Computational Linguistics.
14. Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.
15. Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., & Zettlemoyer, L. (2018). Deep contextualized word representations. arXiv.1802.05365

16. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004 Jan 1;32(Database issue):D267-70. doi: 10.1093/nar/gkh061. PMID: 14681409; PMCID: PMC308795.
17. Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics, 5:135–146.
18. Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.
19. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv. https://doi.org/10.48550/arXiv.1907.11692
20. Gizem Soğancıoğlu, Hakime Öztürk, Arzucan Özgür, BIOSSES: a semantic sentence similarity estimation system for the biomedical domain, Bioinformatics, Volume 33, Issue 14, 15 July 2017, Pages i49–i58, https://doi.org/10.1093/bioinformatics/btx238
21. Baker S, Silins I, Guo Y, Ali I, Högberg J, Stenius U, Korhonen A. Automatic semantic classification of scientific literature according to the hallmarks of cancer. Bioinformatics. 2016 Feb 1;32(3):432-40. doi: 10.1093/bioinformatics/btv585. Epub 2015 Oct 9. PMID: 26454282.
22. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., & Poon, H. (2020). Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. arXiv. https://doi.org/10.1145/3458754
23. Kamal raj Kanakarajan, Bhuvana Kundumani, and Malaikannan Sankarasubbu. 2021. BioELECTRA:Pretrained Biomedical text Encoder using Discriminators. In Proceedings of the 20th Workshop on Biomedical Language Processing, pages 143–154, Online. Association for Computational Linguistics.
24. He, P., Liu, X., Gao, J., & Chen, W. (2020). DeBERTa: Decoding-enhanced BERT with Disentangled Attention. arXiv. https://doi.org/10.48550/arXiv.2006.03654
25. Wenpeng Yin, Dragomir Radev, and Caiming Xiong. 2021. DocNLI: A Large-scale Dataset for Document-level Natural Language Inference. In Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 4913–4922, Online. Association for Computational Linguistics.
26. Dejiao Zhang, Shang-Wen Li, Wei Xiao, Henghui Zhu, Ramesh Nallapati, Andrew O. Arnold, and Bing Xiang. 2021. Pairwise Supervised Contrastive Learning of Sentence Representations. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 5786–5798, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.