# Bf3R at SemEval-2023 Task 7: a text similarity model for textual entailment and evidence retrieval in clinical trials and animal studies

### Mariana Neves

German Centre for the Protection of Laboratory Animals (Bf3R), German Federal Institute for Risk Assessment (BfR), Berlin, Germany

### Abstract

We describe our participation on the Multievidence Natural Language Inference for Clinical Trial Data (NLI4CT) of SemEval'23. The organizers provided a collection of clinical trials as training data and a set of statements, which can be related to either a single trial or to a comparison of two trials. The task consisted of two sub-tasks: (i) textual entailment (Task 1) for predicting whether the statement is supported (Entailment) or not (Contradiction) by the corresponding trial(s); and (ii) evidence retrieval (Task 2) for selecting the evidences (sentences in the trials) that support the decision made for Task 1. We built a model based on a sentence-based BERT similarity model which was pre-trained on ClinicalBERT embeddings. Our best results on the official test sets were f-scores of 0.64 and 0.67 for Tasks 1 and 2, respectively.

## 1 Introduction

The increase on the number of digital resources in biomedicine requires more adequate text mining methods to allow experts to stay up-to-date with the recent findings (DeYoung et al., 2020).

Task 7 of the SemEval'23 (Jullien et al., 2023), namely, Multi-evidence Natural Language Inference for Clinical Trial Data (NLI4CT)<sup>1</sup>, aimed at automatic processing clinical trials for two particular sub-tasks. In the scope of the shared task, the organizers released a collection of clinical trials, each of them composed of four sections, namely, "Intervention", "Eligibility", "Results", and "Adverse Events". Each of these sections contains one or more sentences.

The shared task was based on a set of the socalled "Premise-Statement-Evidence" items, which could be related to one or two clinical trials. If related to only one trial, the item was set as "Single", otherwise it is a comparison between two trials,

<sup>1</sup>https://sites.google.com/view/nli4ct/

i.e., set as "Comparison". Each Premise-Statement-Evidence item refers to only one particular section of the trial and contained a statement that made a claim about one or two trials, and which was the main input to be considered in the tasks. We present two examples in Table 1. Based on these items, two tasks were proposed: "Textual Entailment" (Task 1) and "Evidence retrieval" (Task 2).

The "Text Entailment" task consisted of automatically classifying the relation between the statement and the trial. Two values of labels were possible (cf. Table 1): "Entailment", i.e., the trial supports the claim, or "Contradiction", i.e., the trial contradicts the claim. Both values are possible for either single or comparison items. Only one of the labels is allowed for each item.

The "Evidence retrieval" task consisted of finding evidences that support the claim. The evidences can be one or more sentences from the corresponding section of the clinical trial, as stated in the Premise-Statement-Evidence item. In the case of a comparison, the evidences should come from both clinical trials. Examples of one evidence for the two tasks are shown in Table 1. For the second item (comparison), the evidence from trial 1 clearly shows that the cases of Enterocolitis were low (less than 1%), while the evidences from trial 2 do not cite the adverse effect, meaning that it did not occur at all.

We addressed the task as a similarity problem and trained a sentence BERT-based model on the data that was made available by the organizers. We participated on both tasks and our submissions are under the team name "marianaln". Further, we evaluated our trained models with animals studies (cf. Section 3.2), with similar tasks to the ones proposed for clinical trials in the shared task. We explain details of our methods in the next sections, followed by the results that we obtained in Section 3.

Туре	Section	Statement	Task 1	Task 2
Single	Results	"the primary trial does	Entailment	"Outcome Measurement:
		not report the PFS or ob-		Local Control Using Ip-
		jective response rate of		silateral Breast Tumor
		its patient cohort"		Recurrence Rates, Time
				frame: 2 years after treat-
				ment completion"
Comparison	Adverse Events	"a significant number of	Contradiction	"Enterocolitis 1/167
		the participants in the		(0.60%), Enterocolitis
		secondary trial and the		0/167 (0.00%)" (trial 1)
		primary trial suffered		and "Febrile neutropenia
		from Enterocolitis"		8/458 (1.75%)", Neu-
				tropenia 6/458 (1.31%),
				etc." (trial 2)

Table 1: Examples of the expected output for the two tasks (columns "Task 1" and "Task 2"), for statements which refer to one (Single) or two (Comparison) trials.

# 2 Methods

In this section we describe the details of our approach, including training based on a sentence BERT model, and the pros-processing for deriving the labels for Tasks 1 and 2.

**Sentence-BERT.** We approached the problem as a text similarity task and relied on Sentence-BERT (SBERT)<sup>2</sup> (Reimers and Gurevych, 2019). It is a modification of BERT that relies on siamese and triplet networks, thus being more suitable for text similarity tasks. Our classifier utilizes the ClinicalBERT embeddings<sup>3</sup> (Alsentzer et al., 2019), which is based on electronic health record from the MIMIC III database, and softmax as loss function. Due to time constraints, we only trained using two values of epochs: 1 and 2.

**Datasets.** We present the number of entries that is available for each trial type (Single or Comparison) and label (Entailment or Contradiction) in Table 2. We built one single classifier for all sections and types of statement (i.e., Single or Comparison), and for both tasks (i.e., Task 1 and Task 2). We trained the algorithm based on the available training data, by splitting the training file (train.json) into training (90%) and development (10%) datasets, while keeping the official development data (dev.json) for test purposes. Due to time constraints, we did not perform a 10-fold cross-validation.

<sup>2</sup>https://github.com/UKPLab/ sentence-transformers Pairs of sentences. SBERT receives pairs of sentences and the corresponding label: 0 (Contradiction), 1 (Entailment), and 2 (none). The latter consisted of any sentence not selected as evidence in the Premise-Statement-Evidence item. The pairs are composed of two items: (i) the statement and (ii) one sentence from the primary evidence or from the secondary evidences. Only one sentence from either of the two evidences are considered each time. Since the statements often refer to the trials using the expression such as "primary trial" and "secondary trial", we concatenate the text "primary trial" at the start of the sentences from the primary evidences, and "secondary trial" for sentences from the secondary evidences. SBERT performs label prediction, i.e., "contradiction", "entailment", or "none", for the each pairs, i.e., the statement and one of the sentence in the list of evidences. Based on the confidence score returned for each label, we derive predictions for each of the tasks (cf. below).

**Predictions for Task 1.** We obtain the label for Task 1 (hereafter called TASK1\_LABEL) based on the predictions and confidence scores for the "Contradiction" and "Entailment" labels, as returned by SBERT (cf. above). We simply compute an average of the confidence scores for each label for each sentence in the primary or secondary (if available) trial. The label with the highest score is chosen for Task 1.

**Predictions for Task 2.** We select the primary or secondary (if available) evidences based on the following information: the label which was de-

<sup>&</sup>lt;sup>3</sup>https://huggingface.co/emilyalsentzer/Bio\_ ClinicalBERT

Datasets	Single		Comparison		Total
	Entailment	Contradiction	Entailment	Contradiction	
Training	533	502	317	348	1,700
Development	70	70	30	30	200
Test	229		271		500

Table 2: Statistics for the training, development, and test sets. The labels for "Entailment" and "Contradiction" are still not available for the test set.

cided for Task 1 (TASK1\_LABEL, cf. above), and each evidence's confidence score. A evidence is automatically selected if the label with the highest confidence score is equal to TASK1\_LABEL. Otherwise, if the label with highest confidence score is "none", we compare the confidence scores for the "Contradiction" and "Entailment" labels. For the latter, we select the evidence only in the following situations: (a) the confidence score for the other label (not TASK1\_LABEL) is negative; or (b) both confidence scores are positive, but the score for TASK1\_LABEL is at least twice higher than the other one.

# **3** Results

We present the results that we obtained the test phase of the shared task, as well as for our additional experiments with documents about animal studies.

#### 3.1 Evaluation for NLI4CT

In Table 3, we summarize the results for both tasks, and for both the development and test sets. While training with two epochs slightly improved results for Task 2, these were were much worse for Task 1, for both the development and test sets. We analyzed the mistakes made by our system for the development dataset (cf. Table 4). As far as we know, the gold standard is still not available for the official test set.

For Task 1, the development dataset is well balanced and contains 100 items for each label, i.e., either Entailment or Contradiction. However, our system assigned more labels for Entailment than for Contradiction, namely, 189 vs. 11 for epochs=1 (ep1), and 123 vs. 77 for epochs=2 (ep2). Even though our predictions for ep2 are more balanced, less labels were correct (99 for ep2 vs. 101 for ep1).

For Task 2, the task was to select the sentences that were evidences to support the label of Task 1. We obtained a larger number of true positives (1727

Dev, epoch=1	Task 1	Task 2	
F-score	0.66	0.49	
Precision	0.50	0.41	
Recall	0.95	0.61	
Dev, epoch=2	Task 1	Task 2	
F-score	0.55	0.61	
Precision	0.50	0.46	
Recall	0.61	0.91	
Test, epoch=1	Task 1	Task 2	
Test, epoch=1 F-score	<b>Task 1</b> 0.64 (22)	<b>Task 2</b> 0.66	
Test, epoch=1 F-score Precision	Task 1   0.64 (22)   0.50 (34)	Task 2   0.66   0.58	
Test, epoch=1 F-score Precision Recall	Task 1   0.64 (22)   0.50 (34)   0.90 (6)	Task 2   0.66   0.58   0.76	
Test, epoch=1F-scorePrecisionRecallTest, epoch=2	Task 1   0.64 (22)   0.50 (34)   0.90 (6)   Task 1	Task 2   0.66   0.58   0.76   Task 2	
Test, epoch=1F-scorePrecisionRecallTest, epoch=2F-score	Task 1   0.64 (22)   0.50 (34)   0.90 (6)   Task 1   0.32	Task 2   0.66   0.58   0.76   Task 2   0.67 (22)	
Test, epoch=1F-scorePrecisionRecallTest, epoch=2F-scorePrecision	Task 1   0.64 (22)   0.50 (34)   0.90 (6)   Task 1   0.32   0.41	Task 2   0.66   0.58   0.76   Task 2   0.67 (22)   0.58 (20)	

Table 3: Results for the development and test sets as computed by the Codalab tool. The value in parenthesis is the position of the corresponding score in the leader-board.

vs. 1168) and less false negatives (179 vs. 738) when relying on epochs=2 (ep2). However, the number of false positives was much higher (2001 vs. 1678).

### 3.2 Evaluation on Animal Studies

We evaluated our trained model with text passages from the Animal Study Registry (ASR)<sup>4</sup> (Bert et al., 2019), a database of animal studies, which we host in our institute. This data has a certain similarity with the clinical trials of the NLI4CT task, since it describes procedures that were carried out in animal experiments.

We selected 31 studies which are already out of the embargo, i.e., whose complete text is available for the public. From the "Study Design" section of the study, we selected some passages where the authors describe the experiments that were carried out with the animals. Next, we manually split the text

<sup>&</sup>lt;sup>4</sup>https://www.animalstudyregistry.org

Task 1					
Preds.	Entail.		Co	ntrad.	
ep1	S	C	S	С	Total
TPs	65	30	5	1	101
Errors	5	0	65	29	99
ep2	S	С	S	С	Total
TPs	43	18	26	12	99
Errors	27	12	44	18	101
Task 2					
Preds.	primary		secondary		
ep1	S	C	S	С	Total
TPs	626	284	-	258	1168
FPs	1063	447	-	168	1678
FNs	408	122	-	208	738
ep2	S	С	S	С	Total
TPs	995	397	-	335	1727
FPs	1258	494	-	249	2001
FNs	39	9	-	131	179

Table 4: Overview of the true positives (TPs), false positives (FPs) and false negatives (FNs), for the development set, for Single (S) and Comparison (C) statements, and for values of 1 (ep1) and 2 (ep2) for the epochs.

into sentences, assigned a label of either "Entailment" or "Contradiction" to the study, and wrote a statement appropriate for the label. We started with the "Contradiction" label and alternated between one and the other for the studies. We only addressed Task 1, i.e., we did not identify which sentences support the label, even though the statement was always written based on just a couple of sentences. Further, we only considered one study when preparing the statement, i.e., no comparison between two studies. Table 5 presents one example of each label, derived from the studies asr.0000259<sup>5</sup> (Contradiction) and asr.0000221<sup>6</sup> (Entailment). We publish the data that we created only with the four studies whose license allows the redistribution of the text<sup>7</sup>.

We obtained a total of 16 "Contradiction" and 15 "Entailment" studies, and an average of sentences of 10.6 and 12.5, respectively. The number of sentences varied considerably among the studies, and ranged from only four to 29 sentences.

We obtained a total of 16 and 18 correct labels when relying on the models trained with one or two epochs, respectively. For some studies, we analyzed the confidence scores assigned to each label, in order to check whether sentences on which the statements were based actually obtained a higher score for that particular label. However, this was not the case in most of the times, and there is still much room for improvement for our system.

### 4 Conclusions

We presented our participation in the Clinical Trial Data (NLI4CT) of SemEval'23. We relied on a sentence-based BERT similarity model and utilized a pre-trained domain-specific language model, i.e., a ClinicalBERT embeddings. Our best results for Task 1 and 2 were f-scores of 0.64 and 0.67, respectively. Being Task 1 a classification task with only two labels, either "Entailment" or "Contradiction", our results cannot be considered very significant. However, Task 2 is a much harder task since it consisted of selecting the evidences that supported a certain decision, and we were able to obtain a higher score for this task. Our preliminary evaluation on animal studies shows that the trained models could potentially be used for this task as well.

### Limitations

Due to the time constraints of a shared task, we did not perform a 10-fold cross-validation during the training phase, nor a careful fine tuning of the parameters.

# **Ethics Statement**

Natural language processing aims at supporting experts to better keep up-to-date with the findings, but do not aim at substituting professionals. The clinical trials used in the datasets are publicly available data and do not include personal information about the participants of the study. The animal studies that we used in our additional evaluation are publicly available.

#### References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical BERT embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop, pages 72–78, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

<sup>&</sup>lt;sup>5</sup>https://www.animalstudyregistry.org/10.17590/ asr.0000259

<sup>&</sup>lt;sup>6</sup>https://www.animalstudyregistry.org/10.17590/ asr.0000221

<sup>&</sup>lt;sup>7</sup>https://github.com/mariananeves/nli4asr

Label	Statement	Sentence
Contradiction	The experiment aims to	Specifically, in the second and third group, a low and
	measure the glucose con-	high dose of mixture BPA, PBs (MePB, PrPB, BuPB),
	centration in the animals	TCS, DEHP and glyphosate will be administered,
		while in the fourth and fifth groups a high dose of
		pure and commercial glyphosate, respectively.
Entailment	The study evaluates the	Within our study, we evaluate and compare two anal-
	perception of pain in an-	gesic protocols, Tramadol (0.1 mg/ml) administered
	imals by experimenting	via the drinking water from one day pre-operatively
	with two analgesic drugs.	until 3 days post-operatively and a sustained-release
		depot Buprenorphine (1,2 mg/kg) administered via
		an intra-operative s.c. injection for their efficiency
		and possible side effects on experimental readouts in
		a mouse osteotomy model.

Table 5: Examples of a statement and the corresponding supporting sentence for two animal studies.

- Bettina Bert, Céline Heinl, Justyna Chmielewska, Franziska Schwarz, Barbara Grune, Andreas Hensel, Matthias Greiner, and Gilbert Schönfelder. 2019. Refining animal research: The animal study registry. *PLOS Biology*, 17(10):1–12.
- Jay DeYoung, Eric Lehman, Benjamin Nye, Iain Marshall, and Byron C. Wallace. 2020. Evidence inference 2.0: More data, better models. In *Proceedings* of the 19th SIGBioMed Workshop on Biomedical Language Processing, pages 123–132, Online. Association for Computational Linguistics.
- Mael Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Donal Landers, and André Freitas. 2023. Semeval-2023 task 7: Multi-evidence natural language inference for clinical trial data. In *Proceedings of the 17th International Workshop on Semantic Evaluation*.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.