SemEval 2024 Task 2: Safe Biomedical Natural Language Inference for Clinical Trials *Project BERT*

Raphael Baumann[1][2909063]

Chair of Natural Language Processing
Institute of Computer Science
Faculty of Mathematics and Computer Science
Julius-Maximilians University of Würzburg
Würzburg, Germany
raphael.baumann@stud-mail.uni-wuerzburg.de

Abstract. This paper focuses on the creation and utilization of a specialized dataset for enhancing natural language processing models' comprehension of Clinical Trial Reports (CTR). The resulting dataset, structured as Single-Type and Comparison-Type, serves as the foundation for training and development sets. We explored methodologies, utilizing BERT as the base model, and delves into strategies like Sentence Embedding Architectures, Adapter Tuning, and different Loss Functions. It systematically evaluates factors such as learning rates, sentence permutations, and dataset expansion strategies, to generalize a Model, handling, like a translation task, several Test Entailment Task at once.

Keywords: Natural Language Processing (NLP)· Large Language Models (LLMs) · medical · BERT · Transformer · Machine Learning · Artificial Intelligence

1 Introduction

In clinical studies, Clinical Trial Reports (CTRs) provide detailed information about patients treatments and reactions [6] [18]. However, the increasing volume of CTRs, coupled with a lack of suitable analytical tools, poses challenges for clinicians to deliver personalized, evidence-based care [28][31]. Despite improvements in CTR reliability, there is a notable gap in the ability to analyze and compare them effectively [30]. Natural language inference (NLI) models show promise in deducing logical relationships in texts, but their application in the medical domain presents unique challenges [31][13][26][28][30]. These include the need for domain-specific knowledge due to clinical jargon and the variability in scientific communication. Additionally, distinguishing between non-entailed subsequences further complicates NLI systems [6].

Task: Classification of the relation between CTR premises and a statement, as being Entailed or a Contradiction. Models are expected to predict whether each statement affirms an entailment or forms a contradiction given the associated section from the claimed CTRs.

Recent works are reviewed, highlighting shortcomings in existing systems. Various models are examined, employing strategies such as pipeline concatenation, joint representation, supervised contrastive learning, and role-based enhancement [31][13][30][33]. We are clarifying in Chapter 2 how the CTR Dataset for the SemEval Task 2 is built and how we use them in our Models and Strategies. Also, we are briefly looking into the ranking of several Models, mostly BERT derivatives. Methods, detailed in the Chapter 4, cover loss and metric functions, the use of BERT as the basis, Sentence Embedding architectures like SBERT, and adapter tuning. Experiments in constrained environments explore learning rate effects, token size challenges, and the impact of sentence permutation. Dataset expansion strategies and the effects of different sentence embedding architectures and adapter tuning on model performance are thoroughly evaluated. The experiments (see Chapter 5) collectively shed light on the complexities of clinical trial report analysis, emphasizing the need to address challenges in data processing, model architecture, and training strategies.

2 Datasets

A group of four domain experts, including clinical trial organizers from the Manchester Cancer Institute and the Digital Experimental Cancer Medicine Team (DECMT), participated in an annotation task to generate entailment and contradiction statements for Clinical Trial Reports (CTR). Annotators were tasked with generating non-trivial statements about the contents of primary and secondary trials, encouraging understanding and reasoning. Each CTR was divided into four sections representing facts, and evidence supporting the labeled

Table 1. Two Example Clinical Trial Report taken out of the Dataset to represent the two types Comparison and Single. Each Sample has a Statement, Label, the relevant Section and one or two Premises called Trial [7].

Type	Short	Comparison	Single
Label		Entailment	Contradiction
Statement	stm	The primary trial and the secondary trial both used MRI for their interventions.	More than 1/3 of patients in cohort 1 of the primary trial experienced an adverse event.
Section	sec	Intervention	Adverse Events
Primary Trial	p_1 p_2 ...	INTERVENTION 1: • Letrozole, Breast Enhancement, Safety • Single arm of healthy postmenopausal women to have two breast MRI (baseline and post-treatment). Letrozole of 12.5 mg/day is given for three successive days just prior to the second MRI. • ...	Adverse Events 1: • Total: 69/258 (26.74%) • Anaemia 3/258 (1.16%) • Febrile neutropenia 13/258 (5.04%) • Neutropenia 5/258 (1.94%) • ... Adverse Events 2: • Mitral valve incompetence 0/224 (0.00%) • Pericardial effusion 2/224 (0.89%) • Sinus tachycardia 1/224 (0.45%) • ...
Secondary Trial	s_1 s_2 ...	INTERVENTION 1: • Healthy Volunteers • Healthy women will be screened for Magnetic Resonance Imaging (MRI) contraindications, and then undergo contrast injection, and SWIFT acquisition. • Magnetic resonance imaging: Patients and healthy volunteers will be first screened for MRI contraindications. • ...	

Table 2. Section and Label Distribution of the Training and Validation Dataset, showing an almost equal distribution through the four Sections [7].

Dataset	Label	Section				Σ	Total
		Adverse Events	Eligibility	Intervention	Results		
train	Contradiction	238	241	196	175	850	1700
	Entailment	258	245	200	147	850	
valid	Contradiction	26	28	18	28	100	200
	Entailment	26	28	18	28	100	

statement was selected from these facts. In cases of negation, the full CTR section was provided as evidence. A negative rewriting strategy was employed to create contradictory statements, and evidence contradicting these statements was collected [19] [6]. Each Datapoint in the Training and Development Set have a statement, label being either Entangled or Contradiction, and one or two Prompts being one of four different CTR sections of Adverse Events, Eligibility, Intervention or Results (see Table 1). 60% of the Dataset are Single-Type, which then has only the Primary Prompt as relevant information and the other 40% are Comparison-Type has Primary and Secondary Prompt [6][18][19]. The SemEval Task 2 Dataset consists of 1700 Training and 200 Development samples (see Table 2), distributed equally between the two labels and almost for the 4 kinds of CTR sections. We use the development samples as Validation Dataset to evaluate the performance of the Models.

For our further proceeding we concatenate the parts of a dataset sample as followed in Equation 1. The Single-Type, where there is no Second Trial (see Table 1), $s_1, s_2, ...$ is empty, so the two last $[SEP]$ tokens are directly after another and for Comparison Type are all elements given, therefore the Sentence is exactly:

$$Sentence = [CLS]\ stm\ [SEP]\ sec\ [SEP]\ p_1, p_2, ...\ [SEP]\ s_1, s_2, ...\ [SEP] \qquad (1)$$

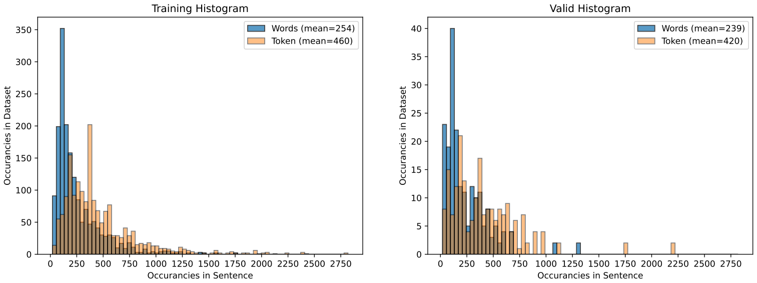


Fig. 1. Histogram of Number of Words from the full Sentence and the tokenized version with BertTokenizer, with 70% are in line of 512 Tokens for both.

As our base Model for our Experiments is BERT and the respective Tokenizer uses 512 token, to encode the Sentence, we have analyzed the length Distribution in a Histogram (see Figure 1). Round about 70% of all Training and Validation Datapoints are in the range of the 512 token length. Visible is also the slight shift on the x-Axis in the tokenized form, since not every word can have a single token representing it.

3 Recent Works

Table 3. Results of the SemEval-2023 Task 7.1 (SemEval-2024 Task 2) [6], with several attempts to use BERT as baseline.

Model/Method	Working Team	F1 Value
BERT	[31]KnowComp [30]Sebis [13]JUST-KM	base: 69.2 large: 70.9 base: 61:0 base: 63.4
DistilBERT	[28]Standford	60.8
BioBERT	[30]Sebis [28]Standford [15]YNU-HPCC	64.5 63.7 67.9
BioClinical-BERT	[31]KnowComp [30]Sebis [28]Standford	65.3 65.7 64.8
GatorTron-BERT	[12]Clemson NLP	70.5
PubMedBERT	[28]Standford	66.0
ALBERT-v2	[31]KnowComp	67.1
BART	[31]KnowComp	base: 67.1 large: 66.9
RoBERTa	[31]KnowComp [13]JUST-KM	base: 70.7 large: 67.6 base: 65.6 large: 66.1 role-based: 67.0
DeBERTa-v3	[31]KnowComp [30]Sebis	base: 75.8 large: 81.5 large: 80.5
ELECTRA	[31]KnowComp [28]Standford	base:70.3 large: 76.1 small: 63.9
GPT2	[31]KnowComp	base: 39.0 medium 44.2 large: 61.5
T5	[26]I2R	base: 62.9 large: 68.3
Flan-T5-xxl	[20]Saama	83.4
MGNet	[33]THiFLY	85.6

The majority of the released systems failed to achieve significantly above the majority-class baseline of 66.7% F1 value (see Table 3) [19]. **Sebis [30]** uses a system concatenating the parts of the CTRs in two different method, called pipeline and joint, where basically the sentence representation includes more $[SEP]$ tokens to dense it up. **KnowComp [31], YNU-HPCC [15], Stanford [28], I2R [26] and Clemson NLP [12]** uses the same strategy as we do (see Equation 1) to feed forward the sentence in several most popular Models. Compared to the others, **YNU-HPCC [15]** utilize Supervised Contrastive Learning with the corresponding loss function, to maximize, if it is contradiction, or minimize, if it is entailed, the spatial representation vector of the two compared inputs. **JUST-KM [13]** Models are enhancing RoBERTa in a role-based approach, where the two RoBERTa-Large Models trained differently, to predict the general outcome. **Saama [20]** finetuned Flan-T5 LLM to SemEval's task-specific data by applying different Instruction Templates. **THiFLY [33]** employ Multigranularity Inference Network, which uses the Equation 1 sentence structure to pass it further to a Joint Semantic Encoder followed by Pooling and Sencence-Level Encoder before Classification.

4 Methods

In this chapter we are clarifying the Loss and Metric Functions, the Architecture BERT used for our Experiments, Ideas principles like Adapter Tuning and Sentence Embedding and how fusing different Dataset together works, which are used in training loop.

4.1 BERT

The Experiments are using BERT, Bidirectional Encoder Representations from Transformers (see Figure 2) [14] developed by Google is based on Transformer architecture introduced by Vaswani et al. [29]. Unlike previous attempts, that process text in a unidirectional way (either left to right or right to left), BERT is designed to understand context bidirectionally as every Token is connected Pathways with every other. A masked language model (MLM) pre-training target is used, where tokens are randomly masked from the input to predict the

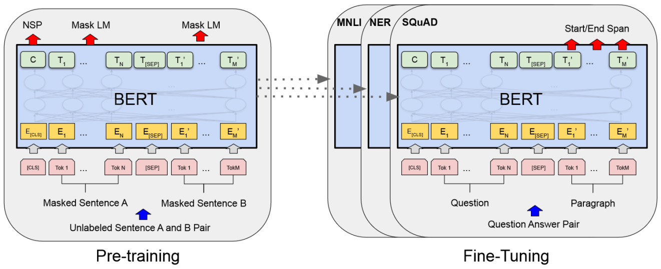


Fig. 2. Bidirectional Encoder Representations from Transformers (BERT) schematic View [14].

original vocabulary IDs. The model can be fine-tuned for specific downstream tasks, such as classification or translation. BERT is available in different sizes like BERT-Base and BERT-Large. There are various implementations such as RoBERTa [23], ALBERT [21], BART [22], DeBERTa [17], which improves BERT architecture differently.

We are using pre-trained BERT-Base Model from Huggingface [11] due to Limitations of the Environment (see Chapter 5.1) and supervised fine-tune it to SemEval.

4.2 Sentence Embedding Architectures

Sentence-BERT (SBERT) [27], a modification of the BERT [14] network, is a strategy designed for semantic similarity tasks, to generate meaningful embeddings. While BERT [14] and RoBERTa [23] excel in sentence-pair regression tasks, they suffer from computational overhead when dealing with large collections of sentences. SBERT addresses this issue by using siamese and triplet network structures to generate semantically meaningful sentence embeddings. Also, a key point is, especially with BERT, the two sentences are passing the Model individual, meaning that the Tokens of each sentence does not interfere with each other. This allows for efficient similarity comparisons, clustering, and semantic search. The authors demonstrate that SBERT significantly reduces the time required for finding the most similar pair in a collection of 10,000 sentences from 65 hours with BERT to about 5 seconds.

We use this Principle and applied it to the SemEval Task (see Figure 3). Therefore, we come up with five different ideas, where they use the Siamese-Architecture as Core, which is sharing the Models Parameters. Version-2 (v2)

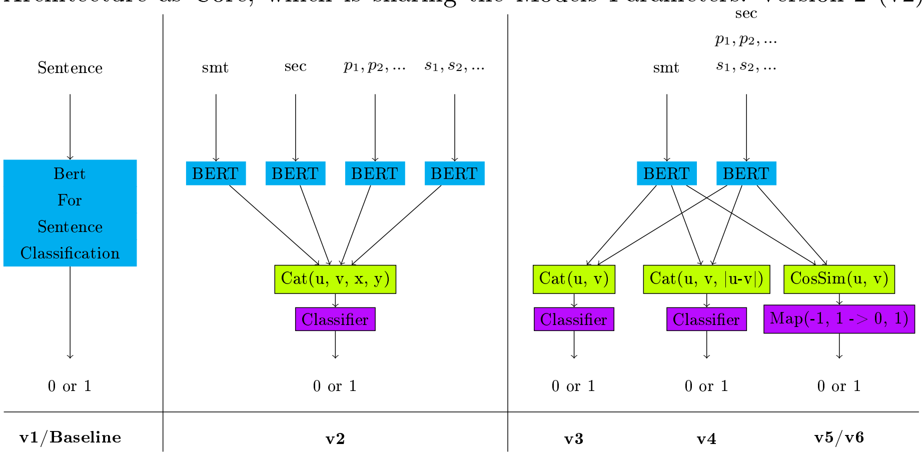


Fig. 3. Sentence Embedding idea is based on BERT Model. Strategy v1 is the Baseline, handling everything in once and every token can contribute to each other, while Strategy v2 split the sentence in four parts, and fed separately to the same model. Strategy v3, v4, v5 and v6 uses the same idea but Concatenating, using only premise and hypothesis and generating the output differently.

has four parts of the defined Sentence (see Equation 1) split at the $[SEP]$ Tokens. Version-3 (v3), Version-4 (v4), Version-5 (v5) and Version-6 (v6) has the split on the first $[SEP]$ Token, to separate Statement from the Hypothesis. Version-3 and Version-4 uses a simple Feed-Forward-Classifier after the Concatenation combined with CrossEntropyLoss. Version-5 and Version-6 uses Cosine Similarity, with a Mapping ensuring Entailment being 1 and Contradiction being -1. Version-5 uses CosineEmbeddingLoss and Version-6 uses MSELoss to minimize, if entailed, and maximize, if contradiction, the spatial distances.

4.3 Adapter Tuning

Adapter Tuning (see Figure 4) is a supervised method, where input, gold label are given and the models parameters are frozen, but adding new fully trainable bottleneck feed-forward networks on each intermediate layer. The objective is to reduce the size of trainable parameters, to gain higher throughput and keeping the pre-trained embeddings[32] [25]. The ultimate goal of adaptation training is to enhance the model's scores on the downstream task, while still benefiting from the broad language understanding gained during the initial pre-training [24]. The effectiveness of adaptation-tuning depends on the similarity between the pre-training task and the target task due to fixed embeddings.

For our Experiments we are comparing supervised fine-tuned BERT, with the SemEval dataset against a Adapter-BERT version, where the pre-trained BERT's Parameters are frozen. Secondly we are taking the BERT Model from SimCSE [16], which applied a Supervised Contrastive Learning idea on SNLI Data, to minimize and maximize the spatial distances, which effects the Output Representations of BERT and using the same principle with the Adapter on this Model.

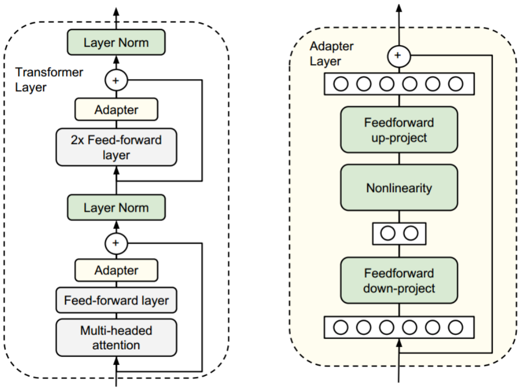


Fig. 4. Basic Structure of Adapter built on top of a Models Architecture (left side), where only the Adapter Layers Parameters (right side) are trainable [32]. In our Case we are using 768 to 512 DOWN-Projector, followed by GELU-Activation and 512 to 768 UP-Projector.

4.4 Loss Functions and Evaluation Metric

We are using 3 commonly used Loss functions for our Training of the different Architectures with x being the Prediction and y being the searched Class. For direct Classification, CrossEntropyLoss 2 is the commonly and most frequently used function. Secondly, as we see later, combined with Cosine Simmilarity, we are using MSELoss 3 or CosineEmbeddingLoss 4 to maximize or minimize the distance between two representations. As Metric the SemEval Task uses F1 Score, which is the harmonic mean between precision and recall defined in Equation 5 [6].

$$Loss_{CE} = -\sum_{i=1}^{M} y_{o,i} \log(x_{o,i}) \tag{2}$$

$$Loss_{MSE} = \sum_{i=1}^{M} (x_i - y_i)^2 \tag{3}$$

$$Loss_{CEB} = \begin{cases} 1 - \cos(x_1, x_2), & \text{if } y = +1 \\ \max(0, \cos(x_1, x_2)), & \text{if } y = -1 \end{cases} \tag{4}$$

$$F_1 = 2\frac{precision \cdot recall}{precision + recall} = \frac{2TP}{2TP + FP + FN} \tag{5}$$

4.5 Fusing different Datasets/Loaders

Normally a Pre-Trained Model is used which is then fine-tuned on the Specific Dataset. The same can be applied to Text Classification. Therefore, the Model is performing only on SemEval very good, while mismatching on generalization on other Text Classification Tasks. To Compensate, two strategies (see Figure 5) are applied to solve this. CombinedLoaders is a strategy, where each Dataset, on which the Model should perform, can contribute equally with the same amount of Datapoints. Another strategy involves the concatenation of all these Datasets to one and then being mixed on the Training. For our Example (see Chapter 5.4 and Figure 6) we are using SNLI [10] which has short sentences for entailment

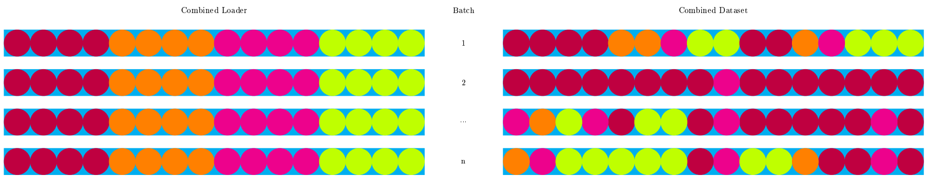
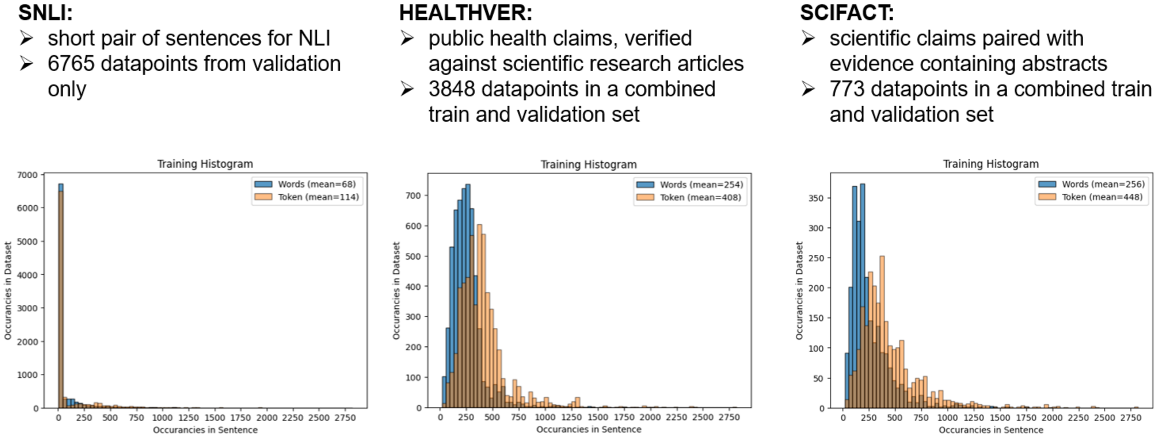


Fig. 5. Combined Dataset and Combined Loader Strategy where 4 different Datasets are getting mixed together. On the Loader-Level for a Batch size of 16, every Dataset is present equally and on the Dataset-Level, they are concatenated and the mixed.



SNLI:

- short pair of sentences for NLI
- 6765 datapoints from validation only

HEALTHVER:

- public health claims, verified against scientific research articles
- 3848 datapoints in a combined train and validation set

SCIFACT:

- scientific claims paired with evidence containing abstracts
- 773 datapoints in a combined train and validation set

Fig. 6. Histogram of how different Datasets combined with SemEval changes the Occurrences in the Bins. HEALTHVER and SCIFACT shows a slight shift on the x-Axis towards right, indicating that the length of the sentences is higher. On the other side SNLI, is dominating, because the sentence structure is rather small.

check, HEALTHVER [2] a Dataset on the medical domain and SCIFACT [1] on the scientific domain, to check the entailment of claim, paper title and abstract. The Evaluation of the Strategies is only applied to SemEval Task Development Dataset not on the chosen ones for expanding. Also, the expected outcome should be lower than the specialized Model on SemEval as it has to generalize the different domains.

5 Experiments

5.1 Environmental Setup

For our Experiments we are bounded by the Environments given from Kaggle [4] and Colab [3], which makes quite complex to find the optimal hyperparameter. Also, there is no possibility to Cache or Precalculate all the Datapoints of Training and Validation without running in to RAM issues. We decided to shift the bottlenecks either in parallel loading the Data and/or running in exceeding GPU RAM due to the amount of parameters.

Colab's Environment:

- CPU: Intel(R) Xeon(R) @ 2.00GHz
- Number of available Cores: 2
- System Ram: 12 GB
- GPU: Nvidia Tesla T4
- GPU Ram: 15 GB
- Time limit: 3-4 Hours a Day/Session

Kaggle's Environment:

- CPU: Intel(R) Xeon(R) @ 2.00GHz
- Number of available Cores: 4
- System Ram: 32 GB
- GPU: Nvidia Tesla P100
- GPU Ram: 16 GB
- Time limit: 30 Hour a Week with 9h each Session

The general handling of the loops is based on Pytorch [8], Lightning AI [5] implementation and for loading the pre-trained BERT Model we are using Huggingface (transformer library) [9]. Therefore, we can focus the Experiments on implementing strategies to enhance BERT's overall performance tested on four different Seeds.

5.2 Learning Rate

In the evaluation of the Baseline BertModelForSentenceClassification, different Learning-Rate were tested, accompanied by variations in token size and the implementation of mixed precision. The tested Range of Learning-Rate (see Table 4) going from $3e-6$ to $7e-6$. All above or below leading that the Model does not learn well or resulting in a rectangle graph, indicating that the gradients are too high or low. The best Learning Rate, based on the Seeds is $6e-6$ with 0.640 ± 0.021.

However, it was observed that altering the token size and introducing mixed precision led to the destruction of tensors within the model leading to re pre-train it with the given Environments this is impossible. This suggests potential challenges and limitations associated with modifying these parameters, emphasizing the need for careful consideration and experimentation when adjusting learning rates and employing mixed precision in the context of sentence classification tasks using the BERT model.

Table 4. F1 Values of the Baseline BertModelForSentenceClassification of the Last (50th) Epoch for different Learning-Rates.

Learning Rate	Seed				mean ± std
	0	42	1998	1M	
3e-6	0.637	0.554	0.623	0.652	0.614 ± 0.034
4e-6	0.589	0.614	0.649	0.615	0.616 ± 0.019
5e-6	0.636	0.640	0.638	0.657	0.630 ± 0.026
6e-6	0.661	0.654	0.602	0.647	**0.640 ± 0.021**
7e-6	0.644	0.563	0.657	0.655	0.621 ± 0.040

5.3 Permutation

$$Sentence = [CLS]\ stm\ [SEP]\ sec\ [SEP]\ \pi(p_1), \pi(p_2), ...\ [SEP]\ \pi(s_1), \pi(s_2), ...\ [SEP] \tag{6}$$

As described above changing Token size does lead to destroying the embeddings due to newly initializing and the fact that the maximum token size is 512, we come up with the idea of permuting the Primary and Secondary Trials (see Equation 6), which can be seen as list of items. The Target to analyze is if the order of the elements is relevant for deciding the Prediction, wich should not be relevant due to BERT's Architecture, and if information after the cutpoint has interesting and relevant facts, which influences the decision making process.

Table 5. F1 Values of the Permutation trial on the Baseline Model from the last (50th) Epoch

Learning Rate	Permutation	Seed				mean ± std
		0	42	1998	1M	
5e-6	No	0.636	0.640	0.638	0.657	0.630 ± 0.026
	Yes	0.633	0.652	0.620	0.643	**0.643 ± 0.016**

With no permutation the Sentence is only 512 tokens long with first-come-first-serve principle. Permutation on indicates that the order is not significant.

The Results (see Table 5) showed that there can be a slight Performance boost and the standard deviation is lower indicating that the fluctuation of different seeds in not that harsh. But, the Runtime cost of permuting the sentences every Batch is for the later Experiments too much. Therefore, we skip it to stay cost-efficient.

5.4 Dataset Expansion

We have tested the two Dataset Expansion Strategies CombinedLoader (CombDL) and CombinedDataset (CombDS) on two phases with the expansion Datasets SNLI, HEALTHVER and SCIFACT. The first Stage is combining one Dataset with SemEval Training and testing the performance only on SemEval Validation. The second Stage including two datasets from the list and doing the same Training and Validation as before.

Table 6. F1 Values of the Dataset Expansion Test of Last (50th) Epoch for the different Strategies and Datasets, which expanding SemEval. The Values are the SemEval only Validation Scores.

Dataset + SemEval	Strategy		Seed				mean ± std
	CombDL	CombDS	0	42	1998	1M	
Baseline			0.636	0.640	0.638	0.657	0.630 ± 0.026
SNLI	✓		0.597	0.613	0.685	0.573	0.617 ± 0.042
		✓	0.619	0.669	0.648	0.634	0.643 ± 0.019
HEALTHVER	✓		0.634	0.614	0.622	0.610	0.620 ± 0.009
		✓	0.637	0.661	0.613	0.631	0.635 ± 0.017
SCIFACT	✓		0.586	0.547	0.603	0.561	0.574 ± 0.022
		✓	0.687	0.619	0.664	0.645	**0.654 ± 0.025**
SCIFACT,HEALTHVER	✓		0.580	0.564	0.551	0.570	0.566 ± 0.010
		✓	0.664	0.637	0.657	0.649	**0.652 ± 0.010**
SCIFACT,SNLI	✓		0.645	0.607	0.545	0.567	0.591 ± 0.039
		✓	0.658	0.574	0.624	0.644	0.625 ± 0.032
HEALTHVER,SNLI	✓		0.615	0.537	0.615	0.658	0.606 ± 0.044
		✓	0.667	0.611	0.664	0.602	0.636 ± 0.030

The Results (see Table 6) showed that the CombinedLoader Strategy is worse than the Baseline, which indicates that the model has generalization issues if the dataset is present with the same amount of Datapoints in one Batch. Also, CombinedDataset Strategy is almost every time better than the Baseline Model, with around 2% Points gain of F1 Value, showing that a generalized Model, handling multiple Text Entailment subtasks on different domain is possible. Secondly it is visible that SCIFACT in combination with SemEval has the highest performance gain, which is the result of being tokenized around the same length of SemEval, therefore the Model has to abstract more the general meaning of the sentences and the Dataset is not dominant, meaning that the size is equally powerful.

5.5 Sentence Embedding Architectures

As Results (see Table 7) of the different Architectures (defined in Chapter 4.2) most of them are equally powerful than the Baseline. Version-3 is 1% Point less than Version-4, which is expected due to we are giving more information to the Feed-Forward-Layer (FNN). Version-2 is better than the Baseline but 4 times expensive in the backward calculations, which makes the performance boost neglectable. Version-6, which is the Siamese-Network combined with Cosine Similarity of the two embeddings of the sentences and MSELoss as Loss function, has a gain of 3.4% F1-Value score, showing that Sentence Embedding can help to improve the Metric-Scores on the Text Entailment Task.

Table 7. F1 Values of the different Architecture from the Last (50th) Epoch, indicating that Sentence Embedding can boost the overall Performance.

Model	Desciption	Seed				mean ± std
		0	42	1998	1M	
Baseline	BertModelForSequenceClassification and CrossEntropyLoss	0.636	0.640	0.638	0.657	0.630 ± 0.026
v2	(u, v, x, y) as input to FFN and CrossEntropyLoss	0.652	0.619	0.604	0.664	0.635 ± 0.024
v3	(u, v) as input to FNN and CrossEntropyLoss	0.708	0.583	0.573	0.611	0.619 ± 0.054
v4	(u, v, |u-v|) as input to FNN and CrossEntropyLoss	0.638	0.631	0.618	0.640	0.632 ± 0.009
v5	CosSim(u, v) and CosineEmbeddingLoss	0.614	0.603	0.637	0.664	0.630 ± 0.023
v6	CosSim(u, v) and MSELoss	0.667	0.675	0.667	0.646	**0.664 ± 0.011**

Table 8. F1 Values of the Adapter Trial from the Last (50th) Epoch for the general BERT Model and the supervised SimCSE trained BERT

Model	Seed				mean ± std
	0	42	1998	1M	
Baseline BERT	0.636	0.640	0.638	0.657	0.630 ± 0.026
Adapter BERT	0.619	0.587	0.602	0.626	0.608 ± 0.015
Baseline SimCSE	0.598	0.667	0.696	0.640	**0.650 ± 0.036**
Adapter SimCSE	0.643	0.673	0.615	0.654	0.646 ± 0.021

5.6 Adapter

As said in Chapter 4.3, we are testing this Principle on SimCSE objective trained BERT Model and on general pre-trained BERT overridden the Bert-ModelForSentenceClassification Class. Firstly the Adapter version of the general pre-trained BERT is around 2% Points worse, which indicates that the remaining trainable Parameters are not enough to get a meaningful representation for the Classification-Layer (see Table 8). On the opposite side, taking the supervised fine-tuned SimCSE Model, which is trained on SNLI data, also a Text Entailment Task, shows a Performance boos from 1.5-2% in F1 Score. Therefore, is important that the task correlate to each other, so the normal Layer of the BERT already have a good representation saved in the Tensor and the Adapter Layer only has to align these.

6 Conclution

In conclusion, this paper presents a thorough investigation into text entailment classification, focusing on Clinical Trial Reports (CTR) annotation. Through careful dataset curation and experimentation using the SemEval Task 2 Dataset, the study explores various approaches, including BERT-based models and adapter tuning. Despite challenges in environmental setup, the findings highlight the importance of learning rate optimization, dataset expansion strategies, and sentence embedding architectures in influencing model performance. The paper contributes valuable insights to natural language processing research, particularly in clinical text analysis, and sets the stage for further exploration in related domains.

References

1. allenai/scifact_entailment · Datasets at Hugging Face, https://huggingface.co/datasets/allenai/scifact_entailment
2. dwadden/healthver_entailment · Datasets at Hugging Face, https://huggingface.co/datasets/dwadden/healthver_entailment
3. Google Colaboratory, https://colab.research.google.com/
4. Kaggle: Your Home for Data Science, https://www.kaggle.com/

5. Lightning AI | Turn ideas into AI, Lightning fast, https://lightning.ai/
6. NLI4CT, https://sites.google.com/view/nli4ct/semeval-2024
7. NLI4CT, https://sites.google.com/view/nli4ct/semeval-2024
8. PyTorch, https://pytorch.org/
9. Hugging Face – The AI community building the future. (Nov 2023), https://huggingface.co/datasets
10. snli · Datasets at Hugging Face (Nov 2023), https://huggingface.co/datasets/snli
11. Hugging Face – The AI community building the future. (Feb 2024), https://huggingface.co/
12. Alameldin, A., Williamson, A.: Clemson NLP at SemEval-2023 Task 7: Applying GatorTron to Multi-Evidence Clinical NLI
13. Alissa, K., Abdullah, M.: JUST-KM at SemEval-2023 Task 7: Multi-evidence Natural Language Inference using Role-based Double Roberta-Large. In: Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023). pp. 447–452. Association for Computational Linguistics, Toronto, Canada (2023). https://doi.org/10.18653/v1/2023.semeval-1.61, https://aclanthology.org/2023.semeval-1.61
14. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding (May 2019), http://arxiv.org/abs/1810.04805, arXiv:1810.04805 [cs]
15. Feng, C., Wang, J., Zhang, X.: YNU-HPCC at SemEval-2023 Task7: Multi-evidence Natural Language Inference for Clinical Trial Data based on a BioBERT Model
16. Gao, T., Yao, X., Chen, D.: SimCSE: Simple Contrastive Learning of Sentence Embeddings (May 2022), http://arxiv.org/abs/2104.08821, arXiv:2104.08821 [cs]
17. He, P., Liu, X., Gao, J., Chen, W.: DEBERTA: DECODING-ENHANCED BERT WITH DIS- ENTANGLED ATTENTION (2021)
18. Jullien, M., Valentino, M., Frost, H., O'Regan, P., Landers, D., Freitas, A.: NLI4CT: Multi-Evidence Natural Language Inference for Clinical Trial Reports (Oct 2023), http://arxiv.org/abs/2305.03598, arXiv:2305.03598 [cs]
19. Jullien, M., Valentino, M., Frost, H., O'Regan, P., Landers, D., Freitas, A.: SemEval-2023 Task 7: Multi-Evidence Natural Language Inference for Clinical Trial Data
20. Kanakarajan, K.R., Sankarasubbu, M.: Saama AI Research at SemEval-2023 Task 7: Exploring the Capabilities of Flan-T5 for Multi-evidence Natural Language Inference in Clinical Trial Data. In: Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023). pp. 995–1003. Association for Computational Linguistics, Toronto, Canada (2023). https://doi.org/10.18653/v1/2023.semeval-1.137, https://aclanthology.org/2023.semeval-1.137
21. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: ALBERT: A LITE BERT FOR SELF-SUPERVISED LEARNING OF LANGUAGE REPRESENTATIONS (2020)
22. Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., Zettlemoyer, L.: BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 7871–7880. Association for Computational Linguistics, Online (2020). https://doi.org/10.18653/v1/2020.acl-main.703, https://www.aclweb.org/anthology/2020.acl-main.703

23. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: RoBERTa: A Robustly Optimized BERT Pretraining Approach (Jul 2019), http://arxiv.org/abs/1907.11692, arXiv:1907.11692 [cs]
24. Manjavacas, E., Fonteyn, L.: Adapting vs. Pre-training Language Models for Historical Languages. Journal of Data Mining & Digital Humanities **NLP4DH**(Digital humanities in...), 9152 (Jun 2022). https://doi.org/10.46298/jdmdh.9152, https://jdmdh.episciences.org/9152
25. Naveed, H., Khan, A.U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., Mian, A.: A Comprehensive Overview of Large Language Models (Oct 2023), http://arxiv.org/abs/2307.06435, arXiv:2307.06435 [cs]
26. Rajamanickam, S., Rajaraman, K.: I2R at SemEval-2023 Task 7: Explanations-driven Ensemble Approach for Natural Language Inference over Clinical Trial Data. In: Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023). pp. 1630–1635. Association for Computational Linguistics, Toronto, Canada (2023). https://doi.org/10.18653/v1/2023.semeval-1.226, https://aclanthology.org/2023.semeval-1.226
27. Reimers, N., Gurevych, I.: Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks (Aug 2019), http://arxiv.org/abs/1908.10084, arXiv:1908.10084 [cs]
28. Takehana, C., Lim, D., Kurtulus, E., Iyer, R., Tanimura, E., Aggarwal, P., Cantillon, M., Yu, A., Khan, S., Chi, N.: Stanford MLab at SemEval 2023 Task 7: Neural Methods for Clinical Trial Report NLI. In: Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023). pp. 1769–1775. Association for Computational Linguistics, Toronto, Canada (2023). https://doi.org/10.18653/v1/2023.semeval-1.245, https://aclanthology.org/2023.semeval-1.245
29. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention Is All You Need (Aug 2023), http://arxiv.org/abs/1706.03762, arXiv:1706.03762 [cs]
30. Vladika, J., Matthes, F.: Sebis at SemEval-2023 Task 7: A Joint System for Natural Language Inference and Evidence Retrieval from Clinical Trial Reports (May 2023), http://arxiv.org/abs/2304.13180, arXiv:2304.13180 [cs]
31. Wang, W., Xu, B., Fang, T., Zhang, L., Song, Y.: KnowComp at SemEval-2023 Task 7: Fine-tuning Pre-trained Language Models for Clinical Trial Entailment Identification. In: Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023). pp. 1–9. Association for Computational Linguistics, Toronto, Canada (2023). https://doi.org/10.18653/v1/2023.semeval-1.1, https://aclanthology.org/2023.semeval-1.1
32. Zheng, H., Shen, L., Tang, A., Luo, Y., Hu, H., Du, B., Tao, D.: Learn From Model Beyond Fine-Tuning: A Survey (Oct 2023), http://arxiv.org/abs/2310.08184, arXiv:2310.08184 [cs]
33. Zhou, Y., Jin, Z., Li, M., Li, M., Liu, X., You, X., Wu, J.: THiFLY Research at SemEval-2023 Task 7: A Multi-granularity System for CTR-based Textual Entailment and Evidence Retrieval (Jun 2023), http://arxiv.org/abs/2306.01245, arXiv:2306.01245 [cs]