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

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. TODO

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

1 Introduction

TODO: Results to last Epoch instead of Best
TODO: Add Cotations

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.

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 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 [16] [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 Adverse Events, Eligibility, Intervention or Results (Table 1). 60% of the Dataset are single type, which has only the Primary Prompt and the other 40% are comparison type has Primary and Secondary Prompt [6][15][16]. The SemEval task 2 dataset consists of 1700 training and 200 development samples (Table 2), distributed equally betwwen the two labels and almost for the 4 kinds of CTR sections. We use the development samples as validation.

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.

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 Dataset

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	

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 remaining are only the two $[SEP]$ tokens and for Comparison Type are all elements given.

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

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.

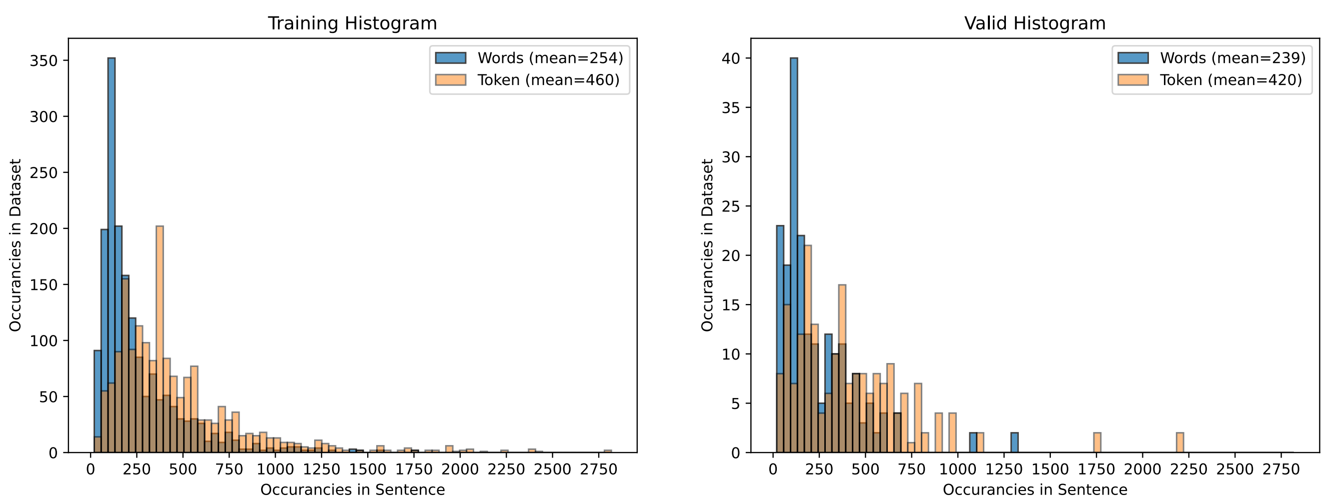

Fig. 1. Histogram of Number of Words from the full Sentence and the tokenized version with BertTokenizer

3 Recent Works

The majority of the released systems failed to achieve significantly above the majority-class baseline of 66.7% F1 value see Table 3 [16]. **Sebis [26]** uses a system concatenating the parts of the CTRs in two diffent method, called pipeline and joint, where basically the sentence represenation includes more $[SEP]$ tokens to dense it up. **KnowComp [27], YNU-HPCC [13] and Stanford [24], I2R [23], Clemson NLP [10]** 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 [13]** utilize Supervised Contrastive Learning with the corresponding loss function. **JUST-KM [11]** Models are enhanching RoBERTa in a role-based approach, where the two RoBERTa-Large Models trained differently, to predict the general outcome. **Saama [17]** finetuned Flan-T5 LLM to SemEval's task-specific data by applying different Instruction Templates. **THiFLY [29]** 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.

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

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

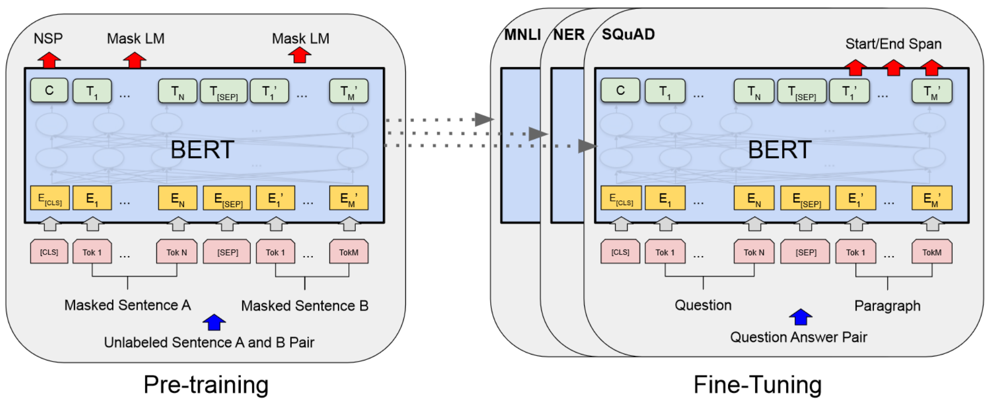

Fig. 2. Bidirectional Encoder Representations from Transformers (BERT) scematic View [12].

The Experiments are using BERT, Bidirectional Encoder Representations from Transformers (see Fig. 3) [12] developed by Google is based on Transformer architecture introduced by Vaswani et. al. [25]. 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 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 [20], ALBERT [18], BART [19], DeBERTa [14], which improves BERT architecture differently.

4.2 Sentence Embedding Architectures

TODO

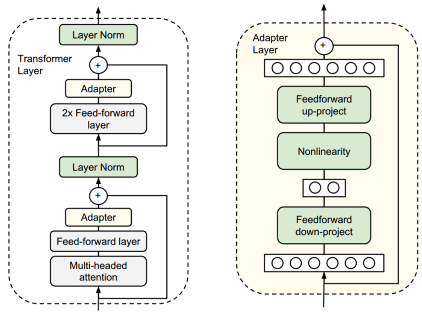

Fig. 3. Basic Structure of Adapter built on top of a Models Architecture, where only the Adapter Layers Parameters are trainable [28].

4.3 Adapter Tuning

Adapter Tuning 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[28] [22]. 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 [21]. The effectiveness of adaptation-tuning depends on the similarity between the pre-training task and the target task due to fixed embeddings.

4.4 Fusing different Datasets/Loaders

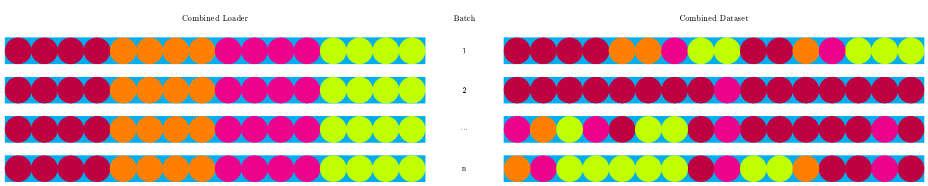

Fig. 4. Combined Dataset and Combined Loader Strategy where 4 different Dataset are getting mixed together. On the Loader-Level for a Batch size of 16, every Dataset is present equally and on the Dataset-Level, they are concatenated and the mixed.

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 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) we are using SNLI [9] which has short sencences for entailment 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 choosen ones for expanding. Also the expected outcome should be lower than the specialized Model on SemEval as it has to generalize the different domains more.

4.5 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].

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

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

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

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

5 Experiments

5.1 Environmental Setup

For our Experiments we are bounded by the Enviroments given from Kaggle [4] and Colab [3], which makes quite complex to find the optimal hyperparameters. 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 bottelnecks 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 [7], Lightning AI [5] implementation and for loading the pretrained BERT Model we are using Huggingface (transfromer libary) [8]. Therefore, we can focus the Experiments on implementing strategies to enhance BERT's overall performance tested on four different Seeds.

5.2 Learning Rate

Learning Rate, Token Length and Mixed-Precision

Table 4. F1 Values of the Baseline BertModelForSentenceClassification of the Last (50th) Epoch

Learning Rate	Seed				mean ± std
	0	42	1998	1M	
3e-6	0.637	0.554	0.663	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	0.578	0.630	0.667	0.621 ± 0.040

5.3 Permutation

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

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**

5.4 Dataset Expansion

Table 6. F1 Values of the Dataset Test of Last (50th) Epoch

Dataset + SemEval	Strategy	Seed				mean ± std
		0	42	1998	1M	
Baseline		0.636	0.640	0.638	0.657	0.630 ± 0.026
SNLI	CombDL	0.597	0.613	0.663	0.573	0.617 ± 0.042
	CombDS	0.619	0.669	0.637	0.664	0.643 ± 0.019
HEALTHVER	CombDL	0.595	0.637	0.602	0.522	0.620 ± 0.009
	CombDS	0.636	0.621	0.631	0.622	0.635 ± 0.017
SCIFACT	CombDL	0.586	0.547	0.602	0.561	0.574 ± 0.022
	CombDS	0.687	0.619	0.664	0.645	0.654 ± 0.025
SCIFACT,HEALTHVER	CombDL	0.580	0.564	0.551	0.570	0.566 ± 0.010
	CombDS	0.664	0.637	0.657	0.649	0.652 ± 0.010
SCIFACT,SNLI	CombDL	0.645	0.607	0.545	0.567	0.591 ± 0.039
	CombDS	0.658	0.574	0.624	0.643	0.625 ± 0.032
HEALTHVER,SNLI	CombDL	0.	0.	0.	0.	0. ± 0.
	CombDS	0.	0.	0.	0.	0. ± 0.

5.5 Architecture

Table 7. F1 Values of the Architecture Test from the Last (50th) Epoch

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	0.629	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	0.632	0.637	0.664	0.630 ± 0.023
v6	CosSim(u, v) and MSELoss	0.667	0.654	0.667	0.646	0.664 ± 0.011

5.6 Adapter

Table 8. F1 Values of the Adapter Trial from the Last (50th) Epoch

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

6 Conclution

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. PyTorch, https://pytorch.org/
8. Hugging Face – The AI community building the future. (Nov 2023), https://huggingface.co/datasets
9. snli · Datasets at Hugging Face (Nov 2023), https://huggingface.co/datasets/snli
10. Alameldin, A., Williamson, A.: Clemson NLP at SemEval-2023 Task 7: Applying GatorTron to Multi-Evidence Clinical NLI
11. 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
12. 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]
13. 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
14. He, P., Liu, X., Gao, J., Chen, W.: DEBERTA: DECODING-ENHANCED BERT WITH DIS- ENTANGLED ATTENTION (2021)
15. 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]
16. 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
17. 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
18. Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.: ALBERT: A LITE BERT FOR SELF-SUPERVISED LEARNING OF LANGUAGE REPRESENTATIONS (2020)

19. 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
20. 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]
21. 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
22. 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]
23. 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
24. 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
25. 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]
26. 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]
27. 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
28. 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]
29. 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]