Team INF-UFRGS at SemEval-2023 Task 7: Supervised Contrastive Learning for Pair-level Sentence Classification and Evidence Retrieval

Abel C. Dias Filipe F. Dias Higor Moreira Viviane P. Moreira João L. D. Comba
Instituto de Informática - UFRGS - Brazil
{abel.correa,ffdias,hmoreira,viviane,comba}@inf.ufrgs.br

Abstract

This paper describes the EvidenceSCL system submitted by our team (INF-UFRGS) to SemEval-2023 Task 7: Multi-Evidence Natural Language Inference for Clinical Trial Data (NLI4CT). NLI4CT is divided into two tasks, one for determining the inference relation between a pair of statements in clinical trials and a second for retrieving a set of supporting facts from the premises necessary to justify the label predicted in the first task. Our approach uses pair-level supervised contrastive learning to classify pairs of sentences. We trained EvidenceSCL on two datasets created from NLI4CT and additional data from other NLI datasets. We show that our approach can address both goals of NLI4CT, and although it reached an intermediate position in the ranking of participating system, there is room for improvement in the technique.

1 Introduction

There has been a significant increase in medical publications in recent years, including clinical trial data. The SemEval-2023 Task 7, called Multi-Evidence Natural Language Inference for Clinical Trial Data (NLI4CT), addresses the problem of large-scale interpretability and evidence retrieval from breast cancer clinical trial reports (Jullien et al., 2023). Currently, there are about 10K reports of breast cancer, making it difficult for clinical practitioners who want to provide care based on reliable clinical evidence to analyze all these reports carefully. Thus, NLI4CT is divided into Task 1 - Textual Entailment and Task 2 - Evidence Retrieval (ER). The goal of Task 1 is to determine the inference relation (contradiction/entailment) between statement pairs (premise and hypothesis), and Task 2 is to output the supporting facts from the premise to justify the label in Task 1.

Recent work leveraged Natural Language Inference (NLI) data to enhance sentence representations, improving results in downstream tasks. PairSupCon (Zhang et al., 2021) adds a discrimination head to separate entailment and contradiction instances while learning high-level semantic information from the sentence pairs and negative examples. A linear classification head enables supporting pairwise entailment and contradiction reasoning. Similarly, PairSCL (Li et al., 2022) adds a cross-attention module to enhance pair-level representation by calculating the token-level co-attention matrix to indicate the relevance of the tokens in the premises and hypotheses.

Our approach explores the cross-attention module in PairSCL for evidence retrieval by separating evidence from non-evidence in the embedding space. **EvidenceSCL** is built on top of PairSCL (Li et al., 2022), employing a supervised contrastive loss in the evidence retrieval objective and a classification head in the textual entailment objective. We also created two datasets combining MedNLI (Romanov and Shivade, 2018), a small part of MultiNLI (Williams et al., 2018), and the NLI4CT dataset. We show that EvidenceSCL addresses the tasks of NLI4CT, and although it achieved an intermediate position in the ranking of participating systems, there is room for improvement.

2 Background

Natural Language Inference (NLI) is the task of determining if two sentences follow each other. We want to determine if a hypothesis can be inferred from a premise. NLI datasets present a set of linguistic phenomena such as negation, modals, quantifiers, pronouns, beliefs, conditionals, tense, and a variety of others. Some phenomena are more frequent in specific genres of text. For instance, conversational genres have the highest percentage of sentence pairs with an occurrence of negation, WH-words, belief-verbs, and time terms. In contrast, the verbatim genre has more sentence pairs with quantifiers and conversational pivots (Williams et al., 2018). However, clinical texts contain specific lin-

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 700–706

guistic phenomena, such as medical terms, abbreviations, or medical concepts written in different forms.

Incorporating domain knowledge has been shown to improve model accuracy in the NLI task for clinical data (Lu et al., 2019). MedNLI (Romanov and Shivade, 2018) is an expert annotated dataset for NLI in the clinical domain. Premises are derived from MIMIC-III (v1.3) (Johnson et al., 2016) (which contains more than 2 million clinical notes written by healthcare professionals in English) and clinicians generated three hypotheses for each. Premises and hypotheses express various medical concepts from UMLS semantic types (Bodenreider, 2004), such as finding, disease or syndrome, sign or symptom, pharmacological substance, and others.

NLI4CT (Jullien et al., 2023) addresses a specific NLI task where multiple premises can be evidence to justify the label assigned to the hypothesis. The NLI4CT data comprises a set of sentences (*i.e.*, premises) from breast cancer clinical trials, statements (*i.e.,* hypotheses), and labels annotated by domain experts. Premises are collected from four sections of the clinical trial reports: eligibility criteria, intervention, results, and adverse events. The sentence pairs may present linguistic phenomena, such as quantifiers, medical concepts, acronyms, drugs, time frames, and dosage. Also, not all premises are relevant to a given hypothesis. Thus, retrieving the correct premises supporting the hypothesis is essential.

Some techniques rely on training sentence embeddings with supervised data to improve accuracy in downstream tasks (Conneau et al., 2017; Lu et al., 2019). Models learned from NLI datasets perform better than supervised tasks or models trained with unlabeled data. The most likely assumption is that models trained on NLI datasets require a high-level understanding of the semantic relationship between tokens in pairs of sentences.

3 EvidenceSCL System

This section details the NLI datasets and the model used in the EvidenceSCL system.

3.1 NLI Datasets

NLI datasets are generated using an anchor text to represent a premise and statements annotated by experts representing assumptions over the anchor (Figure 1). Experts often annotate an accurate assumption, a false assumption, and an assumption that might be true, representing the hypotheses for each label (*i.e.*, entailment, contradiction, or neutral). An instance can be represented as a tuple (p, h, y), where p, h is the premise-hypothesis pair, and y is the label.

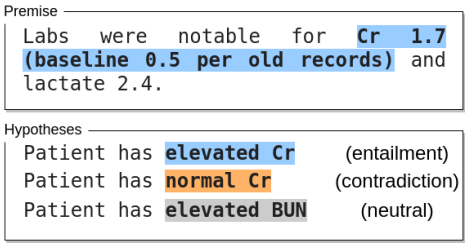

Figure 1: Example of three hypotheses for a single premise from the MedNLI dataset.

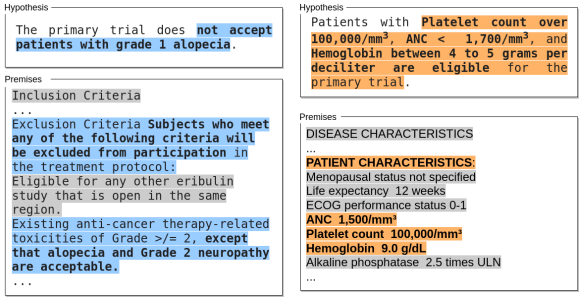

Figure 2: Two instances from the NLI4CT dataset. Entailment on the left, and contradiction on the right. The premises are below each instance. The pieces of evidence are highlighted.

The NLI4CT dataset has multiple candidate pieces of evidence for a given hypothesis (Figure 2). Each sentence in a randomized clinical trial section may be evidence for a hypothesis. Furthermore, the hypothesis may be comparing two clinical trials with premises in different documents. Unlike the MedNLI dataset, no neutral classes are explicitly set in the NLI4CT dataset. The NLI4CT dataset has multiple premises for a single hypothesis, whereas the MedNLI dataset has three hypotheses for a given premise. Figures 1 and 2 illustrate instances from the MedNLI and NLI4CT datasets, respectively. We highlighted the sentences part of the evidence in blue for an entailment instance, orange for contradiction, and gray for neutral.

We took a straightforward approach to construct different datasets to train our method combining MultiNLI, MedNLI, and NLI4CT (Williams et al., 2018; Romanov and Shivade, 2018; Jullien et al., 2023). The first dataset combines MedNLI and NLI4CT to create a three-labeled NLI dataset.

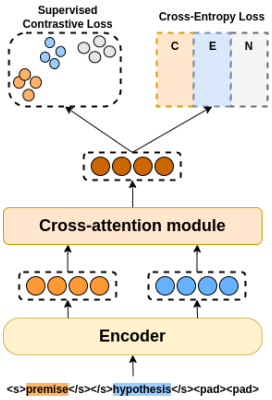

Figure 3: PairSCL components (Li et al., 2022).

Thus, we keep the neutral class from MedNLI, and for the NLI4CT data, we create a neutral instance for a premise-hypothesis pair whose premise is not a piece of evidence. We created a balanced dataset, keeping the exact size of instances for each entailment label. However, the NLI4CT dataset contains different sentence pairs for the same hypothesis with the labels entailment/neutral or contradiction/neutral. In other words, no entailment hypothesis exists for a contradiction or vice-versa, and each hypothesis may have several neutral premises. The second dataset combines MultiNLI, MedNLI, and NLI4CT using only the entailment and contradiction labels.

Additionally, we defined a binary label in both datasets to indicate whether the premise is evidence for the hypothesis in the sentence pairs. Consequently, all the instances of the neutral class were set as a non-evidence label. MultiNLI, MedNLI, and NLI4CT data may help the model learn the NLI objective. However, more neutral pairs in NLI4CT should enforce the separation of entailment and contradiction instances from non-evidence instances, leveraging the evidence retrieval task.

3.2 PairSCL Model

EvidenceSCL uses a modified PairSCL model, which has been shown to perform well on NLI and transfer learning tasks (Li et al., 2022). The model is built on top of PairSCL, starting from pre-trained checkpoints of Biomed RoBERTa. It is explicitly trained on biomedical text data (Gururangan et al., 2020), which allows us to enhance the performance of EvidenceSCL on biomedical NLP tasks, including NLI and evidence retrieval.

PairSCL has three main components: an encoder, a cross-attention module, and a joint-training layer (Figure 3). The encoder computes the sentence representations of the input text, and the cross-attention module augments the sentence representation by concatenating different representations of the sentence pairs. PairSCL is trained with a combined objective. The supervised contrastive loss separates positive and negative instances, bringing latent representations of instances in the same class. On the other hand, a softmax-based cross-entropy loss constitutes the classification objective.

The cross-attention module calculates the co-attention matrix at the token level for the sentence pairs. Its elements represent the semantic relationship between two tokens in a premise-hypothesis pair. The result of the cross-attention module is a pair-level representation $\mathbf{Z}$, which is an aggregation of augmented versions obtained from semantic representations of the premise-hypothesis pair. Equation 1 denotes the concatenation of the augmented semantic representations of the premise, hypothesis, difference, and element-wise product. One can refer to Li et al. (2022) for more details on how these representations are computed.

$$\mathbf{Z} = [\hat{S}^{(p)}; \hat{S}^{(h)}; \hat{S}^{(p)} - \hat{S}^{(h)}; \hat{S}^{(p)} \odot \hat{S}^{(h)}] \quad (1)$$

The supervised contrast loss function for a batch $\mathcal{I}$ of size K is defined in Equation 2. Here, $(X^{(p)}, X^{(h)}, y)_{i \in \mathcal{I}=1,\cdots,K}$ represents the instances in the batch, and $\mathcal{P}$ denotes the set of positive pairs where each p has the same label as i and $p \neq i$. The likelihood $l_{i,p}$ indicates the probability that the pair i is most similar to p than any other pair k in $\mathcal{I}$. The hyperparameter τ controls the temperature of the softmax distribution over the sentence representations. A lower value of τ results in a sharper distribution that assigns higher probabilities to highly similar pairs, thus improving the optimization.

$$l_{i,p} = \frac{\exp(\mathbf{Z}_i \cdot \mathbf{Z}_p/\tau)}{\sum_{k \in \mathcal{I}/i} \exp(\mathbf{Z}_i \cdot \mathbf{Z}_k/\tau)},$$
$$\mathcal{L}_{SCL} = \sum_{i \in \mathcal{I}} -\log \frac{1}{|\mathcal{P}|} \sum_{p \in \mathcal{P}} l_{i,p} \quad (2)$$

For the classification objective, PairSCL adopts a softmax-based cross-entropy loss. Equation 3 presents the cross-entropy loss function, where $\mathbf{W}$ and $\mathbf{b}$ are trainable parameters, $\mathbf{Z}$ is the pair-level representation from the cross-attention module, and y is the label for the pair.

$$\mathcal{L}_{CE} = CrossEntropy(\mathbf{WZ} + \mathbf{b}, y) \quad (3)$$

Equation 4 presents the overall loss, which is the summation of the two losses, with a hyperparameter α to weigh over the cross-entropy loss.

$$\mathcal{L} = \mathcal{L}_{SCL} + \alpha\mathcal{L}_{CE} \quad (4)$$

4 Experimental Setup

We fine-tuned EvidenceSCL starting from the pretrained Biomed RoBERTa checkpoints (Gururangan et al., 2020) for at most 80 epochs using the AdamW optimizer (Loshchilov and Hutter, 2017). We used a batch size of 512[1] and set the temperature parameter τ to 0.1. To combine the two objectives, as suggested by Li et al. (2022), we set α to 1 in our experiments. We performed experiments by setting the maximum sequence length to 128 to assess the performance when information loss occurs. To mitigate overfitting, we added an L1 regularization term to the overall loss, where the L1 coefficient was set to 0.1. We also used weight decay as an L2 regularization factor with a weight decay coefficient of $1e-4$. We implemented an early stopping mechanism to stop the training when the difference between the training and validation losses became negligible. We used different initial values for the learning rate and the cosine annealing warm restart approach to update it following a cosine curve, resetting it each epoch.

4.1 Fine-tuning the EvidenceSCL Model

We fine-tuned PairSCL for Natural Language Inference and Evidence Retrieval tasks using the training data provided, combined with MedNLI and MultiNLI. For NLI4CT data, each sentence in the clinical trial section becomes a premise for a hypothesis, resulting in several pairs for the same hypothesis. During the evaluation, we must work around it to classify all the instances for a single hypothesis and choose the correct label.

An instance in the NLI4CT dataset can be formalized as a tuple $(X^{(p)}, X^{(h)}, e, y)$, where $X^{(p)}$, and $X^{(h)}$ are the tokens in the premise and the hypothesis, respectively. A binary variable e indicates if the premise is evidence for the hypothesis, and y is the textual entailment label (i.e., entailment, neutral, contradiction). As mentioned in the Subsection 3.1, $e = 0$ for all neutral instances.

[1]Batch size is 8 with gradient accumulation for 64 steps.

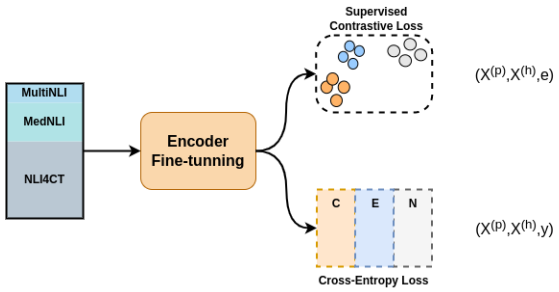

Figure 4: Supervised contrastive loss addresses the evidence retrieval objective and is expected to separate neutral instances from entailment or contradiction. Cross-entropy loss addresses the classification objective.

We tested the following approaches: **a)** training the supervised contrastive loss for separating instances based on the binary label e and the cross-entropy loss to classify instances based on the label y (Figure 4), and **b)** training both losses only on label y. Also, we trained the model using three- and two-labeled versions of the dataset. We explore the neutral class as a bucket for non-evidence pairs in the three-labeled dataset. On the other hand, in the two-labeled dataset, we removed the neutral class. Then, we trained two separated classifiers for evidence retrieval over the label e and textual entailment over the label y.

The encoder training stage aims to fine-tune the model using the combined dataset. Instances from MultiNLI and MedNLI will enforce the model to learn how to separate the textual entailment classes. In contrast, given the high number of neutral classes and the absence of counter-hypotheses (e.g., a contradiction hypothesis for an entailment one for the same set of premises.), NLI4CT will enforce the model to learn how to separate evidence from non-evidence for the hypotheses. To prevent one-class classification, we intentionally reduced the number of neutral instances and shuffled the instances during the encoder fine-tuning.

After training the encoder, we fine-tuned the pretrained model using only the NLI4CT dataset for NLI and evidence retrieval tasks. Since pairs with the same hypothesis belong to the same instance, we grouped all sentence pairs for each hypothesis to measure model accuracy with the NLI4CT dataset and evaluated each pair individually. To obtain the final label, we compared the aggregation of prediction results and a more straightforward approach that considers a sentence pair as a contradiction if all pairs for a single hypothesis are classified as it. We adopted the latter approach be-

cause it provided better results than choosing the majority class. For the evidence retrieval task, we identified as evidence all non-neutral pairs for the three-labeled dataset, and we evaluated a binary classifier for the two-labeled dataset.

4.2 Comparison against the Baseline

We compared EvidenceSCL with the solutions provided in the starter script[2]. For Task 1, we used the TF-IDF entailment prediction baseline. We computed a score based on the average of the cosine distances from all sentence pairs from the relevant sections in the primary and secondary trials. The pair is classified as contradictory if the score passes a threshold t. We had good results with $t = 0.99$. For Task 2, we used the BM25 Okapi baseline from the starter script with no modifications retrieving all the entries with scores higher than 1.

As mentioned in Section 4.1, we tested two different training setups for the model: **a)** training the supervised contrastive loss on the evidence label e and cross-entropy loss on the textual entailment label y, and **b)** training both losses on y. The first approach is called EvidenceSCL, and the second is the original PairSCL. EvidenceSCL-2L and PairSCL-2L (two-labeled) were trained for the NLI task and applied transfer learning to evaluate both approaches for evidence retrieval. The three-labeled versions EvidenceSCL-3L and PairSCL-3L can be employed in both tasks.

4.3 Description of the NLI Datasets

The MedNLI dataset (Shivade, 2019) is available on the PhysioNet website[3] for users with credential access. The user must undergo training on essential aspects of research with human subjects. SemEval 2023: Task 7 - NLI4CT dataset[4] was created from clinical trial reports publicly available on the web and annotated by domain experts.

Table 1 details the composition of the datasets used to train the EvidenceSCL encoder. We excluded test dataset information, as we used all unlabeled instances from the NLI4CT test dataset. Since NLI4CT does not have neutral labels, we took different approaches for the different versions of the EvidenceSCL dataset.

For EvidenceSCL-3L, we set all non-evidence sentence pairs to neutral. For EvidenceSCL-2L, we arbitrarily set all neutral pairs in MedNLI to entailment and set the evidence label to 0. We chose this design because the neutral hypotheses would be closer to the entailment than the contradiction class. Since it is not a contradiction, it is just a non-evidence premise. However, for NLI4CT, we kept the original hypothesis label (contradiction/entailment) and set the evidence label to 0.

The number of sentence pairs from all sections of the clinical trials is higher in NLI4CT than in MedNLI. This results in a highly imbalanced dataset, with over 85% of the samples belonging to the neutral class. On the other hand, the number of neutral pairs is more evenly distributed across the remaining classes in the two-labeled dataset, resulting in a more balanced dataset in terms of class labels but with many non-evidence samples. In the EvidenceSCL datasets, we combined premises from the same hypothesis section, thus reducing the size of the NLI4CT training dataset to 39,935 and the validation dataset to 4,224.

5 Results

We conducted experiments with different hyperparameter settings to evaluate the performance of our model. We used the F1 score, precision, and recall as evaluation metrics. Table 2 shows the best results for the NLI and ER tasks.

We fine-tuned a linear classifier for the ER task to distinguish evidence from non-evidence examples in the two-labeled dataset. In the three-labeled dataset, we labeled non-evidence examples as neutral and used them for evaluation. For the NLI task, we evaluated pairs of sentences individually and aggregated their predictions to obtain a final classification for each sample. We chose the majority

	Training	Validation
MedNLI	11,232	1,395
NLI4CT	146,955	17,584
EvidenceSCL-2L	47,423	5,154
MedNLI	11,232	1,395
NLI4CT	39,935	4,224
EvidenceSCL-3L	40,899	4,218
MedNLI	11,232	1,395
NLI4CT	29,667	17,584

Table 1: Composition of the two- and three-labeled datasets (EvidenceSCL-2L and EvidenceSCL-3L).

[2]Starter script: https://sites.google.com/view/nli4ct/get-data-and-starting-kit.

[3]PhysioNet website: https://physionet.org/content/mednli/1.0.0/

[4]NLI4CT Dataset Description: https://sites.google.com/view/nli4ct/dataset-description

		F1 (dev/test)	Precision (dev/test)	Recall (dev/test)
Task 1 - NLI	TF-IDF Baseline	0.657 / 0.642	0.497 / 0.494	0.970 / 0.920
	EvidenceSCL-2L	0.669 / 0.620	0.533 / 0.496	0.900 / 0.824
	PairSCL-2L	0.488	0.453	0.530
	EvidenceSCL-3L*	0.421 / **0.666 (17)**	0.541 / **0.500 (33)**	0.345 / **0.996 (2)**
	EvidenceSCL-3L	**0.727**	**0.571**	**1.000**
	PairSCL-3L	0.577	0.533	0.629
Task 2 - IR	Okapi BM25	0.322 / 0.350	0.422 / 0.469	0.261 / 0.279
	EvidenceSCL-2L*	0.211 / **0.681 (21)**	0.641 / **0.615 (19)**	0.126 / **0.764 (21)**
	EvidenceSCL-3L	**0.839** / 0.610	**0.907** / 0.517	0.782 / 0.743
	PairSCL-3L	0.660	0.520	**0.903**

Table 2: Official results of EvidenceSCL for tasks 1 and 2 of SemEval 2023 Task 7 - NLI4CT in dev/test datasets are identified with * in bold. The remaining evaluation results were obtained in the post-evaluation phase. The numbers between parentheses indicate the position achieved by our team in the competition.

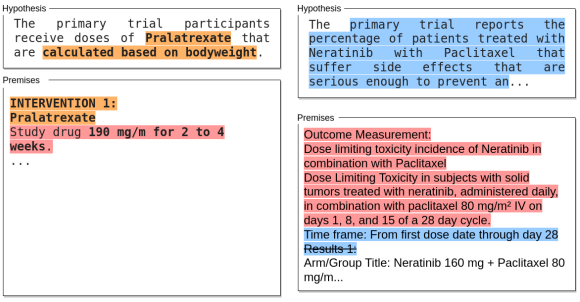

Figure 5: EvidenceSCL-3L classification results for contradiction (orange), entailment (blue), and misclassified instances (sentences in red).

class predicted for the sentence pairs.

EvidenceSCL-3L model outperformed the other models we evaluated in both tasks, with high recall scores but poor precision in the NLI task. We submitted this model to the competition and achieved the 17th position in the ranking. After the competition, we slightly improved its precision by training a new encoder with a learning rate of $1e-6$. Compared to the baseline models in Table 2, EvidenceSCL-2L performed the best on the test dataset for the ER task, while EvidenceSCL-3L achieved the best result on the validation set. We noticed that sometimes the model might suffer from overfitting, and we figured out that it relates to how training and validation stages are performed concerning the NLI4CT dataset.

The EvidenceSCL model performs better on instances with fewer sentence pairs, and its accuracy decreases for instances with many pairs. We ruled out the combination of ER and NLI goals in the loss functions as the cause of misclassification errors, such as the ones observed in Figure 5, and conjectured that the dataset likely caused them.

In our study, we created a dataset by combining samples from MedNLI, MultiNLI, and NLI4CT, although the latter contained sentence pairs labeled as evidence and non-evidence. This design choice may have negatively influenced the model since the same premise can be treated as evidence or non-evidence in different instances, making it challenging for the model to learn when it is one or the other. Additionally, the medical domain presents challenges, such as acronyms or technical terms in the hypotheses not mentioned in the premises, which can affect performance. Improving the mapping of these features could help the model learn to classify a premise as evidence more accurately.

Regarding model parameters, we found that using learning rates of $5e-05$, $1e-06$, and $5e-06$ in models trained between 10 and 15 epochs yielded better results. We also observed that accuracy decreased as the number of training epochs increased.

We noticed that sometimes the model might suffer from overfitting, which relates to how training and validation stages are performed concerning the NLI4CT dataset. We performed new experiments and found that grouping entailment/contradiction instances close to their respective negative examples in the NLI4CT dataset during training improved the accuracy.

6 Conclusion

In this paper, we described our submission to SemEval-2023 Task 7. We used supervised contrastive learning to improve pair-level sentence representations of the Biomed RoBERTa model. We fine-tuned a linear classifier on top of it to figure out

pieces of evidence and perform textual entailment classification on sentence pairs. EvidenceSCL differs from PairSCL by addressing ER and NLI tasks, learning the underlying semantic relationship between terms of the sentence pairs when separating pairs as evidence or non-evidence, and classifying them as entailment or contradiction.

Our experiments trained models with different hyper-parameter settings on two-labeled and three-labeled datasets. EvidenceSCL outperformed the baselines, reaching an F1 score of 0.666 and ranked 17th on the leaderboard for task 1 - textual entailment. For task 2 - evidence retrieval, we reached an F1 score of 0.681 and ranked 21st on the leaderboard. We also showed that the model could be further improved by hyperparameter tuning. We are performing study ablations to analyze the gain in accuracy when fine-tuning the encoder using different combinations of the three datasets. Also, we want to analyze the impact of the hyperparameter tunning on model results, comparing the models on downstream tasks.

References

Olivier Bodenreider. 2004. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database issue).

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. Association for Computational Linguistics.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360. Association for Computational Linguistics.

Alistair E. W. Johnson, Tom J. Pollard, Lu Shen aand Li-Wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3.

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*.

Shuang Li, Xuming Hu, Li Lin, and Lijie Wen. 2022. Pair-Level Supervised Contrastive Learning for Natural Language Inference. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8237–8241.

Ilya Loshchilov and Frank Hutter. 2017. Fixing Weight Decay Regularization in Adam. *CoRR*, abs/1711.05101.

Mingming Lu, Yu Fang, Fengqi Yan, and Maozhen Li. 2019. Incorporating Domain Knowledge into Natural Language Inference on Clinical Texts. *IEEE Access*, 7:57623–57632.

Alexey Romanov and Chaitanya Shivade. 2018. Lessons from Natural Language Inference in the Clinical Domain. *CoRR*, abs/1808.06752.

Chaitanya Shivade. 2019. MedNLI - a Natural Language Inference Dataset for the Clinical Domain (version 1.0.0).

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.

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.