FMI-SU at SemEval-2023 Task 7: Two-level Entailment Classification of Clinical Trials Enhanced by Contextual Data Augmentation

Sylvia Vassileva
FMI, Sofia University
svasileva@fmi.uni-sofia.bg

Georgi Grazhdanski
FMI, Sofia University
ggrazhdans@uni-sofia.bg

Svetla Boytcheva
Ontotext & FMI, Sofia University
svetla.boytcheva@ontotext.com
svetla@uni-sofia.bg

Ivan Koytchev
FMI, Sofia University
koychev@fmi.uni-sofia.bg

Abstract

The paper presents an approach for solving SemEval 2023 Task 7 - identifying the inference relation in a clinical trials dataset. The system has two levels for retrieving relevant clinical trial evidence for a statement and then classifying the inference relation based on the relevant sentences. In the first level, the system classifies the evidence-statement pairs as relevant or not using a BERT-based classifier and contextual data augmentation (subtask 2). Using the relevant parts of the clinical trial from the first level, the system uses an additional BERT-based classifier to determine whether the relation is entailment or contradiction (subtask 1). In both levels, the contextual data augmentation is showing a significant improvement in the F1 score on the test set of 3.7% for subtask 2 and 7.6% for subtask 1, achieving final F1 scores of 82.7% for subtask 2 and 64.4% for subtask 1.

1 Introduction

A huge amount of clinical trial reports are generated annually, and the ability to automatically extract and analyze data from these reports can help healthcare professionals stay abreast of the latest trends and findings from these trials. SemEval 2023 Task 7: Multi-evidence Natural Language Inference for Clinical Trial Data consists of a dataset of breast cancer clinical trial reports and related statements in English (Jullien et al., 2023). The goal is to determine the trial-statement inference relationship (subtask 1) and to extract supporting sentences from the trial text (subtask 2).

The proposed system reverses the task order and redefines subtask 2 as identifying the sentences from the clinical trial text relevant to the statement, similar to Question-Answering Natural Language Inference (QNLI) (Wang et al., 2018). After finding these sentences, they are used as input for classifying the inference relation - entailment or contradiction of the statement with respect to the premise (subtask 1). We evaluated various transformer models based on the original BERT architecture (Devlin et al., 2018) and pre-trained on biomedical/clinical data and a contextual data augmentation approach that improves the model's ability to discriminate related sentences and classify them correctly.

The system showed 82.7% F1 score on the test set and ranked fifth in the leaderboard for subtask 2, 2.6% behind the first-place model. Subtask 1 is more challenging and the system showed 64.4% F1 score on the test set there. The system struggles the most when the statement requires performing quantitative reasoning to determine the relationship.

The code related to this task is available on GitHub [1].

2 Background

The task uses a dataset of breast cancer clinical trial reports, statements, inference relations, and supporting sentences annotated by domain experts. It is separated into two sub-tasks - textual entailment and evidence retrieval (Jullien et al., 2023).

Each clinical trial record consists of 4 standard sections: Eligibility, Intervention, Results, and Adverse Events. And each section in the dataset consists of multiple lines of text with information about the section.

Each statement is a section-specific claim and may be related to one or two clinical trials (primary/secondary). The statement type is single or a comparison depending on whether it refers to one or two clinical trials. Each statement has a label specifying the inference relation - entailment or contradiction - that can be inferred from the clinical trial section. The dataset is split into 3 parts - train/dev/test and the labels are evenly distributed between the two classes for each section in the train and dev sets. The majority of statements are of type

[1] https://github.com/svassileva/semeval-nci4ct

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 1454–1462

single and comparisons are fewer - approximately 39% of the train set and 30% of the dev set.

We have participated in both subtasks:

1. Subtask 1 - Textual entailment, aiming to predict the inference relation of the statement-premise pair;

2. Subtask 2 - Evidence retrieval, aiming to extract the supporting evidence from the premise.

Transformer models have been applied to a wide variety of NLP tasks, including natural language inference. In particular, BERT and BERT-based models have been successfully applied to multiple Natural Language Inference (NLI) datasets and are showing state-of-the-art (SOTA) results on two of the GLUE (Wang et al., 2018) benchmark datasets QNLI corpus (Wang et al., 2018)) - ALBERT (Lan et al., 2019), and WNLI corpus - DeBERTa (He et al., 2021). BERT-based models have a significantly smaller number of parameters than models like T5-11B (Raffel et al., 2020) and PaLM 540B (Chowdhery et al., 2022) which are showing the best results for Multi-Genre Natural Language Inference (MultiNLI), Recognizing Textual Entailment (RTE) (Wang et al., 2018) and CommitmentBank (de Marneffe et al., 2019) datasets [2]. Therefore, BERT-based models can still be useful for NLI tasks and are not as computationally expensive.

3 System overview

In order to determine the inference relationship between premise-sentence pairs of the clinical trial report (CTR), we first identify the parts of a section that contain information related to the statement.

3.1 Subtask 2 Approach

The purpose of the second task is to derive the supporting facts from the premise needed to justify the inference relation label (Jullien et al., 2023). We reformulate the task to extract the premise's relevant parts containing information about the claim.

We approach the task as a task for Question-Answering NLI (NLI (Wang et al., 2018)) where the original dataset consists of question-paragraph pairs where one of the sentences in the paragraph contains the answer to the question. For the second task, we treat each line from the relevant section of the premise as a separate sentence and we use the statement as a question. The goal is to find the lines that contain the information needed to determine the relationship of the inference.

The premises in the dataset consist of well-structured sections with easily observable general subsections. For example, the Eligibility section mostly contains two subsections - Inclusion criteria and Exclusion criteria. Similar subsections can be identified in the other sections and the most common ones are shown in figure 1. The information in each premise line is concise and describes a specific feature of the subsection, for example, one clinical trial eligibility criterion.

Additionally, the claims are analyzed and several common subjects are identified for which the statements are typically related, for example - primary/secondary trial, cohort 1 or 2, etc. Statements sometimes contain references to different features in the different trials, so the model needs to have contextual information about each line of the premise, i.e. which trial and section does the line describe.

3.1.1 Contextual Data Augmentation

In order to improve the ability of the system to distinguish which lines of the premise are relevant to the statement, we augment the data by providing additional context information such as:

- Trial - primary or secondary, based on input data;

- Parent subsection - for example, inclusion criteria. We rely on the well-structured premise and extract subsection headings using the following rule - any line which ends in a colon and is shorter than 30 characters;

- Cohort - when the data refers to a particular cohort number, we add the word "cohort" before the number; for example, "adverse events 1" becomes "adverse events cohort 1".

For example, following the rules above, the evidence sentence *Serious, non-healing wound, ulcer, or bone fracture* from the *Exclusion criteria* subsection of a *primary trial* is transformed into *Primary trial: Exclusion Criteria: Serious, non-healing wound, ulcer, or bone fracture.* Table 1 shows additional examples of contextual data augmentation.

[2] https://paperswithcode.com/task/natural-language-inference

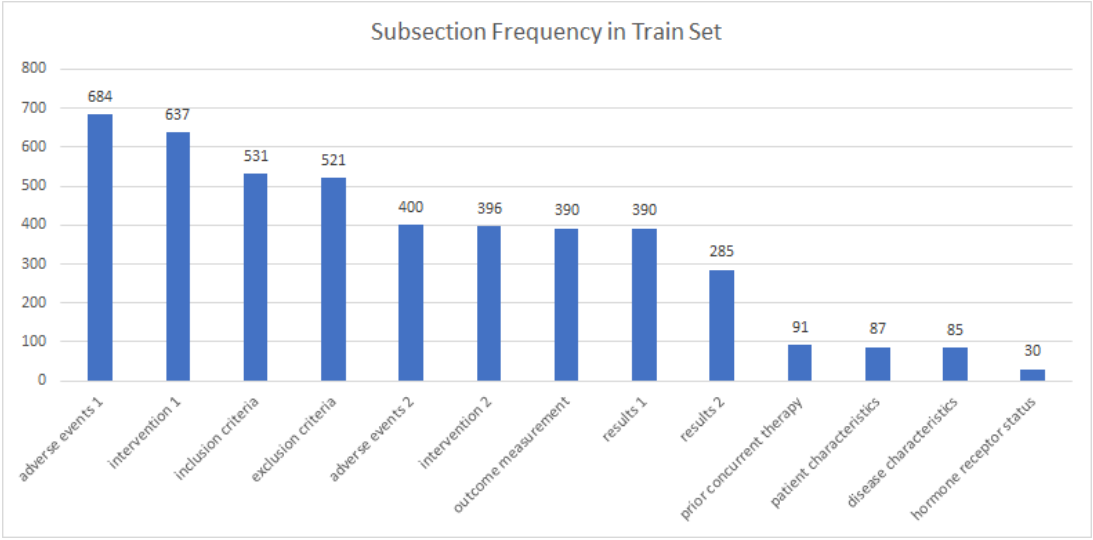

Figure 1: Most common subsections found in the train set and their frequencies.

Original Text	Augmented Text
For patients treated by lumpectomy, whole breast irradiation is required.	**Primary trial: Inclusion Criteria:** For patients treated by lumpectomy, whole breast irradiation is required.
Tamoxifen : 20 mg once daily oral dose	**Primary trial: intervention cohort 2:** Tamoxifen : 20 mg once daily oral dose
Unit of Measure: Participants 2	**Secondary trial: results cohort 2:** Unit of Measure: Participants 2
Adverse events 1: Anaemia 2/752 (0.27%)	**Secondary trial: adverse events cohort 1:** Anaemia 2/752 (0.27%)

Table 1: Examples of contextual data augmentation on the premise.

3.1.2 Subtask 2 Method

Using the augmented dataset, we train a BERT-based classifier that takes each line from the premise, and the corresponding statement and determines whether the information in the line is relevant to the statement. We use the standard BERT for sequence classification architecture (Devlin et al., 2018). Figure 2 shows the pipeline architecture used for subtask 2. We use the augmented train dataset for subtask 2, which consists of the indices of lines in the premise which support the inference relation.

We select several BERT-based models pre-trained on biomedical and/or clinical data as they have been trained using domain-specific terminology and sentence structure. Table 2 shows the different models and data used for their pre-training.

Model	Pre-trained on data
BioM-BERT-PubMed-PMC-Large (Alrowili and Shanker, 2021)	PubMed Abstracts + PMC full article
BioBERT 1.1 (Lee et al., 2020)	PubMed + PMC
Clinical BERT (Alsentzer et al., 2019)	BioBERT + MIMIC notes
BioBERT MNLI (Lee et al., 2020)	BioBERT-Base v1.1
BioMed-RoBERTa-base (Gururangan et al., 2020)	RoBERTa-base + 2.68 million scientific papers
PubMedBERT (Gu et al., 2020)	PubMed Abstracts + PMC full articles

Table 2: Different BERT-based models used for training on subtask 2.

3.2 Subtask 1 Approach

To determine the inference relation (entailment vs contradiction) between clinical trial report (CTR) - statement pairs, we fine-tuned a transformer model with a binary classification head using only statements from the provided evi-

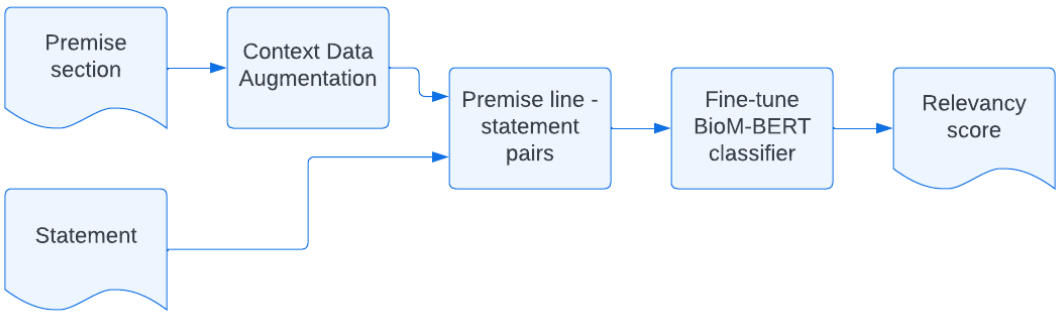

Figure 2: Subtask 2 pipeline architecture - each line of the premise section is first augmented, and then paired with the statement. The pair is passed to the BioM-BERT classifier to determine if the sentence is relevant to the statement.

dence indexes (*Primary_evidence_index* and *Secondary_evidence_index* fields in a CTR file) to represent the premise. The classification of new examples relies on the evidence indexes that the subtask 2 model outputs.

The end-to-end architecture of the system is shown in figure 3.

3.2.1 Classification Model Selection

We experimented with the following models for the classifier:

- BioM-BERT-Large (Alrowili and Shanker, 2021) - a BERT-based model (with ELECTRA architecture) for the biomedical domain, pre-trained on PubMed Abstracts + PMC + general domain vocab (EN Wiki + Books). It achieves state-of-the-art (SOTA) on certain Bio Text Classification Tasks such as ChemProt.

- Clinical Longformer (Li et al., 2023) - a Longformer-based model for the biomedical domain, further pre-trained on MIMIC-III clinical notes. This allowed us to evaluate to a certain extent the effect of long input sequences (more than 512 tokens) on classification results, as some of the statement-premise pairs are truncated when passed to BioM-BERT, although we do not use the full section text for the premise. 8.5-15% of the statement-premise pairs exceed 512 tokens.

3.2.2 Data Normalization

All data is normalized prior to fine-tuning/inference by replacing $<, >, <=, >=, \%$ with the corresponding phrases. Further, we expand some common abbreviations like AEs (Adverse events), PFS (Progression Free Survival), IV (intravenous), PO (orally), QD (every day), which were identified by manual data review.

3.2.3 Classifier Fine-Tuning

The format of the training set proved to have the highest impact on final model performance, as expected. We compared three approaches:

1. **Standard**: each statement-premise pair in the training set consists of the entire statement text and as much of the premise text (concatenated evidence sentences) as possible, up to the maximum input sequence length (512 or 4096 tokens for BioM-BERT and Clinical Longformer, respectively). This is the baseline approach.

2. **Single sentence**: create a statement-premise pair for each evidence sentence. For example, statement *S* and evidence sentences [*E1, E2, E3*] form input statement-premise pairs *S-E1, S-E2,* and *S-E3*.

3. **Single sentence + augmentation**: create a statement-premise pair for each evidence sentence. Prepend trial and subsection information to each premise sentence using the approach from section 3.1.1 (the same approach as in subtask 2).

During evaluation, we always use the entire statement-premise text as input (truncating the premise when necessary).

3.2.4 Challenges

1. Quantitative reasoning - statements requiring quantitative reasoning prove to be quite difficult for the selected models (and large language models (LLMs) in general), as models

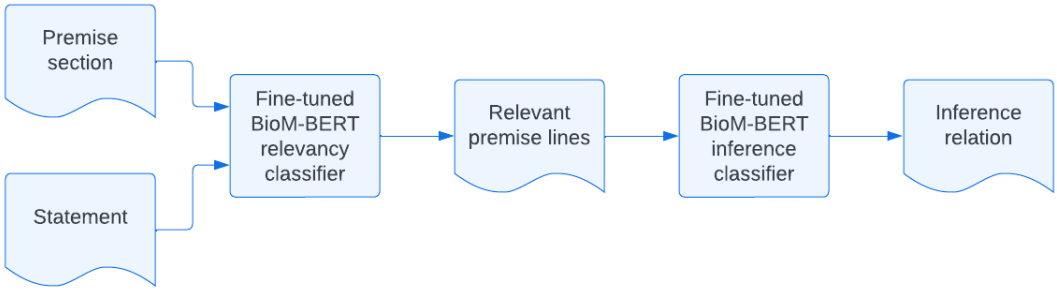

Figure 3: End-to-end architecture of the system - relevant (evidence) sentences found by the task 2 model are used along with the statement as an input to the task 1 classifier that determines the inference relation.

rely on lexical cues for prediction (Ravichander et al., 2019). One such statement is *Febrile Neutropenia was the most common adverse event recorded in the primary trial, affecting more than 5% of patients.* We only attempt to increase the signal for the aforementioned lexical cues by normalizing the data - replacing comparison operators with their corresponding phrases.

2. Clinical trial context - almost 40% of the original examples in the training set compare two clinical trials. Each input prompt must include enough context that clearly signal which part of the premise corresponds to which clinical trial (primary or secondary). We address this by augmenting each premise sentence in the training set with trial and subsection information, and also by testing different augmentation variants for dev and test prompts.

4 Experimental Setup

We perform separate experiments for subtask 2 and subtask 1. For both tasks, we use the train dataset to train different models, evaluate them on the dev set and select the best-performing model. Since subtask 1 uses the output from subtask 2, we select the best-performing model for subtask 2 and feed the output for evaluating subtask 1 models. We optimized subtask 1 hyperparameters using Tune (Liaw et al., 2018).

4.1 Data Splits

We used the entire training set (1700 prompts) for training, formatting it depending on the evaluated pre-processing approach.

For subtask 1, we evaluated several different pre-processing approaches. The standard approach simply uses each of the original examples as a training example. The other two pre-processing approaches produced multiple training prompts from each of the original 1700, resulting in a total of 20,885 prompts. The entire dev and the entire test set were used as intended for dev and testing respectively - some augmentation techniques were tested on them. However, the total number of examples was not affected.

4.2 Tools and Libraries

For subtask 1 and 2 we used Python [3] and the libraries PyTorch [4], Huggingface transformers [5], Ray tune[6], and pandas [7].

4.3 Evaluation Metrics

We used precision, recall, and macro F1 as evaluation metrics for subtask 1 and 2.

We compare the F1 score on the dev set using different models and select the best-performing model to be evaluated on the test set.

5 Results

5.1 Subtask 2 Results

We use train split to train a BERT-based classifier using the standard BERT architecture. We perform training using batch size 16, learning rate 2e-5, and weight decay 0.01 for 3 epochs, taking the epoch showing the best results on the dev set. The same hyperparameter values are used for all of the BERT-based models we compared for subtask 2.

The best performing model we trained was based on BioM-BERT and showed 86.6% F1 score on the dev set and 82.7% F1 score on the test set, ranking fifth in subtask 2.

[3] Python version 3.8.10
[4] Pytorch version 1.13.1+cu116
[5] Transformers version 4.26.1
[6] Ray tune version 2.3.0
[7] Pandas version 1.3.5

Model	Batch size	Learning rate	Adam epsilon	Warmup steps ratio	Epochs
BioM-BERT Large	16	2e-5	1e-8	0.02	3
Clinical Longformer	8	2e-5	1e-8	0.05	3

Table 3: Hyperparameters used for finetuning subtask 1 models.

Table 4 shows the results for the different BERT models used in the experiments. PubMedBERT shows a slightly higher score on the dev set, but the test set performance is quite low - 78.6%, which could signal overfitting to the train/dev sets. The BioBert MNLI model is a close second to the BioM-BERT model with 82% F1 score on the test set.

Model	Dev F1	Test F1
BioM-BERT	0.865	**0.827**
Clinical BERT	0.84	0.79
BioBert MNLI	0.86	0.82
BioBERT	0.846	0.79
BioMed Roberta Base	0.85	0.77
PubMedBERT	**0.87**	0.786

Table 4: Evaluation on dev and test sets of the different BERT-based models trained on subtask 2.

5.1.1 Impact of Contextual Data Augmentation

We also investigated the impact of the contextual data augmentation on the result, by training the same BioM-BERT model using the official training data without augmentation. The model scored 83% F1 on the dev set and 79% F1 on the test set, showing that the augmentation improves the model performance by 3.6% and 3.7% respectively.

5.2 Subtask 1 Results

The best-performing model for subtask 1 that we trained was based on BioM-BERT Large, showing 68.2% F1 on the dev set and 64.4% F1 on the test set. These results were achieved after the competition ended, as our officially submitted result was invalidated due to a bug in our evaluation code, resulting in the model being evaluated only on the 'Comparison' type examples.

5.2.1 Effect of Using Only Evidence Sentences

To investigate the effect of using only evidence sentences (those listed in the primary/secondary evidence indexes in the train/dev sets, and recognized by the task 2 model for the test set), we trained the models using two distinct approaches:

1. **Full section text**: including as much as possible of the full section text in the premise (with truncation). Both models perform poorly in terms of F1, as parts of the context crucial for the correct inference are often truncated, while others that only add noise, are included. BioM-BERT shows better precision while having lower recall. Clinical Longformer achieves a higher recall score, likely due to the larger maximum processable input limit.

2. **Evidence sentences only (Standard)**: using the full statement text and as much as possible of the concatenated evidence sentences text (with truncation) to form statement-premise pairs results in a more robust performance of both models on both the test and the dev set. On the dev set, we observed that Clinical Longformer outperforms BioM-BERT, which was expected, as it was able to process the entire statement-premise pair without loss of context (see Table 6).

5.2.2 Effect of Contextual Data Augmentation

The way the train set was pre-processed had the biggest impact on performance.

1. **Standard**: this is the baseline approach mentioned in the previous section that uses only evidence sentences to form a premise (see Table 6).

2. **Single sentence**: using the full statement text and each evidence sentence (unmodified) to form a statement-premise pair significantly improves the BioM-BERT model performance (from 50.5% F1 to 66.3% on dev, and from 44.3% to 55.6% F1 on test). Performance of Clinical Longformer is comparable to the base case (see Table 7).

3. **Single sentence + augmentation**: extending the second approach, here we prepended information about the trial (*primary* or *secondary*) and the subsection (e.g. *Inclusion criteria*)

Model	Dev Precision	Dev Recall	Dev F1	Test Precision	Test Recall	Test F1
BioM-BERT Large	**0.750**	0.030	0.057	**0.666**	0.024	0.046
Clinical Longformer	0.521	**0.120**	**0.195**	0.490	**0.104**	**0.172**

Table 5: Subtask 1 **Full section text approach**. Results of models, finetuned on full section text.

Model	Dev Precision	Dev Recall	Dev F1	Test Precision	Test Recall	Test F1
BioM-BERT Large	0.510	0.500	0.505	0.502	**0.397**	**0.443**
Clinical Longformer	**0.626**	**0.570**	**0.596**	**0.575**	0.353	0.437

Table 6: Subtask 1 **Evidence sentences only (Standard) approach**. Results of models, finetuned only on the concatenated evidence sentences.

to each evidence sentence prior to pairing it with the full statement text. This additional context leads to performance gains in both models on both train and dev sets, allowing us to achieve our best result of 71% dev F1, and 64.4% test F1 with the BioM-BERT model, which is a 4.7% and 7.6% improvement respectively, compared to the plain Single sentence approach. (see Table 8).

4. Lastly, applying the augmentation technique from the Single sentence + augmentation approach to the full evidence text in the Standard approach results in poor performance by both models, likely due to the increased noise.

In evaluation, using the full statement-premise text as input (truncating the premise when necessary) produced the most robust results on both dev and test set. We explored alternative techniques such as classifying each premise sentence against the given statement and combining the results by, for instance, classifying the input example as Contradiction if enough of its premise sentences contradicted with the statement. They turned out to be quite dependent on the given evaluation set, and often resulted in a majority classifier. Other, more sophisticated approaches may be worth exploring, although they would all suffer from the reduction of context as a result of splitting. We also tested the impact of augmenting the evaluation sets:

1. **Unmodified** - passing the entire unmodified statement+premise pair. This was the baseline approach.

2. **Mark trial** - prepend *Primary trial:* before the block with primary trial sentences in the input premise. Similar for the secondary trial premise sentences block. This does not increase performance and adds noise in some cases.

3. **Full augment** - prepend *Primary trial:* and subsection name to each premise sentence. Similar for the secondary trial. This helped improve the precision of the BioM-BERT model by 1-2% and the F1 by 1% on both dev and train sets. Clinical Longformer performance was not affected.

Overall, when using only evidence sentences as premise, BioM-BERT Large performs better than Clinical Longformer, perhaps due to the input sequence length being significantly reduced - only 8-15% of the statement-premise pairs exceed 512 tokens with the Standard approach. There are no pairs which exceed the limit with the Single sentence or Single sentence + augmentation approach.

5.2.3 Error Analysis

Most of the dev set examples, misclassified by our best model (BioM-BERT Large) require quantitative reasoning. Premises that entail only part of a compound statement while implicitly contradicting with other parts of the statement are incorrectly classified as Entailment. Table 9 provides examples for these two most common types of errors.

6 Conclusion

The proposed two-level system for retrieving relevant clinical trial evidence and classifying entailment shows competitive results for subtask 2 with 82.7% F1 score but could be improved in handling quantitative reasoning. The described approach for augmenting with contextual data improves the performance for both subtasks. As further work, the

Model	Dev Precision	Dev Recall	Dev F1	Test Precision	Test Recall	Test F1
BioM-BERT Large + Unmodified	0.670	0.610	0.638	0.640	0.485	0.552
BioM-BERT Large + Mark trial	**0.677**	0.610	0.642	0.635	0.469	0.540
BioM-BERT Large + Full augment	**0.677**	**0.650**	**0.663**	**0.654**	**0.502**	**0.568**
Clinical Longformer + Unmodified	0.638	0.530	0.579	0.630	0.349	0.449
Clinical Longformer + Mark trial	0.634	0.520	0.571	0.614	0.345	0.442
Clinical Longformer + Full augment	0.629	0.510	0.563	0.623	0.345	0.444

Table 7: Subtask 1 **Single sentence approach**. Results of models, finetuned on statement-premise pairs with a single premise sentence and no augmentation.

Model	Dev Precision	Dev Recall	Dev F1	Test Precision	Test Recall	Test F1
BioM-BERT Large + Unmodified	0.605	**0.860**	**0.710**	0.565	0.726	0.636
BioM-BERT Large + Mark trial	0.579	0.840	0.685	0.566	**0.738**	0.641
BioM-BERT Large + Full augment	0.586	0.850	0.693	0.571	**0.738**	**0.644**
Clinical Longformer + Unmodified	0.612	0.680	0.644	**0.594**	0.554	0.573
Clinical Longformer + Mark trial	**0.640**	0.730	0.682	0.588	0.590	0.589
Clinical Longformer + Full augment	0.625	0.750	0.681	0.586	0.570	0.578

Table 8: Subtask 1 **Single sentence + augmentation approach**. Results of models, finetuned on statement-premise pairs with a single premise sentence and prepended trial + subsection info.

Statement	Premise	Expected Label	Type
in cohort 1 of the primary trial there were more cases of Left ventricular dysfunction than Abdominal pain	Primary trial: Left ventricular dysfunction * 2 cases out of 219 participants (0.91 percent (%)) Abdominal pain * 1 case out of 219 participants (0.46 percent (%))	Entailment	Quantitative reasoning
the primary trial participants are administered Avastin, Bevacizumab and radiotherapy as part of the intervention	Primary trial: INTERVENTION 1: Avastin (Bevacizumab) Plus Hormone All patients received Avastin (Bevacizumab) 15 mg/kg intravenous every three weeks as well as continuing with hormonal therapy they previously were taking.	Contradiction	Implicit contradiction

Table 9: Subtask 1 incorrectly classified dev set examples sample.

system could be enhanced to extract quantitative information and use additional processing to assess whether the statement contradicts the premise.

Limitations

The proposed contextual data augmentation approach relies heavily on the structure of the dataset which includes well-formed sections and subsections. While clinical texts usually have a standard structure, it is not always so uniform, so extracting subsections in any clinical text is more difficult. Due to the limits of the input size of BERT-based models, scaling to longer text remains a challenge for this approach.

Acknowledgements

This research is partially funded by Project UNITe BG05M2OP001-1.001-0004 funded by the OP "Science and Education for Smart Growth", co-funded by the EU through the ESI Funds, and partially financed by the European Union-NextGenerationEU, through the National Recovery and Resilience Plan of the Republic of Bulgaria, project No BG-RRP-2.004-0008

References

Sultan Alrowili and Vijay Shanker. 2021. BioM-transformers: Building large biomedical language models with BERT, ALBERT and ELECTRA. In *Proceedings of the 20th Workshop on Biomedical Language Processing*, pages 221–227, Online. Association for Computational Linguistics.

Emily Alsentzer, John Murphy, William Boag, et al. 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.

A. Chowdhery, S. Narang, J. Devlin, et al. 2022. Palm: Scaling language modeling with pathways.

Marie-Catherine de Marneffe, Mandy Simons, and Judith Tonhauser. 2019. The commitmentbank: Investigating projection in naturally occurring discourse. *Proceedings of Sinn und Bedeutung*, 23(2):107–124.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. Cite arxiv:1810.04805Comment: 13 pages.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020. Domain-specific language model pretraining for biomedical natural language processing.

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 ACL*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing.

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*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Yikuan Li, Ramsey M Wehbe, Faraz S Ahmad, Hanyin Wang, and Yuan Luo. 2023. A comparative study of pretrained language models for long clinical text. *Journal of the American Medical Informatics Association*, 30(2):340–347.

Richard Liaw, Eric Liang, Robert Nishihara, Philipp Moritz, Joseph E Gonzalez, and Ion Stoica. 2018. Tune: A research platform for distributed model selection and training. *arXiv preprint arXiv:1807.05118*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Abhilasha Ravichander, Aakanksha Naik, Carolyn P. Rosé, and Eduard H. Hovy. 2019. EQUATE: A benchmark evaluation framework for quantitative reasoning in natural language inference. *CoRR*, abs/1901.03735.

Alex Wang, Amanpreet Singh, Julian Michael, et al. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.