T5-Medical at SemEval-2024 Task 2: Using T5 Medical Embeddings for Natural Language Inference on Clinical Trial Data

Marco Siino
Department of Electrical, Electronic
and Computer Engineering
University of Catania
Italy
marco.siino@unipa.it

Abstract

In this work, we address the challenge of identifying the inference relation between a plain language statement and Clinical Trial Reports (CTRs) by using a T5-large model embedding. The task, hosted at SemEval-2024, involves the use of the NLI4CT dataset (Jullien et al., 2023a). Each instance in the dataset has one or two CTRs, along with an annotation from domain experts, a section marker, a statement, and an entailment/contradiction label. The goal is to determine if a statement entails or contradicts the given information within a trial description. Our submission consists of a T5-large model pre-trained on the medical domain. Then, the pre-trained model embedding output provides the embedding representation of the text. Eventually, after a fine-tuning phase, the provided embeddings are used to determine the CTRs' and the statements' cosine similarity to perform the classification. On the official test set, our submitted approach is able to reach an F1 score of 0.63, and a faithfulness and consistency score of 0.30 and 0.50 respectively.

1 Introduction

In experimental medicine, clinical trials are essential because they verify the effectiveness and safety of novel treatments (Giaccone, 2002). Clinical Trial Reports (CTRs) are documents that describe the design and outcomes of a clinical trial and are used to direct patient interventions that are specific to them. But with over 400,000 published CTRs and more coming out each year (Bastian et al., 2010), it is not feasible to manually conduct thorough reviews of all the pertinent literature while developing new treatment procedures. For these reasons, the requirement for technologies that can automatically extract and classify information is always expanding.

With the development of machine and deep learning architectures in recent years, there has been a surge in interest in natural language processing, or NLP. Many efforts have gone into creating algorithms that can automatically identify and categorize text information that is accessible on the internet. In the literature, to perform text classification tasks, several strategies have already been proposed. In the last fifteen years, some of the most successful ones have been based on SVM (Colas and Brazdil, 2006; Croce et al., 2022), on Convolutional Neural Network (CNN) (Kim, 2014; Siino et al., 2021), on Graph Neural Network (GNN) (Lomonaco et al., 2022), on ensemble models (Miri et al., 2022; Siino et al., 2022) and, recently, on Transformers (Vaswani et al., 2017; Siino et al., 2022b).

For example, to address the CTR proposed task, and to enable a higher degree of accuracy and efficiency in individualized evidence-based treatment, Natural Language Inference (NLI) (MacCartney, 2009) provides a viable solution for the large-scale interpretation and retrieval of medical evidence (Sutton et al., 2020). SemEval-2024 Task 2 – Multi-Evidence Natural Language Inference for Clinical Trial Data (NLI4CT) (Jullien et al., 2024) – relies on the NLI4CT dataset[1]. The task is to determine the inference relation between a natural language statement, and a CTR. Inference chains in this drop-off range have to be constructed for a significant fraction of the NLI4CT dataset instances. Furthermore, inference on NLI4CT requires quantitative and numerical reasoning. Research has demonstrated that transformer-based models rely on flimsy heuristics for predictions instead of consistently applying this kind of reasoning (Helwe et al., 2021).

To develop our model, we thought of a two-stage architecture. In the first stage, we used a Sentence Transformer specifically trained on the medical domain. On the generated embeddings, we evaluated a cosine similarity to predict the entailment or con-

[1] https://github.com/ai-systems/nli4ct

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 40–46
June 20-21, 2024 ©2024 Association for Computational Linguistics

tradiction relationship between the two sentences analyzed.

The remainder of the paper is structured as follows. We give some background information on Task 2 hosted at SemEval-2024 in Section 2. Section 3 offers an explanation of the submitted approach. We describe the experimental setup to reproduce our work in Section 4. The outcomes of the formal assignment and certain debates are given in Section 5. We provide our conclusion and suggestions for further research in section 6.

We make all the code publicly available and reusable on GitHub[2].

2 Background

We give some background information on Task 2 hosted at SemEval-2024 in this section. The task is predicated on a set of CTRs, statements, labels, and explanations related to breast cancer that have been annotated by domain experts.

The gathered CTRs are compiled into four components for the textual entailment task:

- *Eligibility criteria* — A list of requirements that patients must meet in order to participate in the clinical trial;
- *Intervention* — Details about the type, strength, frequency, and length of the treatments under investigation;
- *Results* — Units, outcome measures, number of trial participants, and results;
- *Adverse events* — These are the symptoms and indicators that the patients had throughout the clinical study.

With an average length of 19.5 tokens, the annotated statements are sentences that make a claim regarding the data presented in one of the CTR premise's sections. The remarks could compare two CTRs or make assertions about a single CTR. Finding the inference relation (entailment vs. contradiction) between CTR is the problem at hand. The training set provided is identical to the training set used in previous tasks (Jullien et al., 2023b), however, the organizers have performed a variety of interventions on the test set and development set statements, either preserving or inverting the entailment relations. The technical details adopted

Figure 1: A sample from the official webpage. Given two trials and a section description, a model has to predict if there is entailment or contradiction with regard to the statement provided.

to perform the interventions were not disclosed, to guarantee fair competition and in the interest of encouraging approaches that are robust and not simply designed to tackle these interventions.

An example is shown in the Figure 1 and is provided in the official task webpage available online[3].

Even if it has already been proved that the Transformers are not necessarily the best option for any text classification task (Siino et al., 2022a), depending on the goal some strategies like domain-specific fine-tuning (Sun et al., 2019; Van Thin et al., 2023), or data augmentation (Lomonaco et al., 2023; Mangione et al., 2022; Siino et al., 2024a) can be beneficial for the considered task.

The training and practice test sets were made available by the task organizers prior to the competition's official commencement. The gold labels were supplied for both sets. Participants could build and test their models during the first phase, called the *practice phase*, by uploading their predictions to CodaLab[4]. The second step, known as the *evaluation phase*, began with the release of the unlabeled test set.

3 System Overview

The rising use of Transformer-based architectures in the literature, has been supported also by several approaches presented at SemEval 2024. These approaches address very different tasks, obtaining interesting results. For example, in the case of the Task 1, where the semantic textual relatedness is evaluated using MPNet (Siino, 2024a), or in the case of the Task 4, where a Mistral 7B model is used for detecting persuasion techniques in meme

[2] https://github.com/marco-siino/SemEval2024/tree/main/Task%202

[3] https://sites.google.com/view/nli4ct/semeval-2024/dataset-description

[4] https://codalab.lisn.upsaclay.fr/competitions/16190

(Siino, 2024c), or, eventually, as in the case of the Task 8, where a DistilBERT model is employed to detect machine-generated text (Siino, 2024b). To develop our model, we also take advantage from a Transformer architecture, creating a two-stage pipeline. In the first stage, we used a *Sentence Transformer* specifically trained on the medical domain. This is a Python framework to create cutting-edge sentence, text, and image embeddings. The initial work is described in (Reimers and Gurevych, 2019). More than 100 languages have sentences and text embeddings that can be computed using this method. Sentences with a similar meaning can subsequently be found by comparing these embeddings, for example, using cosine-similarity. Semantic search, paraphrase mining, and semantic textual similarity can all benefit from this. The framework offers a huge selection of pre-trained models suited for different tasks and is built on PyTorch and Transformers. Moreover, fine-tuning models is also feasible.

The model used as Sentence transformer is T5-large-medical, and it is available on *Hugging Face*[5]. The base model is T5 (Raffel et al., 2020). Specifically, sentences and paragraphs are mapped to a dense vector space of 768 dimensions. PyTorch was used to convert the TensorFlow model st5-large-1 to this one. While the TFHub model and this PyTorch model can provide somewhat different embeddings, they yield the same results when applied to the same benchmarks.

The model was used to map all the words present in the text to the domain-specific embedding. Following the embeddings of the primary section and the statement, the cosine similarity between the two was calculated. In the case of presence of a secondary section, the operation was also carried out between the secondary section and the statement. The cosine similarity between the two embedding vectors is calculated as shown in the Equation 1.

$$\cos(\theta) = \frac{A \cdot B}{\|A\|_2 \, \|B\|_2} \tag{1}$$

In the first case, if the cosine similarity was greater than 0.5, the label of entailment was assigned, vice versa that of contradiction. In the second case, before calculating the cosine similarity, the average between the cosine similarity score between the two sections and the statement was calculated. Our code is available online together with the predictions generated and sent in relation to the test set.

As noted in the recent study by (Siino et al., 2024b), the contribution of preprocessing for text classification tasks is generally not impactful when using Transformers. More specifically, the best combination of preprocessing strategies does not provide relevant improvements compared to not performing any preprocessing when using Transformers. For these reasons, and to keep our system faster and computationally light, we have not performed any preprocessing on the text.

4 Experimental Setup

We implemented our model on Google Colab[6]. The library we used is Sentence Transformer. The library requires Python[7] (>= 3.8) and PyTorch[8] (>=1.11.0). The dataset provided for all the phases are available on the Official Competition page. On the basis of our preliminary experiments, we found beneficial to set the threshold value for the cosine similarity equal to 0.5. We did perform additional fine-tuning on the T5 embedding. To run the experiment, a T4 GPU from Google has been used. After the generation of the predictions, we exported the results on the JSON format required by the organizers. As already mentioned, all of our code is available on GitHub.

5 Results

For the task the official metric used were F1 (also known as balanced F-score or F-measure), Faithfulness and Consistency.

The F1 score can be described as the harmonic mean of the precision and recall, with a maximum score of 1 and a minimum score of 0. Recall and precision both contribute equally to the F1 score in terms of relative importance. Equation 2 shows the formula for the F1 score.

$$F1 = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{2}$$

Faithfulness is a measure of the extent to which a given system arrives at the correct prediction for the correct reason. Intuitively, this is estimated by measuring the ability of a model to correctly change its predictions when exposed to a semantic-altering intervention. Given N statements x_i in the

[5] https://huggingface.co/sentence-transformers/sentence-t5-large

[6] https://colab.research.google.com/

[7] https://www.python.org/

[8] https://pytorch.org/

	F1	Faith	Const
T5-large-medical	0.63	0.30	0.50

Table 1: The suggested method's performance on the test set. In the table, the words *Faith* and *Const* stand out for *Faithfulness* and *Consistency*

contrast set (C), their respective original statements y_i, and model predictions *f()* faithfulness can be computed using Equation 3.

$$Faithfulness = \frac{1}{N}\sum_{n=1}^{N}|f(y_i) - f(x_i)| \quad (3)$$

Consistency is a measure of the extent to which a given system produces the same outputs for semantically equivalent problems. Therefore, consistency is measured as the ability of a system to predict the same label for original statements and contrast statements for semantic preserving interventions. That is, even if the final prediction is incorrect, the representation of the semantic phenomena is consistent across the statements. Given N statements x_i in the contrast set (C), their respective original statements y_i, and model predictions *f()* we compute consistency using Equation 4.

$$Consistency = \frac{1}{N}\sum_{n=1}^{N}1 - |f(y_i) - f(x_i)| \quad (4)$$

In Table 1, the results obtained using the three metrics on the official test set are shown. Considered the very low effort required to run the proposed approach and to generate the predictions, the F1 score of 0.63 appears to be an interesting baseline, while consistency and faithfulness exhibit a very large room for improvements using the proposed approach. It is worth noticing that the approach is a Zero-Shot one with no prior knowledge on the specific task.

In the Table 2, the results obtained by the first three teams and by the last one, as showed on the official CodaLab page, are reported. Compared to the best performing models, our simple approach exhibits some room for improvements. However, it is worth notice that our proposed approach do not require any further pre-training and the computational cost to address the task is manageable with the free online resources offered by Google Colab. We performed few interventions to assess the setup of our approach. For example, we evaluated the number of the epochs to use for fine-tuning the Transformer embedding, the number of warm up steps and the train loss to use. All the details that led our model to reach its final performance, can be deducted from our code available on GitHub.

	F1	Faith	Const
dodoodo (1)	0.78	0.92	0.81
aryopg (2)	0.78	0.95	0.78
jvl (3)	0.78	0.80	0.77
MJ2301 (32)	0.47	0.44	0.47

Table 2: Comparing performance on the test set. In the table are shown the results obtained by the first three users and by the last one. In parentheses is reported the position in the official ranking.

6 Conclusion

This paper presents the application of T5-large model embedding for addressing the Task 2 at SemEval-2024. For our submission we decided to follow an easy Zero-Shot learning approach, employing as-is, an in-domain pre-trained Transformer. After getting the contextual embedding provided by the Sentence Transformer, we made use of a cosine similarity to calculate the similarity between sentences and generate the entailment/contradiction labels. The task is challenging, and there is still opportunity for improvement, as can be noted looking at the final ranking. Possible alternative approaches include utilizing the zero-shot capabilities of models like GPT, increasing the size of the training set by using further data, or directly integrating ontology-based domain knowledge differently than what has been proposed in our work. To assess the effect of biomedical pre-training on MLMs, performance consistency between sections, generalization capacity of models trained on NLI4CT, performance comparability between numerical and biomedical cases and further error analysis is required. Furthermore, given the interesting results recently provided on a plethora of tasks, also few-shot learning (Wang et al., 2023; Maia et al., 2024; Siino et al., 2023; Meng et al., 2024) or data augmentation strategies (Muftie and Haris, 2023; Tapia-Téllez and Escalante, 2020; Siino and Tinnirello, 2023) could be employed to improve the performance. Eventually, an optimal threshold learnt from the validation dataset could be also employed in future works, in place of the fixed one that we used in this study. Compared to the best performing models, our simple approach exhibits some room for improvements. However, it is worth to notice that the proposed approach required no further pre-training and the computational cost to address the task is manageable with the free online resources offered by Google Colab.

Acknowledgments

We would like to thank anonymous reviewers for their comments and suggestions that have helped to improve the presentation of the paper.

References

Hilda Bastian, Paul Glasziou, and Iain Chalmers. 2010. Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS medicine*, 7(9):e1000326.

Fabrice Colas and Pavel Brazdil. 2006. Comparison of svm and some older classification algorithms in text classification tasks. In *IFIP International Conference on Artificial Intelligence in Theory and Practice*, pages 169–178. Springer.

Daniele Croce, Domenico Garlisi, and Marco Siino. 2022. An SVM ensemble approach to detect irony and stereotype spreaders on twitter. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022*, volume 3180 of *CEUR Workshop Proceedings*, pages 2426–2432. CEUR-WS.org.

Giuseppe Giaccone. 2002. Clinical impact of novel treatment strategies. *Oncogene*, 21(45):6970–6981.

Chadi Helwe, Chloé Clavel, and Fabian M. Suchanek. 2021. Reasoning with transformer-based models: Deep learning, but shallow reasoning. In *3rd Conference on Automated Knowledge Base Construction, AKBC 2021, Virtual, October 4-8, 2021*.

Maël Jullien, Marco Valentino, and André Freitas. 2024. SemEval-2024 task 2: Safe biomedical natural language inference for clinical trials. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*. Association for Computational Linguistics.

Maël Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Dónal Landers, and André Freitas. 2023a. NLI4CT: multi-evidence natural language inference for clinical trial reports. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 16745–16764. Association for Computational Linguistics.

Maël Jullien, Marco Valentino, Hannah Frost, Paul O'regan, Donal Landers, and André Freitas. 2023b. SemEval-2023 task 7: Multi-evidence natural language inference for clinical trial data. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 2216–2226, Toronto, Canada. Association for Computational Linguistics.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751. ACL.

Francesco Lomonaco, Gregor Donabauer, and Marco Siino. 2022. COURAGE at checkthat!-2022: Harmful tweet detection using graph neural networks and ELECTRA. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022*, volume 3180 of *CEUR Workshop Proceedings*, pages 573–583. CEUR-WS.org.

Francesco Lomonaco, Marco Siino, and Maurizio Tesconi. 2023. Text enrichment with japanese language to profile cryptocurrency influencers. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023*, volume 3497 of *CEUR Workshop Proceedings*, pages 2708–2716. CEUR-WS.org.

Bill MacCartney. 2009. *Natural language inference*. Ph.D. thesis, Stanford University, USA.

Beatriz Matias Santana Maia, Maria Clara Falcão Ribeiro de Assis, Leandro Muniz de Lima, Matheus Becali Rocha, Humberto Giuri Calente, Maria Luiza Armini Correa, Danielle Resende Camisasca, and Renato Antonio Krohling. 2024. Transformers, convolutional neural networks, and few-shot learning for classification of histopathological images of oral cancer. *Expert Systems with Applications*, 241:122418.

Stefano Mangione, Marco Siino, and Giovanni Garbo. 2022. Improving irony and stereotype spreaders detection using data augmentation and convolutional neural network. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022*, volume 3180 of *CEUR Workshop Proceedings*, pages 2585–2593. CEUR-WS.org.

Zong Meng, Zhaohui Zhang, Yang Guan, Jimeng Li, Lixiao Cao, Meng Zhu, Jingjing Fan, and Fengjie Fan. 2024. A hierarchical transformer-based adaptive metric and joint-learning network for few-shot rolling bearing fault diagnosis. *Measurement Science and Technology*, 35(3).

Mohsen Miri, Mohammad Bagher Dowlatshahi, Amin Hashemi, Marjan Kuchaki Rafsanjani, Brij B Gupta, and W Alhalabi. 2022. Ensemble feature selection for multi-label text classification: An intelligent order statistics approach. *International Journal of Intelligent Systems*, 37(12):11319–11341.

Fuad Muftie and Muhammad Haris. 2023. Indobert based data augmentation for indonesian text classification. In *2023 International Conference on Information Technology Research and Innovation, ICITRI 2023*, page 128 – 132.

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.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Marco Siino. 2024a. All-mpnet at semeval-2024 task 1: Application of mpnet for evaluating semantic textual relatedness. In *Proceedings of the 18th International Workshop on Semantic Evaluation*, SemEval 2024, Mexico City, Mexico.

Marco Siino. 2024b. Badrock at semeval-2024 task 8: Distilbert to detect multigenerator, multidomain and multilingual black-box machine-generated text. In *Proceedings of the 18th International Workshop on Semantic Evaluation*, SemEval 2024, Mexico City, Mexico.

Marco Siino. 2024c. Mcrock at semeval-2024 task 4: Mistral 7b for multilingual detection of persuasion techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation*, SemEval 2024, Mexico City, Mexico.

Marco Siino, Elisa Di Nuovo, Ilenia Tinnirello, and Marco La Cascia. 2021. Detection of hate speech spreaders using convolutional neural networks. In *Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st - to - 24th, 2021*, volume 2936 of *CEUR Workshop Proceedings*, pages 2126–2136. CEUR-WS.org.

Marco Siino, Elisa Di Nuovo, Ilenia Tinnirello, and Marco La Cascia. 2022a. Fake news spreaders detection: Sometimes attention is not all you need. *Information*, 13(9):426.

Marco Siino, Marco La Cascia, and Ilenia Tinnirello. 2022b. Mcrock at semeval-2022 task 4: Patronizing and condescending language detection using multi-channel cnn, hybrid lstm, distilbert and xlnet. In *Proceedings of the 16th International Workshop on Semantic Evaluation, SemEval@NAACL 2022, Seattle, Washington, United States, July 14-15, 2022*, pages 409–417. Association for Computational Linguistics.

Marco Siino, Francesco Lomonaco, and Paolo Rosso. 2024a. Backtranslate what you are saying and i will tell who you are. *Expert Systems*, n/a(n/a):e13568.

Marco Siino, Maurizio Tesconi, and Ilenia Tinnirello. 2023. Profiling cryptocurrency influencers with few-shot learning using data augmentation and ELECTRA. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023*, volume 3497 of *CEUR Workshop Proceedings*, pages 2772–2781. CEUR-WS.org.

Marco Siino and Ilenia Tinnirello. 2023. Xlnet with data augmentation to profile cryptocurrency influencers. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023*, volume 3497 of *CEUR Workshop Proceedings*, pages 2763–2771. CEUR-WS.org.

Marco Siino, Ilenia Tinnirello, and Marco La Cascia. 2022. T100: A modern classic ensemble to profile irony and stereotype spreaders. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022*, volume 3180 of *CEUR Workshop Proceedings*, pages 2666–2674. CEUR-WS.org.

Marco Siino, Ilenia Tinnirello, and Marco La Cascia. 2024b. Is text preprocessing still worth the time? A comparative survey on the influence of popular preprocessing methods on transformers and traditional classifiers. *Information Systems*, 121:102342.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18*, pages 194–206. Springer.

Reed T. Sutton, David Pincock, Daniel C. Baumgart, Daniel C. Sadowski, Richard N. Fedorak, and Karen I. Kroeker. 2020. An overview of clinical decision support systems: benefits, risks, and strategies for success. *npj Digit. Medicine*, 3.

José Medardo Tapia-Téllez and Hugo Jair Escalante. 2020. Data augmentation with transformers for text classification. In *Advances in Computational Intelligence*, pages 247–259, Cham. Springer International Publishing.

Dang Van Thin, Duong Ngoc Hao, and Ngan Luu-Thuy Nguyen. 2023. Vietnamese sentiment analysis: An overview and comparative study of fine-tuning pretrained language models. *ACM Transactions on Asian and Low-Resource Language Information Processing*, 22(6):1–27.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Xixi Wang, Xiao Wang, Bo Jiang, and Bin Luo. 2023. Few-shot learning meets transformer: Unified query-support transformers for few-shot classification. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(12):7789–7802.