YNU-HPCC at SemEval-2023 Task7: Multi-evidence Natural Language Inference for Clinical Trial Data based on a BioBERT Model

Chao Feng, Jin Wang and Xuejie Zhang
School of Information Science and Engineering
Yunnan University
Kunming, China
fengchao@mail.ynu.edu.cn, {wangjin, xjzhang}@ynu.edu.cn

Abstract

This paper describes the system for the YNU-HPCC team in subtask 1 of the SemEval-2023 Task 7: Multi-evidence Natural Language Inference for Clinical Trial Data (NLI4CT). This task requires judging the textual entailment relationship between the given CTR and the statement annotated by the expert annotator. This system is based on the fine-tuned Bi-directional Encoder Representation from Transformers for Biomedical Text Mining (BioBERT) model with supervised contrastive learning and back translation. Supervised contrastive learning is to enhance the classification, and back translation is to enhance the training data. Our system achieved relatively good results on the competition's official leaderboard. The code of this paper is available at https://github.com/facanhe/SemEval-2023-Task7.

1 Introduction

With the massive increase of clinical trial reports (CTRs) publications, there are more than 10000 clinical trial reports for breast cancer at present. Therefore, it is not feasible for clinical practitioners to timely understand all current literature, so as to provide personalized evidence-based nursing (DeYoung et al., 2020). In this context, the application of natural language inference (NLI) can promote the development of large-scale interpretation and retrieval of medical evidence, and the successful development of the system can greatly promote us to combine the latest evidence to support personalized care (Sutton et al., 2020). To this end, the purpose of this task is to use clinical trial data to conduct multiple evidence natural language inference.

The task is based on a group of clinical trial reports of breast cancer, statements, explanations, and labels annotated by domain expert annotators. Each clinical trial report data of patients consists of four sections: (1) Eligibility criteria: a series of conditions that allow patients to participate in the clinical trial. (2) Intervention: information about the type, dose, frequency, and duration of the treatment studied. (3) Results: number of participants in the trial, result measurement, unit, and results. (4) Adverse Events: physical signs and symptoms of patients observed during the clinical trial. An example of this task is shown in Table 1.

The SemEval-2023 shared task7 consists of two subtasks (Jullien et al., 2023):

- subtask 1: judgment of the reasoning relationship between the clinical trial report and the statement (contradiction or entailment);
- subtask 2: given a set of clinical trial reporting premises and statements, extract the supporting facts from the premises;

The main difficulty of subtask 1 is textual entailment. Textual entailment describes the reasoning relationship between two texts, in which text 1 is the premise and text 2 is the hypothesis. If the premise can infer the hypothesis, it is an entailment, otherwise, it is a contradiction. The task of textual entailment can also be regarded as the problem of text classification(Kong et al., 2022). The early task of textual entailment is based on the method of text similarity (Jijkoun et al., 2005) or text alignment (De Marneffe et al., 2008). With the development and expansion of deep learning, convolutional neural networks (CNN) (Kim, 2014) and recurrent neural networks (RNN) (Wang, 2018) have achieved good results in textual entailment. The long short-term memory (LSTM) (Rocktäschel et al., 2016) and application of the attention mechanism have also achieved better results than the CNN and RNN. However, since the introduction of the pre-trained model Bi-directional Encoder Representations from Transformers (BERT) (Devlin et al., 2019), it has raised the task of textual entailment to a new level.

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 664–670

Premise	Hypothesis	Label
A soccer game with multiple males playing	Some men are playing a sport	Entailment
A black race star car starts up in front of a crowd of people	A man is driving down a lonely road	Contradiction

Table 1: Examples of textual entailment.

To investigate the medical entailment relations, this study proposed to use the BioBERT model with supervised contrastive learning for subtask 1. Different from the conventional BERT encoder, BioBERT (Lee et al., 2020) was trained on Biomedical texts, thus can be more beneficial for the task of clinical text analysis. For the limitation of input length, different truncation strategies were conducted. In addition to the conventional cross-entropy loss function, supervised contrastive learning (Gunel et al., 2020) was further introduced to enhance the classification. In each training batch, the samples with the same label are pulled together while the ones with different labels are pushed away in the semantic spaces for representation learning. On the original training data, back translation is used to enhance the data without changing the semantics, through this method, we doubled the training data. In addition, in order to obtain better results, the system explored another medical text processing model Bio_ClinicalBERT (Alsentzer et al., 2019), and the system also explored using the Longformer (Beltagy et al., 2020) model to deal with long text, but unfortunately, the experimental effect did not reach the expected effect, perhaps because the individual did not adjust the appropriate parameters.

Empirical experiments were conducted on the developing set to select the optimal solution for the final submission. The results show that BioBERT + Supervised Contrastive Learning + Back Translation has achieved the best performance of 0.665 in terms of the F_1-$score$. The final submission for the test set has achieved 0.679 and ranked 12^{th} place in subtask 1.

The remainder of this paper is organized as follows. Section 2 describes the model and method used in our system, Section 3 discusses the results of the experiments, and finally, the conclusions are drawn in Section 4.

2 System Description

This section will describe the architecture of the proposed model in detail. There are several components in this section, including the tokenizer, the pre-trained model BioBERT, supervised contrastive learning, and back translation. The system model we proposed is shown in Figure 1.

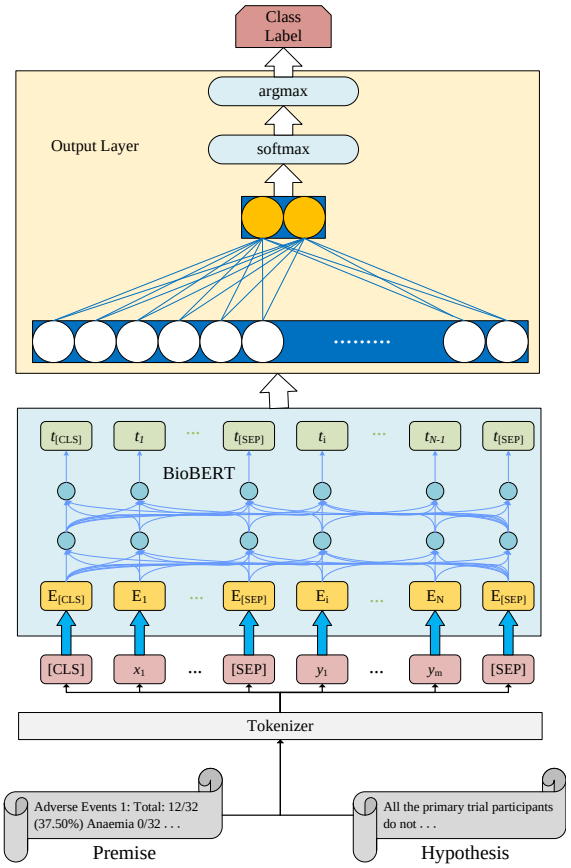

Figure 1: The structure of the system

2.1 Tokenizer

In many NLP tasks, the original text needs to be processed into digital data before it can be processed by computer.Thus, Tokenizer was applied to divide the text into words and convert it into unique coding. In the proposed model, the Bert tokenizer is used to build word vectors with a length of 512, which uses WordPiece to split the text into tokens. The final output X of the tokenizer is denoted as:

$$X = [CLS]x_1x_2...x_n[SEP]y_1y_2...y_m[SEP] \quad (1)$$

where n and m represent the length of the first

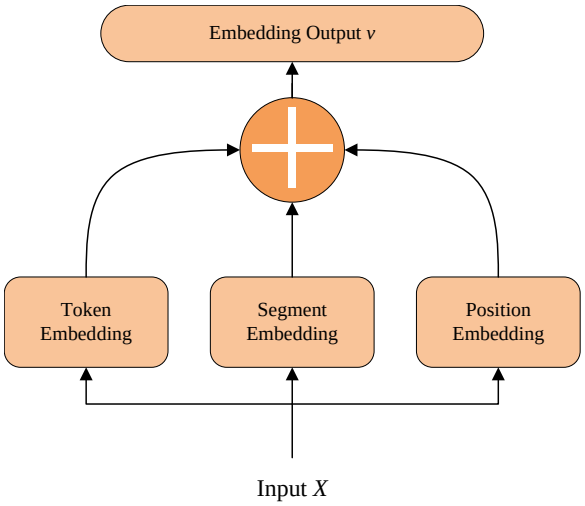

Figure 2: Embedding block

sentence and the second sentence; [CLS] special mark indicating the beginning of the text sequence; [SEP] indicates the separator between text sequences. The Tokenizer has different processing methods for sentences of different lengths. If the total length is less than 512, it uses zero to padding, and if the length is more than 512, it truncates.

2.2 BioBERT Model

BioBERT is a domain-specific BERT model based on the biomedical corpus. It uses the parameter weights of BERT-Base, and also uses PubMed Abstracts and PMC Full-text articles in the biomedical field as training data, which is much better than Bert in various biomedical text processing tasks. Therefore, the pre-trained model BioBERT was applied for subtask 1, it is based on the Transformer library [1]. The structure of BioBERT is basically the same as BERT, which includes two core blocks: Embedded block and TransformerEncoder block. The main parameters of the model used in our experiment: 12 layers, 768 dimensions, 12 self-attention heads, and 109M total parameters.

Embedding block. After the original text is processed, it will first pass through this module. Its structure is shown in Figure 2. It consists of three blocks: Token embedding, which converts words into fixed dimension vectors; Segment embedding to distinguish the block of the current word; Position embedding indicates the absolute position of each word(Zhang et al., 2021).

Transformer Encoder block. This module is composed of multi-layer TransformerEncoderLayer. Each TransformerEncoderLayer is composed of a multi-headed self-attention layer and a feed-forward layer, denoted as:

$$\begin{aligned} &Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \\ &MultiHead(Q,K,V) = Conact(head_1,...,head_h)W^0 \\ &where\, head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \tag{2}$$

where Q, K, V is the embedding multiplication of tokens with three different initializing weight matrices and then initializing different Q, K, V by linear projection, the multi-head results are fused into a multi-head attention layer.

In the BioBERT model, the embedding input representation v is encoded by a multi-layer transformer, and the semantic association between each word in the sentence is fully learned with the help of the self-attention mechanism, and the contextual semantic representation h of the sentence is finally obtained. The embedding input representation v and the contextual semantic representation h are denoted as follows.

$$\begin{aligned} v &= Inputrepresentation(X) \\ h &= BioBERT(v) \end{aligned} \tag{3}$$

2.3 Output Layer

The BERT model has two major pre-training tasks: mask language model (MLM) and next sentence prediction (NSP), and the text implication task usually uses the NSP method to predict, that is, use the hidden layer representation of [CLS] bits to predict the text classification(Ma et al., 2021). [CLS] is the first element of the input sequence, and its hidden layer representation is composed of h_0, which represents the first component of h in context semantics. After obtaining the hidden layer representation h_0 of [CLS] bit, the text label corresponding to the input text is predicted through the fully connected layer. For the probability distribution P of the text label, $W_0 \in \mathrm{R}^{d\times k}$ represents the weight of the fully connected layer, h_0 represents the offset of the fully connected layer, and k represents the number of classification labels. After obtaining the classification probability distribution P, calculate the loss with the real classification label y and learn the model weight. The calculation formula of the probability distribution is as follows.

$$P = Softmax(h_0 W^0 + b^0) \tag{4}$$

[1] https://huggingface.co/dmis-lab/biobert-v1.1

2.4 Methods

Supervised Contrastive Learning (SCL) takes samples of the same kind of labels as positive samples and different ones as negative samples. This idea is used in the fine-tuning part, in addition to using cross-entropy loss, we add contrastive learning loss on the classification task to make the same type of samples as close as possible and different types of samples as far as possible. For a multi-class classification problem with C classes, we use the training examples of size N, $\{\mathrm{x}_i,\ y_i\}_{i=1\ldots N}$, $\Phi(\cdot) \in \mathrm{R}^d$ represents an encoder that outputs l_2 normalized final encoder hidden layer before softmax projection; $y_{i,c}$ is the label, P is the probability that the ith example belongs to class c, τ is the parameter that controls the separation of classes, and α is the scalar-weighted hyperparameter that sets the tuning for downstream tasks. The overall loss is then given in the following:

$$
\begin{aligned}
L_{CE} &= -\tfrac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c}\cdot\log P \\
L_{SCL} &= \sum_{i=1}^{N} -\tfrac{1}{N_{y_i}-1}\sum_{j}^{N} 1_{i\neq j}1_{y_i=y_j} \\
&\quad \cdot\log\frac{\exp(\Phi(x_i)\cdot\Phi(x_j)/\tau)}{\sum_{k=1}^{N}1_{i\neq k}\exp(\Phi(x_i)\cdot\Phi(x_j)/\tau)} \\
L &= (1-\alpha)L_{CE} + \alpha L_{SCL}
\end{aligned}
\tag{5}
$$

3 Experimental Results

Datasets. The NLI4CT shared task data set is composed of normalized JSON data, the size of training set train.json sorted by expert comments is 1700, the size of developing set dev.json is 200, the size of test set test.json is 500, and the size of the used clinical experiment report set CT json is 999. The data part of the expert mark mainly includes Type and Section_id, Primary_id, Secondary_id(comparison), Statement, Label, Primary_evidence_index, and Secondary_evidence_index(comparison). Type is used to indicate the test type (comparison/single); Section_id is used to indicate which part of the CTR is the statement commented by the expert. Primary_id, Secondary_id is used to indicate the CTR of a separate or comparative trial id; Primary_evidence_index, Secondary_evidence_index is used to indicate the evidence index that proves the label is correct.

Evaluation Metrics. The subtasks of the NLI4CT shared task are evaluated using the adopted standard evaluation indicators, including Precision, Recall, and Macro F_1-$score$, the submissions of all teams are ranked according to F_1-$score$. The metrics will be calculated as follows:

$$
\begin{aligned}
Precision &= \frac{true\,positives}{true\,positives+false\,positives} \\
Recall &= \frac{true\,positives}{true\,positives+false\,negatives} \\
F_1\text{-}score &= \frac{2\times Precision\times Recall}{Precision+Recall}
\end{aligned}
\tag{6}
$$

Implementation Details. To facilitate model training and testing, the json data is firstly reorganized into traindata.csv, devdata.csv, and testdata.csv, which include uuid, label, statement, and premise. The premise is to extract the partial text of the corresponding CTR according to Section_id and type (e.g., if Section_id is results and type is single, extract the results text in CTR corresponding to Primary_id). The BERT model is firstly used as the baseline to implement the textual entailment task with processed training and developing sets. However, the result still has a lot of room for improvement. Perhaps because the text is about biomedicine, the model cannot fully capture the text semantics, so the biomedical domain model BioBERT and Bio_ClinicalBERT are used to complete this task After the use of supervised contrastive learning and back translation method and parameter tuning, we obtain better results than baseline.

However, due to the limitation of max position embeddings, the total size of words is limited to 512, including the [CLS] and [SEP]. However, the max length of the premise is 11227, and many of the hypothesis and premise texts add up to a total length of well over 512. In the above models, the length of texts over 512 are all directly truncated, which is simple but results in the loss of some semantics of the texts. In order to overcome this problem, the Longformer model is proposed for this task. However, due to GPU limitations and time constraints, the appropriate parameters were not adjusted to obtain the desired results, and there may be problems due to the wrong personal use method.

Hyper-parameters Fine-tuning. The warmup strategy is used to optimize the learning rate, which is a method for the learning rate mentioned in the ResNet (He et al., 2016) paper. In order to achieve the expected results, we adjusted different learning rates and batchsize to adapt to different models, the parameter tuning process is shown in the following

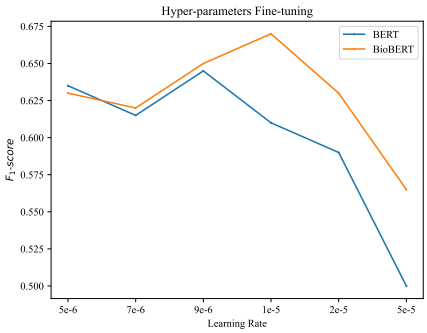

Figure 3: The performance of different learning rates on F_1-$score$

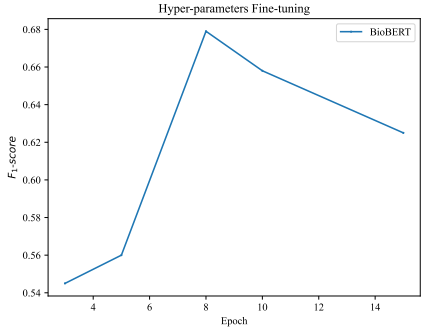

Figure 4: The performance of different epochs on F_1-$score$

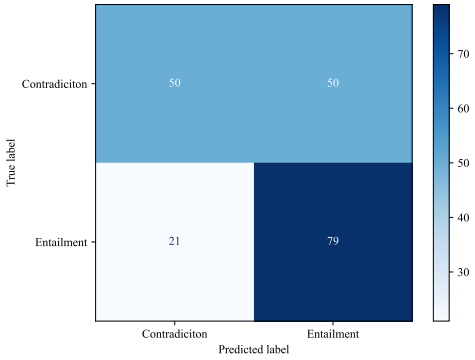

Figure 5: The predicted results of BioBERT on the development set

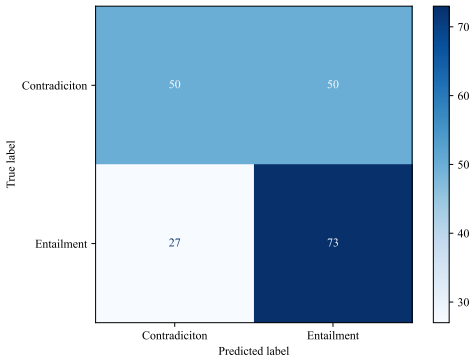

Figure 6: The predicted results of BERT on the development set

Figures 3 and 4, and set parameters in the final submitted results as follows: warmup steps is 500, weight decay is 0.01, the learning rate is 1.5e-5, train batchsize is 8, eval batchsize is 16, and epoch is 8.

Comparative Results and Discussion. The test is first carried out on the development set, whose size is 200. The true and predicted results are represented by the confusion matrix in Figures 5 and 6. Facing the different predicted results of BERT and BioBERT, it is clear that BioBERT performs better. Taking out individual examples and summarizing them in Table 2, we find that the reason for BioBERT's better performance may be word segmentation. BERT used in this experiment is the uncased version, while the BioBERT is the cased version, and the difference is the case sensitivity. For example, the word INTERVENTION is divided by BioBERT and BERT Tokenizer respectively as [' IN ', '##TE', '##R', '##VE', '##NT', '##ION'], [' intervention '], which could affect the predictions. Moreover, the medical terms in the text like the word Metastatic through the different tokenizers will be divided as ['Met', '##ast', '##atic'] and ['meta', '##static']. This has implications for the model to learn the semantics of the text in the next step.

The F_1-$score$ obtained from the experiments of several models and methods are summarized as shown in Table 3, and the result of the final submission is shown in Table 4. The F_1-$score$ of BERT is 0.541, the F_1-$score$ of BERT and Bio_ClinicalBERT using back translation are close, about 0.598, while the F_1-$score$ of BioBERT using back translation is 0.618. BioBERT performed well because it used biomedical corpus for training, while BERT only used general corpus. However, Bio_ClinicalBERT is trained using MIMIC III (a database containing electronic health records from ICU patients at the Beth Israel Hospital in Boston, MA.) on the basis of BioBERT. Its performance is inferior to that of BioBERT, possibly due to poor parameter adjustment, and the corpus is mostly about clinical surgery. The F_1-$score$ of BioBERT and BERT using supervised contrastive learning and back translation achieved 0.679 and 0.648 respectively. Obviously, the addition of supervised contrast learning makes the results of the model

Premise	Hypothesis	BERT Predicted label	BioBERT Predicted label	True label
Participants with HER2+ breast cancer received treatment as follows	Both cohorts of the primary trial undergo a total of 17 cycles, each lasting 3 weeks.	Contradiction	Entailment	Entailment
DISEASE CHARACTERISTICS:T1-3, N0-2, M0	T4 N2 M4 patients are eligible for the primary trial	Entailment	Contradiction	Contradiction
Adverse Events 1: Total: 0/23 (0.00%) Adverse Events 1: Total: 0/655 (0.00%) Adverse Events 2: Total: 0/580 (0.00%)	the primary trial and the secondary trial do not report adverse events	Contradiction	Entailment	Entailment

Table 2: Examples of different models on the dev set.

Model	Loss	F_1-*score*
BERT	CE	0.541
BERT+Back Translation	CE	0.598
BioBERT+Back Translation	CE	0.618
Bio_ClinicalBERT+Back Translation	CE	0.598
BERT+Back Translation	CE+SCL	0.648
BioBERT+Back Translation	CE+SCL	0.679

Table 3: Comparative results of experiments in the test set.

Precision	Recall	F_1-*score*
0.621	0.748	0.679

Table 4: Subtask 1 result.

better. The reason is that supervised contrastive learning brings similar labels closer, and separates different labels, which is more conducive to text classification. However, due to text length limitations, our experiment with Longformer did not work well, and there are still many areas for improvement.

4 Conclusion

In this paper, we present a system submitted in subtask 1 of the SemEval-2023 Task 7, which utilizes the pre-trained model BioBERT to adjust the official baseline model and uses the text classification task to complete the textual entailment task of CTRs. The experimental results show that our proposed system achieves good performance. In addition, in subtask 1, compared with the top-ranked team system, our system still has a lot of room for improvement. In future research, we hope to try other biomedical text-processing models or text-length processing methods to obtain better results.

5 Acknowledgement

This work was supported by the National Natural Science Foundation of China (NSFC) under Grant Nos. 61966038 and 62266051. The authors would like to thank the anonymous reviewers for their constructive comments.

References

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly Available Clinical BERT Embeddings. *arXiv preprint arXiv:1904.03323*.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The Long-Document Transformer. *arXiv preprint arXiv:2004.05150*.

Marie-Catherine De Marneffe, Anna N Rafferty, and Christopher D Manning. 2008. Finding contradictions in text. In *Proceedings of acl-08: Hlt*, pages 1039–1047.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Jay DeYoung, Eric Lehman, Benjamin E. Nye, Iain James Marshall, and Byron C. Wallace. 2020. Evidence inference 2.0: More data, better models. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing, BioNLP 2020, Online, July 9, 2020*, pages 123–132.

Beliz Gunel, Jingfei Du, Alexis Conneau, and Ves Stoyanov. 2020. Supervised Contrastive Learning for Pre-trained Language Model Fine-tuning. *arXiv preprint arXiv:2011.01403*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Valentin Jijkoun, Maarten de Rijke, et al. 2005. Recognizing textual entailment using lexical similarity. In *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*, pages 73–76. Citeseer.

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*.

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.

Jun Kong, Jin Wang, and Xuejie Zhang. 2022. Hierarchical bert with an adaptive fine-tuning strategy for document classification. *Knowledge-Based Systems*, 238:107872.

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. *Bioinform.*, 36(4):1234–1240.

Xinge Ma, Jin Wang, and Xuejie Zhang. 2021. YNU-HPCC at SemEval-2021 task 11: Using a BERT model to extract contributions from NLP scholarly articles. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 478–484, Online.

Tim Rocktäschel, Edward Grefenstette, Karl Moritz Hermann, Tomás Kociský, and Phil Blunsom. 2016. Reasoning about Entailment with Neural Attention. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

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 digital medicine*, 3(1):17.

Baoxin Wang. 2018. Disconnected recurrent neural networks for text categorization. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2311–2320.

You Zhang, Jin Wang, and Xuejie Zhang. 2021. Personalized sentiment classification of customer reviews via an interactive attributes attention model. *Knowledge-Based Systems*, 226:107135.