Enhancing the DeBERTa Transformers Model for Classifying Sentences from Biomedical Abstracts

Abdul Aziz, Md. Akram Hossain, and Abu Nowshed Chy
Department of Computer Science and Engineering
University of Chittagong, Chattogram-4331, Bangladesh
{aziz.abdul.cu, akram.hossain.cse.cu}@gmail.com,
and nowshed@cu.ac.bd

Abstract

Evidence-based medicine (EBM) is defined as making clinical decisions about individual patients based on the best available evidence. It is beneficial for making better clinical decisions, caring for patients, and providing information about the therapy, prognosis, diagnosis, and other health care issues. However, it is a challenging task to build an automatic sentence classifier for EBM owing to a lack of clinical context, uncertainty in medical knowledge, difficulty in finding the best evidence, and domain-specific words in medical articles. To address these challenges, ALTA 2022 introduced a task to build automatic sentence classifiers for EBM that can map the content of biomedical abstracts into a set of pre-defined categories. This paper presents our participation in this task where we propose a transformers-based classification approach to identify the category of the content from biomedical abstracts. We perform fine-tuning on DeBERTa pre-trained transformers model to extract the contextualized features representation. Later, we employ a multi-sample dropout strategy and 5-fold cross-fold training to predict the more accurate class labels. Experimental results show that our proposed method achieved competitive performance among the participants.

1 Introduction

Personalized medicine based on the context of primary clinical evidence has become one of the most engaging and promising tasks in biomedical research. To suggest personalized medicine, practitioners require to study a lot of publications of medical science related to patient diagnosis. This kind of study is known as evidence-based medicine (EBM) (Masic et al., 2008) where the decision is taken based on some control traits and evidence including Population (P), Intervention (I), Comparison (C), and Outcome (O), in short PICO.

To automate the EBM process, (Kim et al., 2011) explored a classification task where the sentences are collected from the medical abstracts. As an expansion of this work, ALTA 2022 [1] organized a shared task where they address the control traits as PIBOSO by the inclusion of three new classes including Background (B), Study Design (S), and Other (O) to improve the search performance. Here, Other (O) refers to the sentence with irrelevant content. To demonstrate a clear view of the task definition, we articulate a few examples in Table 1.

Sentence	Label
The aim of this non-randomized study is to evaluate a group of patients treated by VP and KP procedures and to discuss related risks.	[0 0 1 0 1 0]
We evaluated drug effect through physical examinations and symptom scales.	[0 0 0 0 0 1]

Table 1: Example of ALTA 2022 task . Here, labels are population, intervention, background, outcome, study design, and other. The 0 and 1 in the label field denotes its existence in the corresponding sentence.

However, the ambiguous clinical context, the randomness of the medical events, the uncertainty of medical knowledge, and highly domain-specific term make it difficult to automate the classification process of medical abstracts for EBM. We can consider this task as the predecessor of some other well-defined tasks in biomedical research including automatic question answering (Andrenucci, 2008). Prior work has extensively explored feature-

**The first two authors have equal contributions.

[1] http://www.alta.asn.au/events/sharedtask2022/description.html

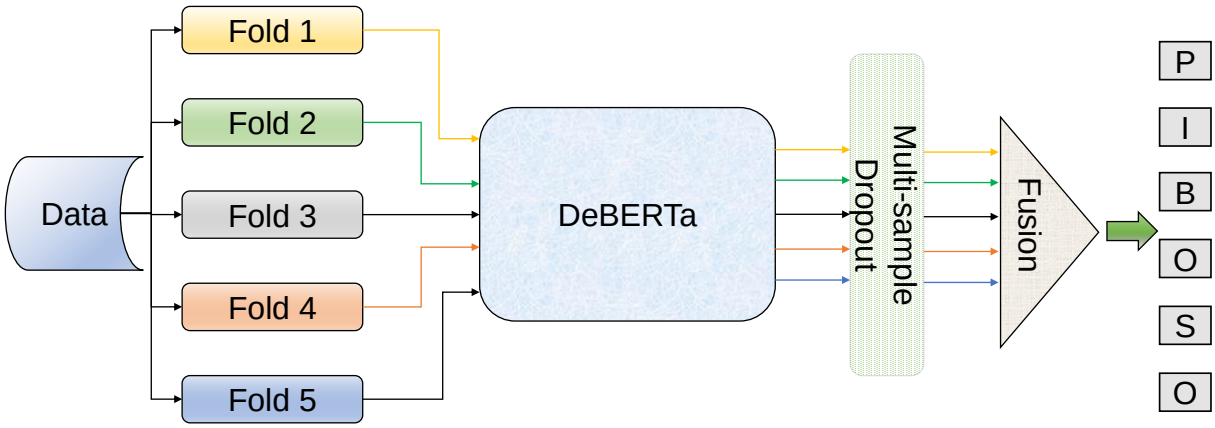

Figure 1: Overview diagram of the proposed system.

based (e.g. lexical and structural features) systems integrated with statistical machine learning (ML) algorithms including support vector machine (SVM), logistic regression, and conditional random fields (CRF) (Amini et al., 2012; Mollá et al., 2012; Sarker et al., 2013). Nevertheless, these approaches are limited to learning complex and ambiguous clinical contexts due to their scattered attention mechanism. Transformer models (Yogarajan et al., 2021) can ameliorate the performance of multi-label natural language processing (NLP) tasks in the medical domain. To overcome the limitations of the prior works and explore the advantages of the transformers model in our proposed system, we fine-tune a SOTA transformers model named DeBERTa (He et al., 2020) integrated with some additional training strategies including the multi-sample dropout and cross-fold training.

We organize the rest of the paper as follows: Section 2 describes our proposed system in the ALTA-2022 automatic labeling medical document abstract into pre-defined classes task whereas, in Section 3, we present our system design with parameter settings along with the results and performance analysis. Finally, we conclude with some future directions in Section 4.

2 Proposed Framework

Transformers models learn the necessary information about the relationship between words effectively. We employed a pre-train transformers model with different training strategies to identify the categories of content from the biomedical abstract. The overview of our proposed transformer-based framework is depicted in Figure 1

For a given biomedical text, we use the DeBERTa transformers model to extract the embedding feature vectors. We fine-tune the DeBERTa model to capture the domain-specific contexts for the biomedical sentence classification task. Later, we apply multi-sample dropout on top of the extracted feature vectors. A classification head averages the feature vectors from multi-sample dropout to predict the confidence of each class. Since a cross-fold training strategy reduce the error rates on class label prediction (Reul et al., 2018; Pikrakis and Theodoridis, 2014), we employ 5-cross-fold training to improve the prediction performances. The predictions obtained from each trained model of each fold are then averaged to determine the final prediction label.

2.1 Transformers Model

Transformers models have the ability to distill long-term dependency and improve the relationship between the words of the sentence. Thus, we fine-tuned the DeBERTa transformers model to extract the contextualized features representation of biomedical sentences.

2.1.1 DeBERTa

DeBERTa (He et al., 2020) stands for decoding-enhanced BERT with disentangled attention. It improves the BERT and RoBERTa models using disentangled attention mechanism and enhanced mask decoder. We used the enhanced version of the DeBERTa model named DeBERTaV3 (He et al., 2021). The DeBERTaV3 model used the ELECTRA style pre-training by replacing mask language modeling (MLM) with the replaced token detection (RTD) strategy where the model is trained as a discriminator to determine whether an input token is either original or replaced by a generator. It also

used the gradient-disentangled embedding sharing (GDES) method that shares the embeddings between the generators and the discriminators. However, this sharing is unidirectional where the generator shares its embeddings with the discriminator but the discriminator is restricted to backpropagating the embeddings. This improved DeBERTa model achieved significant performance on downstream tasks. Motivated by this, we employ Huggingface's (Wolf et al., 2019) implementation of *microsoft/deberta-v3-large* checkpoint to extract the feature representation of the sentences. It is composed of 24 transformer blocks, a hidden size of 1024, and 131M parameters with a vocabulary of 128K tokens in the embedding layer.

2.2 Training Strategies

Prior studies suggested different training strategies to improve the performance of the transformers model (Inoue, 2019). Following this, we use two training strategies including the multi-sample dropout and 5-fold cross-fold training.

2.2.1 Multi-sample Dropout

The multi-sample dropout-based training strategy improves the generalization ability and accelerates the training of the base model, which in turn improves the overall performance of the system (Inoue, 2019). In our proposed transformer-based model, we employ this training strategy where we use five dropout samples. Here, we basically duplicate the features vector of the transformer model after the dropout layer, while sharing the weights among these duplicated fully connected layers. To obtain the final loss, we aggregate the loss obtain from each sample and take their average.

2.2.2 Cross-fold Training

To improve the robustness of our model through reducing the error rates during the model training, we use the stratified cross-fold training strategy (Reul et al., 2018; Pikrakis and Theodoridis, 2014; Sechidis et al., 2011). It maintains the proportion of disjoint groups within a population by using samples taken from these groups. Instead of training a model using the full dataset, it basically creates several folds from the training sample and each fold is then used to train the model. It has a great impact on the hyperparameters tuning phase and effectively captures the diversity of contexts related to the task. We use 5-fold stratified multi-label cross-fold training in our method. Finally, we average the predictions attained from each fold to estimate the final prediction score of each class.

3 Experiment and Evaluation

3.1 Dataset Description

The organizers used a benchmark dataset published in DTMBio-2010 (Kim et al., 2011) to evaluate the performance of the participants' systems at the ALTA-2022 shared task. The dataset statistics are summarized in Table 2. The dataset comprises biomedical sentences taken from 1000 biomedical article abstracts. Each sentence is annotated with six categories including population (P), intervention (I), background (B), outcome (O), study design (S), and other (O).

Category	Data
Train	8216
Dev	459
Test	569
Total	9244

Table 2: The statistics of ALTA 2022 dataset.

3.2 Experimental Settings

We now describe the details of our experimental and hyper-parameter settings along with finetuning strategy that we have employed to design our proposed system for the ALTA 2022 shared task.

Parameter	Optimal Value
Learning rate	3e-5
Max-len	128
Number of epochs	5
Batch size	2
Manual seed	4
Number of fold	5
Dropout	0.1, 0.2,..., 0.5

Table 3: Model settings for ALTA-2022 shared task.

We finetune a state-of-the-art Huggingface transformers model named DeBERTa[2] for this task. We used a CUDA-enabled GPU and set the manual seed = 4 to generate reproducible results. The optimal parameter settings of our proposed model based on the development dataset are presented in

[2]https://huggingface.co/microsoft/deberta-v3-large

Team Name	ROC (micro) Score	Team Rank
CSECU-DSG (ours)	0.968750	2nd
Competitive performance of top ranked methods		
Heatwave	**0.987395**	1st
Michaelibrahim	0.963404	3rd
Necva	0.931843	4th
Dmollaaliod	0.910455	5th

Table 4: Comparative performance of our proposed method along with top-performing participants' method (ROC score; Higher is better.)

Table. We used the default settings for the other parameters. In our multi-sample dropout training, we use the dropout range of 0.1 to 0.5. Later, we concatenate the training and development data during our 5-fold cross-fold training phase.

3.3 Evaluation Measure

The ALTA 2022 shared task organizers employed a standard evaluation metric including the receiver operating characteristic (ROC) score to evaluate the participants' system. They calculate the ROC score utilizing the scikit-learn (Pedregosa et al., 2011) roc_auc_score package with micro averaging for ranking the participants' system.

3.4 Results and Analysis

In this section, we analyze the performance of our proposed CSECU-DSG system in the ALTA-2022 biomedical sentence identification shared task. The comparative performance of our proposed CSECU-DSG system on test data against other top-performing participants' systems in are presented in Table 4.

At first, we presented the result of our proposed method and then we added the system performances of other top-ranked participants. Results showed that our proposed system obtained 2nd position in the ALTA-2022 shared task. The best system Heatwave achieved 0.987395 in terms of the primary evaluation metric receiver operating characteristic (ROC) score. Our proposed system obtained a 0.968750 ROC score in the test set. In our proposed CSECU-DSG system, we perform two training strategies including cross-fold training and multi-sample dropout to train the state-of-the-art DeBERTa transformer model. It helps our proposed model to achieve this score.

To further analyze the performance of our model, we estimate the impact of our used training strategies to train the DeBERTa model. The summarized results regarding this analysis on the validation set are presented in Table 5. Here, we have seen that the multi-sample dropout technique improves the performance of the DeBERTa model by 1% while the cross-fold training improves the performance by 1.2% in terms of ROC score. This validates the effectiveness of these training strategies to improve the overall model performances.

Model	ROC Score
DeBERTa	0.95209
DeBERTa+MSD	0.96112
DeBERTa+MSD+ CFT (Ours)	0.971133

Table 5: Performance analysis of individual model used in our proposed CSECU-DSG system. MSD = Multi-sample Dropout; CFT = Cross Fold Training

4 Conclusion and Future Directions

In this paper, we present an approach to labeling a sentence into six predefined classes in medical abstracts using fine-tuned DeBERTa transformers model with various training strategies including the multi-sample dropout and cross-fold training. Experimental results demonstrated the efficacy of our DeBERTa-based proposed method, where the fusion of cross-fold variants approach helped us to obtain competitive performance and ranked 2nd in the ALTA 2022 shared task.

Further research may focus on other SOTA transformers models and a fusion of multiple models in a unified architecture can also be explored. Since the dataset is imbalanced, exploiting the weighted average fusion strategy on different models may capture better contexts for all PIBOSO classes from medical abstracts.

References

Iman Amini, David Martinez, Diego Molla, et al. 2012. Overview of the alta 2012 shared task.

Andrea Andrenucci. 2008. Automated question-answering techniques and the medical domain. *HEALTHINF (2)*, pages 207–212.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing. *arXiv preprint arXiv:2111.09543*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Hiroshi Inoue. 2019. Multi-sample dropout for accelerated training and better generalization. *arXiv preprint arXiv:1905.09788*.

Su Nam Kim, David Martinez, Lawrence Cavedon, and Lars Yencken. 2011. Automatic classification of sentences to support evidence based medicine. In *BMC bioinformatics*, volume 12, pages 1–10. BioMed Central.

Izet Masic, Milan Miokovic, and Belma Muhamedagic. 2008. Evidence based medicine–new approaches and challenges. *Acta Informatica Medica*, 16(4):219.

Diego Mollá et al. 2012. Experiments with clustering-based features for sentence classification in medical publications: Macquarie test's participation in the alta 2012 shared task.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Aggelos Pikrakis and Sergios Theodoridis. 2014. Speech-music discrimination: A deep learning perspective. In *2014 22nd European signal processing conference (EUSIPCO)*, pages 616–620. IEEE.

Christian Reul, Uwe Springmann, Christoph Wick, and Frank Puppe. 2018. Improving ocr accuracy on early printed books by utilizing cross fold training and voting. In *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, pages 423–428. IEEE.

Abeed Sarker, Diego Mollá, and Cécile Paris. 2013. An approach for automatic multi-label classification of medical sentences. In *Proceedings of the 4th International Louhi Workshop on Health Document Text Mining and Information Analysis. Sydney, NSW, Australia*.

Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. 2011. On the stratification of multi-label data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 145–158. Springer.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *arXiv preprint arXiv:1910.03771*.

Vithya Yogarajan, Jacob Montiel, Tony Smith, and Bernhard Pfahringer. 2021. Transformers for multi-label classification of medical text: an empirical comparison. In *International Conference on Artificial Intelligence in Medicine*, pages 114–123. Springer.