Automatic Classification of Evidence Based Medicine Using Transformers

Necva Bölücü, Pınar Uskaner Hepsağ
Computer Engineering Department
Adana Alparslan Türkeş Science and Technology University
Adana, Turkey
{nbolucu, puskaner}@atu.edu.tr

Abstract

The goal of the shared task is multi-label classification for biomedical records in English used for Evidence-Based Medicine. In this paper, we describe the model based on the Transformer submitted by our team *turkNLP* for the shared task. Our model achieved a Micro ROC score of ≈ 0.93 on the shared task and ranked 5^{th} in the leaderboard.

1 Introduction

The ALTA 2022 shared task[1] is a well-studied Natural Language Processing (NLP) problem which is multi-label sentence classification in biomedical field. The problem is assigning the sentences to one or more labels of the predefined 6 categories for the given dataset which is Evidence-Based Medicine (EBM) dataset presented by Kim et al. (2011).

In this paper, we as a team of *turkNLP* have taken up and proposed a deep learning model based on Transformer (Vaswani et al., 2017) to identify the queries in Evidence-Based Medicine (EBM) presented by Kim et al. (2011) for the ALTA 2022 shared task. Our model concatenates the encoder layer of the Transformer proposed by Vaswani et al (Vaswani et al., 2017) and the embedding of $[CLS]$ token of the BERT model (Kenton and Toutanova, 2019), which is used as the embedding layer of the Transformer model.

The main contribution of this paper is that we investigate the impact of the BERT model on the Transformer for multi-label classification problem. The model has shown an improvement over the Transformer model for multi-label classification, as the concatenation of the embedding of $[CLS]$ token of the BERT model captures the semantic of the whole input, while the Transformer captures the semantic of each word of the input[2].

The rest of the paper is organized as follows: We give the related work on the multi-label classification problem for EBM with the shared task dataset Kim et al. (2011) in Section 2. The proposed model for the problem is given in Section 3 and the experimental setup, results, and detailed analysis of the results are presented in Section 4. Finally, Section 5 concludes the paper with insights on the impact of the proposed model on the multi-label classification problem for EBM and possible future work.

2 Related Work

The first study classifying abstract sentences based on the PIBOSO scheme was conducted by Kim et al. (2011). The NICTA-PIBOSO dataset, the most studied dataset, was also published by Kim et al. (2011). The authors presented a Conditional Random Field (CRF) classifier with lexical (e.g., unigram, bigram), semantic (e.g., metathesaurus), structural (e.g., the position of the words), and sequential (e.g., direct and indirect dependencies on previous sentences) features to assign sentences to predefined labels.

Verbeke et al. (2012) presented a new approach based on $kLog$ (Frasconi et al., 2014), a new language for statistical relational learning with kernels. In the study, the authors extracted features such as PoS tags, lemmas, and dependency labels using BiogaphTA and GENIA dependency parser (Sagae and Tsujii, 2007) and fed them into $kLog$ for the problem. The NICTA-PIBOSO dataset was also the basis of the ALTA 2012 shared task (Amini et al., 2012). Lui (2012) extended the study of Kim et al. (2011) by adding additional features such as PoS n-grams, sentence length etc., and stack the

[1] https://codalab.lisn.upsaclay.fr/competitions/693500

[2] The code is publicly available at https://github.com/adalin16/alta-2022

features with a metalearner to combine multiple feature sets, based on an approach similar to the metalearner of Wolpert (1992). Mollá et al. (2012) presented a model consisting of two stages: (1) using K-means to cluster abstracts according to the actual sentence distribution in the abstract, (2) using clustering results in multi-label classification. Gella and Duong (2012) also used the CRF model with similar features as Kim et al. (2011). The categorization of sentences as structured and unstructured is the main difference compared with previous studies from Kim et al. (2011); Verbeke et al. (2012). If the first sentence in an abstract is a sentence ordering label, the authors categorized the abstract as structured otherwise unstructured. The categorization increased the performance of the problem compared to previous studies.

3 Methodology

Transformer model (Vaswani et al., 2017) is very popular because of its performance in NLP tasks such as sequence tagging (Tsai et al., 2019; He et al., 2020) and machine translation (Wang et al., 2019; Liu et al., 2020). Recently, there are lots of models based on the Transformer (Vaswani et al., 2017) in NLP such as BERT (Kenton and Toutanova, 2019), T5 (Raffel et al., 2020) etc. The success of the Transformer model is processing sequential data in parallel without a recurrent network instead of paying attention to the last state of the encoder, as in Recurrent Neural Networks (RNNs).

In this study, we adopted the encoder of the Transformer model (Vaswani et al., 2017) by extending the model with the pre-trained language models to perform classification by mapping the data to the EBM PIBOSO classes. The architecture of the proposed model is shown in Figure 1.

Let $D = \{S_i, m_i\}_{i=1}^{T}$ denote a set of T samples, where S_i is a sentence and m_i is the corresponding labels "population", "intervention", "background", "outcome", "study design", "other").

The words $\{w_1, w_2, \cdots, w_n\}$ of a sentence are mapped to the corresponding embeddings in the embedding layer, and the positional information E_{pos} is encoded and appended to the text representation and fed into the encoder layer, which consists L identical layers. The output of the Transformer Encoder is the mean of the output of the tokens as given below:

$$T_o = mean(t_1, t_2, \cdots, t_n) \quad (1)$$

We concatenated the output of the Transformer model and the embedding of the $[CLS]$ token as input of the classification layer as defined below.

$$o = T_o \oplus [CLS] \quad (2)$$

In the classification layer, we used the sigmoid function that squeezes the results between 0 and 1, and we used 0.5 as the threshold to convert the probabilities into classes. The formula of the layer is given in Equation 3.

$$\hat{s} = sigmoid(W \cdot o + b) \quad (3)$$

where $\hat{s}$ is the predicted result through the model, W is the weighted matrix, o is the concatenation of the Transformer model and the embedding of the $[CLS]$ token as defined in Equation 2, and b is the bias.

4 Experiments & Results

In this section, we present the details of the dataset, experimental setup, and results.

4.1 Dataset

There are several variants and extensions of the classification PICO (Kim et al., 2011). The dataset called PIBOSO was proposed by Kim et al. (2011), where the tag "comparison" was removed and three new tags "background", "study design", and "other" were added. The PIBOSO dataset has 6 categories namely "background", "population", "Intervention", "outcome", "study design", and "other". Samples taken from train set are given in Table 1.

In the dataset, sentences can have more than one label since it is a multi-label dataset. The train/dev/test size is given in Table 2 with the percentage of sentences annotated with given labels in the train set. The rows sum to more than 100% because a sentence is likely to contain more than one label. Note that "background", "outcome", and "other" received a higher percentage of labels, while "population", "intervention", and "study design" are at least annotated labels in the dataset (Kim et al., 2011).

4.2 Experimental Setting

We implemented the proposed model using the PyTorch library. The Adam optimizer (Kingma and Ba, 2014) was used with an epsilon value of $1e-8$ and the default max grad norm. The BCE loss function was used as the objective function.

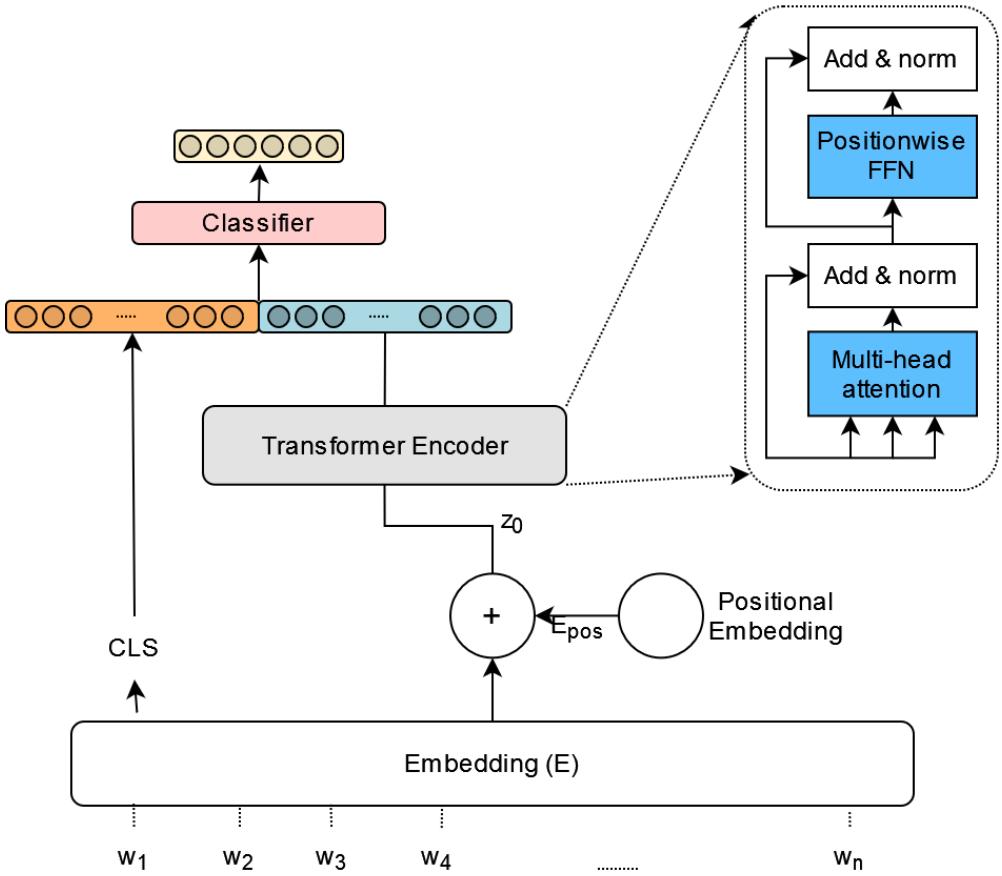

Figure 1: Overview of the architecture for multi-label classification problem

Sentence	population	intervention	background	outcome	study design	other
The response rate was 79.5%.	0	0	0	1	0	0
The average age was 71 years.	1	0	0	0	0	0
This group totaled 410 births.	0	0	0	0	0	1
All of these must be considered.	0	0	1	0	0	0
In an effort to overcome these $\cdots$	0	1	1	0	0	0

Table 1: Samples taken from train set of PIBOSO dataset (Kim et al., 2011)

Set	**Number**
Train	8216
Dev	459
Test	569
Label	**%**
population	7.11
intervention	6.10
background	21.63
outcome	38.85
study design	2.03
other	29.50

Table 2: Percentage of sentences that were annotated with a given label in the dataset

We used BERT (Kenton and Toutanova, 2019) pre-trained language model (SciBERT (Beltagy et al., 2019)[3]) to convert words into embeddings. We finetuned the model using the 0.1 of the train set of the dataset, since the labels of the development set were not revealed in the shared task. The parameters of the model are given in Table 3.

In the shared task, the evaluation metric is the area under the ROC (Receiver Operating Characteristic) curve plotting the fraction of true positives out of positives vs. the fraction of false positives out of the negatives.

4.3 Results

To understand the effect of the concatenation of the embedding of the $[CLS]$ token of the BERT model,

[3] https://huggingface.co/allenai/scibert_scivocab_cased

we conduct experiments with and without it. The Micro ROC scores are given in Table 4. The results show that using the embedding of the $[CLS]$ token improves the results of the Transformer model. The main improvement is due to the fact that embedding of $[CLS]$ token captures the semantic of the entire sentence and provides valuable complementary information for the problem.

HyperParameter	Model
learning rate	1e-4
batch size	16
d_model	1
heads	1
# of layers	1
# of hidden	1
max length	100
dropout	0.1
weight decay	0.1
patience	20

Table 3: Parameter setting of the model

Model	Micro ROC
Transformer	0.87698
Transformer+BERT	0.931843

Table 4: Test Results of the proposed model with base Transformer model

To understand the performance of the model, we generated Precision, Recall, and F_1 scores for each label in the train set of the dataset [4]. The results are shown in Table 5. It can be clearly seen that the result of the label "outcome" which has the best performance of the model. The categories "background" and "other" follow the category "outcome". The categories "population", "intervention", and "study design" show the lowest results of the proposed model. This proves that the model struggles in predicting the "population", "intervention", and "study design" categories. When analyzing the percentage of each categories given in Table 2, there is a correlation between the percentage of the categories and the results.

In Table 6, the results of the proposed model are presented with the results of the teams that participated in the ALTA 2022 shared task[5]. The best

Label	Precision	Recall	F_1
population	0.00	0.00	0.00
intervention	0.00	0.00	0.00
background	0.93	0.19	0.31
outcome	0.84	0.83	0.84
study design	0.00	0.00	0.00
other	0.98	0.60	0.75

Table 5: Precision, Recall and F_1 score for each class in the train set

Micro ROC score was obtained by `heatwave`. Our model couldn't achieve the highest score, but the result of our model is still competitive with the best result.

Team Name	Micro ROC
heatwave-2	**0.987395**
heatwave-1	0.983792
CSECU-DSG	0.968750
michaelibrahim	0.963404
turkNLP (Our model)	0.931843
dmollaaliod	0.910455

Table 6: Test Results of multi-label classification using the proposed model and the best results of the ALTA 2022 shared task

5 Conclusion

In this paper, we presented the model of Transformer model augmented with pre-trained language model (Transformer+BERT) on ALTA 2022 shared task in the English language. Experimental results showed that the Transformer+BERT model outperformed the Transformer model. We found that combining the embedding of $[CLS]$ token of the BERT model helps to capture the semantic of the whole sentence and increase the performance of the model. However, this study has also limitations. Our model couldn't perform on the labels with the lower ratio in the dataset. Labels "population", "intervention", and "study design" are difficult to identify despite the performance of the model.

In the future, further improvements can be made in sampling for multi-label classification to handle the imbalanced dataset problem.

[4]The dev and test set labels are not available, we only calculated the Micro ROC score using the `https://codalab.lisn.upsaclay.fr/competitions/6935`

[5]We couldn't compare the results with previous work (Kim et al., 2011; Gella and Duong, 2012; Lui, 2012; Verbeke et al., 2012), since the evaluation metric is different from the previous ALTA 2012 shared task (Amini et al., 2012).

References

Iman Amini, David Martinez, and Diego Molla. 2012. Overview of the ALTA 2012 Shared Task. In *Proceedings of the Australasian Language Technology Association Workshop 2012*, pages 124–129.

Iz Beltagy, Arman Cohan, and Kyle Lo. 2019. Scibert: Pretrained contextualized embeddings for scientific text. *CoRR*, abs/1903.10676.

Paolo Frasconi, Fabrizio Costa, Luc De Raedt, and Kurt De Grave. 2014. klog: A language for logical and relational learning with kernels. *Artificial Intelligence*, 217:117–143.

Spandana Gella and Long Duong. 2012. Automatic sentence classifier using sentence ordering features for event based medicine: Shared task system description. In *Proceedings of the Australasian Language Technology Association Workshop 2012*, pages 130–133.

Zhiyong He, Zanbo Wang, Wei Wei, Shanshan Feng, Xianling Mao, and Sheng Jiang. 2020. A Survey on Recent Advances in Sequence Labeling from Deep Learning Models. *CoRR*, abs/2011.06727.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

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.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization.

Xiaodong Liu, Kevin Duh, Liyuan Liu, and Jianfeng Gao. 2020. Very Deep Transformers for Neural Machine Translation. *CoRR*, abs/2008.07772.

Marco Lui. 2012. Feature stacking for sentence classification in evidence-based medicine. In *Proceedings of the Australasian Language Technology Association Workshop 2012*, pages 134–138.

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.

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

Kenji Sagae and Jun'ichi Tsujii. 2007. Dependency parsing and domain adaptation with lr models and parser ensembles. In *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL*, pages 1044–1050.

Henry Tsai, Jason Riesa, Melvin Johnson, Naveen Arivazhagan, Xin Li, and Amelia Archer. 2019. Small and Practical BERT Models for Sequence Labeling. *CoRR*, abs/1909.00100.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Mathias Verbeke, Vincent Van Asch, Roser Morante, Paolo Frasconi, Walter Daelemans, and Luc De Raedt. 2012. A statistical relational learning approach to identifying evidence based medicine categories. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 579–589.

Qiang Wang, Bei Li, Tong Xiao, Jingbo Zhu, Changliang Li, Derek F. Wong, and Lidia S. Chao. 2019. Learning Deep Transformer Models for Machine Translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1810–1822, Florence, Italy. Association for Computational Linguistics.

David H Wolpert. 1992. Stacked generalization. *Neural networks*, 5(2):241–259.