Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210125

The Classification of Short Scientific Texts Using Pretrained BERT Model

Gleb DANILOV[a,1], Timur ISHANKULOV[a], Konstantin KOTIK[a], Yuriy ORLOV[b], Mikhail SHIFRIN[a] and Alexander POTAPOV[a]
[a] *Laboratory of Biomedical Informatics and Artificial Intelligence, National Medical Research Center for Neurosurgery named after N.N. Burdenko, Moscow, Russian Federation*
[b] *Keldysh Institute of Applied Mathematics, Russian Academy of Sciences, Moscow, Russian Federation*

Abstract. Automated text classification is a natural language processing (NLP) technology that could significantly facilitate scientific literature selection. A specific topical dataset of 630 article abstracts was obtained from the PubMed database. We proposed 27 parametrized options of PubMedBERT model and 4 ensemble models to solve a binary classification task on that dataset. Three hundred tests with resamples were performed in each classification approach. The best PubMedBERT model demonstrated F1-score = 0.857 while the best ensemble model reached F1-score = 0.853. We concluded that the short scientific texts classification quality might be improved using the latest state-of-art approaches.

Keywords. Text classification, neurosurgery, machine learning, topic modeling, natural language processing, artificial intelligence

1. Introduction

Automated text classification methods could significantly facilitate scientific literature selection by researchers, especially in case of systematic reviews [1–3]. Screening articles related to a specific subject or more stringent inclusion criteria is the most time-consuming stage when analyzing the literature.

The selection of literature for a certain purpose can be addressed as a binary classification task, in which one class is represented by articles of particular interest (e.g., included in a systematic review), and the second - by papers not suitable for a specific task. We have previously used traditional machine-learning methods (logistic regression, support vector machine and random forest) on various vector text representations to solve this task on a small human-prepared dataset and achieved the best F1-score = 0.78 [4].

In the last years, pretrained large neural language models provided impressive gains in many natural language processing (NLP) tasks. This study aimed to assess the classification quality on our previously used dataset using Bidirectional Encoder Representations from Transformers (BERT) technology as the state-of-the-art approach in many NLP applications.

[1] Corresponding author, Gleb Danilov, N.N. Burdenko Neurosurgery Center, 4th Tverskaya-Yamskaya str. 16, Moscow 125047, Russian Federation; E-mail: glebda@yandex.ru.

2. Methods

The dataset for experiments was initially obtained from the PubMed search engine while performing a systematic review of artificial intelligence (AI) applications in neurosurgery in July 2019 [5,6]. The exact search query to PubMed implied a broad definition of artificial intelligence: ("neurosurgical procedures" OR "neurosurgery") AND ("artificial intelligence" OR "machine learning" OR "natural language processing" OR "NLP" OR "text mining" OR "fuzzy logic" OR "data mining" OR "big data" OR "topic model"). The publications obtained with that query were manually divided into two classes. An article was assigned to the first class (to be included in the review) in accordance with the following criteria:

- original research peer-reviewed article;
- abstract in English was available;
- the pathology/treatments discussed in the article were directly related to neurosurgery;
- the paper reported the results of AI assessment in diagnosis, treatment, prognosis, rehabilitation, or prevention.

The second class contained all other publications, which did not meet these requirements.

To accomplish our classification task we used a pre-trained PubMedBERT model modified with its parameters in 27 variants [7]. PubMedBERT model was trained from scratch using 14 million abstracts from PubMed. In our experiments an untrained layer of neurons was added upon the upper layer of the above-mentioned model and the newly shaped model was fine-tuned to classify our dataset.

We varied global parameter sets to perform fine-tuning of PubMedBERT models. The quality was evaluated after each parameter was altered. Such parameters included the number of epochs, the dropout value for the upper layer, the maximum length of PubMedBERT vectors and the batch size, decision-making threshold for the output values of PubMedBERT model, the minimum acceptable accuracy assessed on the training sample, enabling the model to be used for verification on the test sample.

The training subset was randomly sampled as 80% of the initial dataset, while the remaining 20% were divided evenly on validation and test subsets. We used automated stratification sampling provided by the sklearn python package to keep the subsets class-balanced. All the textual data were transformed into vector representations using PubMedBERT tokenizer. We applied a validation after each training epoch to evaluate model's quality. Such an experiment was repeated three hundred times to estimate the average quality of every classification approach with its own parameter set on the validation and testing subsamples.

In addition to PubMedBERT as a standalone solution we tested four ensemble models. The models in ensembles included PubMedBERT, with three others: logistic regression (LR), random forest (RF) and support vector machine (SVC). The latest three models were trained on vector representations based on counts, term frequency-inverse document frequency (TF-IDF) statistic, Word2Vec vectorizer with weighting by TF-IDF.

Totally 10 models were included in the basic ensemble, in which PubMedBERT was assigned with the maximal weight of 3/12 while the remaining nine models gained the weight of 1/12. Each model in the ensemble had two numbers in the output: the first one indicated the probability of first class assignment to a document, the second one — to the second class accordingly. We tested 4 ensemble models different in the way they summarize the output results (voting).

The first ensemble model used soft voting as a decision-making function. The soft voting algorithm calculated weighted arithmetic mean of the probabilities for each class among all the models. The highest number between two average probabilities (for two classes) indicated the final decision of the ensemble.

The second ensemble model used hard voting classifier. Hard voting changed the output probabilities of every single model to 1 if the probability was greater than 0.5 and to 0 if the probability was less or equal to 0.5. Then the algorithm calculated the sum of the scores for each class and compared them. The greater value determined the class assignment.

For the third and the fourth ensemble models, a new LR model was trained over the outputs of the models within the ensemble. The LR model in the third ensemble was trained on validation outputs of all models, while the LR model in the fourth ensemble was trained on training outputs. We considered that LR training over the ensemble outputs received on the validation dataset might provide better results than learning from the ensemble obtained on the original training set. A 5-fold cross-validation was used in these two LR models to estimate the prediction quality. Following the training and validating the results, each LR model predicted classes on the testing subset. The quality metrics of LR models' predictions were used for the evaluation of the ensembles.

3. Results

A total of 630 articles were manually assigned to the first (n = 323) and the second (n = 307) class prior to experiments. The results within each series of 300 tests were averaged for each of 27 PubMedBERT models. The average metrics (mostly routine for deep learning models) are demonstrated in Table 1. The number of epochs is shown in the "Eps" column. The dropout layer value is notated as "Dropout". Maximum vector length and batch size were changed in the only two experiments (shown in the "Len" and "Batch" columns, respectively).

The "Threshold" indicates the method and cutoff value for decision making at the prediction stage. Setting it to "torch.max" implied using the maximum value of PubMedBERT outputs for prediction. If set to numeric, then PubMedBERT outputs were compared with the threshold value and predicted the class for which the output value was higher. "TAC" ("Training accuracy") column indicates the classification accuracy on a training subset reaching the minimum value of 0.99. Thus, if the training accuracy was less than 0.99, the model was re-initialized for up to 5 times until the training accuracy exceeded 0.99. The experiments with uncontrolled training accuracy are designated by a dash.

Validation accuracy (accuracy measured on validation dataset), F1-score and the area under the receiver operating characteristic curve (ROC AUC) are referred in the columns "VAC", "F1" and "AUC" accordingly.

Table 1. Classification results of PubMedBERT with various parameter sets averaged for 300 test series.

#	Eps	Dropout	Len	Batch	Threshold	TAC	VAC	F1	AUC
1	10	0.3	512	16	2.0	>0.99	0.859	0.857	0.857
2	10	0.3	512	16	1.5	>0.99	0.852	0.854	0.855
3	10	0.3	512	16	0.7	>0.99	0.849	0.853	0.854
4	10	0.3	512	16	0.6	>0.99	0.849	0.853	0.853
5	10	0.3	512	16	2.5	>0.99	0.857	0.853	0.856
6	10	0.3	512	16	0.9	>0.99	0.852	0.852	0.852
7	10	0.3	512	16	0.8	>0.99	0.849	0.850	0.850

8	10	0.3	512	16	1.0	>0.99	0.854	0.850	0.851
9	10	0.3	512	16	0.5	>0.99	0.850	0.848	0.848
10	10	0.3	512	16	torch.max	>0.99	0.849	0.847	0.847
11	10	0.3	512	16	0.5	—	0.844	0.844	0.844
12	12	0.3	512	16	torch.max	—	0.840	0.837	0.838
13	10	0.3	512	16	torch.max	—	0.834	0.836	0.836
14	10	0.3	512	16	3.0	>0.99	0.836	0.835	0.845
15	8	0.4	512	16	torch.max	—	0.828	0.833	0.834
16	6	0.3	512	16	torch.max	—	0.835	0.832	0.832
17	4	0.3	512	16	0.8	—	0.839	0.832	0.833
18	8	0.3	512	16	torch.max	—	0.838	0.831	0.833
19	10	0.4	512	16	torch.max	—	0.834	0.830	0.831
20	4	0.3	512	16	0.5	—	0.835	0.829	0.830
21	5	0.3	512	16	torch.max	—	0.833	0.824	0.825
22	4	0.2	512	16	torch.max	—	0.825	0.822	0.824
23	3	0.2	512	16	torch.max	—	0.814	0.822	0.823
24	4	0.3	512	16	1.0	—	0.820	0.821	0.825
25	4	0.3	512	16	torch.max	—	0.826	0.818	0.819
26	4	0.3	512	32	torch.max	—	0.815	0.815	0.815
27	4	0.3	256	16	torch.max	—	0.806	0.813	0.815

The results of ensemble modeling are shown in Table 2. The validation accuracy was calculated only for ensembles with LR trained on the output data (soft-voting and hard-voting were not considered). The validation accuracy is shown in the "VAC" column. F1-score and ROC AUC score obtained on the testing subset are referred accordingly. The results within each series of 300 tests were averaged for each ensemble modeling approach. The maximum F1-score (0.853) obtained from ensemble models and reached by soft voting was less than the maximum F1-score for a single PubMedBERT model (0.857).

Table 2. Classification results by four ensemble models averaged within each test series.

#	Ensemble model	VAC	F1-score	AUC
1	Soft voting	—	0.853	0.853
2	Hard voting	—	0.805	0.806
3	LR trained on validation subset	0.841	0.846	0.846
4	LR trained on training subset	0.999	0.833	0.833

4. Discussion

The results of this study demonstrate that the quality of automated classification of scientific abstracts into user-defined classes might be improved using the latest state-of-art language models, such as pre-trained BERT [7,8]. Comparing the results of our work to those of other authors, we note a high classification quality achieved (i.e., accuracy = 0.830 reported by Simon C. et al. (2019); accuracy = 0.743 showed by Chen J. et al. (2019) on the binary classification task) [9,11].

Ensemble models with the inclusion of PubMedBERT demonstrate worse results compared to the best PubMedBERT model. Further research with other datasets could provide more comprehension on ensembles efficiency. Adding new machine learning (ML), e.g., deep learning (DL) models, to ensembles may possibly lead to better performance.

In our study, we applied the hard-set threshold values for the classification based on BERT output values. We hypothesized that a broader spectrum of decision-making rules may be applied to improve the classification quality [10]. It may be presumably effective

to train other types of models over the BERT output values. Additional statistical text features could be added to our models to enhance classification quality [12].

A crucial limitation of our work was the relatively small document size used for classifications. This study was also limited to one user-generated class-balanced dataset. Our future work will be focused on applying the above-described methods to the new user-generated datasets to estimate models' robustness. Besides, we will proceed to complement the ensemble models with new ML and DL algorithms and extend our approaches to the articles' full texts.

5. Conclusion

The classification of scientific publications by their abstracts might be to a certain extent technically solvable and provide a basis for literature tracking in user-defined tasks. The classification quality might be improved using the latest state-of-art approaches.

The research was supported by the Russian Foundation for Basic Research grant 19-29-01174.

References

[1] Buchlak QD, Esmaili N, Leveque JC, Farrokhi F, Bennett C, Piccardi M and Sethi RK. Machine learning applications to clinical decision support in neurosurgery: an artificial intelligence augmented systematic review. Neurosurg. Rev. 2019.

[2] Rios A and Kavuluru R. Convolutional neural networks for biomedical text classification: Application in indexing biomedical articles, in: BCB 2015 - 6th ACM Conf. Bioinformatics, Comput. Biol. Heal. Informatics, 2015; p. 258–267.

[3] Ambalavanan AK and Devarakonda M V. Using the contextual language model BERT for multi-criteria classification of scientific articles. J. Biomed. Inform. 2020.

[4] Danilov G, Ishankulov T, Orlov Y, Shifrin M, Kotik K and Potapov A. The classification of scientific literature for ts topical tracking on a small human-prepared dataset, in: Stud. Health Technol. Inform., 2020; p. 191–194.

[5] Danilov GV, Shifrin MA, Kotik KV, Ishankulov TA, Orlov YN, Kulikov AS and Potapov AA. Artificial intelligence in neurosurgery: A systematic review using topic modeling. part i: Major research areas. Sovrem. Tehnol. v Med. 2020;12:106–113.

[6] Danilov GV, Shifrin MA, Kotik KV, Ishankulov TA, Orlov YN, Kulikov AS and Potapov AA. Artificial intelligence technologies in neurosurgery: A systematic literature review using topic modeling. Part II: Research objectives and perspectives. Sovrem. Tehnol. v Med. 2020;12:111–118.

[7] Gu Y, Tinn R, Cheng H, Lucas M, Usuyama N, Liu X, Naumann T, Gao J and Poon H. Domain-specific language model pretraining for biomedical natural language processing. ArXiv 2020.

[8] Lee J, Yoon W, Kim S, Kim D, Kim S, So CH and Kang J. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36:1234–1240.

[9] Simon C, Davidsen K, Hansen C, Seymour E, Barnkob MB and Olsen LR. BioReader: A text mining tool for performing classification of biomedical literature. BMC Bioinformatics 2019;19.

[10] Oleynik M, Kugic A, Kasáč Z and Kreuzthaler M. Evaluating shallow and deep learning strategies for the 2018 n2c2 shared task on clinical text classification. J. Am. Med. Informatics Assoc. 2019;26:1247–1254.

[11] Chen J, Hu Y, Liu J, Xiao Y and Jiang H. Deep short text classification with knowledge powered attention. ArXiv 2019.

[12] Cheng CH and Chen HH. Sentimental text mining based on an additional features method for text classification. PLoS One 2019;14.