ISSN 2224-087X. Електроніка та інформаційні технології. 2022. Випуск 20. С. 34–44
Electronics and information technologies. 2022. Issue 20. P. 34–44

UDC 004.8 DOI: https://doi.org/10.30970/eli.21.4

# INFLUENCE OF LEARNING RATE PARAMETER IN DIFFERENT TRANSFORMER MODELS IN TEXT CLASSIFICATION TASK

B. Pavlyshenko, M. Stasiuk

*Ivan Franko National University of Lviv,*
*50 Drahomanova St., 79005 Lviv, Ukraine*
*bohdan.pavlyshenko@lnu.edu.ua, mykola.stasiuk@lnu.edu.ua*

This article is focused on the influence of the learning rate parameter on the training results of pre-trained transformer models: BERT, DistilBERT, ALBERT, and XLM-RoBERTa. As data for models training and testing dataset from HuggingFace portal is used. This dataset contains labeled data for both testing and training purposes. Moreover, it contains unlabeled data for unsupervised models and algorithms. Instead of direct training and testing, Trainer and TrainingArgument classes from the HuggingFace portal were used. For batch formation, DataColator class was utilized. Different metrics of model training efficiency were considered: learning time, the output of validation and training loss functions. Work result allows comparing the efficiency of every observed model in binary text classification tasks standalone or in assembly with other models.

*Keywords*: transformers, binary text classification, BERT, ALBERT, DistilBERT, XLM-RoBERTa.

## Introduction.

In the modern world, computers are used in every field of people’s life. Starting with computers, which are used in everyday life, continuing with devices that are utilized in modern vehicles to the big industrial companies, everywhere much automation is present. And all those devices are communicating with controlling computers or among themselves using data. Furthermore, with all new technologies being invented the amount of generated data is increasing dramatically. According to IDC [1], it was estimated that overall digital data had reached around 100 zettabytes in 2023 and will reach around 173 zettabytes by 2025. Thus the interest in algorithms and mechanisms for data analysis is increasing. In addition, it was estimated [2] that only around 20% of data is structured, while nearly 80% of data is unstructured, with text being one of the most common types of this kind of data.

In recent years different methods for data analysis have been created, beginning with simple linear models to complicated multi-layers neural networks [3, 4]. All of them have advantages and disadvantages and are utilized according to the task. For text analysis [5-13] many methods were researched as well, but game-changer technology was introduced by Ashish Vaswani et. al. in the [14]. In this work, transformer architecture was proposed and the attention mechanism was reimplemented: instead of using an RNN-based encoder-decoder mechanism, it was implemented by dispensing with recurrence and convolutions or relying solely on a self-attention mechanism. Based on this paper new methods for Natural Language Processing (NLP) [15] were created. One of the most important was the introduction of the BERT [16] model.

BERT or Bidirectional Encoder Representation from Transformers is an open-sourced NLP pre-trained model. This model is the first deeply bidirectional, unsupervised language representation that was pre-trained using only a plain text corpus of unlabeled text. A large amount of research was conducted about the BERT model itself, and many new models were created based on BERT. RoBERTa [17] is a robustly optimized BERT approach. Both RoBERTa and BERT use masked language models, but they utilize them in different ways. With BERT, masking is performed only once during the preparation, and it masks each sentence in 10 different ways. With RoBERTa masking is done dynamically during training when a sentence is incorporated in a batch, so the number of different masked versions is not bounded as in BERT. DistilBERT [18] is a distilled version of BERT. It uses roughly the same architecture as BERT but with some changes, like fewer encoder blocks or removed token-type embeddings and the pooling functionalities. The aim of DistilBERT is to be as much as possible efficient as BERT, but with a smaller model and with greater training speed. ALBERT [19] is a Lite BERT. It was introduced at around the same time as DistilBERT and likewise it has a smaller model and can be trained faster. But these gainings are not obtained by cutting the performance of the model, unlike DistilBERT. The difference between them is in the way both models are structured.

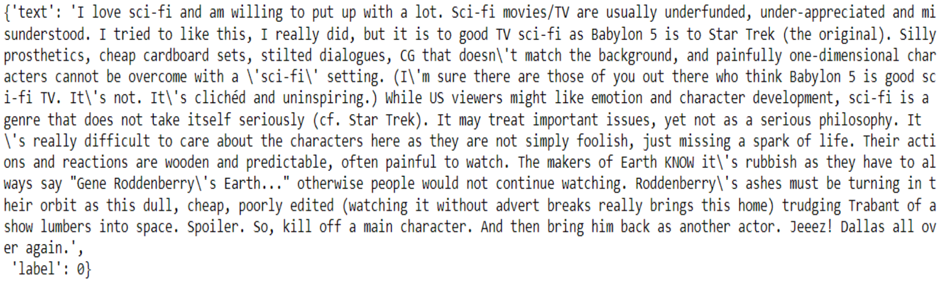
Described models are pre-trained meaning that they were developed and trained on large datasets to solve a specific task by another person or group of people. Pre-trained transformer usage has many benefits. Such models reduce computation costs, training time is decreased, and they allow usage of the state-of-art models without the necessity of creating it all by yourself. Moreover, they outperformed recurrent neural networks in NLP tasks. Although, in particular cases with additional adjustments, other models, like convolutional neural networks [20] can do certain tasks better than transformers. This is the reason why for some tasks it is recommended to train models with a dataset specific to the issue. This process is also known as fine-tuning. When a model is being tuned, it is trained with a number of different parameters that might have an influence on performance, training time, model size, required memory for training, etc.

To evaluate the efficiency of transformer models many measurement metrics can be used: accuracy, precision, F1-score, recall, etc. In this paper two metrics are used: validation and training loss. The training loss metric assesses how well the model fits training data. It is computed after each batch of data. On the contrary, the validation loss metric indicates the performance of trained models on data that the model has never seen before. Another important metric is the time that is consumed by a model to be trained. Validation loss, in contrast to training loss, is computed after each epoch. Thus, it demonstrates whether the model requires additional fine-tuning or not. Combined, these metrics indicate which aspects of the model might require additional training, and avoid training issues such as overfitting. Model training is expensive both in money and computation cost, hence the less time is consumed, the cheaper and more profitable it is.

The purpose of this paper is to analyze the influence of the learning rate hyperparameter on the training and validation loss, and the time consumed for model training with all described transformers in binary text classification tasks.

**Methods and materials.**

As data for the research, the dataset [21] was used from the HuggingFace portal [22]. To conduct experiments with the same data across all the models only the first 20% of the dataset, or 5000 records, were used because of the limited amount of memory in the GPU. Only the labeled part was used for the experiment. Figure 1 demonstrates that labeled records are stored as a dictionary with two pairs of values.



```
{'text': 'I love sci-fi and am willing to put up with a lot. Sci-fi movies/TV are usually underfunded, under-appreciated and mi
sunderstood. I tried to like this, I really did, but it is to good TV sci-fi as Babylon 5 is to Star Trek (the original). Silly
prosthetics, cheap cardboard sets, stilted dialogues, CG that doesn\'t match the background, and painfully one-dimensional char
acters cannot be overcome with a \'sci-fi\' setting. (I\'m sure there are those of you out there who think Babylon 5 is good sc
i-fi TV. It\'s not. It\'s clichéd and uninspiring.) While US viewers might like emotion and character development, sci-fi is a
genre that does not take itself seriously (cf. Star Trek). It may treat important issues, yet not as a serious philosophy. It
\'s really difficult to care about the characters here as they are not simply foolish, just missing a spark of life. Their acti
ons and reactions are wooden and predictable, often painful to watch. The makers of Earth KNOW it\'s rubbish as they have to al
ways say "Gene Roddenberry\'s Earth..." otherwise people would not continue watching. Roddenberry\'s ashes must be turning in t
heir orbit as this dull, cheap, poorly edited (watching it without advert breaks really brings this home) trudging Trabant of a
show lumbers into space. Spoiler. So, kill off a main character. And then bring him back as another actor. Jeeez! Dallas all ov
er again.',
 'label': 0}
```

Fig. 1. Example of a labeled record in the dataset.

The experiments were conducted on the NVIDIA GeForce GTX 1080 Ti GPU. All used models and tokenizers are available on the HiggingFace portal and are accessible by the next names:

- bert-base-uncased;
- distilbert-base-uncased;
- xlm-roberta-base;
- albert-base-v1.

For batch formation, DataColator class from the HuggingFace portal was used. Trainer and TrainingArgumets classes from the same portal were utilized for easier feature-complete training.

**Results and discussion.**

All models were trained in 3 epochs, with a static value of training and evaluation batch size of 8, weight decay was set to 0.01, evaluation and save strategies were set to epoch.

Table 1. Time of model's training.

| **Transf ormer model** | **Learning rate** | | | |
|---|---|---|---|---|
| | $1e^{-4}$ | $1e^{-5}$ | $1e^{-6}$ | $1e^{-7}$ |
| | **Time, m:s** | | | |
| BERT | 16:16 | 16:37 | 16:19 | 15:21 |
| RoBE RTa | 18:44 | 19:00 | 27:56 | 18:52 |
| DistilB ERT | 09:03 | 8:25 | 8:21 | 8:22 |
| ALBE RT | 16:34 | 16:21 | 16:36 | 15:36 |

Table 1 presents the difference in time consumption between different models and with different learning rate parameter values. We can see that DistilBERT shows the least training time for most learning rates, being less than the second-best model nearly two times. The model that took the longest to be trained is ALBERT. Also, it is clear that the learning rate parameter has an influence on the training time for most models. All except DistilBERT have a difference in time from 40 seconds to 9 minutes. The consumed time for the model's training was only under 30 minutes in that case, but with a bigger model, the difference in training time with various learning rate parameters can be days or more. However, it doesn’t show any particular dependency pattern of training time to learning rate value.

Table 2. Influence of learning rate parameter on training and validation losses for BERT transformer model.

| **poch** | **Learning rate** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $1e^{-4}$ | | $1e^{-5}$ | | $1e^{-6}$ | | $1e^{-7}$ | |
| | L | L | L | L | L | L | L | L |
| | .0014 | .0004 | .0128 | .0001 | .075 | .005 | .6049 | .202 |
| | .0 | .0002 | .0001 | .0005 | .051 | .002 | .1945 | .1328 |
| | .0 | .0001 | .0001 | .0005 | .0028 | .0016 | .1413 | .1234 |

The task that is being done with the models is a binary text classification by the sentiments. All models are trained on the same data with identical learning rate parameter.  To evaluate the influence of the learning rate parameter value on the training process two parameters were considered: Training Loss (TL) and Validation Loss(VL).

Table 3. Influence of learning rate parameter on training and validation losses for RoBERTa transformer model.

| **poch** | **Learning rate** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $1e^{-4}$ | | $1e^{-5}$ | | $1e^{-6}$ | | $1e^{-7}$ | |
| | L | L | L | L | L | L | L | L |
| | .0047 | .0 | .006 | .0001 | .138 | .0018 | .6675 | .362 |
| | .0 | .0 | .0001 | .0001 | .003 | .0008 | .348 | .211 |
| | .0 | .0 | .0 | .0001 | .002 | .0006 | .256 | .168 |

Table 4. Influence of learning rate parameter on training and validation losses for DistilBERT transformer model.

| **poch** | **Learning rate** | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $1e^{-4}$ | | $1e^{-5}$ | | $1e^{-6}$ | | $1e^{-7}$ | |
| | L | L | L | L | L | L | L | L |
| | .0033 | .0 | .0213 | .0002 | .137 | .007 | .575 | .464 |
| | .0 | .0 | .0002 | .0001 | .007 | .003 | .443 | .335 |
| | .0 | .0 | .0001 | .0000 | .004 | .002 | .347 | .294 |

Table 5. Influence of learning rate parameter on training and validation losses for ALBERT transformer model.

| poch | Learning rate | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | $1e^{-4}$ | | $1e^{-5}$ | | $1e^{-6}$ | | $1e^{-7}$ | |
| | L | L | L | L | L | L | L | L |
| | .0018 | .0003 | .0121 | .0004 | .0647 | .0015 | .3758 | .1943 |
| | .0 | .0001 | .0045 | .0001 | .0014 | .0006 | .171 | .0929 |
| | .0 | .0001 | .0 | .0001 | .0007 | .0005 | .0981 | .0734 |

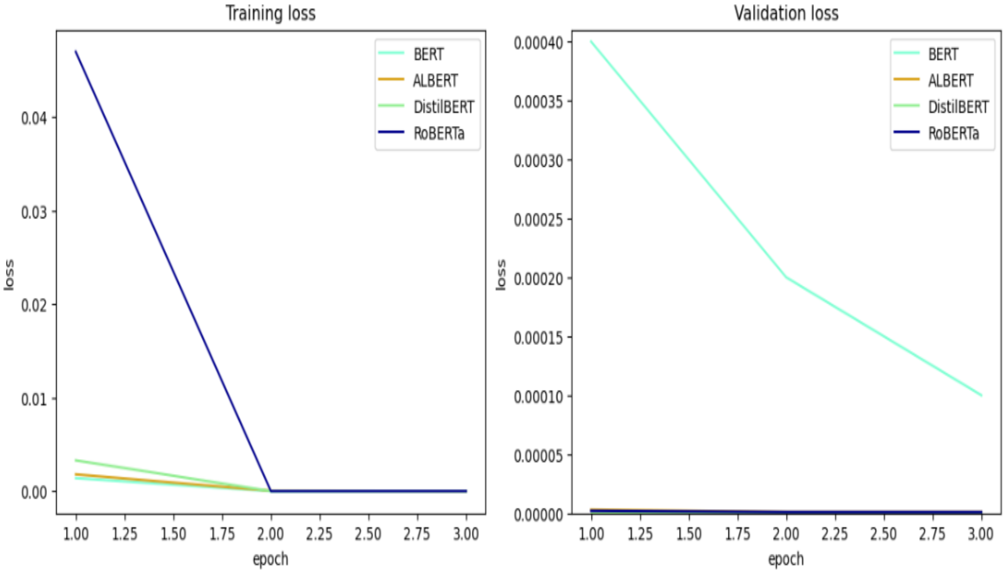


Fig. 2. Training and validation loss for learning rate $1e^{-4}$.

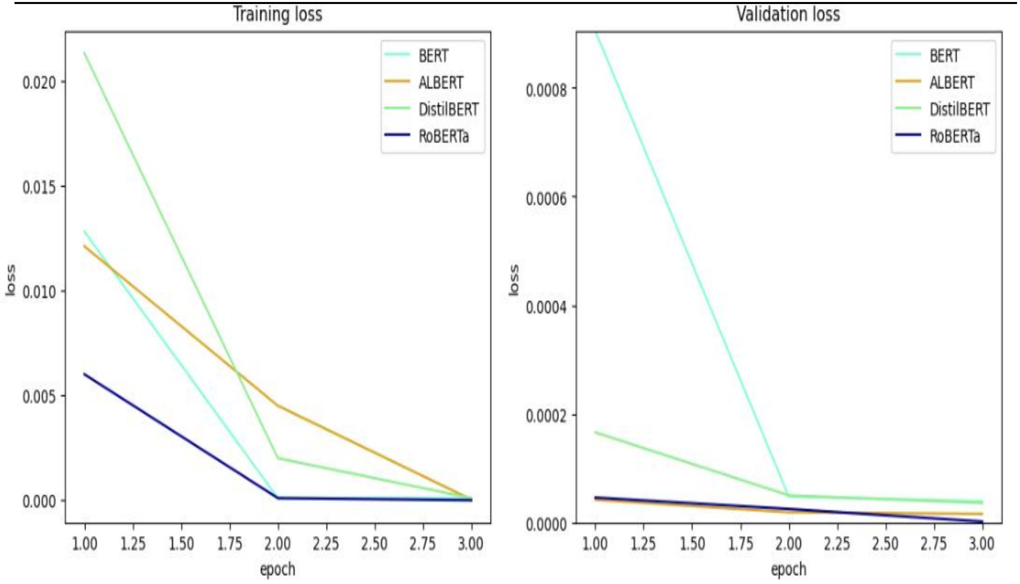


Fig. 3. Training and validation loss for learning rate $1e^{-5}$.

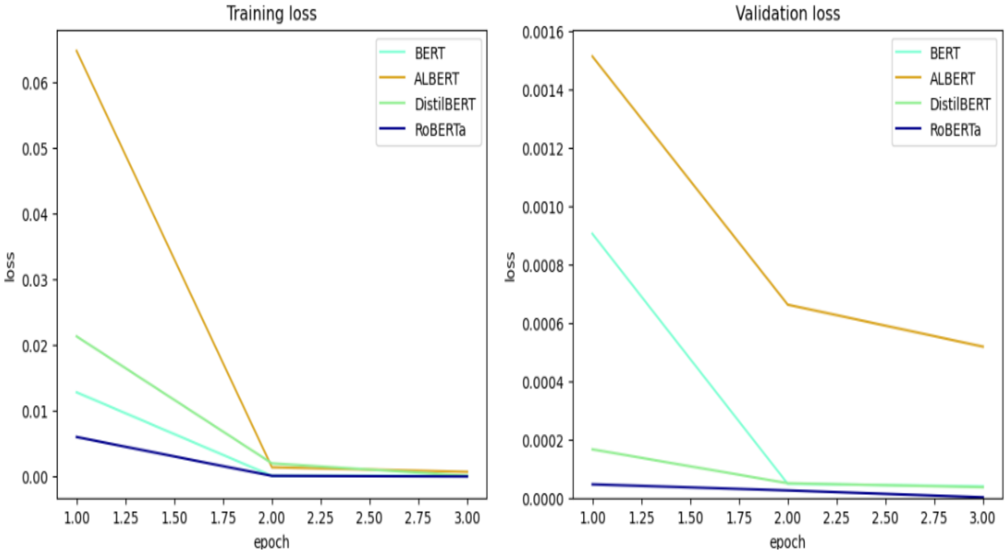


Fig. 4. Training and validation loss for learning rate $1e^{-6}$.

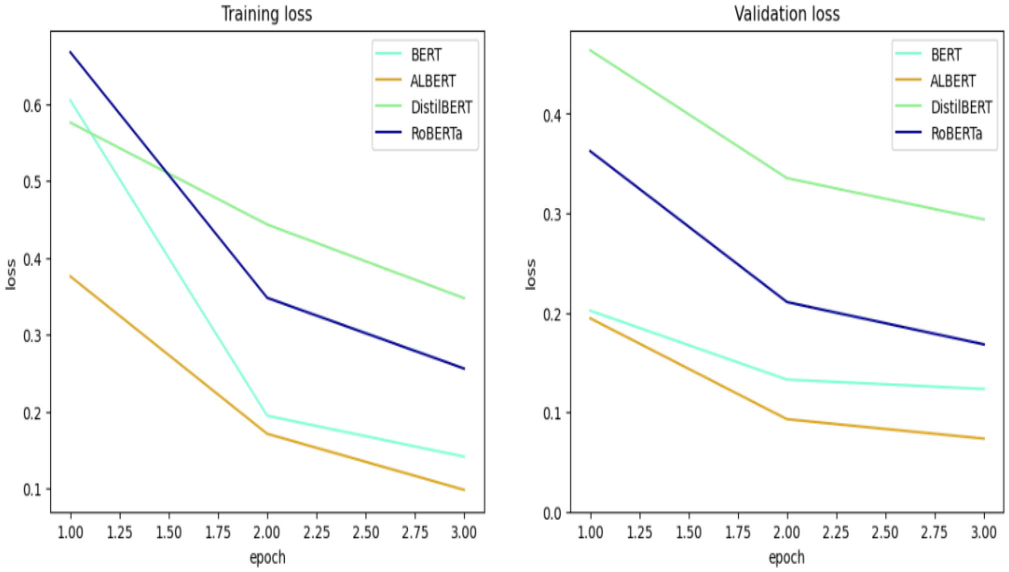


Fig. 5. Training and validation loss for learning rate $1e^{-7}$.

N.B. For numbers that were received after the experiment results, in case the first nonzero digit value after the decimal point was on the fifth or further position, this number was specified as zero in a table. But on the graphs, those numbers were used as were received.

From the results (Table 2-5 and Figure 2-5) we can observe that for all models the best results were received when the learning rate value was set to $1e^{-4}$. With that value, training and validation losses for all models were the smallest among all received results and were near 0 after the second epoch was completed. Furthermore, for the BERT model, an additional epoch can be considered as validation loss continues to descend. For the ALBERT model there where only one case, when received losses were bigger than in other models after training was completed and it was with the learning rate set to $1e^{-6}$, hence it showed the worst results. All other learning rate values, that were experimented with, presented the best results using the ALBERT model. An additional training epoch can be considered for all models with all learning rate values except $1e^{-4}$. Moreover, for all contemplated models, learning rate values smaller than $1e^{-7}$ causes the validation and training losses to increase dramatically, even though both learning and training losses functions continued to decrease after training was finished.

**Conclusion.**

In the paper, the influence of the learning rate parameter on the training time and training and evaluation losses was studied in BERT, RoBERTa, DistilBERT, and ALBERT transformer models. The optimal value of the learning rate parameter in the binary text classification task for those models was considered.

In most cases, the ALBERT model showed the smallest validation loss value, but it took almost the longest to train. DistilBERT was the fastest to train but it showed the biggest validation losses among all four models. RoBERTa showed second-best results in validation loss output, but it took the longest to train. The original BERT model was not the best both in the validation loss and in consumed training time, which was expected, as all other modes are enhanced versions of the BERT.

Those models were tested with a certain dataset for a particular task. In general, optimal parameters should be considered for the task and validated and tested with the dataset, with which the model will be used in the future.

The results of the research can be used in tasks of binary text classification with one of the considered transformer models.

## References.

[1] IDC. *The digitization of the World from Edge to Cor*e. 2018.

[2] Eberendu, Adanma Cecilia. "Unstructured Data: an overview of the data of Big Data." *International Journal of Computer Trends and Technology* 38, no. 1 (2016): 46-50.

[3] Ott, R. Lyman, and Micheal T. Longnecker. *An introduction to statistical methods and data analysis*. Cengage Learning, 2015.

[4] Anderson, James A. *An introduction to neural networks*. MIT press, 1995.

[5] Pavlyshenko, Bohdan. "Classification analysis of authorship fiction texts in the space of semantic fields." *Journal of Quantitative Linguistics* 20, no. 3 (2013): 218-226.

[6] Pavlyshenko, Bohdan. "The Model of Semantic Concepts Lattice For Data Mining Of Microblogs." *arXiv preprint arXiv:1210.7917* (2012).

[7] Pavlyshenko, Bohdan. "Clustering of authors' texts of english fiction in the vector space of semantic fields." *Cybernetics and Information Technologies* 14, no. 3 (2014): 25-36.

[8] Pavlyshenko, Bohdan. "The Distribution of Semantic Fields in Author's Texts." *Cybernetics and Information Technologies* 16, no. 3 (2016): 195-204.

[9] Pavlyshenko, Bohdan M. "Methods of Informational Trends Analytics and Fake News Detection on Twitter." *arXiv preprint arXiv:2204.04891* (2022).

[10] Pavlyshenko, Bohdan. "Genetic optimization of keyword subsets in the classification analysis of authorship of texts." *Journal of Quantitative Linguistics* 21, no. 4 (2014): 341-349.

[11] AminiMotlagh, Masoud, HadiShahriar Shahhoseini, and Nina Fatehi. "A reliable sentiment analysis for classification of tweets in social networks." *Social Network Analysis and Mining* 13, no. 1 (2022): 7.

[12] Wang, Shaonan, Yunhao Zhang, Weiting Shi, Guangyao Zhang, Jiajun Zhang, Nan Lin, and Chengqing Zong. "A large dataset of semantic ratings and its computational extension." *Scientific Data* 10, no. 1 (2023): 106.

[13] Cho, Kyunghyun, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. "Learning phrase representations using RNN encoder-decoder for statistical machine translation." *arXiv preprint arXiv:1406.1078* (2014).

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

[15] Khurana, Diksha, Aditya Koli, Kiran Khatter, and Sukhdev Singh. "Natural language processing: State of the art, current trends and challenges." *Multimedia tools and applications* 82, no. 3 (2023): 3713-3744.

[16] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

[17] Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. "Roberta: A robustly optimized bert pretraining approach." *arXiv preprint arXiv:1907.11692* (2019).

[18] Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." *arXiv preprint arXiv:1910.01108* (2019).

[19] Lan, Zhenzhong, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. "Albert: A lite bert for self-supervised learning of language representations." *arXiv preprint arXiv:1909.11942* (2019).

[20] Tay, Yi, Mostafa Dehghani, Jai Gupta, Dara Bahri, Vamsi Aribandi, Zhen Qin, and Donald Metzler. "Are pre-trained convolutions better than pre-trained transformers?." *arXiv preprint arXiv:2105.03322* (2021).

[21] Maas, Andrew, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. "Learning word vectors for sentiment analysis." In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies*, pp. 142-150. 2011.

[22] HuggingFace. URL: https://huggingface.co/ (accessed on April 19, 2023).

# ВПЛИВ ПАРАМЕТРУ ШВИДКОСТІ НАВЧАННЯ В РІЗНИХ МОДЕЛЯХ ТРАНСФОРМЕРІВ У ЗАДАЧІ КЛАСИФІКАЦІЇ ТЕКСТУ

Б. Павлишенко, М. Стасюк

*Львівський національний університет імені Івана Франка,*
*вул. Драгоманова 50, 79005 Львів, Україна*
*bohdan.pavlyshenko@lnu.edu.ua*, *mykola.stasiuk@lnu.edu.ua*

У сучасному світі спостерігається швидкий ріст обсягу даних, причому значна їхня частка виявляється неструктурованою, включаючи великі набори текстів. Через це, попит на методи для обробки та аналізу даних тільки збільшується. Визначним моментом у розробці моделей штучного інтелекту для роботи з текстовими даними було переосмислення механізму уваги та створення архітектури трансформерів. Ці моделі ефективно використовуються для вирішення широкого спектру завдань Обробки Природного Мовлення (ОПМ), таких як: класифікація текстів, семантичний пошук та призначення ролей, розпізнавання іменованих сутностей тощо. Для тренування або налаштування трансформерів існує велика кількість гіперпараметрів, кожен з яких впливає на результати роботи моделей такого виду. Для прикладу, швидкість навчання, кількість епох тренування, величина зменшення ваги, стратегії оцінки і збереження та інші. Однак, досі недостатньо вивчено вплив окремих гіперпараметрів на тривалість навчання та ефективність роботи трансформерів.

У роботі досліджено вплив параметру швидкості навчання на результати тренування попередньо навчених моделей трансформерів: BERT, DistilBERT, ALBERT, та XLM-RoBERTa. У якості даних для тренування та тестування моделей використано набір даних з порталу HuggingFace. У цьому наборі є дані для тестування та тренування. Також, у ньому містяться дані без міток для моделей та алгоритмів без вчителя. Замість прямого тестування та тренування використані класи Trainer та TrainingArgument з порталу HuggingFace. Для формування батчів використано клас DataColator. Досліджено час тренування та такі метрики ефективності тренування як: втрати тренування та втрати оцінки. Результати роботи дозволяють порівняти ефективність кожної розглянутої моделі у задачах бінарної класифікації тексту окремо, або у ансамблях з іншими моделями.

*Ключові слова:* трансформери, бінарна класифікація тексту, BERT, ALBERT, DistilBERT, XLM-RoBERTa.

*Стаття надійшла до редакції 26.04.2023*
*Прийнята до друку 03.05.2023*