International Journal of Differential Equations and Applications
Volume **23**, No. 1 (2024), pages: 141-149
ISSN (Print): 1311-2872; ISSN (Online): 1314-6084;
url: https://www.ijdea.eu

BGT5 – PRE-TRAINED T5 MODEL ON BULGARIAN DATA

Simeon Monov[1], Nikolay Pavlov [2], Detelinka Trifonova[3]
[1,2,3]Facaulty of Mathematics and Informatics
University of Plovdiv Paisii Hilendarskiy
236, Bulgaria Blvd., 4027 Plovdiv, BULGARIA

ABSTRACT: In recent years many resources were created in Natural Language Processing (NLP) but most of them are available in other languages than Bulgarian. With the exception of few multilingual models, which support Bulgarian language, there are not many others. In this work, we collect Bulgarian data from different sources and pre-train a T5 model on this data. We further evaluate its performance against other multilingual models on different Natural language processing tasks.

Key Words: NLP, transformer model, pre-training, large language models, monolingual language models, Bulgarian language

Received: October 11, 2024 **Revised:** December 21, 2024
Published: December 28, 2024 **doi:** 10.12732/ijdea.v23i1.11

1. INTRODUCTION

Pre-trained Language Models have been used in various Natural Language Processing (NLP) tasks [1][2][3]. Most of these models are pre-trained on English language but some multilingual and monolingual models in other languages exist [4][5][6][7], too. Multilingual models show very good performance on cross-lingual tasks but the monolingual counterparts perform better on language-specific tasks [7][8][6][9][10]. In [11] the authors make a comprehensive empirical comparison between pre-trained multilingual models versus their monolingual counterparts in 9 different languages. Very few works explore pre-training language models on Bulgarian data and most of them use BERT [1] or similar based models [12].

Text-to-Text transformer model T5[3] is one of the most popular models used when text generation tasks are needed such as summarization, abstractive question answering, question generation and translation. The original T5 model was trained on English data only but the same architecture was used to produce a multilingual version of it mT5 [4], which supports more than 100 languages. For the reasons mentioned in the previous paragraph, multiple monolingual T5 models were recently pre-trained to support different languages [8][9][10]. These models improve performance over their multilingual baseline on tasks such as language understanding, generation tasks and natural language inference.

In this work, we pre-train a T5 model in a Bulgarian dataset and test and evaluate it on various Bulgarian natural language processing tasks. We make the following contributions:

- We create an extensive dataset with unstructured Bulgarian data, collected from different sources on internet.
- We introduce BGT5 models - pre-train versions from mT5 language model on the above data using unsupervised learning.
- We evaluate our models on bgGLUE [14] benchmarks, which demonstrate better performance over their multilingual baselines.

2. DATA.

To train the model we created an extensive dataset, collected from various authoritative sources, including student textbooks, lectures, scientific literature, articles, books, and data from the Bulgarian National Corpus of the Bulgarian Academy of Sciences [13]. To these we added texts from Craw Wikipedia in Bulgarian, as well as various educational materials from different fields. After a process of careful filtering, the final dataset includes approximately 5.5 billion words and 235 million sentences with a total volume of 58 GB.

DATA PROCESSING AND CLEANING

To ensure data quality, a thorough cleaning and structuring process was applied. The raw data was split into sentences. Sentences longer than 512 tokens were sliced by dot character. Sentences which could not be sliced to a maximum of 512 tokens were

disregarded. Inappropriate elements such as special characters that do not contribute to the context of the language tasks (e.g. remainder %, ampersand &, etc.) were removed in order to improve the accuracy of model training. As a result, the created dataset unifies and standardizes all sources, providing a quality training base.

3. MODEL AND TRAINING PARAMETERS.

mT5 model was chosen as a base model for our training. We used the vocabulary from the mT5 model. We trained small and base sizes of the model on a single NVIDIA V100 GPU with 32GB VRAM. The small model was trained for 2 epochs and the base model for 1 epoch with a fixed learning rate of 0,0003. We used pytorch_lightning library with cross entropy loss and RAdam [18] optimizer. The training took about 30 hours for the small and 50 hours for the base model.

3.1. MODEL PRE-TRAINING

Unsupervised pre-training approach was used with a span-mask denoising objective. In this setup, we used a predefined probability (15%) for input tokens to be replaced with a sentinel token (a.k.a. unique mask token) within a sentence. The masked tokens are represented as ¡extra_id_0¿, ¡extra_id,1¿ ,. . . , ¡extra_id_100¿. Then the model is fed the masked sequence and trained to predict the original sentence.

For example, if the original sentence is:

"Като използва машинно обучение и обработка на естествен език, редакторът прави предложения

the generated masked input sequence can be:

"Като използва <extra_id_0> и обработка на естествен език, редакторът прави <extra_id_1>"

and the label / output sequence provided to train the model is a concatenation of the mask tokens and the original tokens:

«extra_id_0> машинно обучение <extra_id_1> предложения <extra_id_2>"

3.2. EVALUATION

The new models were evaluated using the “bgGLUE: Bulgarian as a General Language Comprehension Assessment Benchmark” [14]. bgGLUE is a benchmark for evaluating language models on Natural Language Understanding (NLU) tasks in Bulgarian. It

includes nine datasets which can be used to evaluate models against nine different NLU tasks. For evaluating our models, we used three tasks included in the bgGLUE benchmark:

- **Cinexio Movie Reviews (Cinexio) - Sentiment Analysis.** Cinexio [17] metric is focused on sentiment analysis of movie reviews. Its goal is to automatically assess the sentiment expressed in movie reviews, typically classifying them as positive, neutral, or negative.
- **CheckThat! (2021), Task 1A (CT21.T1) - Check-Worthiness assessment task [15].** This task focuses on assessing the importance of claims in texts to determine whether they require fact-checking. In such tasks, algorithms are trained to assess the relevance and significance of statements, which is crucial in areas such as journalism and content moderation to combat misinformation.
- **Cross-lingual Natural Language Inference (XLNI) – Natural language inference.** XNLI [16] is a subset of several thousand examples from MNLI, which have been translated into 14 different languages. In other words, XNLI is the multilingual version of MNLI. As with MNLI, the goal is to predict the textual consequence (does sentence A imply/contradict any sentence B) and is a classification task (given two sentences, predict one of three labels).

Our models were compared to the following other models:

- **mT5-small** – the small size model of mT5 [4] - multilingual variant of T5 [3], pre-trained on mC4 corpus [4], covering 101 languages.
- **mT5-base** – the base size model of mT5 [4] - multilingual variant of T5 [3], pre-trained on mC4 corpus, covering 101 languages.
- **flan-T5-small** – small size model of FLAN-T5 [19] - enhanced version of T5 [3] model.
- **flan-T5-base** – base size model of FLAN-T5 [19] - enhanced version of T5 [2] model.
- **xlm-roberta-base** – base size of XLM-RoBERTa [5] - a multilingual version of RoBERTa [24] - a variant of BERT [1] model, developer by the researchers at Facebook AI.

Figure 1, 2 and 3 show graphs with the train evaluation results between the different models on the three NLP tasks.

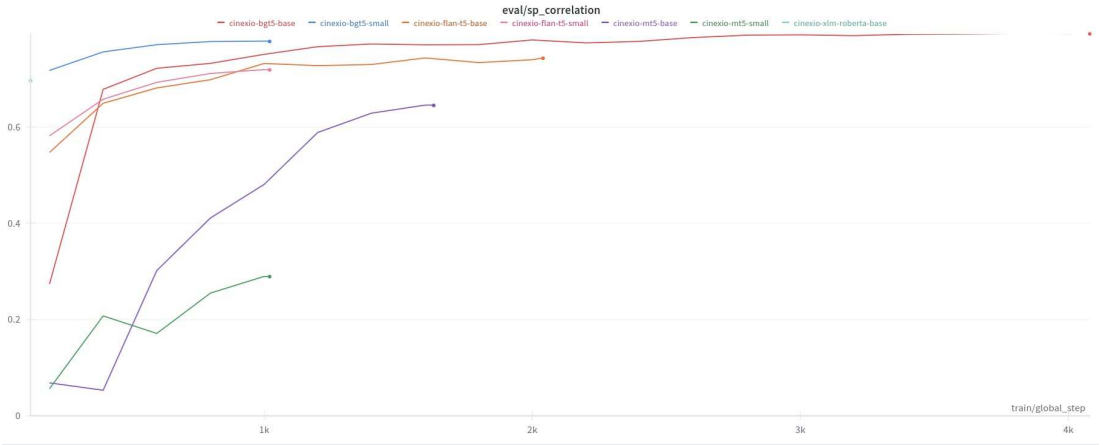

Figure 1. Pearson-Spearman Corr results for Cinexio task

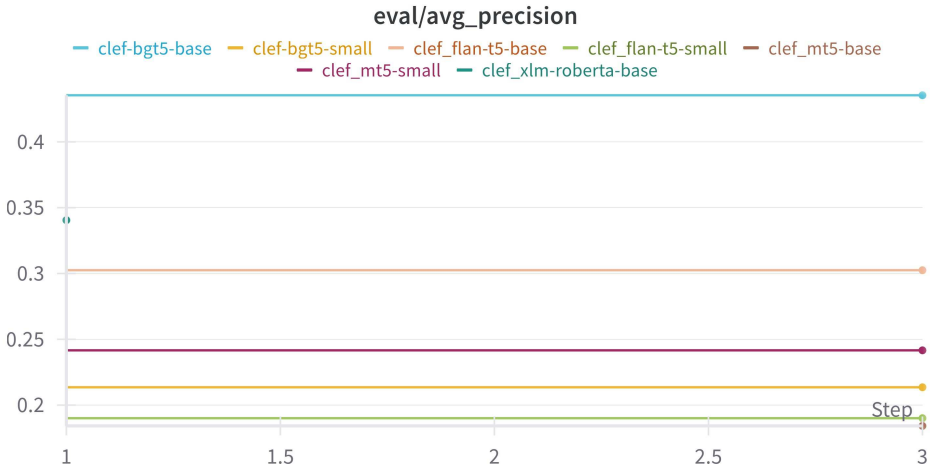

Figure 2. Average Precision results for CheckThat! (2021), Task 1A task

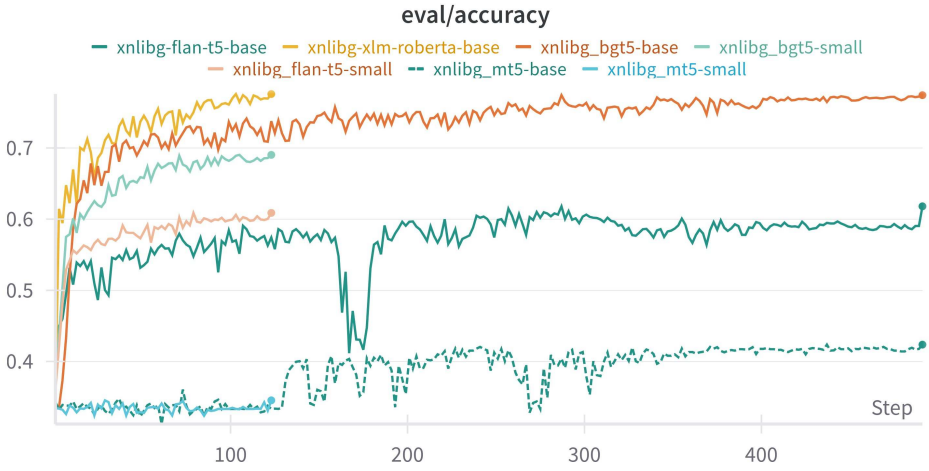

Figure 3. Accuracy results for XNLIBG task

Table 1 shows summarized results from the different evaluation tests.

Table 1: Summarized results of the model evaluation

Model name	**Cinexio sp_correlation**	**CheckThat!'21 Task 1 avg_precision**	**XLNIBG accuracy**
bgt5-small	77.78	51.19	69.38
mt5-small	28.91	18.86	31.92
flan-t5-small	71.86	34.36	61.4
xlm-roberta-base	69.65	34	**78.28**
bgt5-base	**79.38**	**60.43**	78.14
mt5-base	64.49	19.75	43.63
flan-t5-base	74.28	40	61.81

The results demonstrate that our models outperform their baseline model mT5 and achieve very good results on the evaluation tasks.

4. CONCLUSIONS.

In this paper we introduced small and base sizes of a pre-trained mT5 [4] model on Bulgarian data. The resulting models achieved better performance than the original

mT5 on three different tasks from bgGLUE [14] benchmark. This proves that monolingual models can perform better on language specific tasks versus their multilingual counterparts. Our models can be utilized in different use cases and have big potential to improve the performance of different tasks in Bulgarian language [20][21][22][23].

5. ACKNOWLEDGMENT.

This work is partially supported by the project MUPD23-FMI-009 of the Scientific Fund of the Paisii Hilendarski University of Plovdiv, Bulgaria.

REFERENCES

[1] Devlin, J., Chang, M., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv: 1810.04805.

[2] Wang, H., Li, J., Wu, H., Hovy, E. and Sun, Y., 2023. Pre-trained language models and their applications. *Engineering*, **25**, pp.51-65.

[3] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W. and Liu, P.J., 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, **21**(140), pp.1-67.

[4] Xue, L., 2020. mt5: A massively multilingual pre-trained text-to-text transformer. arXiv preprint arXiv:2010.11934.

[5] Conneau, A., 2019. Unsupervised cross-lingual representation learning at scale. arXiv preprint arXiv:1911.02116.

[6] Delobelle, P., Winters, T. and Berendt, B., 2020. Robbert: a dutch roberta-based language model. arXiv preprint arXiv:2001.06286.

[7] Straka, M., Náplava, J., Straková, J. and Samuel, D., 2021. RobeCzech: Czech RoBERTa, a monolingual contextualized language representation model. *In Text, Speech, and Dialogue: 24th International Conference*, TSD 2021, Olomouc, Czech Republic, September 6–9, 2021, Proceedings **24** (pp. 197-209). Springer International Publishing.

[8] Phan, L., Tran, H., Nguyen, H. and Trinh, T.H., 2022. Vit5: Pretrained text-to-text transformer for vietnamese language generation. arXiv preprint arXiv: 2205.06457.

[9] Carmo, D., Piau, M., Campiotti, I., Nogueira, R. and Lotufo, R., 2020. Ptt5: Pretraining and validating the t5 model on brazilian portuguese data. arXiv preprint arXiv:2008.09144.

[10] Sarti, G. and Nissim, M., 2022. IT5: Text-to-text Pretraining for Italian Language Understanding and Generation. arXiv preprint arXiv:2203.03759.

[11] Rust, P., Pfeiffer, J., Vulić, I., Ruder, S. and Gurevych, I., 2020. How good is your tokenizer? on the monolingual performance of multilingual language models. arXiv preprint arXiv:2012.15613.

[12] Marinova, I., Simov, K. and Osenova, P., 2023, September. Transformer-based language models for bulgarian. *In Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing* (pp. 712-720).

[13] Koeva, S., Blagoeva, D. and Kolkovska, S., 2010. Bulgarian National Corpus Project. *Politics*, **207**, pp.2-2.

[14] Hardalov, M., Atanasova, P., Mihaylov, T., Angelova, G., Simov, K., Osenova, P., Stoyanov, V., Koychev, I., Nakov, P. and Radev, D., 2023. bgGLUE: A Bulgarian general language understanding evaluation benchmark. arXiv preprint arXiv:2306.02349

[15] Nakov, P., Da San Martino, G., Elsayed, T., Barrón-Cedeno, A., Míguez, R., Shaar, S., Alam, F., Haouari, F., Hasanain, M., Babulkov, N. and Nikolov, A., 2021. The CLEF-2021 CheckThat! lab on detecting check-worthy claims, previously fact-checked claims, and fake news. *In Advances in Information Retrieval: 43rd European Conference on IR Research*, ECIR 2021, Virtual Event, March 28–April 1, 2021, Proceedings, Part II **43** (pp. 639-649). Springer International Publishing.

[16] Conneau, A., Lample, G., Rinott, R., Williams, A., Bowman, S.R., Schwenk, H. and Stoyanov, V., 2018. XNLI: Evaluating cross-lingual sentence representations. arXiv preprint arXiv:1809.05053.

[17] Kapukaranov, B. and Nakov, P., 2015, September. Fine-grained sentiment analysis for movie reviews in Bulgarian. *In Proceedings of the international conference recent advances in natural language processing* (pp. 266-274).

[18] Liu, L., Jiang, H., He, P., Chen, W., Liu, X., Gao, J. and Han, J., 2019. On the variance of the adaptive learning rate and beyond. arXiv preprint arXiv:1908.03265.

[19] Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S. and Webson, A., 2024. Scaling instruction-finetuned

language models. *Journal of Machine Learning Research*, **25**(70), pp.1-53.

[20] Monov, S., Pavlov, N., Trifonova, D., 2024. Automatic translation of SQuAD and RACE Question Answering datasets in Bulgarian language, *International Journal of Differential Equations and Applications*, **23**(1), pp.83-95. doi: 10.12732/ijdea.v23i1.7

[21] Pavlov, N., Iliev, A., Rahnev, A. and Kyurkchiev, N., 2018. Some deterministic growth curves with applications to software reliability analysis. *Int. J. of Pure and Appl. Math*, **119**(2). doi: 10.12732/ijpam.v119i2.8

[22] Kyurkchiev, V., Pavlov, N. and Rahnev, A., 2018. Cloud-based architecture of DisPeL. *International Journal of Pure and Applied Mathematics*, **120**(4), pp.573-581. doi: 10.12732/ijpam.v120i4.8

[23] Terzieva, T., Rahneva, O. and Dilyanov, V., 2021. Pedagogical strategies for development of cognitive skills in a digital environment. *International Journal of Differential Equations and Applications*, **20**(2), pp.251-261.

[24] Liu, Y., 2019. Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692, 364.