Comparison of Czech Transformers on Text Classification Tasks

Jan Lehečka[(✉)] and Jan Švec

Department of Cybernetics, University of West Bohemia in Pilsen,
Univerzitní 2732/8, 301 00 Pilsen, Czech Republic
{jlehecka,honzas}@kky.zcu.cz

Abstract. In this paper, we present our progress in pre-training monolingual Transformers for Czech and contribute to the research community by releasing our models for public. The need for such models emerged from our effort to employ Transformers in our language-specific tasks, but we found the performance of the published multilingual models to be very limited. Since the multilingual models are usually pre-trained from 100+ languages, most of low-resourced languages (including Czech) are under-represented in these models. At the same time, there is a huge amount of monolingual training data available in web archives like Common Crawl. We have pre-trained and publicly released two monolingual Czech Transformers and compared them with relevant public models, trained (at least partially) for Czech. The paper presents the Transformers pre-training procedure as well as a comparison of pre-trained models on text classification task from various domains.

Keywords: Text categorization and summarization · Monolingual transformers · Sentiment analysis · Multi-label topic identification

1 Introduction

In the last few years, deep neural networks based on Transformers [4] have dominated the research field of Natural Language Processing (NLP) and Natural Language Understanding (NLU). Self-attention [19] Transformers and especially the self-supervised-trained variants known as BERT (Bidirectional Encoder Representations from Transformers) models, have achieved amazing results in many tasks, including text classification [1,9,17].

Because the self-supervised pre-training of Transformers is computationally very costly, researchers around the world publish their pre-trained models in order to compete on a large variety of NLP and NLU tasks, however the majority of these models are monolingual English models or multilingual models including many languages at once. Since the multilingual models are usually pre-trained from Wikipedia dumps or large web archives, most of low-resourced languages are under-represented in these models. This makes practical use of such models in non-English languages very limited. Therefore in present days, we observe

a significant trend in pre-training also monolingual transformers for other languages, e.g. CamemBERT for French [13], Finnish BERT [20] and very recently also Czech models Czert [15] and RobeCzech [16].

The trend in pre-training monolingual Transformers is boosted by the accessibility of powerful hardware devices (especially graphical and text processing units, GPUs and TPUs) and by the availability of huge text corpora even for languages with relatively small number of native speakers. For example, many researchers use data from the Common Crawl project[1] which is a huge public web archive consisting of petabytes of crawled web pages.

In this paper, we present our effort to pre-train two monolingual Czech Transformers (BERT and RoBERTa) from different datasets and we compare the performance with published multilingual and monolingual Czech models on downstream text classification tasks. We release both our pre-trained models publicly on HuggingFace hub[2].

2   Related Work

Aside from the famous English BERT-base model [4], also its multilingual variant has been published by Google researchers. It has been trained from Wikipedia dumps of 104 languages, including Czech. We denote this model *MultiBERT* in this paper.

BERT model trained from Wikipedia dumps of only four Slavic languages (Russian, Bulgarian, Czech and Polish) has been presented in [2]. The model was not pre-trained from the scratch, but MultiBERT weights were used to initialize the model. We denote this model *SlavicBERT* in this paper.

Researchers from Facebook have published multilingual *XLM-RoBERTa* model [3] pre-trained on one hundred languages (including Czech), using more than two terabytes of filtered Common Crawl data. Two sizes of the model have been published: base model with 270 million parameters and a large one with 550 million parameters.

There are also two recent papers presenting monolingual Czech models. *Czert* [15] is a BERT model pre-trained from a mixture of Czech national corpus [7], Wikipedia pages and self-crawled news dataset. The second one, *RobeCzech* [16], is a RoBERTa [11] model pre-trained from Czech national corpus [7], Wikipedia pages, a collection of Czech newspaper and magazine articles and a Czech part of the web corpus W2C [12]. However, these models were pre-trained from a standard (literary) Czech, which makes them unsuitable for tasks where texts are often written in a common (spoken) Czech, such as social media posts or product reviews. We provide the basic information about described Transformers in the Table 2.

---

[1] https://commoncrawl.org.
[2] https://huggingface.co/fav-kky.

3 Datasets

3.1 Pre-training Datasets

Since self-supervised pre-training of Transformers requires a huge amount of unlabeled text, our aim was to get as much cleaned Czech text as possible. In the following paragraphs, we provide a description of datasets we were experimenting with.

News Corpus. We used Czech news dataset from our web-mining framework [18]. The dataset consists mainly of web-crawled news articles and transcripts of selected TV and radio shows. In this dataset, we cleaned boilerplate and HTML tags from original documents and thoroughly processed each document with normalization, true-casing and rule-based word substitution dealing with typos, multiwords and word-form unification. See [18] for more details. The corpus contains 3.3 billion words (20.5 GB of cleaned text).

C5. Czech Colossal Clean Crawled Corpus (C5) generated from the Common Crawl project is a Czech mutation of the English C4 dataset [14]. We used the language information provided in the index files to select Czech records only. The corresponding plain texts (stored in WET archives) were downloaded. To clean the data, we followed almost the same rules which were used to pre-process the C4 dataset, i.e.:

– We only retained lines that ended with a terminal punctuation mark (".", "?" or "!").
– We removed lines containing "javascript" or "cookies".
– We removed web pages containing offensive words or strings "lorem ipsum" or "{".
– We retained only pages classified as Czech with probability of at least 0.99 according to `langdetect`[3] tool.
– To deduplicate the dataset, we discarded all but one line occurring more than once in the data set.
– We only retained lines with at least 3 words and pages with at least 5 sentences.

This simple yet rigorous cleaning process removed about 98% of downloaded plain texts from the dataset, mainly due to the deduplication (the more data we downloaded, the harder it was to find a new unobserved line).

We downloaded and cleaned recent crawls up to the crawl from August 2018, which is the first one providing the language column in its index. Together, we processed 25 crawls (from August 2018 to October 2020) and the resulted dataset contains almost 13 billion words (93 GB of cleaned text[4]).

---

[3] https://pypi.org/project/langdetect/.

[4] To get a picture how huge the Common Crawl archive is, consider that this 93 GB of text is (due to rigorous cleaning rules) just about 2% of all Czech texts we downloaded and Czech records occupy just about 1% of the full archive.

3.2   Text Classification Datasets

We decided to fine-tune and evaluate Transformers on a text classification task. Specifically, we fine-tuned models for sentiment analysis and multi-label topic identification tasks as these tasks are in our main focus. To keep our experiments reproducible, we experimented mainly with publicly available datasets. In the following paragraphs, we describe five Czech datasets used in the experiments together with the basic statistics of the datasets in Table 1. The first three datasets are for sentiment analysis task. All of them were created in [5] and contain text samples with positive, neutral or negative class labels. The last two datasets are for multi-label topic identification task. Both of them are from news domain and contain crawled news articles, each with at least one topic label.

Table 1. Czech text classification datasets statistics. The table shows numbers of documents in train, development and test datasets, number of classes and the label cardinality (i.e. the average number of labels per document) of the datasets.

Dataset	Train	Devel	Test	Classes	LabCard
CSFD	91 381	–	–	3	1
MALL	145 307	–	–	3	1
FCB	9 752	–	–	3	1
CN	184 313	20 479	43 762	577	3.06
CTDC	11 955	2 538	–	37	2.55

CSFD. Movie review dataset consists of 91 thousand movie reviews from the Czech-Slovak Movie Database (ČSFD)[5]. Based on movie ratings, each review was classified into positive, neutral or negative sentiment class.

MALL. Product review dataset consists of 145 thousand posts from a large Czech e-shop Mall.cz[6] which offers a wide range of products. Similar to CSFD dataset, each post was classified into positive, neutral or negative sentiment class based on the user's ratings.

FCB. The Facebook dataset contains 10 thousand posts selected from several Facebook pages with a large Czech fan base. The sentiment of individual posts was assigned manually by multiple annotators. We ignored the 248 bipolar posts and trained models only with three sentiment classes as suggested by authors of the dataset [5].

---

[5] https://www.csfd.cz.
[6] https://www.mall.cz.

CTDC. Czech Text Document Corpus (CTDC)[7] is a news dataset described in [6]. The dataset consists of 12 thousand news articles annotated with category labels. In order to keep our results comparable with other papers (such as [10, 15]), we used the same setup and selected only 37 (out of 60) most frequent labels.

CN. Many news servers publish articles along with topic labels assigned by journalists. We keep this information in our framework [18] and map the most frequent labels into a standardized IPTC news topic tree[8]. We created a CN dataset by selecting articles with assigned IPTC topic labels from one news server (`ceskenoviny.cz`, hence the name CN) which we believe to be labeled thoroughly and consistently. The dataset contains 250 thousand articles from 10-years epoch. Each article has assigned one or more topic labels. The total number of labels in the dataset is 577. Unfortunately, we do not have licence to publish this dataset. It is, however, the only private dataset appearing in this paper.

4   Models

In this paper, we present two new Transformers. We named our models Flexible Embedding Representation NETwork (FERNET). Our aim was to pre-train one general model suitable for a wide range of domains and one single-domain model specifically for the news domain, which is in our main focus. After we had pre-trained the first BERT model, we reflected the change of paradigm and switched to the RoBERTa architecture when pre-training the second one. We tabulate basic summary of our models along with all other models we have experimented with in Table 2, and describe them in the following paragraphs.

FERNET-C5. This is a BERT model trained from scratch using the C5 dataset. The model has the same architecture as BERT-base model [4], i.e. 12 transformation blocks, 12 attention heads and the hidden size of 768 neurons.

In contrast to Google's BERT models, we used SentencePiece tokenization [8] instead of the Google's internal WordPiece tokenization. We trained a BPE SentencePiece model from the underlying dataset with vocabulary size set to 100 thousands. Since the datasets contains a small portion of non-Czech text fragments, we carefully tuned the character coverage parameter to fully cover all Czech characters and only reasonable number of graphemes from other languages. We kept original casing of the dataset.

FERNET-News. This is a RoBERTa model trained from scratch using our thoroughly pre-processed News dataset. The model has the same architecture as

---

[7] http://ctdc.kiv.zcu.cz.
[8] http://cv.iptc.org/newscodes/mediatopic.

Table 2. Summary of pre-trained Transformers trained at least partially on Czech. The first four models are multilingual, the next two models are already published monolingual Czech Transformers. The two models at the bottom of the table are our new models presented in this paper. The sizes of vocabularies are in thousands (k) and number of trainable parameters are in millions (M). Reported numbers of parameters include language model head on top of the model. SPM stands for SentencePiece tokenization [8], BBPE for Byte-level Byte-Pair Encoding and CC for Common Crawl.

Model	Architecture	Tokenization	Vocab	Pre-train data	Params
MultiBERT [4]	BERT-base	WordPiece	120 k	Wiki, 104 langs	179 M
SlavicBERT [2]	BERT-base	subword-nmt	120 k	Wiki, 4 langs	179 M
XLM-RoBERTa-base [3]	RoBERTa-base	SPM	250 k	CC, 100 langs (2 TB)	278 M
XLM-RoBERTa-large [3]	RoBERTa-large	SPM	250 k	CC, 100 langs (2 TB)	560 M
Czert [15]	BERT-base	WordPiece	40 k	Nat+Wiki+News (37 GB)	110 M
RobeCzech [16]	RoBERTa-base	BBPE	52 k	Nat+Wiki+Czes+W2C	126 M
FERNET-C5	BERT-base	SPM	100 k	C5 (93 GB)	164 M
FERNET-News	RoBERTa-base	BBPE	50 k	News Corpus (21 GB)	124 M

RoBERTa-base model [11], i.e. 12 transformation blocks, 12 attention heads and the hidden size of 768 neurons. To tokenize input texts, we trained byte-level BPE tokenizer with the vocabulary size 50 000 tokens, as suggested in the paper.

4.1   Pre-training

Our BERT model was pre-trained using two standard tasks: masked language modeling (MLM), which is a fill-in-the-blank task, where the model is taught to predict the masked words, and next sentence prediction (NSP) task, which is the classification task of distinguishing between inputs with two consequent sentences and inputs with two randomly chosen sentences from a corpus. We pre-trained the model for 6.5 million gradient steps in total (3.5 M steps with 128-long inputs and batch size 256, and 3 M steps with 512-long inputs and batch size 128). For both input variants, we used whole word masking, duplication factor of 5 and the learning rate warmed up over the first 100 000 steps to a peak value at 1×10^{-4}. To pre-train the model, we used the software provided by Google researchers[9] and to speed up the training, we used one 8-core TPU with 128 GB of memory. The pre-training took about two months.

Our RoBERTa model was pre-trained using MLM task only. We trained the model for 600 thousand gradient steps with batch size 2048: first 500 thousand steps with 128-long inputs, 24 000 warmup steps and learning rate 4×10^{-4}, and the following 100 thousand steps with 512-long inputs, 2 000 warmup steps and learning rate 1×10^{-4}. Since the news dataset is rather small (but very clean), the model iterated about 40-times over the whole training corpus during the pre-training. However, due to dynamic language model masking, the model saw differently masked input samples in each epoch. To pre-train the model, we

---

[9] https://github.com/google-research/bert.

used the HuggingFace's Transformers software[10] and to speed up the training, we used two Nvidia A100 GPUs. The pre-training took about one month.

4.2   Fine-Tuning

We run fine-tuning experiments on single GPUs using HuggingFace's Transformers software.

To fine-tune models for sentiment analysis, we fixed the peak learning rate to 1×10^{-5}, maximum sequence length 128 tokens, batch size 64, 10 training epochs and 100 warmup steps. We were also experimenting with other hyperparameters setting, but this one gave consistently the best results across all models and datasets. After each training epoch, we evaluated the model on the development dataset (if not present in the dataset, we randomly held out 10% of the training samples as a development data) and at the end, we used the best model for evaluation on the test dataset.

To fine-tune models for multi-label topic identification task, we fixed the peak learning rate to 5×10^{-5}, maximum sequence length 512 tokens and batch size 64. We trained on the CTDC dataset for 20 epochs whereas to train on CN dataset, 10 epochs was enough, because the dataset is much larger.

4.3   Evaluation

For both BERT and RoBERTa model architectures, we predicted classes using standard pooling layers for sequence classification implemented in the Transformers library along with the corresponding model types. This pooling layer uses the last hidden state of the first input token (`[CLS]` or `<s>`) to make predictions.

In the case of sentiment analysis task, we assigned sentiment class with the highest predicted value for each document and evaluated F1 score.

In the case of multi-label text classification tasks, we added a sigmoid activation function to the final output layer of each model, thus the output for each document was a vector of per-label scores $s = (s_1, s_2, ..., s_K)$, $s_i \in (0, 1)$, where K is total number of labels. We converted this "soft" predictions into binary "hard" predictions by following thresholding strategy: assign i-th label, if

$$\frac{s_i}{\max(s_1, s_2, ..., s_K)} \geq 0.5. \tag{1}$$

This strategy gave consistently better results than simple thresholding sigmoid outputs on 0.5. After thresholding, we evaluated sample-averaged F1 score.

For datasets which do not contain separated test dataset for evaluation (all but CN), we performed 10-fold cross-validation. For sentiment analysis datasets, we split the data while preserving the percentage of samples for each class (i.e. stratified cross-validation), for topic identification datasets, we split the data

---

[10] https://huggingface.co/transformers.

randomly. For each pre-trained model and each dataset, we fine-tuned 10 different models (one for each fold), averaged the results and reported the mean and standard deviation.

5   Results

We summarize the experimental results in Table 3 (sentiment analysis task) and Table 4 (multi-label topic identification task). In both tables, we first report "base-model-size" category results as these models have more or less comparable number of training parameters (see Table 2 for details). In the last row, we report also results with XLM-RoBERTa-large model, which has, however, several-times more parameters.

Table 3. Table of results for sentiment analysis task. We report comparison of Transformers in the mean of F1 score on three Czech datasets. Aside from base-sized models, we also report comparison with XLM-RoBERTa-large, which is a much larger model.

Model	CSFD	MALL	FCB
MultiBERT [4]	81.04 ($\pm$ 0.31)	76.91 ($\pm$ 0.47)	73.24 ($\pm$ 2.37)
SlavicBERT [2]	82.14 ($\pm$ 0.33)	78.17 ($\pm$ 0.39)	74.19 ($\pm$ 1.85)
XLM-RoBERTa-base [3]	83.74 ($\pm$ 0.39)	77.03 ($\pm$ 0.61)	78.92 ($\pm$ 1.71)
Czert [15]	83.76 ($\pm$ 0.32)	79.35 ($\pm$ 0.52)	78.34 ($\pm$ 1.59)
RobeCzech [16]	85.01 ($\pm$ 0.43)	78.18 ($\pm$ 0.36)	79.12 ($\pm$ 1.39)
FERNET-C5	**85.36** ($\pm$ 0.30)	**79.75** ($\pm$ 0.37)	**81.07** ($\pm$ 1.63)
FERNET-News	84.21 ($\pm$ 0.41)	77.43 ($\pm$ 0.49)	75.94 ($\pm$ 1.85)
XLM-RoBERTa-large [3]	86.03 ($\pm$ 0.32)	79.94 ($\pm$ 0.48)	82.23 ($\pm$ 1.29)

From the sentiment analysis results (Table 3), we can see that among base-sized models, our FERNET-C5 model scored best across all datasets. We attribute this superiority to the underlying pre-training data of the model. In all three fine-tuning datasets, we are dealing with text samples written by common people, often under a strong emotions (e.g. angry about a defective product they just bought, delighted about a movie they just saw etc.). It is also a frequent phenomenon (especially in social media domain), that people are writing posts in a common (spoken) Czech, which is extraordinarily different from literary (standard) Czech, used e.g. by journalists. Moreover, diacritical and punctuation marks are often missing in user's posts. Since the C5 dataset was crawled from miscellaneous Czech web pages, all mentioned phenomenons are covered richly in this dataset. Other popular pre-training data sources (Wikipedia, news articles, books etc.) contain mainly standard Czech, which makes models pre-trained from such sources less suitable for this task. Good results scored by multilingual XLM-RoBERTa models, which were trained from Common Crawl data as well, support this hypothesis.

Table 4. Table of results for multi-label topic identification task. We report comparison of Transformers in the mean of sample-averaged F1 score on two Czech datasets. Aside from base-sized models, we also report comparison with XLM-RoBERTa-large, which is a much larger model.

Model	CTDC	CN
MultiBERT [4]	89.07 ($\pm$ 0.63)	79.83
SlavicBERT [2]	89.59 ($\pm$ 0.48)	80.46
XLM-RoBERTa-base [3]	88.76 ($\pm$ 0.63)	80.35
Czert [15]	90.23 ($\pm$ 0.49)	81.50
RobeCzech [16]	90.47 ($\pm$ 0.53)	81.20
FERNET-C5	**91.25** ($\pm$ 0.38)	82.13
FERNET-News	90.85 ($\pm$ 0.47)	**82.29**
XLM-RoBERTa-large [3]	91.18 ($\pm$ 0.53)	82.78

From the multi-label topic identification results (Table 4), we can see that both our models scored better than other base-sized models. Since our RoBERTa model was pre-trained from in-domain data for this task (news) and RoBERTa is known to perform better than BERT models, we expected the FERNET-News model to score significantly better than other models. However, we were surprised by the universality of the FERNET-C5 model, which performed comparably well, and on the CTDC dataset, the FERNET-C5 model even outperformed in-domain FERNET-News model. Moreover, FERNET-C5 outperformed even the 5-times larger XML-RoBERTa-large model, however the difference is not statistically significant. We attribute such good results of the FERNET-C5 model to the significantly larger pre-training dataset.

To summarize our results, our models outperformed all public multilingual and monolingual Transformers in the base-model-size category on the text classification tasks. The model pre-trained from 93 GB of filtered Czech Common Crawl dataset (FERNET-C5) demonstrated surprising universality across all domains we were experimenting with. Results achieved by our base-sized models are also comparable (although slightly worse in most cases) with results scored by 5-times larger model XLM-RoBERTa-large.

6   Conclusions

In this paper, we used large Czech text datasets to pre-train two monolingual Transformers, one from 93 GB of filtered Czech Common Crawl dataset (FERNET-C5) and one from 21 GB of thoroughly cleaned self-crawled news corpus (FERNET-News). The paper describes the dataset preprocessing as well as the models pre-training procedure and evaluates the models on five different Czech text-classification tasks. Our models outperformed all published multilingual and monolingual Transformers in the base-model-size category. The model pre-trained from Common Crawl (FERNET-C5) demonstrated surprising universality across all domains we were experimenting with.

Since Czech is a special language due to its significant differences between spoken and written form, and since internet users often tend to use spoken form with many grammatical errors to write down a content to be classified (social media posts, reviews etc.), it could be very beneficial to cover these phenomena also in the pre-training corpus. Our results showed that model trained from Common Crawl has universal usage across all domains, whereas models pre-trained from popular literary Czech datasets like Wikipedia or news articles, has limited usage for literary datasets only. For example, our FERNET-News pre-trained from thoroughly cleaned news corpus failed to classify correct sentiment of Facebook posts (F1 75.94 vs. 81.07).

Our models are publicly available for research purposes on HuggingFace hub (See footnote 2).

References

1. Adhikari, A., Ram, A., Tang, R., Lin, J.: DocBERT: BERT for document classification. arXiv preprint arXiv:1904.08398 (2019)
2. Arkhipov, M., Trofimova, M., Kuratov, Y., Sorokin, A.: Tuning multilingual transformers for language-specific named entity recognition. In: Proceedings of the 7th Workshop on Balto-Slavic Natural Language Processing, pp. 89–93 (2019)
3. Conneau, A., et al.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 8440–8451 (2020)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics, Minneapolis (2019)
5. Habernal, I., Ptáček, T., Steinberger, J.: Sentiment analysis in Czech social media using supervised machine learning. In: Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 65–74 (2013)
6. Kral, P., Lenc, L.: Czech text document corpus v 2.0. In: Calzolari, N., et al. (eds.) Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). European Language Resources Association (ELRA), Paris, France, May 2018
7. Křen, M., et al.: SYN v4: large corpus of written Czech (2016). http://hdl.handle.net/11234/1-1846. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University
8. Kudo, T., Richardson, J.: SentencePiece: a simple and language independent subword tokenizer and detokenizer for neural text processing. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 66–71. Association for Computational Linguistics, Brussels, November 2018. https://doi.org/10.18653/v1/D18-2012. https://www.aclweb.org/anthology/D18-2012

9. Lehečka, J., Švec, J., Ircing, P., Šmídl, L.: Adjusting BERT's pooling layer for large-scale multi-label text classification. In: Sojka, P., Kopeček, I., Pala, K., Horák, A. (eds.) TSD 2020. LNCS (LNAI), vol. 12284, pp. 214–221. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58323-1_23

10. Lenc, L., Král, P.: Deep neural networks for Czech multi-label document classification. In: Gelbukh, A. (ed.) CICLing 2016. LNCS, vol. 9624, pp. 460–471. Springer, Cham (2018). https://doi.org/10.1007/978-3-319-75487-1_36

11. Liu, Y., et al.: RoBERTa: a robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692 (2019)

12. Majliš, M.: W2C - web to corpus - corpora (2011). http://hdl.handle.net/11858/00-097C-0000-0022-6133-9. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University

13. Martin, L., et al.: CamemBERT: a tasty French language model. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 7203–7219. Association for Computational Linguistics, July 2020. https://doi.org/10.18653/v1/2020.acl-main.645. https://www.aclweb.org/anthology/2020.acl-main.645

14. Raffel, C., et al.: Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res. **21**(140), 1–67 (2020). http://jmlr.org/papers/v21/20-074.html

15. Sido, J., Pražák, O., Přibáň, P., Pašek, J., Seják, M., Konopík, M.: Czert-Czech BERT-like model for language representation. arXiv preprint arXiv:2103.13031 (2021)

16. Straka, M., Náplava, J., Straková, J., Samuel, D.: RobeCzech: Czech RoBERTa, a monolingual contextualized language representation model. arXiv preprint arXiv:2105.11314 (2021)

17. Sun, C., Qiu, X., Xu, Y., Huang, X.: How to fine-tune BERT for text classification? In: Sun, M., Huang, X., Ji, H., Liu, Z., Liu, Y. (eds.) CCL 2019. LNCS (LNAI), vol. 11856, pp. 194–206. Springer, Cham (2019). https://doi.org/10.1007/978-3-030-32381-3_16

18. Švec, J., et al.: General framework for mining, processing and storing large amounts of electronic texts for language modeling purposes. Lang. Resour. Eval. **48**(2), 227–248 (2014)

19. Vaswani, A., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems, pp. 5998–6008 (2017)

20. Virtanen, A., et al.: Multilingual is not enough: BERT for finnish. arXiv preprint arXiv:1912.07076 (2019)