SlamaTrain – Representative Training Dataset for Slavonic Large Language Models

Marek Medveď[1,2], Radoslav Sabol[1], and Aleš Horák[1]

[1] Masaryk University, Faculty of Informatics
Brno, Czechia,
[2] Lexical Computing
xmedved1@fi.muni.cz, xsabol.fi.muni.cz, hales.fi.muni.cz

Abstract. The Slama project focuses on building a series of foundational language models for Slavonic languages. Even though the latest development yields a number of new large pre-trained and fine-tuned models, the main data source came from English-written websites. Therefore the majority of the training data that is used for language model development consists of the English language. Multilingual language models like Llama, GPT-4o, mT5, etc. are also predominantly (around 80%) trained on the English language, even though they capture the structure of dozens of languages.
In this paper, we detail the process of acquiring one of the largest training datasets for Czech, Slovak and other Slavonic languages. We started with huge multi-lingual datasets, extracted the mono-lingual data and joined them with other sources. The combined mono-lingual datasets were then cleaned, deduplicated and filtered for adult content. As a result, we have obtained 71 billion tokens for the Czech and Slovak languages suitable for the Slama language models training.

Keywords: Slama models, LLM, large language models, training, dataset

1 Introduction

Large language models (LLMs) require extensive and good-quality collections of texts for the causal language modeling task. First LLMs, such as BERT [3] or RoBERTa [9] have been trained purely on English collections of Wikipedia articles or book texts. Current LLMs, such as Meta Llama [6] or Mistral Nemo [10] are built from several orders of magnitude larger networks and for maximum exploitation of the encoded model "memory", they need appropriately larger text collections in tens of languages. However, most of the largest collections are based on Common Crawl [12] that, naturally reflecting the proportions of texts available on-line, are imbalanced in favor of English and other mainstream languages.

A multilingual LLM can process input in most of the languages contained in its training data, however, its internal semantic representations incline to follow

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2024, pp. 25–33, 2024.

the word and phrase senses of the biggest languages. In the following sections, we present SlamaTrain,[3] a new representative training dataset for Czech, Slovak and other Slavonic languages.

2 The Dataset Building Process

The Slama dataset focuses on combining large language datasets into one source that is more focused towards the Slavonic language family. This requirement affects the dataset in favor of Slavic languages, which make up the core of the Slama dataset. The language proportions of the dataset are thus intentionally skewed so that the standard mainstream languages do not outnumber the Slavonic parts. Furthermore, the processing pipeline ensures data de-duplication, non-Latin-script removal and adult content filtering to prevent Large Language models from generating unwanted texts.

2.1 Sources

The Slama dataset consists of Czech, Slovak, Polish, Slovene, Croatian, English, French, Italian, German and Spanish texts. Even though the Slama project primarily focuses on Slavonic languages the dataset also contains selected European mainstream languages that are frequently used (English, German, etc.) and present in Slavonic texts, however, their proportions are reduced.

The texts of the Slama dataset come form several sources:

- CulruraX dataset[4] [11]: this multilingual dataset consisting of 6.3 trillion tokens and 167 languages was developed for training LLM models. Due to its size, data cleanliness that undergoes multiple stages and MinHash at document level de-duplication, it constitutes the main part of Slama dataset.
- HPLT dataset [7]: for the languages in focus (Czech, Slovak, etc.), the HPLT 1.2 dataset served as the second big source. The monolingual part of HPLT includes 75 languages in this release which resolved in 11 TB of de-duplicated files and 8.4 TB of clean files. In the Slama dataset, we include the Czech, Slovak and Slovene part of HPLT 1.2 dataset.
- TenTen corpora [8]: is a family of web corpora created by the Sketch Engine team. Each monolingual corpus usually contains at least 1 billion tokens in the target language. The Czech and Slovak TenTen corpora are included in the Slama dataset.
- Aranea corpora [2]: is a family of Slovak-centric corpora, de-duplicated on the document level and harvested from the web using the SpiderLing tool [15]. This source also expands the Slovak and Czech parts of the Slama dataset.
- czes corpus[5]: is a Czech corpus consisting of newspaper and magazine articles from 1995–1998 and 2002.

[3] Here "Slama" stands for *Slavonic Large Foundational Language Model for AI*.
[4] https://huggingface.co/datasets/uonlp/CulturaX
[5] https://www.sketchengine.eu/czes-corpus/

Table 1: The Czech part of the Slama dataset. The sizes are in millions of words.

Corpus	Source	Latin	in %	Deduplication	in %	Adult filter	in %
Araneum Maius (CS)	1 224	1 223	99.9	1 192	97.3	1 168	95.4
cstenten23	5 747	5 732	99.7	3 248	56.5	3 085	53.7
cstenten_all_mj2	14 278	14 249	99.8	13 450	94.2	12 778	89.5
czes2	455	455	100.0	286	62.8	280	61.5
CulturaX	35 617	35 480	99.6	16 835	47.3	15 825	44.4
HPLT	22 713	22 470	98.9	3 106	13.7	2 889	12.7
SUM	80 037	79 611	99.5	38 120	47.6	36 027	45.0

Table 2: The Slovak part of the Slama dataset. The sizes are in millions of words.

Corpus	Source	Latin	in %	Deduplication	in %	Adult filter	in %
CulturaX	10 323	10 286	99.6	5 911	57.3	5 408	52.4
skTenTen21	1 198	1 196	99.8	1 194	99.7	1 166	97.4
HPLT	5 913	5 854	99.0	2 332	39.5	1 859	31.4
Araneum Maius (SK)	1 244	1 243	99.9	521	41.9	509	40.9
skTenTen2	866	865	99.9	458	52.9	430	49.7
SUM	19 545	19 446	99.5	10 418	53.3	9 375	48.0

– MaCoCu corpora [1]: were created by crawling internet top-level domains from 2021 to 2022. The data underwent boilerplate removing, deduplication, very short texts and non-targeted language removal. The Slama dataset includes the Slovenian part of the MaCoCu corpora.

2.2 Filtering

The sources of the Slama dataset described in the previous section additionally go through a filtering process that exclude texts that contain scripts other than the Latin script, contain adult content and are de-duplicated as a whole collection as the individual sources can contain the same texts.

The original sizes for the Czech and Slovak parts are in Tables 1 and 2. The other languages are omitted as the sources are bigger than required and the resource of this data is usually filtered in their original datasets.

Latin-script Filtering The Slama dataset primarily focuses on Latin-script Slavonic languages, therefore every sentence containing characters outside the

Slovami jedného z diskutujúcich pod článkom "Клин клином вышибают, но на Украине это поняли слишком буквально - дыры дырами латают..."

Fig. 1: Example of the Latin-script filter.

Table 3: Number of both positive and negative instances of all presented adult filtering datasets

Dataset	Adult Documents	Regular Documents
BUT-LCC (CS)	203	2504
Rebalanced (CS)	2264	2504
Rebalanced (SK)	2467	3507

Latin-script and emojis is discarded from the source. An example of such sentence is present in Figure 1. For the Slovak and Czech languages, the resulting Latin filtered data are displayed in Tables 1 and 2.

2.3 Paragraph De-duplication

For Czech and Slovak, we also employ data de-duplication on the paragraph level using the Onion system [13] that removes all duplicate paragraphs seen in previously observed data. Resulting data sizes after de-duplication are present in Tables 1 and 2.

2.4 Adult Content Filtering

The adult content filtering procedure was based on a trained document classifier that provided a real valued score for the input text. The original adult content classification training dataset is a manually annotated subset from the Brno University of Technology Large Czech Collection (BUT-LCC) [5,4]. The classification dataset contains 2,707 samples, where 203 are labeled as adult content. As roughly 7.5% of the samples are represented by adult data, the classification algorithms may become biased towards generic documents that are in of lesser importance.

Privatportal.sk si zamiluje každý pán, ktorý nie je spokojný so svojim sexuálnym životom. Vyskúšajte niečo nové, netradičné a vzrušujúce vďaka sex ponukám Martin. Zažite strhujúce erotické zážitky s príťažlivými slečnami, ktoré ponúkajú svoje erotické služby v Martine. Uprednostňujete štíhle slečny, ktoré sa s vami nežne pohrajú alebo sú vám bližšie ženy plnších tvarov, ktoré vedia s chlapom zatočiť? Z každého rožka troška, nájdete medzi sex inzerátmi Martin. Sex Martin ponúka inzeráty na erotické služby. Dokonalý prehľad kde v meste sa nachádzajú sex priváty Martin poskytujúce rôzne sex praktiky a sex ponuky za peniaze. Inzercia tiež zahŕňa erotické priváty ponúkajúce cez amatérky alebo profesionálky Zafo služby erotické masáže Martin. Vybrať si môžete aj dievča na sex formou Escort služby v Martine na celú noc.Jednoducho si vyber ženu podľa predstáv a rýchlo si s ňou dohodni stretnutie.

Fig. 2: Example of adult content with a borderline score in the Slovak Slama part `sk1552704`. Source url: `privatportal.sk/sex-ponuky/martin`

Table 4: Tokenizer vocabulary sizes for various well-known language models

Tokenizer	Vocabulary Size
GPT-4o	200 000
mT5	250 112
Llama 3	128 256
Llama 2	32 000
Mistral 7B	32 000

To address this issue, we have created silver labels from unlabeled documents to balance the existing dataset using a classifier trained on the original data. The classifier is based on Support Vector Machines (SVM) trained on top of Bag-of-Words (BoW) document representation. The BoW has a maximum vector length of 10,000, where positions encode the TF-IDF of word 1–3-grams. The accents are stripped beforehand, and only n-grams with maximum relative document frequency of 95 % and minimum absolute document frequency of 2 are accepted. Finally, an SVM classifier in the default scikit-learn configuration with radial basis kernel is trained.

Instead of directly predicting classes for each document, we use a score-based approach to compute the TF-IDF vector distance from the separating hyperplane of SVM. Higher scores indicate a higher probability that the selected document is adult content and thus is unsuitable for LLM pretraining.

Using the final adult content classifier for Czech, we have selected 2,061 samples from the SlamaTrain with highest rankings to create a new rebalanced adult content filtering dataset. Finally, we have trained a new SVM to rank the entire Czech corpus.

For Slovak adult content filtering, we have machine-translated the augmented adult content filtering dataset for Czech. We have also added more genuine examples to further improve classification accuracy. For the negative document class, we have added 1,003 new documents mostly from the online video sharing domains, as these are common cases of false positives. The positive class

Table 5: Token/word ratio statistics for each tokenizer. Gray fields indicate that the tokenizer was not trained on the corresponding languages

Tokenizer	cs	sk	pl	en	sl	hr	fr	it	de	es
GPT-4o (Tiktoken)	1.98	1.95	2.26	1.11	1.87	1.92	1.24	1.41	1.49	1.25
Slama 32k (HF)	1.95	1.77	2.06	1.41	1.76	2.12	1.47	1.48	1.77	1.50
Slama 52k (HF)	1.85	1.68	1.91	1.33	1.66	2.04	1.38	1.39	1.64	1.39
Slama 100k (HF)	1.72	1.55	1.74	1.24	1.51	1.93	1.28	1.28	1.51	1.28
Slama 200k (HF)	1.58	1.43	1.58	1.17	1.39	1.82	1.20	1.21	1.38	1.20
Slama 52k (hftoks)	1.75	1.71	1.74	1.33	1.64	2.04	1.34	1.35	1.62	1.35
Slama 100k (hftoks)	1.60	1.55	1.53	1.22	1.46	1.93	1.23	1.22	1.45	1.23
Slama 200k (hftoks)	1.45	1.42	1.36	1.14	1.32	1.83	1.14	1.14	1.33	1.16
Western Slama 100k	1.59	1.47	1.55	1.19	1.94	2.14	1.84	2.22	1.84	2.00

Table 6: Tokenization comparison between GPT-4o and the Slama 200k tokenizer

Tokenizer	Tokens	#tokens
GPT-4o 200k	|Je| lep|ší| být| krás|ná| než| chy|tr|á|,| protože| pr|ům|ěr|ný| muž| lé|pe| vid|í|,| než| mys|lí|.|	26
Slama 200k	|Je| lepší| být| krásná| než| chyt|rá|,| protože| průměrný| muž| lépe| vidí|,| než| myslí|.|	17
GPT-4o 200k	Ve| tř|íd|ě| je| tř|ic|et| d|ětí|.| Ve| škole| se| uč|íme| ří|kat| ří|k|anky|.|	23
Slama 200k	|Ve| třídě| je| třicet| dětí|.| Ve| škole| se| u|číme| říkat| ří|kan|ky|.|	16

was enriched with 203 documents from Slovak adult websites that have reached a threshold score of 1.0 on a model trained on the original machine-translated data. The final proportions of newly-created datasets can be observed in Table 3.

An example of document containing a adult content is present in Figure 2. This example represents a borderline score document that was still assessed as unwanted text and removed from the dataset.

From the final text the adult-content filtering process removes all documents exceeding a threshold determined from manual data examination. Additionally if a major part of the web domain was removed with the first step, in the second step the whole domain is removed. Resulting sizes for the Slovak and Czech Slama parts are again present in Tables 1 and 2.

2.5 Dataset Tokenization

The dataset preparation phase includes decisions related to converting the text to tokenized versions. The baseline for comparison of the proposed tokenizers is the latest GPT-4o tokenizer, where OpenAI claims improved compute performance and lower output token lengths for mid to low-resourced languages.

Slama tokenizer (HF) is a GPT2-style byte-level version of the Byte-Pair Encoding (BPE). We have experimented with vocabulary sizes of 200k, 100k, 52k, and 32k. The choices for vocabulary sizes were made according to the sizes of well recognized language models as present in Table 4. Contrary to GPT-2, we add the End of Sequence token (EOS) at the beginning of each prompt. Western Slama Tokenizer (HF) uses the same setting as the Slama tokenizer with a lower vocabulary size due to lesser language diversity. Each of the tokenizers were trained on 10 million documents per each language.

High-Frequency Tokenizer (HFT) [14] was also tested, as a recent subword tokenization algorithm that addresses the problems translating low frequency tokens in neural machine translation. The goal is to produce a vocabulary of tokens with as high frequency representation in the data as possible.

Evaluation Method We have selected a sample of 10,000 documents for each language that is disjoint from the training data. For each tokenizer, we divide the

Table 7: Dataset token size statistics

Dataset	Disk Usage	# data shards	# tokens (B)
Slama	688 GB	44 914	358 B
Western Slama	301 GB	19 494	166 B

Table 8: The final Slama dataset sizes

Language	Size in words	Size in tokens
Czech	36 027 958 429	57 284 453 902
Slovak	9 375 238 405	13 969 105 223
Polish	35 000 000 354	53 550 000 542
Slovene	10 650 799 895	14 698 103 855
Croatian	1 221 882 240	2 223 825 677
Spanish	35 000 000 136	42 000 000 163
English	35 000 000 481	41 650 000 572
French	35 000 000 332	42 000 000 398
Italian	35 000 000 427	42 350 000 517
German	35 000 000 114	48 650 000 158
SUM	267 275 880 813	358 375 491 007

number of tokens produced with BPE by the number of words (tokens created by Unitok tokenizer), resulting in token/word ratios.

According to Table 5, HFtoks provides the most compact representation of the input data. However, it is still in a prototype stage where some implementation details and compute performance need to be improved. Until then, HuggingFace BPE implementation is deemed as a more suitable choice.

Typically, higher vocabulary sizes help with creating shorter token representations. However, even with the smallest vocabulary sizes, Slama tokenizer outperforms GPT-4o tokenizer in both Czech and Slovak languages.

The last step of data processing is a conversion to the MosaicML streaming dataset format ready for use in LLM training. The format is designed to make training on large datasets fast and scalable in a distributed setting. We have converted the data on a single machine, which took approximately a week for Western Slama and two weeks for the whole Slama dataset. The resulting dataset token size statistics can be seen in Table 7, where *n* data shards denotes number of files stored in the filesystem. Each shard is a 64MB portion of the dataset. When compressed via ZSTD, the size of each shard reduces approximately to 17MB, which is almost 26 % of the original size.

3 Conclusions and Future Directions

We have presented the details of the first phase of training new large generative models oriented to Slavonic languages, so called Slama models. The quality of model internal knowledge representation depends on the size and quality of

the training datasets. We have thus identified all large sources of texts for Czech, Slovak, Polish and other Slavonic languages that are based on Latin-script. The data was then merged, de-duplicated, cleaned and filtered for unwanted content. The resulting dataset consists of more than 71 billion tokens for the Czech and Slovak languages, making it one of the largest cleaned datasets for them, and 358 billion tokens for the complete dataset with 10 languages.

In the coming months, the SlamaTrain dataset is being used in training a series of new Slama generative language models and their evaluation.

Acknowledgements. This work has been partly supported by the Ministry of Education, Youth and Sports of the Czech Republic within the LINDAT-CLARIAH-CZ project LM2023062.

The authors acknowledge the OSCARS project, which has received funding from the European Commission's Horizon Europe Research and Innovation programme under grant agreement No. 101129751.

Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

References

1. Bañón, M., et al.: MaCoCu: Massive collection and curation of monolingual and bilingual data: focus on under-resourced languages. In: Proceedings of the 23rd Annual Conference of the European Association for Machine Translation. pp. 303–304. European Association for Machine Translation, Ghent, Belgium (Jun 2022), https://aclanthology.org/2022.eamt-1.41
2. Benko, V.: Aranea: Yet another family of (comparable) web corpora. In: Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds.) Text, Speech and Dialogue. pp. 247–256. Springer International Publishing, Cham (2014)
3. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of NAACL-HLT. pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota, USA (2019)
4. Doležal, J.: Adult content classifier dataset (2024), https://huggingface.co/datasets/BUT-FIT/adult_content_classifier_dataset
5. Doležal, J., Dočkal, M., Fajčík, M., Kišš, M., Beneš, K., Ondřej, K., Hradiš, M.: Brno University of Technology Large Czech Collection (2024), https://huggingface.co/datasets/BUT-FIT/BUT-LCC
6. Dubey, A., et al.: The Llama 3 Herd of Models. arXiv preprint arXiv:2407.21783 (2024)
7. de Gibert, O., Nail, G., Arefyev, N., Bañón, M., van der Linde, J., Ji, S., Zaragoza-Bernabeu, J., Aulamo, M., Ramírez-Sánchez, G., Kutuzov, A., Pyysalo, S., Oepen, S., Tiedemann, J.: A new massive multilingual dataset for high-performance language technologies. In: Calzolari, N., Kan, M.Y., Hoste, V., Lenci, A., Sakti, S., Xue, N. (eds.) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 1116–1128. ELRA and ICCL, Torino, Italia (May 2024), https://aclanthology.org/2024.lrec-main.100

8. Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., Suchomel, V.: The tenten corpus family. In: 7th International Corpus Linguistics Conference CL 2013. pp. 125–127. Lancaster (2013), http://ucrel.lancs.ac.uk/cl2013/
9. Liu, Y., et al.: RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692 (2019)
10. Mistral AI team: Mistral NeMo (2024), https://mistral.ai/news/mistral-nemo/
11. Nguyen, T., Nguyen, C.V., Lai, V.D., Man, H., Ngo, N.T., Dernoncourt, F., Rossi, R.A., Nguyen, T.H.: Culturax: A cleaned, enormous, and multilingual dataset for large language models in 167 languages (2023)
12. Patel, J.M.: Introduction to Common Crawl Datasets, pp. 277–324. Apress, Berkeley, CA (2020). https://doi.org/10.1007/978-1-4842-6576-5_6, https://doi.org/10.1007/978-1-4842-6576-5_6
13. Pomikálek, J.: Removing boilerplate and duplicate content from web corpora. Ph.D. thesis, Masaryk university, Faculty of informatics, Brno, Czech Republic (2011)
14. Signoroni, E., Rychlý, P.: HFT: High frequency tokens for low-resource NMT. In: Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022). pp. 56–63. Association for Computational Linguistics, Gyeongju, Republic of Korea (Oct 2022), https://aclanthology.org/2022.loresmt-1.8
15. Suchomel, V., Pomikálek, J.: Efficient web crawling for large text corpora. In: Proceedings of the seventh Web as Corpus Workshop (WAC7). pp. 39–43 (2012)