Filtering of Czech Texts

Jan Brichta

Faculty of Informatics
Masaryk University, Czech Republic
492814@mail.muni.cz

Abstract. This paper compares several approaches to text filtering for the Czech language. Two manually created gold standard datasets were prepared for evaluating filter performance. We will also examine what characteristics could contribute to lowering of a text quality. Models are evaluated on data with gradually increasing levels of text noise to assess their robustness and consistency. In addition, basic model reasoning and processing speed are analyzed.

Keywords: Text Filtering; Filter Evaluation; Text Classification; Czech; LLMs

1 Introduction

High quality texts are crucial for modern LLMs and neural machine translation models. Both benefit from large scale corpora, but their performance depends strongly on the cleanliness of the data used. Web sources such as Common Crawl provide enormous quantities of text, yet they contain a large amount of noise, which can degrade performance unless appropriately filtered. [11,6].

This paper investigates several text filters for the Czech language. To evaluate robustness, model performance is measured on datasets with progressively increasing noise levels and a manually curated gold standard test set is used for quantitative comparisons.

Contributions of this work are:

- A comparative evaluation of several text filtering methods applied to Czech.
- Two Czech human made gold standard datasets for filter evaluation, including experiments across controlled noise levels.
- An exploration of factors influencing filtering quality, together with observations on model reasoning behavior and computation time.

2 Datasets

Several datasets were used in this work. These include Czech texts filtered from Reddit [1], Twitter [2], and the mtdata dataset collection used in the WMT24 shared task [12].

A. Horák, P. Rychlý, A. Rambousek (eds.): Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2025, pp. 127–136, 2025.

2.1 Human Annotated Gold Standards

Two manually annotated datasets were created to serve as gold standards for evaluating filter performance.

Reddit Annotated Dataset. This dataset consists of 2,000 Czech Reddit sentences, each labeled by a human as `GOOD` or `BAD` according to overall sentence quality. The annotation criteria included grammatical correctness, sentence completeness, and overall clarity of the message.

Mixed Annotated Dataset. To achieve a more balanced and diverse evaluation sample, a second dataset was compiled from Reddit, Twitter, and mtdata for human evaluation. It contains 209 sentences, each annotated with a quality label and a short justification describing the reasoning behind the decision. The reasoning part of the dataset was originally intended to be used for future fine tuning of models to increase filtering performance.

3 Good vs. Bad Sentences

Determining what is a *good* or *bad* sentence is not straightforward. A good sentence can be defined linguistically (grammatical, complete) or functionally (useful for improving model performance). In this work, a good sentence is considered to be a well formed sentence in Czech. This section outlines common sources of low quality sentences observed during annotation and illustrates them with representative examples.

It should be noted that the presence of a few of the identified attributes, like a single emoticon or an occasional typo does not necessarily imply poor quality. In fact, a limited amount of noise could improve model robustness and generalization [5].

Non-Sentences: These seem to come from incorrect text extraction from crawled webs, as names of chapters, misaligned sentence segmentation or sometimes ASCII-art (meaning images created from characters):

```
Blbosti.cz | O nás | Kontakty | Nápověda | Mapa stránek
```

Non-Czech: Despite prior language filtering, some sentences still contain foreign text or mixed language content. The next sentence is from `Statmt--commoncrawl-wmt22-ces.ces` from the WMT shared task and the second is from Twitter data filtered for Czech language:

```
Consulta la disponibilidad
I'm at Restarace Čtyřlístek in Ústecký https://t.co/TSWbnMMUNc
```

Emoticons: Although emoticons can convey the user's emotion, they can make the text less readable when used in excess:

`Billa je ještě dost v pohodě oproti Penny`💀💀💀💀💀💀

Tags and Links: Often appearing in texts from Twitter that reference websites and users with # and `@` or as a web link:

```
Dnes mi v @nutricCZ udělali fakt radost.
  #mademyday https://t.co/EbUxf2TZbr
```

Informal Speech, Slang and Grammatical Errors: These could be missing interpunction, missing capitalization or usage of slang words:

`místo lolka bych dal valo`

Very Short Sentences: Some sentences are too short to provide meaningful information. These could again be due to incorrect sentence segmentation, or simply a reply in a conversation:

`zatím... zatím... snad.`

4 Filtering Models and Configuration

This section describes the models and methods used to identify low quality sentences in the datasets.

4.1 Language Classifier as Filter

For this model, two language identification tools, `ftlangdetect`[4,3] and `langid`[8] were used as a fast indication of text quality. In addition, lines shorter than 20 characters were automatically discarded, as they rarely convey meaningful information. In total, text is considered high quality if both models label it as Czech and the text is also longer than 20 characters.

4.2 Large Language Models

Three large language models were used for sentence classification:

- `qwen3-30b-a3b-instruct-2507-q6_k_xl` [10], from now on referred to as qwen3.
- `gpt-oss-120b` [7], referred to as gpt-oss.
- `glm4.6-iq4_k` [9], referred to as glm4.6.

These models were selected based on their availability on the servers that the experiments were run on.

Each model received prompts containing sentence examples and was asked to label them as `GOOD`, `NOT SURE` or `BAD` according to the previously described criteria. For simplicity of evaluation, anything that is not `GOOD` is considered `BAD`.

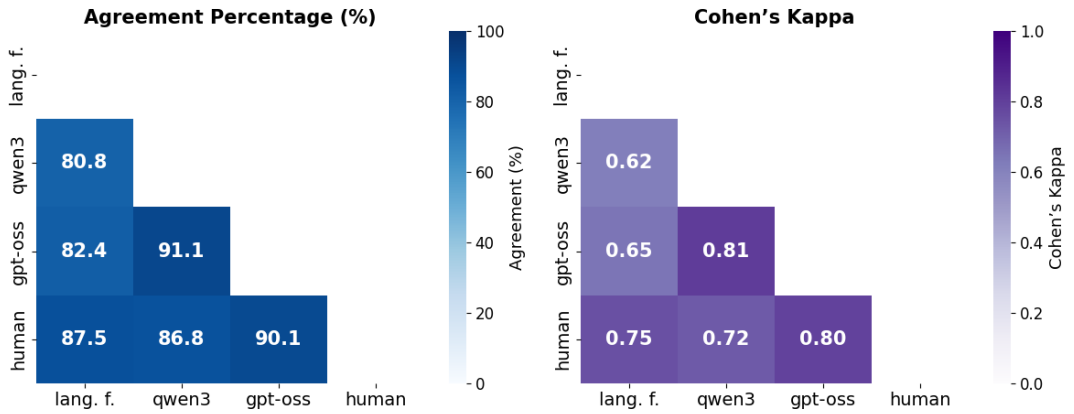

Fig. 1: Heatmap showing the percentage of agreement on Reddit gold standard dataset, along with Cohen's kappa for the same pairs.

4.3 Ability to Follow Instructions

During the evaluation of LLMs, it was observed that the model gpt-oss consistently replaced the ASCII sequence '->' with the arrow symbol '→', despite explicit instructions not to do so. This behavior required minor adjustments to the answer parsing process to ensure correct interpretation of model outputs.

4.4 Few Shot Learning in Prompts

No complex prompt optimization was performed in this work. However, to improve reliability and reduce undesired rejections, a few shot learning was applied.

Three representative examples were added to the prompt to demonstrate the expected reasoning and labeling format. This aimed to make the model more tolerant of minor grammatical errors and to encourage consistent labeling:

```
SENTENCE: A 7% šance, že je to vágus hledající odemčený auto.
REASON: Proper and understandable Czech sentence with slight
  slang (vágus).
-> GOOD
```

5 Evaluation of Models

This section evaluates the performance of the filtering models described in Section 4 using the datasets introduced earlier along with other metrics, like average time to classify a line.

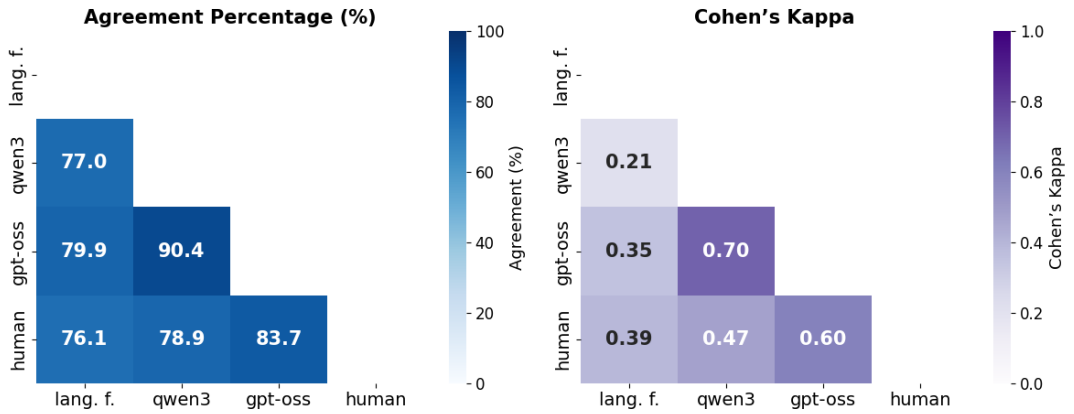

Fig. 2: Heatmap showing the percentage of agreement on the mixed gold standard dataset, along with Cohen's kappa for the same pairs.

5.1 Model Comparison on Gold Standard Data

Figures 1 and 2 show the pairwise agreement between the evaluated models and the human annotations. Agreement values represent the percentage of matching `GOOD`/`BAD` decisions and Cohen's kappa measures the agreement beyond what would be expected by chance. The goal is to measure how closely model classifications align with human judgments of sentence quality.

On the Reddit gold standard, the highest agreement with human annotation was achieved by gpt-oss, the largest of the evaluated models. Overall agreement scores ranged between roughly 80–90%, and rather high Cohen's kappa between 0.6 to 0.8 indicating high consistency across models.

It is also worth noting that the Czech Reddit data was obtained by using `langid` which is part of the language filters model. However the overall model of both language filters labels around 50% of the Reddit gold standard dataset as `GOOD`.

As shown in Figure 2, overall agreement levels in the mixed gold standard were lower than the Reddit gold standard (Figure 1). Here gpt-oss again showed the highest alignment with human judgments, although all models had reduced scores when compared to the Reddit evaluation.

It is interesting to note that the most severely impacted values from changing the dataset are agreement percentages in the comparisons with the human labeling, while the most severely impacted Cohen's kappas are from the language filter model. And like in Figure 1, Qwen3 and gpt-oss maintain a very high agreement in both the percentage and the Cohen's kappa.

5.2 Evaluation by Incrementing Text Noise

To further examine the models, different types of noise are gradually introduced to a clean sentence to explore how it affects their classification. Though the

noised up sentences are not present in the data, it is a good illustration of the behavior of the models.

Starting from a sentence that all models classified as `GOOD`, the following noise types corresponding to the error classes defined in Section 3 are successively introduced: grammatical errors, informal speech, emoticons, # and @ tags, web links, non-Czech words, and sentence shortening. After each modification, the sentence was re-evaluated by all models.

The original clean sentence was:

```
Domnívám se, že zločin proti solidaritě, který byl spáchán, jak
  dokazuje tento proces, by měl Parlament odsoudit.
```

After adding several noise types (grammar errors, informal words, and an emoticon), the corrupted version becomes:

```
Domnívám se, že zločin proti solidaritě kterej byl spáchánej jak
   dokazuje tento proces by mněl parlament odsoudit😕
```

Finally, adding tags, links, foreign words, and truncation gives us:

```
Crimes against solidarity: #soud zločin proti solidaritě kterej byl
  spáchánej jak dokazuje tento proces by mněl @parlament😕
  https://pocasi.seznam.cz/radar?lat=49.686&lon=15.926&z=5
```

Table 1 summarizes how many modification iterations each model tolerated before it stopped classifying the sentence as `GOOD`. A higher number than indicates greater robustness to accumulated noise, at least for this sentence.

Table 1: Iterations until sentence stops being labeled `GOOD`.

Filter Name	language filters	Qwen3	gpt-oss	glm4.6
Number of Iterations	all good	7	8	5

Another approach is to apply noise only from a given class. For example only adding more grammatical errors to a sentence and seeing how the results change. The last iteration of such an example being:

```
domnivam se ze zlocin proty solidarite ktery bil spachan jak
  dokazuje tento proces bi mnel parlament odsoudit
```

Table 2: Iterations until sentence is not considered GOOD, with classes separately.

Noise Type	language filters	Qwen3	gpt-oss	glm4.6
# and @ Tags	all good	4	3*	3*
Links	all good	4	5	3
Grammar Errors	all good	all good	all good	all good
Non-Czech	5	1	1	1
Sentence Shortening	5	all good	2*	4*
Emoticons	all good	all good	all good	all good

* classifier later accepts some sentences after more noise is applied.

Table 2 indicates that the simple language filters, with the exception of short sentences, are more lenient compared to others, only rejecting non-Czech and short sentences. In contrast, gpt-oss and glm4.6 are the strictest. Though none of the filters seemed to consider any of the grammar error noise to be severe enough to not be accepted. This could be caused by insufficient problem specification in the prompt.

5.3 Model Speed

Although the evaluated models were executed on different hardware (language filters on CPUs and LLMs on GPUs), it is still informative to compare their relative processing times for the specific servers the experiments were run on.

The speed measurements were obtained using the same noised up text used in the previous experiments. While only 45 sentences long, the differences in processing time are large enough to show clear performance differences between models.

Table 3: Average time to classify a line.

filter name	language filters	Qwen3	gpt-oss-120b	glm4.6
average time	18.8 ms	3.4 s	20.6 s	153.6 s

The results in Table 3 show that traditional language filters are orders of magnitude faster than LLM based models. Among the LLMs, Qwen3 was the most efficient, on average processing a line more than forty times faster than glm4.6.

Due to its long inference time, glm4.6 was excluded from the gold standard evaluation experiments.

5.4 Reasoning Examples of Qwen3 and gpt-oss

In addition to quantitative evaluations, we will now inspect a subset of the reasoning outputs of Qwen3 and gpt-oss. These examples are an attempt to illustrate the behavior of the model in an explainable way. For clarity, only the model response to each sentence is shown without the full prompt.

Qwen3 Example.

```
request took 0.9083759784698486s
INPUT: Manžel je proti Markéta Reinischová, Elmer Carvalho
OUTPUT: REASON: The sentence starts with a proper Czech subject
 "Manžel je proti" but then shifts to foreign names "Markéta
  Reinischová, Elmer Carvalho" without context or proper grammatical
  structure. It lacks a complete thought or meaningful sentence in
  Czech. The names are not integrated into a coherent Czech sentence.
-> BAD
```

Here the model correctly identify that the structure of the sentence is incorrect.

gpt-oss Example.

```
request took 11.366015434265137s
INPUT: "Maminko, vyzapněte prosím ten osobní počítač!"
  "Čeho díš, synáčku můj milovaný?"
  "Prostě přespusťte prováděcí zřízení!"
  "Však jsem se jej jala otočiti již třikráte a stále se neche
  nazout zavaděč vykreslovací desky!"
OUTPUT: REASON: The text is Czech-looking but contains
  many misspellings, non-standard words and largely nonsensical
  phrasing, making the meaning unclear. The high amount of errors
  lowers its quality despite being in Czech.
-> BAD
```

In this case, the model considers the use of archaic language and unusual phrasing to be a grammar error. Though the text quality might be debatable, it does not contain ”many misspellings” like the model claims.

Overall, the manual inspection suggests that both LLMs provide understandable natural language reasoning for their classifications.

6 Known Issues

There are multiple limitations that were identified during this work:

Attribute Representation. It is not known whether all the mentioned cases of bad sentence attributes (e.g.: links, short sentences) are equally important. In fact the noised sentence evaluation suggests that they might not be.

Subjectivity of Sentence Quality. Sentence quality is inherently subjective. Even a single annotator may disagree with their own earlier judgment when re-evaluating the same sentence. This variability was observed during the creation of the gold standard datasets and is a limitation of human evaluation.

Dataset Size and Coverage. The gold standard data is relatively small due to the manual annotation effort required. This limits the ability to draw strong conclusions on model performance.

Model and Hardware Constraints. Model evaluations were performed under differing hardware conditions (CPUs for language filters, GPUs for LLMs). Execution speed and model size therefore limit the scalability of experiments. Specifically, the glm4.6 model was excluded from the gold standard experiments due to its prohibitive computational cost.

Evaluation Metrics. The analysis in this paper relies on human comparisons and agreement scores. More quantitative measures (e.g., training and evaluation of LLMs or machine translation models) are needed to better estimate model quality beyond human interpretation.

7 Future Work

Future work will aim to overcome some of the aforementioned limitations.

- Creating training data obtained by using gpt-oss as a filter to train a FastText and neural network classifiers in a similar way to FineWeb2-HQ [6]
- And Secondly, a machine translation model using data from the filters in order to obtain a more practical and objective evaluation of data quality on metrics like COMET or SacreBLEU scores.

8 Conclusions

The experiments demonstrate that LLMs can identify low quality Czech sentences with high agreement to human annotations, but at a large computational cost.

Among the tested models, gpt-oss achieved the highest alignment with human labels, while Qwen3 offered a reasonable trade off between speed and accuracy, along with a sensible reasoning for its answers.

Traditional language filter based on `langid` and `ftlangdetect` were significantly faster but had lower agreement with other models.

And lastly, the noised text experiments showed that models respond differently to specific noise types: Specifically, the models were more lenient toward grammar errors and emoticon noise than other types of noise.

Acknowledgements. The described study comes from the project "On our own: Opportunities and Risks in the Individualization of Society (PRINS) CZ.02.01.01/00/23_025/0008710" (PRINS), which is co-financed by the European Union. DeepSeek LLM was used to correct this paper's style and grammar mistakes.[1]

References

1. academictorrents.com: Reddit data. https://academictorrents.com/details/ba051999301b109eab37d16f027b3f49ade2de13, accessed: 2025-11-08
2. huggingface.co: Twitter data. https://huggingface.co/datasets/enryu43/twitter100m_tweets, accessed: 2025-11-08
3. Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., Mikolov, T.: Fasttext.zip: Compressing text classification models. arXiv preprint arXiv:1612.03651 (2016), specifically from https://pypi.org/project/fasttext-langdetect/
4. Joulin, A., Grave, E., Bojanowski, P., Mikolov, T.: Bag of tricks for efficient text classification (2016), specifically from https://pypi.org/project/fasttext-langdetect/
5. Karpukhin, V., Levy, O., Eisenstein, J., Ghazvininejad, M.: Training on synthetic noise improves robustness to natural noise in machine translation (2019), https://arxiv.org/abs/1902.01509
6. Messmer, B., Sabolčec, V., Jaggi, M.: Enhancing multilingual llm pretraining with model-based data selection (2025), https://arxiv.org/abs/2502.10361
7. OpenAI: gpt-oss-120b & gpt-oss-20b model card (2025), https://arxiv.org/abs/2508.10925, model specifically from https://huggingface.co/openai/gpt-oss-120b
8. pypi.org: langid python library. https://pypi.org/project/langid/, specifically from https://pypi.org/project/langid/1.1.5/
9. Team, ., Zeng, A., Lv, X., Zheng, Q., et al.: Glm-4.5: Agentic, reasoning, and coding (arc) foundation models (2025), https://arxiv.org/abs/2508.06471, model specifically from https://huggingface.co/Thireus/GLM-4.6-THIREUS-IQ4_K-SPECIAL_SPLIT
10. Team, Q.: Qwen3 technical report (2025), https://arxiv.org/abs/2505.09388, model specifically from https://huggingface.co/unsloth/Qwen3-30B-A3B-Instruct-2507-GGUF/
11. Wenzek, G., Lachaux, M.A., Conneau, A., Chaudhary, V., Guzmán, F., Joulin, A., Grave, E.: Ccnet: Extracting high quality monolingual datasets from web crawl data (2019), https://arxiv.org/abs/1911.00359
12. www2.statmt.org: Wmt24 shared task data collection. https://www2.statmt.org/wmt24/mtdata/, accessed: 2025-11-08

[1] https://chat.deepseek.com