Bootstrapping a Sentence-Level Corpus Quality Classifier for Web Text using Active Learning

Maximilian Bley[1], Thomas Eckart[2], and Christopher Schröder[1,3]

[1]Institute for Applied Informatics (InfAI) at Leipzig University
[2]Saxon Academy of Sciences and Humanities, Leipzig
[3]Center for Scalable Data Analytics and Artificial Intelligence (ScaDS.AI), Dresden/Leipzig

Abstract

The quality of training data is an essential factor for training large language models (LLMs) as it directly impacts their performance. While high-quality data is crucial for training competitive LLMs, existing preprocessing pipelines still partly rely on rules, which are computationally cheap but also inherently limited to simpler patterns. Model-based filtering on the other hand, is more flexible and can detect finer-grained patterns and semantics, but often requires substantial amounts of labeled data. While there are models for common problems (such as toxicity classification), this is often only the case for resource-rich languages and well-studied problems—leaving gaps in coverage for other languages, problems, or combinations thereof. In this work, we investigate the feasibility of model-based preprocessing despite the absence of labeled data. We use active learning to bootstrap a sentence-level multi-label classifier that detects textual problems of traditional text cleaning approaches. With only 498 examples, the final classifier reaches macro- and micro-F$_1$ scores of 0.80 and 0.84, making it suitable for practical use. Moreover, we find that it captured subtle errors compared to a rule-based baseline. We publish the training code, a labeled corpus quality classification dataset, and the resulting classifier[1].

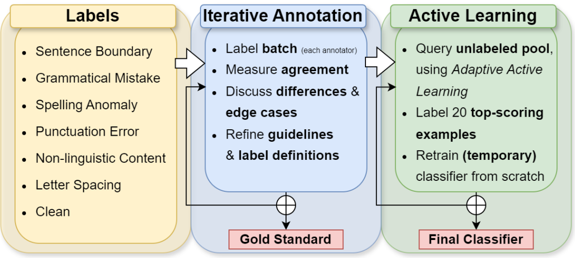
Labels
Sentence Boundary
Grammatical Mistake
Spelling Anomaly
Punctuation Error
Non-linguistic Content
Letter Spacing
Clean
Iterative Annotation
Label **batch** (each annotator)
Measure **agreement**
Discuss **differences** & **edge cases**
Refine **guidelines** & **label definitions**
Gold Standard
Active Learning
Query **unlabeled pool**, using *Adaptive Active Learning*
Label 20 **top-scoring examples**
Retrain **(temporary)** classifier from scratch
Final Classifier

Figure 1: The development process of our approach. We begin by defining seven corpus quality labels along with their annotation guidelines, then we annotate a gold standard for evaluation, and finally train a corpus quality classifier with active learning.

1 Introduction

Pre-training large language models (LLMs) requires not only vast amounts of textual data but also high-quality content, as recent studies show the impact of data quality on downstream performance (Raffel et al., 2020; Penedo et al., 2023; Longpre et al., 2024; Li et al., 2024).

While there have been many efforts to curate and clean LLM pre-training corpora, only some of the possible steps use model-based approaches such as language identification (Joulin et al., 2016; Grave et al., 2018), perplexity-based filtering (Ankner et al., 2025; Thrush et al., 2025), predicting similarity to reference text (Li et al., 2024), toxic or adult content detection (Soldaini et al., 2024), or the targeted search for certain contents such as educational texts (Wettig et al., 2024). However, there still are various preprocessing steps that resemble traditional text cleaning. They target noise that usually results from text extraction artifacts in web corpora such as, among others, incorrectly formatted text, non-linguistic content, random word sequences, letter spacing, encoding errors, or repeating characters, and are still predominantly rule-based (Albalak et al., 2024). Developing such rules is known to be time-consuming, often highly tailored to specific domains and languages, and may fail to capture more subtle issues compared to supervised models (Laurençon et al., 2022; Longpre et al., 2024; Henriksson et al., 2025).

Supervised learning, however, requires a substantial amount of training data. While there are some existing datasets for tasks such as toxic or adult content detection, they only cover a limited number of languages, and moreover, to the best of our knowledge, none of them address the problems that are usually handled by rule-based filtering. With

[1] https://github.com/maximilian-bley/german-webtext-quality-classification

Proceedings of the Workshop on Beyond English: NLP for all Languages in an Era of LLM from Texts associated with RANLP 2025, pages 98–109, Varna, Bulgaria, Sep 12, 2025.
https://doi.org/10.26615/978-954-452-105-9-012

smaller[2] language models[3] becoming increasingly effective, we argue that it is time to widen the scope of model-based preprocessing.

To investigate the feasibility of model-based preprocessing, we train a supervised corpus quality classifier with seven classes: six representing distinct types of textual deficiencies and one capturing the absence of those.[4] Since no training data exists for this particular task, we apply *active learning*, an iterative approach that aims to minimize annotation effort. We begin by developing a classification scheme and corresponding annotation guidelines. To evaluate the resulting classifier, we create a gold standard through iterative annotation, optimizing label agreement among three annotators (Pustejovsky and Stubbs, 2012; Klie et al., 2024). Figure 1 summarizes our three-staged development process.

We investigate the following research questions:

RQ1 How effective is a sentence-level classifier in recognizing several text quality classes, given an active learning scenario with a budget of an 8-hour day of annotation?

RQ2 How does the resulting classifier that has been trained only on few examples perform in comparison to a rule-based approach?

Contributions (1) We develop and refine a classification scheme and corresponding guidelines to obtain a gold dataset for evaluation. (2) We perform an active learning experiment investigating the feasibility of a sentence-level classifier, built for a specific domain and language in under 8 hours. (3) We compare our approach to a rule-based one.

Results The final classifier shows reasonable performance despite being trained on only 498 examples, reaching macro- and micro-F_1 scores of 0.80 and 0.84 respectively. Compared to a rule-based baseline, our approach achieves improvements of four to five percentage points in F_1 and captures certain types of errors, often more subtle, that the rule-based system tends to miss. We publish the training code, a labeled corpus, and the classifier.

2 Related Work

Preprocessing of Web Data The web has long been used as an important source of text data in natural language processing (Kilgarriff and Grefenstette, 2003), but requires cleaning procedures to remove noisy parts such as boilerplate code, encoding errors, non-linguistic content, or broken text. In the context of LLM training, text cleaning has gained renewed attention, since carefully-curated high-quality data is the currently best known recipe for training strong models (Penedo et al., 2023, 2024). Some preprocessing steps involve handcrafted (often language-specific) rules that have been developed in and adopted from previous work such as the C4 (Raffel et al., 2020) and ROOTS (Laurençon et al., 2022) corpora. Similar (or partly even identical) heuristics have been confirmed in follow-up work and are still used in more recent datasets such as FineWeb (Penedo et al., 2024). Notably, while some preprocessing steps, such as language identification, are realized using models, many cleaning steps still rely on rules.

Active Learning Transformer-based language models (Vaswani et al., 2017; Devlin et al., 2019) have shown considerable results in the context of active learning for text classification using only a small amount of data (Ein-Dor et al., 2020; Margatina et al., 2021; Schröder et al., 2022), encouraging this work where a lack of training data is a severe obstacle. With language models continuously increasing in size, some recent approaches even attempt to replace the human annotator with an LLM (Xiao et al., 2023; Kholodna et al., 2024). Many contemporary corpora are, however, very large, and computational costs are still an obstacle for practical active learning (Romberg et al., 2025), therefore we opt to use small language models, which have shown remarkable effectiveness (Nachtegael et al., 2023; Schröder and Heyer, 2024; Gonsior et al., 2025), while at the same time allowing us to process larger volumes of data.

The majority of the recent work at the intersection of language models, active learning, and text classification revolves around single-label classification (among others in the works of Ein-Dor et al. (2020) and Lesci and Vlachos (2024)), while studies focusing on multi-label active learning are rare (e.g., Wertz et al. (2022a,b, 2023) and Wang and Liu (2023)). Moreover, active learning research is often operationalized through simulated

[2] The distinction of what is considered a *small* model is evolving, but the important aspect is that at the current time larger models quickly render computation efforts infeasible, while small models can process large amounts of data despite of limited compute resources.

[3] We rely on the definition of Rogers and Luccioni (2024) for LLMs, which includes encoder models.

[4] We use *model* to refer to the base architecture (e.g., BERT (Devlin et al., 2019)), and *classifier* to denote the model including a task-specific classification head.

experiments (Margatina and Aletras, 2023). Therefore, practical multi-label active learning applications are highly important to investigate the effectiveness of contemporary active learning.

3 Quality Criteria and Gold Standard

Our approach is not limited to a specific corpus or language. The following work is conducted at the example of German web text, which is reflected in the class descriptions and textual examples.

3.1 Quality Criteria Labels

We define *low-quality* labels to capture visible deficiencies that interrupt the flow of a text on the lexical and syntactical level. Conversely, text without such interruptions is considered *high-quality* (or *clean*). These labels are inspired mostly by rules from related work (e.g., by Raffel et al., 2020; Kreutzer et al., 2022; Laurençon et al., 2022) and from a field-tested rule-based approach, developed for the same kind of data (Goldhahn et al., 2012).[5]

To provide examples to the reader, exemplary sentences with their corresponding label sets are presented in Table 1, where one label is highlighted for each example. The respective classes are defined in the following:

Sentence Boundary Sentence boundary errors occur if the start or ending of a sentence is malformed. This is the case if it begins with a lower case letter or an atypical character, or lacks a proper terminal punctuation mark (e.g., period, exclamation mark, or question mark).

Grammar Mistake Grammar mistakes are any grammatical errors such as incorrect articles, cases, word order and incorrect use or absence of words. Moreover, random-looking sequences of words, usually series of nouns, should be tagged. In most cases where this label is applicable, the sentence' comprehensibility or message is impaired.

Spelling Anomaly A spelling anomaly is tagged when a word does not correspond to German spelling. This includes typos and incorrect capitalization (e.g. "all caps" or lower-case nouns). Spelling anomalies are irregularities that occur within the word boundary, meaning here text between two whitespaces. In particular, individual letters or nonsensical word fragments are also tagged.

Punctuation Error Punctuation errors are tagged if a punctuation symbol has been placed incorrectly or is missing in the intended place. This includes comma errors, missing quotation marks or parentheses, periods instead of question marks or incorrect or missing dashes or hyphens.

Non-linguistic Content Non-linguistic content includes all types of encoding errors, language-atypical occurrences of numbers and characters (e.g. random sequences of characters or letters), code (remnants), URLs, hashtags and emoticons.

Letter Spacing Letter spacings are deliberately inserted spaces between the characters of a word.

Clean Assigned if none of the other labels apply.

3.2 Active Learning

To overcome the lack of labeled data, we aim to use active learning (Lewis and Gale, 1994), an iterative approach whose goal is to maximize model performance while minimizing human annotation effort. During each iteration, a so-called *query strategy* selects examples, which are labeled by a human annotator. The model is then retrained on all data labeled so far, and the process repeats in the next iteration.

3.3 Gold Standard and Annotation

While this work is not limited to a specific corpus, we need to evaluate the targeted corpus quality classifier. For this reason, we introduce a dataset, which will be used as a gold standard. *This considerable effort is only conducted to enable an experimental evaluation.*

Data Through a direct request to the Leipzig Corpora Collection[6] (Goldhahn et al., 2012) we obtained 165 M sentences (∼ 4 B tokens) of German web text. The resulting text originates from various crawls from 2018. The data is already preprocessed (through text extraction from HTML, sentence splitting, and deduplication). In the following, we operate on the resulting sentences.

Annotation Process We rely on agile annotation (Alex et al., 2010; Pustejovsky and Stubbs, 2012; Klie et al., 2024), to iteratively annotate the gold standard over three rounds. During each round, all three annotators (the authors of this work) label a set of given sentences independently. Inter-annotator agreement (IAA) is then assessed using

[5] https://github.com/Leipzig-Corpora-Collection/sentencecleaner

[6] https://wortschatz-leipzig.de/en

Example sentence	Labels
© zhu difeng | Visionen zum intelligenten Zuhause gibt es schon lange, und teilweise sind sie sehr ambitioniert. **EN:** *© zhu difeng | Visions of the intelligent home have been around for a long time, and some of them are very ambitious.*	Sentence Boundary, Grammar Mistake, Non-linguistic Content
Medisana Luftbefeuchter Ultrabreeze zusätzlichem Nachtlicht **EN:** *Medisana humidifier Ultrabreeze additional night light*	Grammar Mistake, Sentence Boundary
Wie viel Geld wollen wir fÃ1/4r den Kalender ausgeben? **EN:** *How much money do we want to spend fÃ1/4r the calendar?*	Spelling Anomaly, Non-linguistic Content
Pegasus Solero SL 28 Zoll 58cm Schwarz.. **EN:** *Pegasus Solero SL 28 Inches 58cm Black..*	Punctuation Error, Sentence Boundary, Grammar Mistake
Zweitens: Ich L I E B E Beeren < 3 In jeglicher Form, Art und GrÃ¶ÃŸe. **EN:** *Second: I L O V E berries < 3 In all shapes and siÃ¶ÃŸs.*	Non-linguistic Content, Grammar Mistake, Spelling Anomaly, Letter Spacing
V O R T R A G u n d G E S P R Ä C H **EN:** *T A L K a n d D I S C U S S I O N*	Letter Spacing, Sentence Boundary, Grammar Mistake, Spelling Anomaly
Die Spiel- und Lernstube ist Kontakt- und Anlaufstelle für Kinder, Jugendliche, Eltern und Bewohner im Stadtteil. **EN:** *The play and learning center is a point of contact and a drop-in center for children, adolescents, parents, and residents in the neighborhood.*	Clean

Table 1: Exemplary sentences (in German with an English translation below) and their respective gold labels.

Cohen's Kappa for each pair of annotators and each label. The score is analyzed and shortcomings of the guidelines or difficult edge cases are discussed. After this, class definitions or the guidelines are adjusted (e.g., by adding new positive or negative examples) and the sentences are relabeled.

Since we follow an iterative approach, any time we revise the guidelines for *all* classes, we would need to relabel every sentence in the batch to reflect the updated definitions. To keep the effort manageable, we relabeled the entire batch only in the first round. In the subsequent two rounds, we focused on specific classes that showed significant discrepancies between annotators.

The first batch of examples (460 in total) was collected using multi-label Adaptive-Active-Learning (Li and Guo, 2013) to primarily identify error cases. The second batch consisted of 600 randomly selected examples to increase text diversity, while the third batch comprised 275 manually collected examples aimed to cover previously underrepresented classes.

We report agreement scores of each batch of the initially labeled version in comparison with the final version in Table 2. We see the largest improvements in the first batch. This can be attributed to the initially low inter-annotator agreement, which prompted a thorough discussion, followed by a complete relabeling of the whole batch. We repeated this process two times. After that, we only selected examples from low-performing classes. We saw a moderate increase in IAA in the third batch, but not in the second one. Although there were clear improvements in the initial IAA values for batches 2 and 3, the final IAA value of 0.74 for the first batch could not be reached.

Batch	Initial IAA	Final IAA	Size
First batch	0.54	0.74	460
Second batch	0.71	0.71	600
Third batch	0.72	0.75	275

Table 2: IAA (Cohen's Kappa) between three coders for the iterative labeling approach over three iterations.

Final Dataset The three batches are combined and a majority voting is used to merge the labels. We had to discard 16 sentences which contained harmful content or Personally Identifiable Infor-

Label	IAA	# Examples
Sentence Boundary	0.86	439
Grammar Mistake	0.76	594
Spelling Anomaly	0.61	290
Punctuation Error	0.41	78
Non-linguistic Content	0.75	147
Letter Spacing	0.96	25
Clean	0.80	577
Avg/Total	0.74	2150

Table 3: Class-wise and averaged inter-annotator agreement, class distribution and number of class examples (2150 labels in 1319 sentences) of our gold standard.

mation. The final inter-annotator agreement and additional dataset statistics are shown in Table 3. According to the Kappa interpretation of Landis and Koch (1977), with an average of 0.74 we reach a substantial agreement level (0.6–0.8).

4 Experiments

In this experiment, we examine the feasibility of detecting the proposed text quality classes, in a scenario where training data and annotation time are severely limited (**RQ 1**). For this purpose, we bootstrap a classifier with active learning that is evaluated against the annotated gold standard. To further assess the effectiveness of the resulting classifier, we compare it to a rule-based baseline, which detects similar textual issues (**RQ 2**).

4.1 Experimental Setup

To reflect realistic constraints, we simulate the scenario of a small team facing large volumes of unlabeled data with a limited annotation budget by imposing a time budget of one working day (8 hours). Active learning is warm-started with an initial training pool of 70 hand-picked examples ($\sim$ 10 examples per class). In each round, the query strategy returns a batch of 20 sentences to a human annotator. To improve the fine-tuning stability, we train the classifier from scratch, e.g., from the pre-trained base model, after every batch.

Data A new dataset is used, which was created as described in Section 3, based on more recent crawling data of the same project, crawled in 2022 with 136 M extracted sentences ($\sim$ 3.4 B tokens).

Classification For classification, we use SetFit (Tunstall et al., 2022), an efficient fine-tuning paradigm that leverages contrastive learning. Using a sampling strategy, it generates similar and dissimilar sentence pairs which are used to train a siamese network. In the multi-label setting, sentences are considered similar (positive pair), if they have a label in common, and dissimilar otherwise (negative pair). While there are variations to SetFit, we stayed close to the original version in which a Sentence Transformer (Reimers and Gurevych, 2019) is fine-tuned and the classification operates on the resulting embeddings. Instead of a logistic regression head, however, we opted for a neural network head, which is faster for even a moderate number of classes at a similar classification performance.

Base Model As the base model, we create a Sentence Transformer (Reimers and Gurevych, 2019) by mean pooling over the output layer from multilingual DistilBERT (Sanh et al., 2019) (135 M parameters). Compared to BERT, DistilBERT contains only half the number of layers and is therefore more efficient regarding training and inference.

Query Strategy We use multi-label Adaptive-Active-Learning (AAL; Li and Guo, 2013) as the query strategy, which balances two scores to find informative samples: (1) Max-Margin Uncertainty Sampling (MMUS) and (2) Label-Cardinality-Inconsistency (LCI). MMUS calculates the distance between the maximum of the predicted negative labels and the minimum of the predicted positive labels, according to a fixed threshold (e.g., 0.5). If the distance is small, the sample is considered highly informative. LCI assumes that multi-label instances often have a similar label count. It computes the deviation of the predicted label count from the average in the so far annotated data (for details, see Section 4 in Li and Guo, 2013).

One limitation of selecting data based on predictions is that the data has to be passed forward through the classifier before any selection criteria can be applied. To make this step feasible, during every round we subsample 10 K unlabeled sentences before applying the query strategy. A batch of the 20 highest-scoring samples is selected.

Implementation The implementation for the active learning routine and query strategies are based on `small-text`[7] (Schröder et al., 2023), an active learning library specialized in text classification, with integrations for transformers and SetFit.

[7] https://github.com/webis-de/small-text

Metric	**Value**
$\mathbf{F}_1^{macro}$	0.80
$\mathbf{F}_1^{micro}$	0.84
Subset Acc	0.67

Table 4: Active learning results for 498 examples.

Class	**F_1**	**# Count**
Sentence Boundary	0.96	169
Grammar Mistake	0.86	256
Spelling Anomaly	0.57	158
Punctuation Error	0.62	77
Non-linguistic Content	0.77	110
Letter Spacing	0.94	11
Clean	0.86	145

Table 5: Class-wise active learning results with the number of training examples per class.

To ease the process for the annotator, we connected the annotation tool `argilla`[8] to our backend.

5 Results

5.1 Active Learning Experiment

The experiment took 7 hours and 50 minutes in total, during which 22 batches with 20 examples each were processed. Re-training from scratch with every newly annotated batch required overall 5 hours on one Nvidia Tesla A30 (24 GB), querying in total ∼ 1 hour, labeling less than 2 hours. The human annotator in this experiment was the first author of this paper. During the annotation process 12 samples had to be discarded due to the problems mentioned above (see Section 3.3).

In Table 4, we report F_1 and subset accuracy of the last active learning round on our gold standard, which used 498 examples for the training (428 samples + 70 initial examples). The classifier achieves average scores of $\mathrm{F}_1^{macro} = 0.80$ and $\mathrm{F}_1^{micro} = 0.84$. The subset accuracy of 0.67 is sufficiently high, considering that only exactly matching label combinations are considered correct. The class scores vary considerably, ranging from 0.57 to 0.96 (see Table 5). Every second sentence was annotated with the label *"Grammar Mistake"* (256 examples), followed by *"Sentence Boundary"* (169 examples) in terms of frequency (last column of Table 5). When comparing F_1 values, there is no indication that a higher number of training examples always results in higher scores (e.g., when comparing *"Grammar Mistake"* $= 0.86$ and *"Sentence Boundary"* $= 0.96$). This can also be shown with other low-quality classes, notably including *"Letter Spacing"* that only required 11 examples to achieve a score of 0.94.

To further investigate the active learning process, we reproduce the classifier's progression during the experiment by training checkpoints with different seeds at every two batches of training data and plot the results (see Figure 2). For example, we train five times with all the training data, which was sampled until batch four (70 initial and 80 queried examples), then train five times on batch six, and so on. Figure 2 shows improvements across all classes, with steeper increases initially that gradually level off over the course of the experiment, albeit at different rates. Although the macro F_1 curve shows signs of stagnation during the last two batches, increasing the annotation budget may yield further improvements. However, the point at which performance would begin to decline remains unclear. One approach would be to proceed cautiously by reducing the active learning batch size.

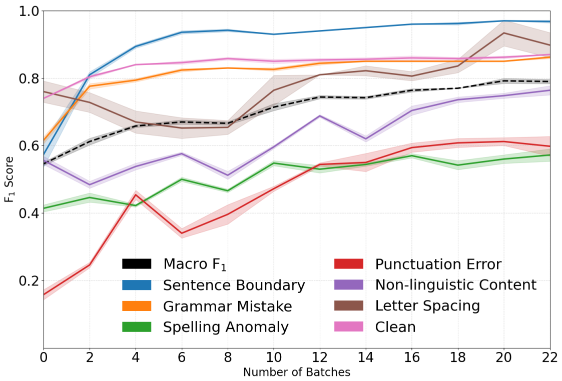

Figure 2: Macro and class-wise F_1 in relation to train examples per batch, showing classification progress during the active learning cycle. Re-trained with 5 seeds at every second batch. Batch 0 are the 70 initial examples.

5.2 Comparison to Rule-Based Filtering

To further investigate the classifier's performance, we compare it against a rule-based baseline, Sentencecleaner, which was developed within the context of the project through which our dataset was obtained (Goldhahn et al., 2012). This tool applies a set of 40 rules to filter out low-quality sentences and is typically used on web crawling data. These

[8] https://github.com/argilla-io/argilla

	Not-Clean	Clean
Sentencecleaner	0.8181	0.8150
Corpus Quality Classifier	**0.8673**	**0.8603**

Table 6: Comparison in F_1 score between our supervised approach and a rule-based baseline, both developed for the same task and dataset. Evaluated against *not-clean* and *clean* sentences from our gold standard.

rules include checks for character or symbol ratios, letter spacing and invalid sentence boundaries.[9]

Quantitative Comparison We apply a black-box evaluation, comparing both methods solely based on their outputs against our gold standard, focusing on their ability to detect impaired and clean sentences. Classifications are considered *not-clean* when the classifier identifies any low-quality label or when any Sentencecleaner rule applies, whereas classifications are considered *clean* when the class *"Clean"* is correctly predicted or no Sentencecleaner rules are applicable. Table 6 shows that our approach outperforms by 4.92 (not-clean) and 4.53 (clean) percentage points in F_1, demonstrating both better error detection and clean text recognition.

Qualitative Comparison To have a better understanding of which *additional* patterns the classifier can find, we perform a brief qualitative comparison.

We first review examples that were incorrectly filtered out by the rule-based baseline, but correctly retained by our approach. Among those 39 sentences, 12 could be captured with simple rule adjustments. This would be feasible, for example, for the rule according to which sentences are marked as not clean if they begin with a number that is not part of a valid date format. In addition, 26 sentences were filtered out by a rule prohibiting a sentence length of more than 255 characters which is a good example of how difficult it is to find reasonable thresholds. Of the 56 sentences that triggered this rule, half were true positives and half were false positives, yielding a precision of just 50%.

We also look at the 107 sentences, which our classifier is correctly predicting as not clean and the rule-based approach missed. Among these, it is notable that the majority (92 sentences) contain *"Grammar Mistake"* in their label set, which covers all sorts of violations that affect the comprehensibility of a sentence. To further investigate the error patterns, we grouped them into different subcategories and briefly describe them (see Table 7). There are 51 cases where a finite verb form is missing (*"Missing Predicate"*), e.g. headlines (news, e-commerce, advertisement, etc.), product descriptions or bullet points. They all have typical characteristics of well-formed sentences, like starting with a capital lettered word, ending with a punctuation mark while not containing any misplaced or random symbols. The second largest group contains 28 cases with foreign language parts ($\sim$ 50% non-German text), which are, according to our definition, grammar violations (*"Language Mixing"*). The remaining cases comprise various textual anomalies, including incoherent sentences, missing word boundaries causing lexical merging, and sentences that appear truncated (*"Gibberish"*, *"Merged Words"* and *"Truncation"*).

To assess the severity of the overlooked errors, We also examined the classifier's limitations, specifically the 104 sentences it mistakenly identified as clean. When looking at the examples and their gold labels, the two most common label sets are the single labels *"Spelling Anomaly"* (40 cases) and *"Grammar Mistake"* (31 cases). Single-label occurrences often reflect subtle errors, which could be confirmed by examining the actual text content.

6 Discussion

Considering the total amount of time (8h) and training data (498 examples), we argue that our proposed setup worked sufficiently well to build a classifier for text cleaning and could serve as a blueprint for data efficient training. Although this has been demonstrated on German web crawls, our pipeline is agnostic to language and domain: only the annotation scheme and seed examples would need to be adapted.

Without further experiments, however, it is not clear how these methods will perform in comparison to traditional supervised-learning using random data points. Nevertheless, during the annotation of the second batch of the gold standard—600 random examples—we observed that $\sim$ 50% of sentences were clean. In contrast, within the active learning training data, the class *"Clean"* was sampled only 145 times (29%), thereby focusing annotation effort on noisy examples. This suggests that the traditional supervised classifier will likely be trained on fewer error cases compared to using ac-

[9] https://github.com/Leipzig-Corpora-Collection/sentencecleaner

Pattern	Exemplary sentence
Missing Predicate	*Große Abgeschlagenheit und Trägheit des Körpers.* **EN:** *Great fatigue and sluggishness of the body.*
Language Mixing	*Nun, das lässt sich übertragen: What is a school but the people?* **EN:** *Well, that can be transferred: What is a school but the people?*
Gibberish	*Ist dort Folklore, war schon der 16. Angriff.* **EN:** *Is folklore there, was already the 16th attack.*
Merged Words	*Erdäpfelgulasch - Der SpeisenzustellerEs befinden sich keine Produkte im Warenkorb.* **EN:** *Potato goulash - The food delivererThere are no products in your basket.*
Truncation	*Wie in anderen Bundesländern muss auch in.* **EN:** *As in other federal states, this must also be done in.*

Table 7: Various low-quality patterns, the classifier **additionally** found, in comparison to a rule-based approach.

tive learning, which will reduce performance of our low-quality classes, but may improve on *"Clean"*.

When looking at *additional* low-quality patterns that our approach identifies (Table 7), we find various textual problems, some of which are less obvious to recognize by looking at the surface structure alone. One could argue that certain *"Missing Predicate"* instances, such as headlines that only lack finite verbs, do not constitute low-quality text. While this does make sense at the document level, where the text might serve its function as a summarizing heading, our sentence-level approach assesses quality focused on syntactically valid sentences.

The comparison also demonstrates the inherent problem in selecting suitable threshold values in rule-based approaches, as can be seen with the imprecise sentence length heuristic.

It is worth noting that our seven-class schema represents only a *first effort* to define web text quality and does not fully capture what constitutes low (or high) quality sentences. This work focused on data efficient training methods rather than the development of a comprehensive taxonomy.

To obtain a rough estimate of GPU requirements for corpus preprocessing, we processed 1 M sentences ($\sim$25 M tokens) on a Nvidia H100 (80GB), which took $\sim$123.10 s. Extrapolating to 1 T corresponds to about 1388 GPU hours.[10] While this constitutes a significant resource demand, scaling across multiple GPUs or nodes would render even corpora an order of magnitude larger computationally feasible. Moreover, this estimate reflects the contemporary throughput, but as GPU capabilities and computational speed continue to advance, the boundary of what is feasible will steadily expand.

7 Conclusions

In this work, we proposed a labeling scheme for corpus quality classification, provided a gold standard of 1,319 annotated sentences for German web data, and applied active learning to bootstrap a classifier that predicts corpus quality indicators. For evaluation purposes, we created a gold standard using an iterative annotation process, which yielded a corpus with substantial inter-annotator agreement (with a Cohen's Kappa of 0.74), making it suitable for further use.

Using a multi-label active learning setup, we trained a classifier that predicts the defined quality labels for German language with a macro F_1 score of 0.80 and micro F_1 of 0.84 despite using only 498 training examples in total, labeled over the course of 8 hours. We showed that our supervised approach outperforms a rule-based one developed for the same task. Additionally, we find that the classifier is able to capture error types, particularly those involving the comprehensibility of a sentence, which the rule-based baseline tends to miss.

This work demonstrates a successful proof of concept for enabling model-based filtering through LLM-based active learning for text classification in resource-constrained scenarios. As capabilities of LLMs grow and computational costs decline, preprocessing of larger volumes becomes increasingly feasible, and as a result we predict that preprocessing will shift towards small efficient models, making preprocessing for specific languages and domains increasingly prevalent.

Limitations

We did not continue training a pre-trained Sentence Transformer (ST) model for SetFit, but boot-

[10] We use vanilla inference using the SetFit library, but we observed that the throughput plateaued beyond a certain batch size, even though GPU memory was not saturated. We suspect that with code optimizations, the runtime could be further reduced, so the reported number serves as a lower bound.

strapped one from a vanilla transformer model (due to a lack of a comparable German ST model), which may produce suboptimal sentence-level embeddings compared to one whose representations have been pre-trained on sentence pairs and cosine similarity loss. We encourage exploring the use of a pre-trained ST, as this could further improve performance. While we were not aware of a suitable model for German, the multi-lingual model from Reimers and Gurevych (2019) is a promising candidate for further investigation.

During the creation of the gold standard, we discovered a bug in the query strategy used to select data for the first batch. We assume these sentences were drawn roughly randomly like the second batch, but still covered more error cases.

Ethical Statement

The dataset annotations may exhibit bias reflecting the perspectives of the annotators, who are all computer science researchers, potentially limiting the diversity of opinions represented in our quality assessments. Additionally, our definition of high quality correlates strongly with standard German grammar, which may inadvertently exclude dialectal variations or linguistic phenomena such as code-switching. This presents a particular concern given that LLM pre-training corpora should ideally be as representative as possible of natural language variation. To address these limitations, we will release our dataset and model to enable further investigation of these problems.

Acknowledgments

This work has been partially funded by the German Federal Ministry of Research, Technology, and Space (BMFTR) under the grant numbers 01IS24077A and 01IS24037B. This work has been partially supported by the ScaDS.AI Center for Scalable Data Analytics and Artificial Intelligence, project identification number: ScaDS.AI.

Computations for this work were done (in part) using resources of the Leipzig University Computing Center.

References

Alon Albalak, Yanai Elazar, Sang Michael Xie, Shayne Longpre, Nathan Lambert, Xinyi Wang, Niklas Muennighoff, Bairu Hou, Liangming Pan, Haewon Jeong, et al. 2024. A survey on data selection for language models. *arXiv preprint arXiv:2402.16827*.

Bea Alex, Claire Grover, Rongzhou Shen, and Mijail Kabadjov. 2010. Agile corpus annotation in practice: An overview of manual and automatic annotation of CVs. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 29–37, Uppsala, Sweden. Association for Computational Linguistics.

Zachary Ankner, Cody Blakeney, Kartik Sreenivasan, Max Marion, Matthew L. Leavitt, and Mansheej Paul. 2025. Perplexed by perplexity: Perplexity-based data pruning with small reference models. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Liat Ein-Dor, Alon Halfon, Ariel Gera, Eyal Shnarch, Lena Dankin, Leshem Choshen, Marina Danilevsky, Ranit Aharonov, Yoav Katz, and Noam Slonim. 2020. Active Learning for BERT: An Empirical Study. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7949–7962, Online. Association for Computational Linguistics.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the Leipzig corpora collection: From 100 to 200 languages. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 759–765, Istanbul, Turkey. European Language Resources Association (ELRA).

Julius Gonsior, Tim Rieß, Anja Reusch, Claudio Hartmann, Maik Thiele, and Wolfgang Lehner. 2025. Survey of active learning hyperparameters: Insights from a large-scale experimental grid.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Erik Henriksson, Otto Tarkka, and Filip Ginter. 2025. FinerWeb-10BT: Refining web data with LLM-based line-level filtering. In *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, pages 258–268, Tallinn, Estonia. University of Tartu Library.

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov.

2016. Fasttext. zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651*.

Nataliia Kholodna, Sahib Julka, Mohammad Khodadadi, Muhammed Nurullah Gumus, and Michael Granitzer. 2024. Llms in the loop: Leveraging large language model annotations for active learning in low-resource languages. In *Machine Learning and Knowledge Discovery in Databases. Applied Data Science Track: European Conference, ECML PKDD 2024, Vilnius, Lithuania, September 9–13, 2024, Proceedings, Part X*, page 397–412, Berlin, Heidelberg. Springer-Verlag.

Adam Kilgarriff and Gregory Grefenstette. 2003. Introduction to the special issue on the web as corpus. *Computational Linguistics*, 29(3):333–348.

Jan-Christoph Klie, Richard Eckart de Castilho, and Iryna Gurevych. 2024. Analyzing dataset annotation quality management in the wild. *Computational Linguistics*, 50(3):817–866.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, Monang Setyawan, Supheakmungkol Sarin, Sokhar Samb, Benoît Sagot, Clara Rivera, Annette Rios, Isabel Papadimitriou, Salomey Osei, Pedro Ortiz Suarez, Iroro Orife, Kelechi Ogueji, Andre Niyongabo Rubungo, Toan Q. Nguyen, Mathias Müller, André Müller, Shamsuddeen Hassan Muhammad, Nanda Muhammad, Ayanda Mnyakeni, Jamshidbek Mirzakhalov, Tapiwanashe Matangira, Colin Leong, Nze Lawson, Sneha Kudugunta, Yacine Jernite, Mathias Jenny, Orhan Firat, Bonaventure F. P. Dossou, Sakhile Dlamini, Nisansa de Silva, Sakine Çabuk Ballı, Stella Biderman, Alessia Battisti, Ahmed Baruwa, Ankur Bapna, Pallavi Baljekar, Israel Abebe Azime, Ayodele Awokoya, Duygu Ataman, Orevaoghene Ahia, Oghenefego Ahia, Sweta Agrawal, and Mofetoluwa Adeyemi. 2022. Quality at a glance: An audit of web-crawled multilingual datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gerard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Minh Chien Vu, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Alexandra Luccioni, and Yacine Jernite. 2022. The bigscience roots corpus: A 1.6tb composite multilingual dataset. In *Advances in Neural Information Processing Systems*, volume 35, pages 31809–31826. Curran Associates, Inc.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. 2022. The bigscience roots corpus: A 1.6 tb composite multilingual dataset. *Advances in Neural Information Processing Systems*, 35:31809–31826.

Pietro Lesci and Andreas Vlachos. 2024. AnchorAL: Computationally efficient active learning for large and imbalanced datasets. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8445–8464, Mexico City, Mexico. Association for Computational Linguistics.

David D. Lewis and William A. Gale. 1994. A sequential algorithm for training text classifiers. In *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval*, pages 3–12. Springer, ACM/Springer.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Guha, Sedrick Scott Keh, Kushal Arora, et al. 2024. Datacomp-lm: In search of the next generation of training sets for language models. *Advances in Neural Information Processing Systems*, 37:14200–14282.

Xin Li and Yuhong Guo. 2013. Active learning with multi-label SVM classification. In *IJCAI 2013, Proceedings of the 23rd International Joint Conference on Artificial Intelligence, Beijing, China, August 3-9, 2013*, pages 1479–1485. IJCAI/AAAI.

Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, and Daphne Ippolito. 2024. A pretrainer‘s guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3245–3276, Mexico City, Mexico. Association for Computational Linguistics.

Katerina Margatina and Nikolaos Aletras. 2023. On the limitations of simulating active learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4402–4419, Toronto, Canada. Association for Computational Linguistics.

Katerina Margatina, Giorgos Vernikos, Loïc Barrault, and Nikolaos Aletras. 2021. Active learning by acquiring contrastive examples. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 650–663, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Charlotte Nachtegael, Jacopo De Stefani, and Tom Lenaerts. 2023. A study of deep active learning methods to reduce labelling efforts in biomedical relation extraction. *PLOS ONE*, 18(12):1–23.

Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin A Raffel, Leandro Von Werra, Thomas Wolf, et al. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for falcon LLM: outperforming curated corpora with web data only. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning: A Guide to Corpus-Building for Applications*. O'Reilly Media.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Anna Rogers and Sasha Luccioni. 2024. Position: Key claims in LLM research have a long tail of footnotes. In *Forty-first International Conference on Machine Learning*.

Julia Romberg, Christopher Schröder, Julius Gonsior, Katrin Tomanek, and Fredrik Olsson. 2025. Have LLMs made active learning obsolete? Surveying the NLP community. *arXiv preprint arXiv:2503.09701*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Christopher Schröder and Gerhard Heyer. 2024. Self-training for sample-efficient active learning for text classification with pre-trained language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11987–12004, Miami, Florida, USA. Association for Computational Linguistics.

Christopher Schröder, Lydia Müller, Andreas Niekler, and Martin Potthast. 2023. Small-text: Active learning for text classification in python. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 84–95, Dubrovnik, Croatia. Association for Computational Linguistics.

Christopher Schröder, Andreas Niekler, and Martin Potthast. 2022. Revisiting uncertainty-based query strategies for active learning with transformers. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2194–2203, Dublin, Ireland. Association for Computational Linguistics.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics.

Tristan Thrush, Christopher Potts, and Tatsunori Hashimoto. 2025. Improving pretraining data using perplexity correlations. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Lewis Tunstall, Nils Reimers, Unso Eun Seo Jo, Luke Bates, Daniel Korat, Moshe Wasserblat, and Oren Pereg. 2022. Efficient few-shot learning without prompts. *arXiv preprint arXiv:2209.11055*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Mengqi Wang and Ming Liu. 2023. An empirical study on active learning for multi-label text classification. In *Proceedings of the Fourth Workshop on Insights from Negative Results in NLP*, pages 94–102, Dubrovnik, Croatia. Association for Computational Linguistics.

Lukas Wertz, Jasmina Bogojeska, Katsiaryna Mirylenka, and Jonas Kuhn. 2022a. Evaluating pre-trained sentence-BERT with class embeddings in active learning for multi-label text classification. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 366–372, Online only. Association for Computational Linguistics.

Lukas Wertz, Jasmina Bogojeska, Katsiaryna Mirylenka, and Jonas Kuhn. 2023. Reinforced active learning for low-resource, domain-specific, multi-label text classification. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10959–10977, Toronto, Canada. Association for Computational Linguistics.

Lukas Wertz, Katsiaryna Mirylenka, Jonas Kuhn, and Jasmina Bogojeska. 2022b. Investigating active learning sampling strategies for extreme multi label text classification. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4597–4605, Marseille, France. European Language Resources Association.

Alexander Wettig, Aatmik Gupta, Saumya Malik, and Danqi Chen. 2024. Qurating: Selecting high-quality data for training language models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 52915–52971. PMLR.

Ruixuan Xiao, Yiwen Dong, Junbo Zhao, Runze Wu, Minmin Lin, Gang Chen, and Haobo Wang. 2023. FreeAL: Towards human-free active learning in the era of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14520–14535, Singapore. Association for Computational Linguistics.