Unsupervised Outlier Detection for Language-Independent Text Quality Filtering

Jón Friðrik Daðason, Hrafn Loftsson
Department of Computer Science
Reykjavik University, Iceland
{jond19, hrafn}@ru.is

Abstract

Web-crawled corpora offer an abundant source of training data for language models. However, they are generally noisy and are typically filtered using heuristic rules or classifiers. These methods require careful tuning or labeling by fluent speakers. In this paper, we assess the effectiveness of commonly applied rules on TQ-IS, a manually labeled text quality dataset for Icelandic. Additionally, we advocate for the utilization of unsupervised clustering and outlier detection algorithms for filtering. These algorithms are language-independent, computationally efficient and do not require language expertise. Using grid search, we find the optimal configuration for every combination of rules, optimizing for F_1 score on TQ-IS. For a rule-based approach, we discover that optimal results can be achieved with only a small subset of the full ruleset. Using five rules, we obtain an F_1 score of 98.2%. We then evaluate three unsupervised algorithms, i.e., Gaussian Mixture Models (GMMs), Isolation Forests and One-Class SVMs. Our findings reveal that unsupervised algorithms perform well on the TQ-IS dataset, with GMMs obtaining the best results, comparable to those obtained with the rule-based approach. Finally, we show that unsupervised methods appear to be equally suitable for languages other than Icelandic, including Estonian and Basque.

Keywords: Text quality, text filtering, language modeling

1. Introduction

Researchers increasingly rely on vast amounts of web-crawled text in order to pre-train language models. Although a valuable resource, web-crawled corpora are often noisy, containing a large number of low-quality documents that, in sufficient quantities, can degrade downstream performance (Kreutzer et al., 2022; Muennighoff et al., 2023). This includes text that may be poorly machine-translated, error-prone, corrupted or incoherent.

The exact definition of "noisy" or "low-quality" text varies and is subject to interpretation. However, it is well established that filtering web-crawled corpora can significantly improve the downstream performance of pre-trained language models (Wenzek et al., 2020; Brown et al., 2020; Raffel et al., 2020; Muennighoff et al., 2023). Filtering is typically performed using classifiers or threshold-based rules. In the rule-based approach, documents are filtered out if certain metrics, such as their mean word length, fall outside a predefined acceptable range (Rae et al., 2022). Alternatively, a classifier may be used to label or score documents based on their quality. This includes supervised classifiers, trained on a manually labeled text quality dataset (Wu et al., 2021), and self-supervised classifiers, trained to distinguish between documents from a high-quality, curated corpus and a noisy, web-crawled corpus (Brown et al., 2020). The effectiveness of these approaches depends heavily on the choice of metrics and thresholds for the rule-based approach, and features, parameters, training data and model type for the classifier-based approach. Moreover, accurate evaluation can only be achieved with the help of fluent speakers.

There is no standardized approach to rule-based text quality filtering. Some corpora are filtered based on only a single metric (Wenzek et al., 2020; Muennighoff et al., 2023), while others combine as many as 15 distinct rules (Öhman et al., 2023). As the size of the ruleset increases, it can become more difficult to determine the impact that individual rules might have on the overall effectiveness of the filtering process, whether positive or negative. Rules that may be effective when evaluated individually can become redundant as more rules are added to the ruleset. Conversely, a rule that appears ineffective on its own may become more useful when applied in conjunction with other rules. Using TQ-IS (Daðason, 2024), a manually labeled text quality dataset for Icelandic, we perform experiments to better understand how commonly applied rules interact with one another.

A review of the current literature on text quality filtering reveals two prevailing strategies for selecting either threshold values for rules, or parameters for classifiers. For rules, thresholds may simply be selected based on linguistic intuition (Rae et al., 2022; Laurençon et al., 2022; Öhman et al., 2023). Alternatively, parameters or thresholds may be chosen through statistical analysis, such as aligning the distribution of the filtered corpus with that of a known high-quality corpus, or by selecting thresholds that

SIGUL2024 Workshop, pages 383–393
21-22 May, 2024.

discard a certain proportion of the documents, effectively filtering out outliers (Brown et al., 2020; Muennighoff et al., 2023; Nguyen et al., 2023). In either case, the quality of the chosen thresholds or parameters can only be assessed through empirical validation. In practice, this may involve either manually labeling a portion of the target corpus for evaluation (Wu et al., 2021), or comparing the downstream performance of language models that have been pre-trained on filtered and unfiltered versions of the corpus (Raffel et al., 2020).

In this paper, we analyze several unfiltered web-crawled corpora, visualizing the distribution of their documents based on metrics that are commonly employed in a rule-based approach. In each corpus, we find that there exists a distinct, large and well-defined cluster of high-quality documents. In contrast, low-quality documents appear as outliers in these distributions. We find that in TQ-IS, the boundaries of these high-quality clusters align closely to optimal threshold values discovered through exhaustive grid search. On the basis of these findings, we also describe a novel text quality classifier by reframing the task as an outlier detection problem. We evaluate three types of clustering and outlier detection algorithms on TQ-IS, the main benefit of which is their unsupervised nature and explainability. This allows their few parameters to be quickly tuned through iterative experimentation and visualization of their decision boundaries, without the need for fluency in the target language.

The main contributions of our work are the following:

- A thorough evaluation of the effectiveness of commonly used text filtering rules on a manually labeled text quality dataset. We demonstrate that only a few rules are needed to obtain optimal results. Furthermore, we show that visualizing documents in a web-crawled corpus based on the metrics targeted by the rules reveals a large, well-defined cluster of high-quality documents, and that close to optimal threshold values can be found at the edges of this cluster.
- An exploration of how well unsupervised clustering and outlier detection algorithms perform on the task of text quality filtering. We find that they can obtain comparable results to a rule-based approach, without requiring fluency in the target language or time-consuming parameter optimization.

The rest of this paper is organized as follows. In Section 2, we discuss related work, and in Section 3, the Icelandic Text Quality Dataset. Commonly employed document-level rules are presented in Section 4, and three types of outlier detection algorithms in Section 5. The experimental setup and our results are presented in Sections 6 and 7, respectively. Finally, we conclude in Section 8.

2. Related Work

Common Crawl (CC) is an organization that maintains a massive repository of data crawled from over 25 billion websites.[1] There are many web-crawled corpora that are derived from the CC dataset, such as the Multilingual Colossal Clean Crawled Corpus (mC4), which consists of 6.3T tokens in 101 languages (Xue et al., 2021). The mC4 corpus has only been lightly filtered with regard to text quality. A language classifier was used to identify the primary language of each document, duplicate occurrences of three line spans were discarded, and lines that did not end on a terminal punctuation mark were removed.

MassiveText is an English-language corpus consisting of 2.35 trillion tokens, created for pre-training the Gopher language model (Rae et al., 2022). It is composed of several curated and web-crawled corpora. One of the web-crawled subcorpora is MassiveWeb, which contains 506 billion tokens, collected using a custom HTML scraper. It was filtered using a set of seven heuristic rules. These rules include discarding documents if their mean word length falls outside a specified range or if they do not contain a minimum number of unique stop words. The authors find that the filtering results in a lower validation loss when pre-training a 1.5B parameter version of the Gopher model.

ROOTS is a large, multilingual text corpus spanning 46 natural languages, combined from a collection of mono- and multilingual language resources, both curated and web-crawled (Laurençon et al., 2022). The corpus was filtered using a set of seven heuristic rules which, for example, enforce a maximum perplexity score, a maximum word repetition ratio and a minimum language classification confidence. The thresholds for the rules were determined by fluent speakers for each language. ROOTS has been used to pre-train language models such as BLOOM (Scao et al., 2023).

CulturaX (Nguyen et al., 2023) is a web-crawled corpus that was obtained by combining multiple web-crawled corpora, all of which are derived from Common Crawl. It consists of 6.3 trillion tokens in 167 languages and is filtered using the same rules as were used for the ROOTS corpus. For each language, the authors apply a variant of the interquartile range (IQR) method (Dekking et al., 2005) by considering the distribution of each metric and setting minimum thresholds at the 10th percentile and maximum thresholds at the 90th percentile. In total, about 39% of the documents are discarded using these settings.

[1] https://commoncrawl.org/about/

Young et al. (2024) combine heuristic rules, classifiers, and unsupervised semantic clustering to filter a large, web-crawled corpus consisting of documents in Chinese and English. The rules are used to discard documents based on their length, ratio of special symbols, ratio of short, incomplete or consecutive sentences, and other metrics. The thresholds for the rules are determined using the IQR method described above. Classifiers are used to filter documents based on their perplexity as well as quality, coherence, and safety scores. Finally, documents in the corpus are grouped by semantic similarity and each cluster is annotated with a quality label. The effectiveness of these filters is not reported.

We have previously evaluated several text quality classifiers on web-crawled corpora in Icelandic, Estonian and Basque (Daðason and Loftsson, 2024). We found that the classifiers performed well on the TQ-IS dataset, with a supervised classifier obtaining an F_1 score of 99.01%. However, for all three languages, we observed only a very modest benefit to downstream performance after filtering the web-crawled corpora, potentially owing to their relatively small size. For this reason, we omit an evaluation on downstream tasks in this work.

3. TQ-IS

TQ-IS (Daðason, 2024) is a dataset that consists of 2,000 unique documents that were sampled from several web-crawled corpora, such as the Icelandic Crawled Corpus (Daðason, 2021) and the Icelandic subset of the mC4 dataset. Each document contains between 50 to 500 space-delimited tokens. The source corpora have primarily been filtered using language classifiers and by enforcing a minimum token or character count, but have otherwise undergone minimal filtering with regard to text quality. Each document in TQ-IS was manually labeled as either high or low-quality, based on specific annotation guidelines presented in (Daðason and Loftsson, 2024). The two categories are equally represented in the dataset.

There is no precise definition of what constitutes a high or low-quality document when it comes to pre-training language models, beyond the impact (positive or negative) that it may have on the model with regard to downstream performance. It is difficult to know where exactly the line between these two categories of documents lies. Therefore, TQ-IS only includes documents that were considered to be clear-cut examples of each category. Documents were labeled as high-quality if they primarily consist of running text in the form of sequences of full, grammatically structured sentences that are connected in a meaningful and coherent way. High-quality documents contains few errors, if any, and the text is properly capitalized and punctuated. Documents that are disjointed, incoherent, error-prone, repetitive, or largely consist of non-Icelandic, non-running, or non-linguistic text were classified as low-quality. For a more detailed overview of what we consider to be low or high-quality text, we refer to the TQ-IS annotation guidelines.

4. Rules

Rules are typically applied on the token, line, sentence, paragraph, or document level. More granular filtering methods can result in more text being preserved, but this may come at the cost of making filtered documents less coherent. Furthermore, tokenization and sentence and paragraph segmentation errors may degrade the quality of filters that rely on their accuracy, especially in noisy corpora. For this reason, we only consider document-level filtering in this paper. We describe 12 document-level rules that were used to filter the ROOTS and MassiveWeb corpora, and propose one additional rule based on our analysis of low-quality documents in the TQ-IS dataset. All 13 rules, described in this section, are included in our experiments.

4.1. ROOTS

In our experiments, we evaluate several rules that were used to filter the ROOTS corpus. We omit one rule that discards documents if they contain too many sexually explicit words, as such word lists are not readily available for all languages. We also exclude a rule that discards documents containing too many or too few words, as documents in TQ-IS are already limited to between 50 and 500 space-delimited tokens in length.

Perplexity A language model is used to calculate the perplexity score of a document, giving an estimate of how likely it is that the model could generate the same text. The less predictable the text is, the higher its perplexity score will be. A high perplexity score means that the document differs from the language model's training corpus in some respect. When used to discriminate between low and high-quality documents, perplexity is usually calculated using a language model that has been trained on a curated corpus containing minimal noise. This ensures that low-quality documents should tend to receive higher perplexity scores than high-quality documents. Documents with a perplexity score above a certain threshold are discarded.

Character Repetition Ratio This rule targets documents that have a high proportion of repeated character n-grams. This ratio is calculated as the number of frequently occurring character n-grams

divided by the total number of character n-grams. A high ratio can be indicative of a document that largely consists of automatically generated text (e.g., log files) or text-based visuals (e.g., ASCII art). If the character repetition ratio exceeds a maximum threshold, it is discarded.

Word Repetition Ratio Similarly, the word repetition ratio of a document is calculated by dividing the number of frequently repeated words by the total number of words it contains. A high word repetition ratio may suggest that a document contains a large amount of spam or content intended for search engine optimization (e.g., keywords that are repeated in an effort to increase search rankings) or automatically generated text. Documents with a high word repetition ratio are discarded.

Special Character Ratio Documents that contain a large proportion of non-alphabetic characters, such as emojis, Unicode symbols, digits and punctuation marks may be corrupted (e.g., due to incorrect character encoding) or otherwise contain a limited amount of natural language text. If the special character ratio within a document exceeds a certain maximum threshold, it is discarded.

Stop Word Ratio In the context of text quality filtering, stop words generally consist of common function words, i.e., words that serve a syntactically and grammatically important purpose, but lack any significant meaning on their own. This generally includes word classes such as conjunctions, prepositions, pronouns and articles. A document that has a very low ratio of stop words is unlikely to contain coherent, running text in a natural language.

Language Confidence Score A language classifier is used to determine the primary language of each document. If the primary language is not targeted for inclusion in the corpus, or if the confidence falls below a certain threshold, the document is discarded.

4.2. MassiveWeb

We also consider the rules that were used to filter the MassiveWeb corpus. We omit one rule that enforces a minimum and maximum word length for documents.

Mean Word Length If the mean word length within a document falls outside an expected range, it could suggest that the document is malformed (e.g., poorly digitized text where spaces have been frequently inserted or removed) or does not contain text in a natural language. Only documents with a mean word length within a specified range are retained.

Symbol to Word Ratio If a document contains a high ratio of hashtag or ellipsis characters to words, it may suggest that the documents consists in large part of keywords or text that has been truncated. If this ratio exceeds a maximum threshold, the document is discarded.

Initial Bullet Point Ratio Documents that contain a large number of lines beginning with a bullet point likely consist primarily of itemized lists rather than running text. If the ratio of such lines is too high, the document is discarded.

Trailing Ellipsis Ratio If a large proportion of lines in a document end with an ellipsis, it may suggest that it contains a large amount of truncated text. This indicates that the text in the document may be incoherent. If this ratio exceeds a maximum threshold, the line is discarded.

Alphabetic Character Ratio A low ratio of tokens containing at least one alphabetic character within a document may suggest that the text is primarily non-linguistic. If the ratio falls below a minimum threshold, the document is discarded.

Stop Word Count If the document does not contain at least two unique stop words, it is discarded.

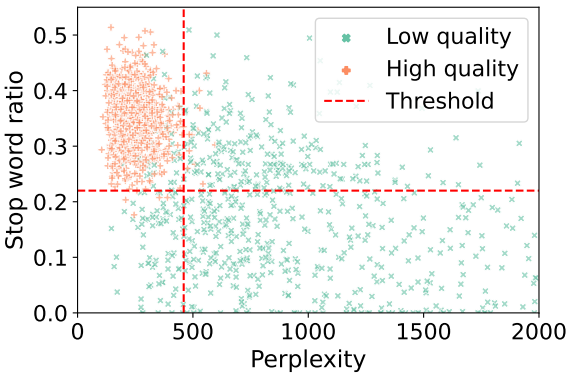

Figure 1: The distribution of documents in the TQ-IS dataset based on their perplexity score and their stop word ratio. High-quality documents form a single, dense cluster with a large number of low-quality outliers. The red, dashed line shows the optimal perplexity and stop word ratio thresholds that were found using grid search.

4.3. Other Rules

Finally, we propose one additional rule based on our observations on the TQ-IS dataset.

Mean Subword Length Subword tokenizers process out-of-vocabulary tokens by breaking them down into sequences of known subwords (Wu et al., 2016). When documents contain a large amount of foreign words, numbers, URLs, or other tokens that might not exist in the tokenizer's vocabulary, they tend to get broken down into many, short subwords. We propose a new rule that discards documents with a mean subword length (i.e., average number of characters per subword) that falls below a minimum threshold.

5. Outlier Detection

A visualization of feature pairs in TQ-IS, shown in Figure 1, reveals that high-quality documents form a single, dense and well-defined cluster. Low-quality documents, on the other hand, are most densely distributed in areas around the high-quality cluster, growing more sparse the further away they are. This suggests that it may be possible to accurately classify documents as low or high-quality using unsupervised clustering or outlier detection algorithms. We evaluate three such algorithms which are described in the following sections. For these algorithms, we use the same features that were used for the rule-based approach (e.g., perplexity, character repetition ratio, word repetition ratio, and so on).

5.1. Gaussian Mixture Model

A Gaussian Mixture Model (GMM) is a probabilistic model that can be used to estimate the parameters (means, covariances and mixture weights) of Gaussian distributions within a dataset. It can be used as a clustering algorithm under the assumption that each Gaussian distribution corresponds to a distinct cluster. Unlike density-based clustering algorithms, GMM is parametric and offers a soft clustering approach. This means that it can be fitted to one dataset and then used to probabilistically assign each data point in another dataset to these clusters.

5.2. Outlier Detection Algorithms

We also evaluate One-Class Support Vector Machines (OCSVM) (Schölkopf et al., 2001) and Isolation Forests (Liu et al., 2008), two outlier detection algorithms that are based on fundamentally different strategies. OCSVMs map the dataset into a higher-dimensional feature space using a kernel function. They then attempt to find the smallest possible boundary that encapsulates the densest region of the data, while maximizing the distance between the boundary and the feature space's origin. Data points that fall outside this boundary are considered to be outliers.

Isolation Forests generate an ensemble of binary trees (i.e., a forest) for a dataset, by repeatedly and randomly splitting the data until all data points have been isolated. Each data point is scored based on the average number of splits required to isolate it across all trees. A data point with a low average score is regarded as an outlier under the assumption that outliers are few and different.

6. Experimental Setup

In this section, we describe how we extract certain metrics from documents, our choice of languages for evaluation, how we apply grid search to optimize the thresholds for the rule-based approach, and how we tune the parameters of the clustering and outlier detection algorithms. We release the code used for our experiments with an open license.[2]

6.1. Feature Extraction

Extracting features from a document is usually a straightforward process, although some features require additional considerations. In order to calculate perplexity, we follow the general approach described by Guillaume et al. (2020), where the curated corpus is first processed by a subword tokenizer and an n-gram model is trained on the processed corpus. We choose to use a bigram model and a byte-pair encoding tokenizer with a vocabulary of 32k, following the results obtained by Daðason and Loftsson (2024). We use the same tokenizer to calculate the mean subword length of a document.

Character and word repetition ratios are calculated based on the proportion of recurring n-grams. We evaluate character n-gram sizes between 2 and 20 and word n-gram sizes between 2 and 10. For each rule, we choose whichever size yields the highest F_1 score when applied to the TQ-IS corpus in conjunction with other rules. While the optimal threshold value varies with n-gram size, the overall impact of both rules remains consistent. Based on our experiments, we calculate 5-gram word and 10-gram character repetition ratios.

We use the langid.py library for Python (Lui and Baldwin, 2012) to calculate a language confidence score for each document in TQ-IS. For documents where the primary language is not Icelandic, we set the confidence score to zero.

6.2. Language Selection

We evaluate the methods on a selection of three languages: Icelandic, Estonian and Basque. All three languages are reasonably well represented in

[2]The code used for our experiments is available at https://github.com/jonfd/tq-is.

Language	Curated (tokens)	mC4 (tokens)
Icelandic	1.7B	1.1B
Estonian	505M	3.0B
Basque	288M	576M

Table 1: The number of space-delimited tokens in the curated and web-crawled corpora for each language.

the mC4 corpus and, for each language, there exists a publicly available, high-quality curated corpus. Additionally, for Icelandic, TQ-IS (see Section 3) allows us to accurately assess the effectiveness of different text filtering approaches. Each language belongs to a different language family, with Icelandic being Indo-European, Estonian being Finno-Ugric and Basque being a language isolate. This represents a diverse selection of morphologically rich languages that should present a significant test for the robustness of any text filtering technique.

These three languages can hardly be categorized as under-resourced languages anymore. National Language Technology (LT) Programmes have been established both for Icelandic (Nikulásdóttir et al., 2020; Nikulásdóttir et al., 2022) and Estonian (Vider et al., 2012), and the development of LT in Basque Country has quite a long history (Alegria and Sarasola, 2017). However, as shown in Section 7.5, our results indicate that the unsupervised methods proposed in this paper should be applicable to under-resourced languages.

6.3. Corpora

We derive all web-crawled corpora from the mC4 corpus (Xue et al., 2020). For the curated corpora, which are used to learn the vocabulary for the subword tokenizer and to train the n-gram language model for calculating perplexity, we use the Icelandic Gigaword Corpus (IGC) for Icelandic (Barkarson et al., 2022) described in Steingrímsson et al. (2018), the Estonian National Corpus (ENC) for Estonian (Koppel and Kallas, 2022a), described in Koppel and Kallas (2022b), and Euscrawl for Basque (Artetxe et al., 2022a), described in Artetxe et al. (2022b). For each corpus, we do not include any subcorpora that were obtained from noisy web-crawled sources, such as Common Crawl. The total size of each corpus is shown in Table 1.

6.4. Threshold Optimization

To optimize the F_1 score on the TQ-IS dataset, we conduct a grid search with 10-fold cross-validation to determine the best combination of rules and thresholds. For each rule, we consider a range of values starting just before the point where the first false negative is produced (i.e., high-quality document misclassified as low-quality) and extending to where an F_1 score of 95% becomes unattainable.

Given the large search space for the full ruleset, we initially focus on individual rules, finding the threshold that optimizes their F_1 score. We select the highest-scoring rule and then determine optimal thresholds and F_1 scores for all possible pairings with the remaining rules. We then select the rule that yields the largest improvement to the F_1 score. We repeat this process iteratively until all available rules have been selected, or the F_1 score cannot be improved further.

6.5. Outlier Detection

For Icelandic, we optimize the parameters of each algorithm to achieve the highest possible F_1 score on the TQ-IS dataset. For Estonian and Basque, we use the optimal parameters for Icelandic as a starting point, iteratively adjusting them, if needed, by visual inspection until we deem their predictions to be subjectively satisfactory.

For the three clustering and outlier detection algorithms, we use the implementation from the scikit-learn library for Python (Pedregosa et al., 2011). As OCSVM is sensitive to the presence of extreme outliers, we scale the features using scikit-learn's robust scaler. For GMM, we instead trim the training set by discarding any document with a perplexity value of 4,000 or higher. We find that this produces better results than using the robust scaler.

Our experiments show that, when measured in terms of optimal F_1 scores, GMM models perform best when trained on a noisy, web-crawled corpus, while OCSVM and Isolation Forest models achieve better results when trained on a high-quality corpus. Therefore, to obtain the optimal parameters for Icelandic, we fit a GMM model to the Icelandic subset of the mC4 corpus, and the OCSVM and Isolation Forest models to the IGC. We train each model on a sample of 50,000 documents, as we find that larger training sets do not yield improved results. We then create a stratified 10-fold split of TQ-IS, in each fold using 90% of the documents as a validation set and the remaining 10% as a test set. We select the parameters that obtain the highest average F_1 score on the validation sets.

7. Results

In this section, we detail the results of our experiments with heuristic rules as well as the clustering and outlier detection algorithms. For each approach, we report F_1 scores that were obtained on the TQ-IS dataset and visualize the predictions made by the best performing algorithm on the Icelandic, Estonian and Basque subsets of the mC4

corpus.

7.1. Rule-based Approach

When performing a grid search on the TQ-IS dataset, our results show that perplexity is the single most effective feature when it comes to discerning between low and high-quality documents. When evaluated individually, we find the optimal maximum perplexity threshold to be 400, which yields an average F_1 score of 94.58%. We observe that the optimal threshold is relaxed significantly when other rules are included in the grid search, rising to 460 for the optimal ruleset.

For TQ-IS, we find that the optimal F_1 score is obtained when applying a combination of five rules, leaving eight rules unused. This includes all six rules that were used to filter the MassiveWeb corpus (see Section 4.2), as well as the character repetition ratio and language confidence rules used for the ROOTS corpus. The rules and their overall impact are shown in Table 2.

Metric	Ratio	F_1 score
Perplexity	44.85%	94.06%
+ Stop word ratio	35.25%	97.48%
+ Mean subword length	40.50%	97.86%
+ Word repetition ratio	5.80%	98.15%
+ Special character ratio	13.60%	98.20%

Table 2: Optimal ruleset and thresholds obtained for the TQ-IS dataset using cross-validated grid search. The rules appear in decreasing order of impact. The table shows the F_1 score of each rule when applied in conjunction with the rules above it, and the ratio of documents that fall outside the optimal threshold for each metric. In total, 50.2% of the documents are filtered with these rules.

Method	Features	F_1 score
GMM	PPL/SWR/MSL	98.32%
OCSVM	PPL/SWR	96.40%
Isolation Forest	PPL/SWR/MSL	97.52%

Table 3: F_1 scores obtained on TQ-IS using outlier detection models with optimized parameters (as described in Section 6.5). The GMM and Isolation Forest models obtained the best results using perplexity (PPL), stop word ratio (SWR) and mean subword length (MSL) as features, while OCSVM performed best using only perplexity and stop word ratio.

If we do not consider rules that require additional resources beyond a high-quality corpus (e.g., the stop word ratio) or additional tuning (e.g., character and word repetition ratios, which are n-gram based), we obtain an optimal F_1 score of 97.43% using only rules for perplexity and mean subword length. This may prove to be a reasonable approach for large, multilingual corpora, given the relatively low penalty that is incurred to the F_1 score.

7.2. Interquartile Range

We also evaluate the IQR method for selecting minimum and maximum thresholds, as described by Nguyen et al. (2023) (see Section 2). In this approach, all thresholds are configured to discard the exact same proportion of documents. For example, we might set the maximum perplexity, word repetition and special character ratio thresholds to the 90th percentile, and minimum stop word and mean subword length thresholds to the 10th percentile. Using the IQR method, we find the optimal ratio for the five rules shown in Table 2 to be 27%, which results in an F_1 score of only 91.53%, a notably lower score than was obtained through grid search. Having each rule discard the same proportion of documents results in some rules being underutilized (e.g., perplexity and mean subword length) and others being applied much too aggressively (e.g., word repetition ratio). Table 2 shows that under optimal settings, each rule classifies between 5.8% to 44.9% of the documents as low quality. Choosing a threshold somewhere in between leads to poor overall results. We therefore conclude that IQR is not an ideal approach to approximating optimal thresholds for text quality filtering.

7.3. Outlier Detection

The results for the three clustering and outlier detection algorithms are shown in Table 3. We observe that the optimal set of features for all three methods is smaller than the number of metrics used for the optimal rule-based approach, with OCSVM using only two features. This may be explained, in part, by the fact that the modest benefit to F_1 score offered by some rules, such as word repetition ratio (+0.29%) and special character ratio (+0.05%), may not make up for the cost of increasing the dimensionality of the data by adding a new feature.

7.4. Gaussian Mixture Model Visualization

We have shown that clustering and outlier detection algorithms obtain good results on the TQ-IS dataset. In order to determine whether the same holds true for larger, web-crawled corpora in other languages, we train GMMs on the Icelandic, Estonian and Basque subsets of the mC4 corpus and visualize the predictions they make. The results can be seen in Figure 2.

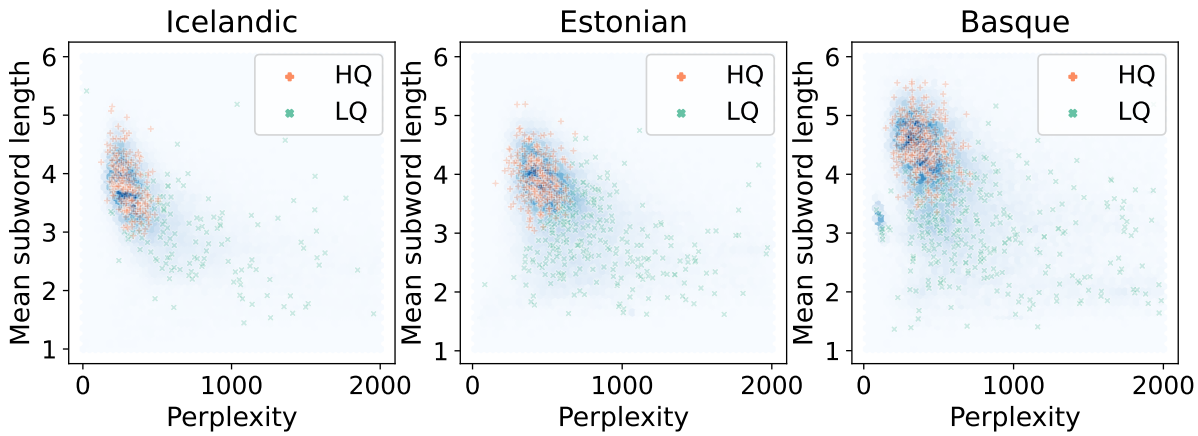

Figure 2: A visualization of the predictions made by GMMs on the Icelandic, Estonian and Basque subsets of the mC4 corpus. A scatter plot showing approximately 1,000 predictions made by each model is overlaid on a hexbin plot which depicts the distribution of documents in mC4 based on their perplexity and mean subword length.

First, we note that all three subsets share the same characteristics, having a single, large, elliptical cluster, surrounded by outliers that become more sparse the further away they are from the cluster. The distribution of the documents largely matches what we observed in TQ-IS, as shown in Figure 1. With a low perplexity value and a high mean subword length, it is easy to conclude that all three clusters consist primarily of high-quality documents. The predictions made by the GMMs for each language fully agree with our evaluation. While we lack text quality datasets for Estonian and Basque, we feel that this visualization is a strong indicator that clustering and outlier detection algorithms are well suited for text quality filtering in most languages.

7.5. Impact of Training Set Size

To determine the impact of training set size on the performance of the three models, we evaluate them on a variety of training set sizes, ranging from 100 to 30,000 documents. For each size, we sample ten distinct training sets from the appropriate corpus (mC4 for GMM and IGC for OCSVM and Isolation Forests) and report the average F_1 score obtained on TQ-IS.

As Figure 3 shows, we observe significantly diminished returns for all three methods after increasing the training set size to around 5,000 documents. Notably, the GMM model appears to be the most robust of the three, maintaining the most stable score and exhibiting the smallest standard deviation. These results indicate that the methods are likely to be effective even for under-resourced languages where web-crawled text may be limited.

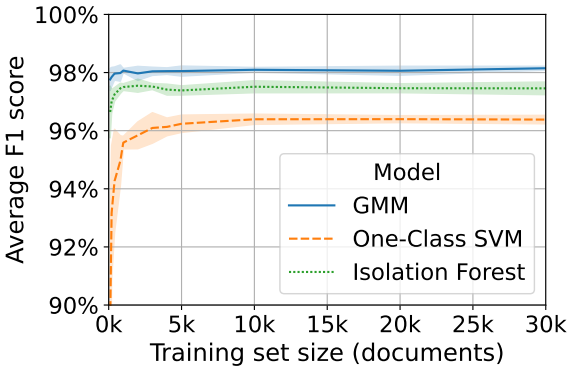

Figure 3: Average F_1 scores obtained by the three clustering and outlier detection algorithms on TQ-IS. The results show that a GMM performs very well even when fitted only to a handful of web-crawled documents, and that OCSVM and Isolation Forest models only require a small number of high-quality documents to be able to effectively identify low-quality outliers.

8. Conclusion

In this paper, we have evaluated the effectiveness of a large number of commonly applied heuristic rules for text quality filtering, both individually and when applied in conjunction with one another. We have demonstrated that perplexity is the most effective metric, by far, when it comes to discerning between low and high-quality documents. We have also shown that optimal results can be obtained with only the use of a handful of rules. Optimal rulesets and thresholds may differ between corpora and languages depending on their characteristics. However, we have shown that visualizing the distribution of documents within a corpus based on target met-

rics can reveal close to optimal threshold values in an intuitive manner, avoiding time-consuming analysis, manual labeling or guesswork.

Furthermore, we have proposed a novel approach to text quality filtering based on clustering and outlier detection algorithms. In particular, we find that the results obtained by a GMM-based approach can match those obtained with a rule-based approach, where the optimal set of rules and thresholds have been derived from a manually labeled dataset. The key benefits of this approach is that it does not require time-consuming feature engineering or threshold or parameter optimization, the creation of any manually labeled data or language expertise for the languages that are being filtered. Finally, our experiments indicate that the clustering and outlier detection algorithms are likely to be effective for under-resourced languages.

For future work, we intend to investigate how different categories of low-quality text impact the quality of pre-trained language models, particularly with regard to downstream performance. By answering these questions, we hope to gain a better understanding of how to improve text quality datasets such as TQ-IS, or construct them for other languages.

9. Limitations

As we lack document-level text quality datasets other than TQ-IS, we cannot empirically validate the effectiveness of clustering or outlier detection algorithms on languages other than Icelandic. However, as demonstrated in Figure 2, we have shown that relatively unfiltered web-crawled corpora in several languages have the same characteristics that make these methods so effective on TQ-IS (i.e., containing a single well-defined cluster of what the metrics strongly indicate to be high-quality documents).

10. Bibliographical References

Iñaki Alegria and Kepa Sarasola. 2017. Language Technology for Language Communities: An Overview based on Our Experience. In *Communities in Control: Learning tools and strategies for multilingual endangered language communities", CinC*, pages 19–21.

Mikel Artetxe, Itziar Aldabe, Rodrigo Agerri, Olatz Perez-de Viñaspre, and Aitor Soroa. 2022b. Does Corpus Quality Really Matter for Low-Resource Languages? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7383–7390, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models are Few-Shot Learners. *arXiv preprint arXiv:2005.14165*.

Jón Friðrik Daðason and Hrafn Loftsson. 2024. Text Filtering Classifiers for Medium-Resource Languages. In *The 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Torino, Italy. Forthcoming.

Frederik Michel Dekking, Cornelis Kraaikamp, Hendrik Paul Lopuhaä, and Ludolf Erwin Meester. 2005. *A Modern Introduction to Probability and Statistics: Understanding Why and How*, 1 edition. Springer Texts in Statistics. Springer London.

Mohamed Abdel Fattah and Fuji Ren. 2009. GA, MR, FFNN, PNN and GMM based models for automatic text summarization. *Computer Speech & Language*, 23(1):126–144.

Kristina Koppel and Jelena Kallas. 2022b. Eesti keele ühendkorpuste sari 2013–2021: mahukaim eestikeelsete digitekstide kogu. *Eesti Rakenduslingvistika Ühingu aastaraamat*, 18:207–228.

Julia Kreutzer, Isaac Caswell, Lisa Wang, Ahsan Wahab, Daan van Esch, Nasanbayar Ulzii-Orshikh, Allahsera Tapo, Nishant Subramani, Artem Sokolov, Claytone Sikasote, et al. 2022. Quality at a Glance: An Audit of Web-Crawled Multilingual Datasets. *Transactions of the Association for Computational Linguistics*, 10:50–72.

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 1.6TB Composite Multilingual Dataset. In *Advances in Neural Information Processing Systems*, volume 35, pages 31809–31826. Curran Associates, Inc.

Fei Tony Liu, Kai Ming Ting, and Zhi-Hua Zhou. 2008. Isolation Forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422.

Marco Lui and Timothy Baldwin. 2012. langid.py: An Off-the-shelf Language Identification Tool. In *Proceedings of the ACL 2012 System Demonstrations*, pages 25–30, Jeju Island, Korea. Association for Computational Linguistics.

Niklas Muennighoff, Alexander M. Rush, Boaz Barak, Teven Le Scao, Aleksandra Piktus, Nouamane Tazi, Sampo Pyysalo, Thomas Wolf, and Colin Raffel. 2023. Scaling Data-Constrained Language Models. *arXiv preprint arXiv:2305.16264*.

Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2023. CulturaX: A Cleaned, Enormous, and Multilingual Dataset for Large Language Models in 167 Languages. *arXiv preprint arXiv:2309.09400*.

Anna Nikulásdóttir, Jón Guðnason, Anton Karl Ingason, Hrafn Loftsson, Eiríkur Rögnvaldsson, Einar Freyr Sigurðsson, and Steinþór Steingrímsson. 2020. Language Technology Programme for Icelandic 2019-2023. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 3414–3422, Marseille, France. European Language Resources Association.

Anna Björk Nikulásdóttir, Þórunn Arnardóttir, Jón Guðnason, Þorsteinn Daði Gunnarsson, Anton Karl Ingason, Haukur Páll Jónsson, Hrafn Loftsson, Hulda Óladóttir, Einar Freyr Sigurðsson, Atli Þór Sigurgeirsson, Vésteinn Snæbjarnarson, and Steinþór Steingrímsson. 2022. Help Yourself from the Buffet: National Language Technology Infrastructure Initiative on CLARIN-IS. In *Selected Papers from the CLARIN Annual Conference 2021*, pages 109–125.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. 2019. Asynchronous Pipeline for Processing Huge Corpora on Medium to Low Resource Infrastructures. In *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7) 2019. Cardiff, 22nd July 2019*, pages 9 – 16, Mannheim. Leibniz-Institut für Deutsche Sprache.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. 2022. Scaling Language Models: Methods, Analysis & Insights from Training Gopher. *arXiv preprint arXiv:2112.11446*.

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.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *arXiv preprint arXiv:2211.05100*.

Bernhard Schölkopf, John C. Platt, John Shawe-Taylor, Alex J. Smola, and Robert C. Williamson. 2001. Estimating the Support of a High-Dimensional Distribution. *Neural Computation*, 13(7):1443–1471.

Steinþór Steingrímsson, Sigrún Helgadóttir, Eiríkur Rögnvaldsson, Starkaður Barkarson, and Jón Guðnason. 2018. Risamálheild: A Very Large Icelandic Text Corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Kadri Vider, Krista Liin, and Neeme Kahusk. 2012. Strategic Importance of Language Technology in Estonia. In *Human Language Technologies — The Baltic Perspective*. IOS Press.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. 2020. CCNet: Extracting High Quality Monolingual Datasets from Web Crawl Data. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France. European Language Resources Association.

Shaohua Wu, Xudong Zhao, Tong Yu, Rongguo Zhang, Chong Shen, Hongli Liu, Feng Li, Hong Zhu, Jiangang Luo, Liang Xu, and Xuanwei Zhang. 2021. Yuan 1.0: Large-Scale Pre-trained Language Model in Zero-Shot and Few-Shot Learning. *arXiv preprint arXiv:2110.04725*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. *arXiv preprint arXiv:1609.08144*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya

Barua, and Colin Raffel. 2021. mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Lan You, Qingxi Peng, Zenggang Xiong, Du He, Meikang Qiu, and Xuemin Zhang. 2020. Integrating aspect analysis and local outlier factor for intelligent review spam detection. *Future Generation Computer Systems*, 102:163–172.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. 2024. Yi: Open Foundation Models by 01.AI. *arXiv preprint arXiv:2403.04652*.

Joey Öhman, Severine Verlinden, Ariel Ekgren, Amaru Cuba Gyllensten, Tim Isbister, Evangelia Gogoulou, Fredrik Carlsson, and Magnus Sahlgren. 2023. The Nordic Pile: A 1.2TB Nordic Dataset for Language Modeling. *arXiv preprint arXiv:2303.17183*.

11. Language Resource References

Mikel Artetxe and Itziar Aldabe and Rodrigo Agerri and Olatz Perez-de-Viñaspre and Aitor Soroa. 2022a. *EusCrawl*. Ixa Group.

Starkaður Barkarson and Steinþór Steingrímsson and Þórdís Dröfn Andrésdóttir and Hildur Hafsteinsdóttir and Finnur Ágúst Ingimundarson and Árni Davíð Magnússon. 2022. *Icelandic Gigaword Corpus (IGC-2022) - unannotated version*. CLARIN-IS.

Jón Friðrik Daðason. 2021. *The Icelandic Crawled Corpus*. Hugging Face.

Jón Friðrik Daðason. 2024. *TQ-IS: A Text Quality Dataset for Icelandic (forthcoming)*. CLARIN-IS.

Kristina Koppel and Jelena Kallas. 2022a. *Estonian National Corpus 2021*. META-SHARE.

Linting Xue and Noah Constant and Adam Roberts and Mihir Kale and Rami Al-Rfou and Aditya Siddhant and Aditya Barua and Colin Raffel. 2020. *Multilingual Colossal Clean Crawled Corpus (mC4)*. Hugging Face.