Text Filtering Classifiers for Medium-Resource Languages

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

Abstract

Web-crawled corpora are essential resources for linguistic and NLP research, offering far more data than is available from curated corpora. However, they often contain a great deal of low-quality texts which can complicate research and degrade the quality of pre-trained language models. Therefore, they are typically filtered, e.g. by applying rules or classifiers. In this paper, we compare the effectiveness of various text filtering classifiers and measure their impact on language model performance for three medium-resource languages. We present TQ-IS, an Icelandic text quality dataset consisting of 2,000 web-crawled documents, in which spans of low-quality text have been manually identified and labeled. We then evaluate a perplexity-based classifier, a supervised classifier trained on TQ-IS, and a self-supervised classifier trained to discern between documents from curated and web-crawled corpora on Icelandic, Estonian and Basque. We find that these classifiers obtain F_1 scores of 94.48%, 99.01% and 93.40%, respectively, when evaluated on the TQ-IS dataset. Furthermore, our results show that while adding filtered web-crawled text to a pre-training corpus can improve downstream performance for pre-trained language models, any improvement is likely to remain modest unless the web-crawled corpus is significantly larger in size.

Keywords: Text quality, text filtering, language modeling

1. Introduction

Early Transformer-based language models, such as GPT (Radford et al., 2018) and BERT (Devlin et al., 2019), were typically pre-trained on curated corpora consisting of up to several billion words. It is now well established that increasing the size of pre-training corpora can significantly improve the downstream performance of such models (Liu et al., 2019). For this reason, it has become common practice to supplement high-quality pre-training corpora with, or even to rely exclusively on, documents scraped from online sources (Brown et al., 2020; Raffel et al., 2020; Xue et al., 2021b; Wu et al., 2021). Online texts are often obtained from large datasets, such as those published by Common Crawl (CC).[1] The GPT-3 model was pre-trained on 499B tokens, 410B of which were obtained from CC (Brown et al., 2020), while the T5 model was pre-trained on one trillion tokens from CC (Raffel et al., 2020).

While web-crawled corpora offer researchers the chance to dramatically increase the size of their datasets, such texts are typically quite noisy, often containing significant amounts of low-quality text, such as HTML tags, JavaScript code, navigation menus, headers, footers, boilerplate text, text in unwanted languages, or text that is otherwise incoherent or incomplete. An audit of 205 web-crawled corpora revealed that the ratio of usable text was below 50% in 87 of them, while 15 corpora contained no usable text at all (Kreutzer et al., 2022).

Due to the inherent quality issues in many web-crawled corpora, an initial filtering step is often performed to minimize the amount of low-quality text (i.e., noise) they contain. This often involves deduplication (i.e., the removal of duplicate text segments), applying rule-based filters (e.g., discarding documents with a high proportion of non-alphabetic characters), and using classifiers to identify and remove low-quality text. Several studies have demonstrated that text filtering can significantly improve the quality of pre-trained embeddings as well as the downstream performance of pre-trained language models (Wenzek et al., 2020; Brown et al., 2020; Raffel et al., 2020; Muennighoff et al., 2023).

Although text filtering is standard practice when relying on web-crawled text, there have been few in-depth experiments that include a fine-grained evaluation of the impact of individual filters. When evaluations are performed, they are frequently limited to measuring the overall impact of the text filters, which often comprise multiple heuristic rules, sometimes combined with text quality classifiers (e.g., the work by Raffel et al. (2020); Rae et al. (2022)), as opposed to evaluating the impact of each individual filter. Furthermore, detailed information on training data and experimental settings for text quality classifiers is sometimes omitted (e.g., the work by Wu et al. (2021)).

In this paper, we describe TQ-IS, a new, manually annotated text quality dataset for Icelandic. We describe three text quality classifiers, based on previously described approaches, and evaluate them on the TQ-IS dataset. We also use these classifiers to filter Icelandic, Estonian and Basque

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

LREC-COLING 2024, pages 15789–15801
20-25 May, 2024.

web-crawled corpora and evaluate how the filtering step impacts the downstream performance of pre-trained language models for those languages. Our evaluations show that all three classifiers are well suited for the task of text quality classification. Furthermore, we find that while supplementing a high-quality corpus with filtered web-crawled text can yield statistically significant improvements for downstream performance on some tasks, the benefit is often modest, even when the web-crawled corpus is several times larger in size.

2. Related Work

As mentioned earlier, it has been conclusively demonstrated that noisy text in pre-training corpora can degrade the quality of pre-trained language models. However, there is no clear definition of what exactly constitutes noisy or low-quality text, or how noisy a document must be in order to negatively impact downstream performance. For example, JavaScript code is often considered to be undesirable and is specifically targeted with rule-based filters for some corpora (Raffel et al., 2020). On the other hand, Muennighoff et al. (2023) find that augmenting a monolingual pre-training corpus with Python code, thereby doubling the size of the corpus, can lead to improved results on downstream tasks. Furthermore, noisy, web-crawled pre-training corpora can sometimes outperform high-quality, curated corpora of a similar size, as demonstrated by Artetxe et al. (2022b). Even though web-crawled corpora tend to be noisy, they may still be more balanced and representative than curated corpora with regard to downstream datasets. Additionally, some level of noise in the training data can have a regularizing effect during pre-training, which may improve the quality of the model.

Various metrics have been used to estimate the quality and readability of texts for tasks such as automated essay scoring and evaluating the output of generative language models (Attali and Burstein, 2006; Mathias and Bhattacharyya, 2018; Zhang et al., 2020b). One common metric is perplexity, which measures how well a probabilistic model can predict a given text segment (the lower the perplexity, the higher the probability of the segment). Perplexity has often been used in intrinsic evaluation of language models, for example by calculating the perplexity of a held-out portion of a pre-training corpus (Devlin et al., 2019; Radford et al., 2019; Zhang et al., 2020a).

Wenzek et al. (2020) propose classifying the quality of web-crawled documents based on their perplexity. They train a 5-gram model on a curated corpus that has been processed with a subword tokenizer and use this model to compute the perplexity of all documents within a web-crawled corpus. Next, they divide the corpus into three parts by perplexity, i.e., creating low, medium and high perplexity segments. For English and Polish, they train word embeddings on each segment and compare them on a selection of NLP tasks, finding that the quality of the embeddings degrades as the perplexity of the training data increases. The authors use this method to filter monolingual web-crawled corpora for English, Russian, Chinese and Urdu. For each of these languages, they find that a language model that has been pre-trained on a filtered web-crawled corpus obtains better downstream results than a language model pre-trained on a smaller, curated corpus. Their results suggest that this method generalizes well across different languages. Muennighoff et al. (2023) use the same approach to filter two English-language corpora that were derived from CC, and find that it improves downstream results by a statistically significant margin.

Wu et al. (2021) describe a pre-trained language model which has been fine-tuned on a dataset comprising articles that have been labeled as high-quality, low-quality, or advertisements. The model is then used to filter out low-quality documents and advertisements from a large web-crawled corpus. After manually reviewing a sample of documents labeled as advertisements (which account for half of the discarded documents), the authors report that less than 2% are false positives.

Brown et al. (2020) describe a logistic regression classifier with bag-of-word representations to distinguish between documents from curated corpora, on the one hand, and noisy web-crawled corpora, on the other. For each document in the web-crawled corpus, the authors use the classifier to estimate the probability that it originated from a curated corpus. This approach builds on the assumption that the lower the probability, the more likely it is that the document contains low-quality text. The authors select which documents to keep using a probability distribution that favors documents with high probabilities, but also includes some out-of-distribution documents. However, the authors do not evaluate the effectiveness of this classifier or its impact on downstream performance.

3. Icelandic Text Quality Dataset

In this section, we describe TQ-IS, a new text quality dataset for Icelandic, consisting of documents from a number of different web-crawled corpora. We performed a fine-grained analysis of the dataset, identifying and labeling low-quality text spans within each document. We then assigned each document an overall label of “low-quality” if low-quality text spans account for more than a third of the characters in the document, or “high-quality” if they account for less than 10%. Our annotation

guidelines are described in detail in Section 3.2. We release the TQ-IS dataset with an open license.[2]

3.1. Source Corpora

The documents in TQ-IS were obtained by sampling from the three corpora described in Sections 3.1.1–3.1.3, until a total of 1,000 documents of each category had been annotated. We limit the length of each document to 50 to 500 space-delimited tokens.

3.1.1. The Icelandic Crawled Corpus

The Icelandic Crawled Corpus (ICC) (Daðason, 2021) was created by scraping documents from a selection of Icelandic websites. The ICC consists of approximately 930M tokens, mostly from online forum posts, sports and municipal news articles, and adjudications from various governmental agencies. The websites were scraped using ad-hoc crawlers, specifically targeting items of interest while avoiding or discarding boilerplate text, headers, footers, metadata and other unrelated elements. Duplicate documents were discarded from the corpus, but no filtering was performed otherwise.

3.1.2. The Multilingual Colossal Clean Crawled Corpus

The Multilingual Colossal Clean Crawled Corpus (mC4), (Xue et al., 2021a), described in Xue et al. (2021b), consists of documents that were extracted from the entire CC dataset and classified with regard to their primary language. In total, mC4 contains subsets for 108 different languages, including 1.1B tokens for Icelandic. In an effort to remove low-quality text and duplicate content, the authors removed any duplicate occurrences of three line spans and discarded lines that do not end with punctuation marks.

3.1.3. The Icelandic Common Crawl Corpus

The Icelandic Common Crawl Corpus (IC3) (Snæbjarnarson, 2022), described in Snæbjarnarson (2021), is derived from websites with an Icelandic top-level domain (.is) within the CC dataset. A language classifier was used to remove pages with a primary language other than Icelandic. Similarly to mC4, duplicate occurrences of three line spans were discarded.

3.2. Annotation Guidelines

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 during training. It is impossible to know where exactly the line between these two categories of documents lies. Therefore, when creating TQ-IS, we chose to only include documents that we felt were clear-cut examples of each category. We considered the ideal high-quality document to primarily consist of running text in the form of sequences of full, grammatically structured sentences that are connected in a meaningful and coherent way. The text should contain few errors, if any, and be properly capitalized and punctuated. Documents that are disjointed, incoherent, error-prone, highly repetitive, or largely consist of foreign text, non-running text or non-linguistic data were classified as low-quality.

Specifically, we considered the following categories of text to be of low quality:

- **Foreign text**: Text where the primary language is not Icelandic.
- **Non-standard spelling**: Icelandic text that does not conform to modern standards of spelling or grammar.
- **Corrupted text**: Icelandic text that contains character encoding errors (e.g., “Reykjav??k”), HTML character entities (e.g., “"”), soft hyphens and escaped characters (e.g., “\n” and “\u266c”).
- **Run-on text**: Icelandic text that contains a large number of run-on sentences or words.
- **OCR text**: Digitized Icelandic text that contains a large number of errors and flaws caused by the optical character recognition (OCR) process (e.g., misrecognized characters or text columns appearing out of order).
- **Non-linguistic text**: Text with no apparent meaning (e.g., seemingly random sequences of symbols and numbers).
- **Incoherent text**: An apparently meaningless sequence of Icelandic words.
- **Code**: Text that consists primarily of code, such as HTML or JavaScript.
- **Non-content text**: Icelandic text that doesn’t contribute to the main subject of the document (e.g., boilerplate text, headers, footers, metadata and navigational elements).
- **Non-running text**: Icelandic text that is relevant to the main subject of the document, but isn’t in the form of full, grammatically structured sentences or breaks the flow of the document (e.g., lists, bullet points, tabulated data and image captions).

[2] https://github.com/jonfd/tq-is

- **Fragmented text**: Icelandic text that lacks flow or continuity (e.g., a list of headlines from news article or a sequence of short, truncated previews from unrelated blog posts).
- **Low-quality translations**: Icelandic text that has clearly been translated from another language with subpar results.
- **Repetitive text**: Icelandic text that has occurred elsewhere in the document (e.g., a short excerpt directly from an article which appears before the article itself).

The above categories are specifically intended to identify text that may degrade the quality of monolingual pre-trained language models. However, we intend for TQ-IS to be used for a variety of NLP tasks that may have different definitions of what constitutes low-quality text. Since we assigned fine-grained labels to text spans, different categories can be ignored depending on the task at hand.

We manually identified low-quality text spans in each document and annotated them according to the categories listed above. If at least a third of the document consists of low-quality text, the document itself was classified as being low-quality. However, if 10% or less of a document consists of low-quality spans, we instead labeled it as high-quality.

In general, we observe that documents in TQ-IS either consist primarily of low-quality spans or contain no low-quality spans at all. We find that in 80% of low-quality documents, low-quality spans account for 90% or more of the text, whereas 93% of high-quality documents contain no low-quality spans whatsoever.

4. Classifiers

In this section, we describe three types of text classifiers that we use for classifying documents as either low or high-quality.

4.1. Perplexity-Based Classifier

We implement a classifier that uses an n-gram model to calculate the perplexity of a given document, which is then labeled as high or low-quality based on a predetermined threshold. Like Wenzek et al. (2020), we train our n-gram models using the KenLM library (Heafield, 2011). However, instead of using a 5-gram model, we measure the effectiveness of a variety of n-gram orders (2-grams, 3-grams and 4-grams). We also compare several different vocabulary sizes (8k, 16k and 32k) for a WordPiece tokenizer (Wu et al., 2016). In order to determine the optimal n-gram order and vocabulary size, we train one model for each combination of these two settings on the Icelandic Gigaword Corpus (IGC) (Steingrímsson et al., 2018). We then create a stratified 10-fold split of the TQ-IS dataset, ensuring that each split contains an equal ratio of low and high-quality documents. With each n-gram model, we calculate the perplexity of all documents in the dataset. We then sort the values and generate candidate thresholds from the values between every two consecutive perplexity scores. The threshold that yields the highest F_1 score on the training set is then evaluated on the test set. We use the optimal threshold we find for the TQ-IS dataset to guide our choice of threshold for the other languages in our evaluation (see Section 5.1).

4.2. Supervised Classifier

Secondly, for Icelandic, we consider a document classifier which is trained on the TQ-IS dataset. We use a language model that has been pre-trained on the IGC (see Section 5.2) and fine-tune it on the dataset, similarly to Wu et al. (2021). Since our experiments are performed using small language models where the maximum sequence length is limited, we classify each document using a sliding window with a size of 128 tokens and a stride of 64. Each window is labeled according to the ratio of low quality text it contains, according to the annotated text spans. If at least a third of the text in a given window consists of low-quality text, that window is labeled as low-quality, or high-quality otherwise. In total, we extract 5,194 low-quality samples and 3,626 high-quality samples from the TQ-IS dataset.

The supervised classifier is evaluated using 10-fold cross-validation and a hold-out test set, which consists of 10% of the documents in the dataset. In this stage, the model is trained and evaluated on classifying relatively short text windows, rather than full documents. In order to classify a document as either low or high-quality, we first use the classifier to calculate the ratio of high-quality windows it contains. If the ratio is above a certain threshold, we classify the document as high-quality, or low-quality otherwise. In order to find this threshold, we first select the model that obtained the highest F_1 score during cross-validation. We then use this model to classify all windows within each document in the validation set and calculate the ratio of high-quality windows that each document contains. Next, we find the threshold which optimizes the document-level F_1 score using the same approach as for the perplexity-based classifier. Finally, we evaluate the model using this threshold on the hold-out test set.

4.3. Self-Supervised Classifier

Finally, we consider the approach described by Brown et al. (2020), where a classifier is trained to discern between documents from a high-quality curated corpus and a noisy web-crawled corpus.

The major benefit of this method is that it does not require a manually labeled dataset to train the classifier, since the label for each document is simply derived from its source. Rather than using a logistic regression classifier, we instead fine-tune a language model that was pre-trained on a high-quality corpus (see Section 5.1), similarly to the supervised classifier. As the two classifiers share the same model architecture, it will be easier to compare the effectiveness of each approach.

To train the classifier, we sample 50,000 documents from a curated corpus and another 50,000 from a web-crawled corpus. We then create a hold-out test set consisting of 10% of the sampled documents, equally split between the two corpora. Next, we extract 100,000 windows from each corpus from the remaining 90% of the documents, using the same sliding window approach as for the supervised classifier. This gives us a total of 200,000 training samples.

Once fine-tuned, the classifier can be used to estimate the probability that a given window originated from a web-crawled corpus. If this probability is extremely high, the window is assumed to be of low quality. In order to perform document-level classification, we take the same general approach as for the supervised classifier. For Icelandic, we use 80% of the TQ-IS dataset to find the optimal threshold value for classifying both windows and documents. In both cases, we find the threshold that optimizes the F_1 score of the classifier. Finally, we evaluate the classifier on the remaining 20% of the dataset. For the other languages in our evaluation, we use the same document-level threshold as we obtained for TQ-IS, but adjust the window-level threshold to match the ratio of tokens that are discarded from the Icelandic web-crawled corpus.

5. Experimental Setup

In this section, we detail our choice of languages, pre-training corpora, downstream tasks, datasets, and training and evaluation settings.

5.1. Language Selection

We evaluate our classifiers on a selection of three medium-resource languages: Icelandic, Estonian and Basque. Due to the fact that 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), we categorize these three languages as medium-resource.

For all three languages, there exists an openly available, high-quality curated corpus suitable for training language models. Additionally, they are all represented in mC4, which has not been extensively filtered with regard to text quality. Using mC4 as the source for our web-crawled corpora means that they will all be comparable with regard to how the documents were collected and processed. Finally, there are at least two openly available and reasonably sized datasets that can be used to evaluate pre-trained language models for each language. For even lower resource languages than the ones we selected, we note that a more practical approach may be to create ad-hoc scrapers for websites and domains that contain a significant amount of text in that language, since their number is likely to be relatively small. This would allow for high-quality text to be specifically targeted and eliminate the need for extensive filtering. We describe the corpora used for each language in the following sections.

Language	HQ (tokens)	LQ (tokens)
Icelandic	1.7B	1.1B
Estonian	505M	3.0B
Basque	288M	576M

Table 1: The number of space-delimited tokens in the high and low-quality corpus for each language.

5.1.1. Icelandic

For high-quality text in Icelandic, we use the 2022 version of the IGC (Barkarson et al., 2022), described in Steingrímsson et al. (2018). It contains 2.4B running words from genres such as news articles, parliamentary speeches, and adjudications. The 2022 version of IGC includes a large number of sentences that have been collected from Twitter and online forums, which we discard, leaving approximately 1.7B tokens of high-quality text. For low-quality text, we extract documents from the Icelandic subset of the mC4 corpus (described in Section 3), consisting of 1.1B tokens.

5.1.2. Estonian

Our curated Estonian corpus is based on the 2021 version of the Estonian National Corpus (ENC) (Koppel and Kallas, 2022a), described in Koppel and Kallas (2022b). It consists of a total of 2.9B tokens, largely composed of documents that have been collected from the web, which we discard, leaving 505M tokens. The remaining documents are from a variety of genres, such as news articles, fiction, and scientific literature. The low-quality corpus consists of documents extracted from the Estonian subset of the mC4 corpus, amounting to a total of 3.0B tokens.

5.1.3. Basque

For high-quality Basque text, we use the EusCrawl corpus (Artetxe et al., 2022a), described in Artetxe et al. (2022b). It contains 288 million tokens that were collected from high-quality websites using ad-hoc scrapers. EusCrawl consists primarily of news articles. For the web-crawled corpus, we use the Basque subset of the mC4 corpus, which consists of approximately 576M tokens.

5.2. Pre-training

For each language, we pre-train a language model on the high-quality corpus alone, the high-quality corpus supplemented with the unfiltered web-crawled corpus, and the the high-quality corpus combined with the different filtered versions of the web-crawled corpus. Each model is then evaluated on a selection of downstream tasks.

For these experiments, we pre-train ELECTRA-Small models, which consist of 14M parameters and take approximately 8 hours to pre-train using a TPU v3 accelerator (Clark et al., 2020). ELECTRA models are pre-trained using the replaced token detection (RTD) task, where tokens in a training sample are randomly replaced and the model attempts to determine which tokens are original and which are not. It has been suggested that the RTD task is highly data efficient, since it allows the model to learn from each token in a training sample, rather than a small portion as is often the case for other tasks such as masked language modeling (Wu and Dredze, 2020; Pyysalo et al., 2021).

5.3. Fine-Tuning and Downstream Tasks

We fine-tune and evaluate each pre-trained language model on a selection of downstream tasks in order to estimate the effectiveness of each classifier. Specifically, we evaluate each model on part-of-speech (PoS) tagging, named entity recognition (NER), and dependency parsing (DP). The following sections detail the datasets that are used for each language.

5.3.1. Icelandic

For PoS tagging, we fine-tune the models on MIM-GOLD (Barkarson et al., 2021), which consists of one million tokens that have been semi-automatically labeled with PoS tags (Loftsson et al., 2010). For NER, we use MIM-GOLD-NER (Ingólfsdóttir et al., 2022a), a version of MIM-GOLD that has been manually annotated with named entities (Ingólfsdóttir et al., 2020b). For DP, we fine-tune our models on the Icelandic Parsed Historical Corpus (IcePaHC) (Arnardóttir et al., 2023), which is a collection of one million tokens from the 12^{th} century to modern times that have been manually annotated with constituents (Rögnvaldsson et al., 2012). Specifically, we use the Universal Dependencies (UD) conversion of IcePaHC (Nivre et al., 2016).

5.3.2. Estonian

For PoS tagging and DP, we use the UD version of the Estonian Dependency Treebank (EDT) (Muischnek et al., 2023), which contains approximately 440 thousand tokens that have been annotated with universal PoS tags and constituents (Muischnek et al., 2016). For NER, we fine-tune our models on the EstNER corpus (Sirts, 2023a), which contains 184,638 tokens that were manually annotated with named entities (Sirts, 2023b).

5.3.3. Basque

For PoS tagging and DP, we fine-tune our models on the UD version of the Basque Dependency Treebank (BDT) (Aranzabe et al., 2023), which consists of 121,443 tokens that were annotated with morphosyntactic features and constituents (Aranzabe et al., 2015). For NER, we use the EIEC corpus (Alegria et al., 2004a), which contains 59,759 tokens which were semi-automatically labeled with named entities (Alegria et al., 2004b).

5.3.4. Settings

When fine-tuning the language models, we generally follow the same settings as Daðason and Loftsson (2022). For NER, we fine-tune models for 10 epochs with a learning rate of 5e-5 and a batch size of 16 and report entity-level F_1 sores. For PoS tagging, we use the same learning rate and batch size, but fine-tune the models for 20 epochs, and report tagging accuracy. For both tasks, we fine-tune models using the token classification training scripts that are included with the Transformers library (Wolf et al., 2020). For NER and PoS datasets that consist of fewer than 250,000 tokens (EstNER, EIEC and BDT), we use a weight decay of 1e-2 to reduce overfitting. For PoS tagging on the EDT dataset, we only train the model for 10 epochs, since it is both fairly large and we are only predicting the UPOS (word class) of each token.

For DP, we fine-tune models using DiaParser, a biaffine dependency parser that can extract contextual word embeddings from Transformer-based language models (Attardi et al., 2021). We fine-tune models for 200 epochs, otherwise using default settings, and report the labeled attachment score (LAS) for the model that obtained the best results on the validation set.

For MIM-GOLD and MIM-GOLD-NER, we perform 10-fold cross-validation using the same splits that were used by Daðason and Loftsson (2022).

For IcePaHC, EDT and BDT, we use the standard training, validation and test splits and report average results across five runs with different random seeds. For EstNER, since it is not distributed with any splits, we generate a 10-fold split and perform cross-validation.

We fine-tune the supervised classifier for 5 epochs using a batch size of 32 and a learning rate of 3e-4. We fine-tune the self-supervised classifier using the same settings, except that we find that a slightly lower learning rate of 1e-4 yields better results on the validation set. For both classifiers, we using a warmup ratio of 0.1 and a weight decay of 1e-2 to reduce overfitting. We perform 10-fold cross-validation, and for each fold, we use the model that obtained the highest F_1 score on the validation set.

6. Results

In this section, we report the results of our experiments. First, we evaluate each classifier on the TQ-IS dataset. Next, we determine how filtering noisy pre-training corpora for all three languages impacts the downstream performance of pre-trained language models.

6.1. Classifier Performance

We evaluate each of the three classifiers on TQ-IS. An overview of the results can be seen in Table 2.

Classifier	F_1 score
Supervised classifier	99.01%
Perplexity-based classifier	94.48%
Self-supervised classifier	93.40%

Table 2: F_1 test scores on the TQ-IS dataset for the three classifiers.

6.1.1. Perplexity-Based Classifier

When evaluating the Icelandic n-gram models on the TQ-IS dataset, as described in Section 4.1, we find that the best results are obtained using a bigram model with a vocabulary size of 32k. This model outperforms all other models but one by a statistically significant margin. The average F_1 cross-validation score of each model is shown in Table 3. Our results suggest that lower order models may be more suitable for text filtering, at least for medium-resource languages, where available corpora may not be large enough to sufficiently train higher order n-gram models. Furthermore, we find that a larger vocabulary size can improve the capabilities of the model, especially for lower order n-gram models.

Vocab.	2-gram	3-gram	4-gram
8k	93.18%	93.30%	92.85%
16k	93.62%	93.16%	92.44%
32k	**94.48%**	**94.06%**	92.90%

Table 3: These results show the average F_1 score for each n-gram model, obtained using 10-fold cross validation on the TQ-IS dataset. For each fold, we use the training set to find the optimal threshold value and evaluate it on the validation set. Scores in bold are statistically indistinguishable from the best result (paired t-test; $p < 0.05$).

We obtain an average F_1 test score of 94.48% using a bigram model with a vocabulary size of 32k. Proportionally, we find the greatest source of errors to consist of low-quality documents that contain non-running and fragmented text. This is likely due to short context length of the n-gram model.

Based on these results, we also train bigram models with a vocabulary size of 32k for Estonian and Basque. For all three languages, we use these models to calculate the perplexity of all documents in the mC4 corpus. The distribution of perplexity values for all three subsets of the mC4 corpus is shown in Figure 1.

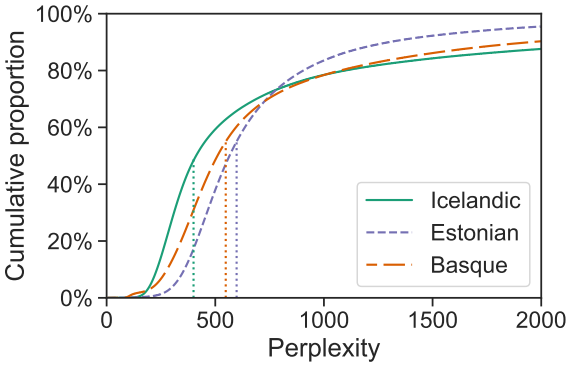

Figure 1: The cumulative distribution of perplexity values for documents in the Icelandic, Estonian and Basque subsets of the mC4 corpus. Dotted lines represent perplexity thresholds for document classification.

We next find the optimal threshold value for the Icelandic n-gram model using the TQ-IS dataset, as described in Section 4.1. With this threshold (399.99), classifying documents in the Icelandic subset of the mC4 corpus results in a total of 499M tokens being discarded, reducing the size of the corpus from 1.1B to 625M tokens, which is approximately a 45% reduction. Lacking a manually labeled text quality dataset for Estonian and Basque, we choose threshold values which discard the same ratio of tokens as for Icelandic. We find there is a noticeable increase in the amount of documents

which are clearly of low quality (e.g., containing foreign text, character encoding errors and code) around the chosen threshold.

6.1.2. Supervised Classifier

For the supervised classifier, we first pre-train an ELECTRA-Small model on the IGC and then fine-tune and evaluate it on the TQ-IS dataset using stratified 10-fold cross-validation, reserving 10% of the documents for a hold-out test set. We find that the fine-tuned models obtain an average F_1 score of 96.80% on the validation set.

Next, we find the optimal threshold for classifying document-level text quality based on the ratio of windows within each document that were predicted to be of high quality. We take the model that obtained the highest F_1 score during cross-validation (97.77%) and use it to score all windows within each document in the appropriate validation set. Once we have established the ratio of high-quality windows within each document in the validation set, we search for the optimal threshold using the same approach as before. We find that the threshold that maximizes the document-level F_1 score is 66.67%, meaning that at least two thirds of the windows within a document must be classified as high-quality for the document itself to also be classified as high-quality (or low-quality, otherwise). This is consistent with our annotation guidelines, according to which a document is considered low-quality if at least one third of its contents are of low quality. Using this threshold, the classifier correctly labels all documents in the validation set and obtains an F_1 score of 99.01% on documents in the test set.

6.1.3. Self-Supervised Classifier

For each language, we sample documents and training examples from the curated and web-crawled corpora as described in Section 4.3. For Icelandic, we perform 10-fold cross-validation in the same manner as for the supervised classifier, obtaining an average F_1 score of 90.26% when predicting the origin of the text windows. The quality of each window is then determined based on the score returned by the classifier. For Icelandic, we use the TQ-IS dataset to find the optimal thresholds for both window and document-level classification. We use 80% of the documents in TQ-IS to determine both thresholds and evaluate them on the remaining 20%.

We select the model which obtained the highest F_1 score during cross-validation (90.48%) and use it to score every window in the TQ-IS dataset. We find that the optimal threshold for classifying the quality of the windows is quite low, as expected, with a value of 0.00113, which results in a window-level F_1 score of 88.98%. This supports the theory that when the classifier is extremely confident that a text segment originates from a web-crawled corpus, it most likely contains low-quality text. We find that the optimal threshold for classifying documents based on the ratio of high to low-quality windows is 50%, meaning that at least half of the windows have to be classified as high-quality for the document to be labeled as high-quality as well. This results in a document-level F_1 score of 94.34%. When we use these thresholds to classify the remaining 20% of documents in the TQ-IS corpus, we obtain an F_1 score of 93.40%.

Unlike the perplexity-based classifier, the self-supervised classifier proves more capable when it comes to detecting documents with fragmented text. However, it also struggles with non-running text. This may be due to the fact that non-running text is represented to some degree in IGC. Additionally, we find that it does not perform as well as the perplexity-based classifier when it comes to identifying documents that contain a large number of OCR errors.

Filtering the Icelandic subset of the mC4 corpus using these thresholds results in 47% of the tokens being discarded, which is a similar ratio to the supervised classifier. For Estonian and Basque, we also use a threshold of 50% for document-level classifications, but tune the window-level threshold so that the same ratio of tokens are discarded as for Icelandic.

6.2. Language Model Performance

Our evaluation shows that pre-trained language models appear to be surprisingly resilient to noisy text. Only four out of the nine datasets we evaluate show a statistically significant decline in performance when a high-quality corpus is supplemented with an unfiltered web-crawled corpus. However, if the web-crawled corpus is first filtered using the perplexity-based classifier, we observe that we get identical or improved results for eight of the nine datasets. Our results also suggest that when supplementing a curated corpus with web-crawled text, the increase in size must be substantial in order to see a meaningful improvement in downstream performance. For Icelandic and Basque, where supplementing a curated corpus with mC4 increases the size of the pre-training corpus by 36% and 123%, respectively (using perplexity filtering), we only observe a statistically significant improvement in one out of three tasks for both languages. However, for Estonian, where the pre-training corpus is enlarged by 548%, we find that two out of three tasks see a statistically significant improvement. The overall downstream performance of the language models that were pre-trained on filtered and unfiltered corpora can be seen in Table 4.

Overall, filtering low-quality documents with the

	PoS			NER			DP		
Corpora	**IS**	**ET**	**EU**	**IS**	**ET**	**EU**	**IS**	**ET**	**EU**
HQ	**96.95**	97.93	**96.88**	**91.30**	**91.36**	**83.16**	84.79	88.38	84.31
+ mC4	96.80	**97.93**	96.82	**91.08**	91.14	81.32	**84.89**	**88.66**	85.13
+ mC4-PPL	**96.90**	**97.95**	96.84	**91.39**	**91.70**	**82.66**	84.75	**88.75**	**85.27**
+ mC4-SC	96.86	-	-	**91.42**	-	-	**84.79**	-	-
+ mC4-SSC	96.85	**97.96**	**96.92**	**91.27**	91.04	**83.01**	**84.82**	88.44	85.03

Table 4: Downstream performance of Icelandic (IS), Estonian (ET), and Basque (EU) language models pre-trained on the curated corpus alone (HQ), and with different versions of the web-crawled corpus: unfiltered (mC4), filtered with the perplexity (mC4-PPL), supervised (mc4-SC) and self-supervised classifiers (mC4-SSC). Results in bold are statistically indistinguishable from the best score for each task (paired t-test; $p < 0.05$). We note that models that obtain identical average results for the same task can still be statistically distinguishable from one another depending on the variability of the results obtained during cross-validation.

perplexity-based classifier appears to yield the greatest improvement in downstream performance, with the supervised classifier obtaining similar results for Icelandic. However, while our results show that filtering web-crawled corpora can yield better results on downstream tasks, a substantial increase is unlikely unless the web-crawled corpus adds a significant amount of text.

7. Conclusions

We have presented a new text quality dataset for Icelandic, TQ-IS, which consists of 2,000 documents that have been manually annotated, both with regard to overall document quality as well as by identifying and labeling low-quality text segments within each document. We evaluated three different text quality classifiers on three medium-resource languages. We evaluated all three classifiers on the TQ-IS dataset and found them all to be well suited for the task of text quality classification. We further evaluated the classifiers by using them to filter low-quality documents from web-crawled corpora. To measure the effectiveness of the filtering, we compared the downstream performance of language models that were pre-trained on filtered and unfiltered versions of the same corpora. Our experiments showed that filtering results in similar or improved performance on all downstream tasks.

Although we did not observe a substantial improvement in downstream performance, we note that filtering the web-crawled corpora reduced their size roughly by half. This could proportionally reduce computational costs and training time when pre-training for a certain number of epochs. In general, it is reasonable to assume that removing unhelpful or harmful pre-training examples should improve efficiency during pre-training, whether targeting a specific number of epochs or steps. Thus, even in cases where filtering might not significantly improve downstream performance, the potential benefits with regard to efficiency should likely make it worthwhile to filter noisy pre-training corpora.

In this paper, we have evaluated the capabilities of three text quality classifiers on three medium-resource languages. The supervised classifier proved to be the most effective of the three when evaluated on TQ-IS. We have shown that a very small language model with only 14M parameters can be fine-tuned to detect a wide range of low-quality text categories with near perfect accuracy when trained on a small, manually labeled sample from the corpus that is to be filtered. Our results also agree with previous findings that show perplexity to be a highly useful proxy measure of document quality, as long as the language model has been trained on a high-quality, representative corpus. Finally, we find self-supervised classifiers, trained to discern whether documents originate from high or low-quality corpus, to perform well on the task of text quality classification, though not quite as effective as the other two classifiers.

In the future, we intend to experiment with other approaches to text quality filtering, such as comparing the effectiveness of text quality classifiers to commonly used rule-based filters, training a sequence labeling classifier on the TQ-IS dataset and evaluating zero-shot classifiers. We plan to include web-crawled corpora for high-resource languages, such as English, in these experiments. Furthermore, we will investigate how the size and diversity of the corpora used to train text quality classifiers affect their performance.

Acknowledgements

This research was supported with Cloud TPUs from Google's TPU Research Cloud (TRC).

8. Limitations

While our evaluation extends to several languages, we do not consider how linguistic idiosyncrasies might affect the downstream performance of the models, for example in relation to the size or quality of the pre-training corpus. Furthermore, for some languages, such as Icelandic, the amount of web-crawled text is quite limited compared to what is already available in existing high-quality corpora, especially after having been filtered. In such circumstances, there is unlikely to be significant benefit from supplementing a high-quality corpus with web-crawled text. Our experiments show how much improvement can realistically be expected for the languages included in our evaluation, but they do not reveal how much larger the web-crawled corpus has to be for downstream performance to improve by a significant margin. We leave this as a potential avenue for future work.

9. 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.

Iñaki Alegria, Olatz Arregi, Irene Balza, Nerea Ezeiza, Izaskun Fernandez, and Ruben Urizar. 2004b. Design and Development of a Named Entity Recognizer for an Agglutinative Language. In *First International Joint Conference on NLP (IJCNLP-04). Workshop on Named Entity Recognition*.

Maria Jesús Aranzabe, Aitziber Atutxa, Kepa Bengoetxea, Arantza Diaz de Ilarraza, Iakes Goenaga, Koldo Gojenola, and Larraitz Uria. 2015. Automatic Conversion of the Basque Dependency Treebank to Universal Dependencies. In *Proceedings of the Fourteenth International Workshop on Treebanks and Linguistic Theories (TLT14)*, pages 233–241, Warszawa, Poland. Institute of Computer Science of the Polish Academy of Sciences.

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.

Yigal Attali and Jill Burstein. 2006. Automated Essay Scoring With e-rater® V.2. *The Journal of Technology, Learning and Assessment*, 4(3).

Giuseppe Attardi, Daniele Sartiano, and Maria Simi. 2021. Biaffine Dependency and Semantic Graph Parsing for Enhanced Universal Dependencies. In *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 184–188, Online. 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*.

Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. In *8th International Conference on Learning Representations, ICLR 2020*, Addis Ababa, Ethiopia.

Jón Friðrik Daðason and Hrafn Loftsson. 2022. Pre-training and Evaluating Transformer-based Language Models for Icelandic. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7386–7391, Marseille, France. European Language Resources Association.

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.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, pages 187–197, Edinburgh, Scotland. Association for Computational Linguistics.

Svanhvít L. Ingólfsdóttir, Ásmundur A. Guðjónsson, and Hrafn Loftsson. 2020b. Named Entity Recognition for Icelandic: Annotated Corpus and Models. In *Proceedings of the 8th International Conference on Statistical Language and Speech Processing (SLSP 2020)*, pages 46–57, Cardiff, United Kingdom.

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.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

Hrafn Loftsson, Jökull H. Yngvason, Sigrún Helgadóttir, and Eiríkur Rögnvaldsson. 2010. Developing a PoS-tagged corpus using existing tools. In *Proceedings of 7th SaLTMiL Workshop on Creation and Use of Basic Lexical Resources for Less-Resourced Languages*, LREC 2010, Valetta, Malta.

Sandeep Mathias and Pushpak Bhattacharyya. 2018. Thank "Goodness"! A Way to Measure Style in Student Essays. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 35–41, Melbourne, Australia. 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*.

Kadri Muischnek, Kaili Müürisep, and Tiina Puolakainen. 2016. Estonian Dependency Treebank: from Constraint Grammar tagset to Universal Dependencies. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1558–1565, Portorož, Slovenia. European Language Resources Association (ELRA).

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.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Sampo Pyysalo, Jenna Kanerva, Antti Virtanen, and Filip Ginter. 2021. WikiBERT Models: Deep Transfer Learning for Many Languages. In *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 1–10, Reykjavik, Iceland (Online). Linköping University Electronic Press, Sweden.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

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.

Eiríkur Rögnvaldsson, Anton Karl Ingason, Einar Freyr Sigurðsson, and Joel Wallenberg. 2012. The Icelandic Parsed Historical Corpus (IcePaHC). In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1977–1984, Istanbul, Turkey. European Language Resources Association (ELRA).

Kairit Sirts. 2023b. Estonian Named Entity Recognition: New Datasets and Models. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 752–761, Tórshavn, Faroe Islands. University of Tartu Library.

Vésteinn Snæbjarnarson. 2021. Automated methods for Question-Answering in Icelandic. Master's thesis, University of Iceland.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, et al. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

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*.

Shijie Wu and Mark Dredze. 2020. Are All Languages Created Equal in Multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

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. 2021b. 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.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. 2020a. PEGASUS: Pre-training with Extracted Gap-sentences for Abstractive Summarization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 11328–11339. PMLR.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.

10. Language Resource References

Iñaki Alegria and Olatz Arregi and Irene Balza and Nerea Ezeiza and Izaskun Fernandez and Ruben Urizar. 2004a. *Basque Named Entities Corpus (EIEC)*. Ixa Group.

Maria Jesus Aranzabe and Aitziber Atutxa and Kepa Bengoetxea and Arantza Diaz de Ilarraza and Iakes Goenaga and Koldo Gojenola and Larraitz Uria. 2023. *UD Basque Dependency Treebank (BDT)*. LINDAT/CLARIAH-CZ.

Þórunn Arnardóttir and Hinrik Hafsteinsson and Einar Freyr Sigurðsson, Hildur Jónsdóttir and Kristín Bjarnadóttir and Anton Karl Ingason and Kristján Rúnarsson and Steinþór Steingrímsson and Joel C. Wallenberg and Eiríkur Rögnvaldsson. 2023. *UD Icelandic Parsed Historical Corpus (IcePaHC)*. LINDAT/CLARIAH-CZ.

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

Starkaður Barkarson and Þórdís Dröfn Andrésdóttir and Hildur Hafsteindóttir and Árni Davíð Magnússon and Kristján Rúnarsson and Steinþór Steingrímsson and Haukur Páll Jónsson and

Hrafn Loftsson and Einar Freyr Sigurðsson and Eiríkur Rögnvaldsson and Sigrún Helgadóttir. 2021. *MIM-GOLD 21.05 - train/test*. CLARIN-IS.

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.

Svanhvít Lilja Ingólfsdóttir and Ásmundur Alma Guðjónsson and Hrafn Loftsson. 2022a. *MIM-GOLD-NER 2.0 – named entity recognition corpus (22.06) (2022-06-10)*. CLARIN-IS.

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

Kadri Muischnek and Kaili Müürisep and Tiina Puolakainen and Andriela Rääbis and Liisi Torga. 2023. *UD Estonian Dependency Treebank (EDT)*. LINDAT/CLARIAH-CZ.

Kairit Sirts. 2023a. *New Estonian NER Dataset (New EstNER)*. Github.

Vésteinn Snæbjarnarson. 2022. *The Icelandic Crawled Corpus*. Hugging Face.

Linting Xue and Noah Constant and Adam Roberts and Mihir Kale and Rami Al-Rfou and Aditya Siddhant and Aditya Barua and Colin Raffel. 2021a. *Multilingual Colossal Clean Crawled Corpus (mC4)*. Hugging Face.