

HPLT3-PerRef: A Refined Large-Scale Persian Web Corpus via Supervised Document Quality Filtering

Alireza Razzaghi, Tomasz Galica, Erik Henriksson, Veronika Laippala

TurkuNLP, University of Turku

{alireza.razzaghi, tomasz.galica, erik.henriksson, mavela}@utu.fi

Abstract

Large-scale web corpora are valuable resources for natural language processing (NLP), linguistic research, and language model training, particularly for digitally low-resource languages such as Persian. However, web-derived text often contains document-level noise, even after quality-filtering pipelines. We present HPLT3 Persian Refined (HPLT3-PerRef), a refined Persian web corpus containing 105.9 billion tokens and 102.37 million documents, obtained by applying a supervised quality-filtering pipeline to the Persian subset of HPLT3. We manually annotated 6,144 Persian web documents and trained a multilingual transformer-based classifier to distinguish high- and low-quality documents. The classifier removed approximately 20% of the original corpus while preserving broad domain coverage and complements HPLT3’s heuristic-based Web Document Score (WDS) by identifying additional low-quality documents. We publicly release the binary annotated corpus, the trained classifier, the HPLT3-PerRef, and the complete source code on [GitHub](#), [Hugging Face \(Data\)](#) and [Hugging Face \(Quality classifier\)](#).

1 Introduction

The web has become a valuable source of textual data for linguistic research and natural language processing (NLP), providing large-scale, diverse, and continuously updated language data that are difficult to obtain from curated resources alone. This is particularly important for digitally low-resource languages such as Persian, where high-quality text collections remain limited. However, web-derived corpora often contain document-level noise that can affect linguistic analyses and language model training, motivating continued efforts to improve corpus quality.

Several monolingual Persian (Farsi) corpora have been developed in recent years ([Hosseinbeigi](#)

[et al., 2025](#); [Khansari and Shamsfard, 2021](#); [Ayoubi, 2022](#); [Sarmadi et al., 2025](#); [Sabouri et al., 2024](#); [Ayoubi, 2021](#); [Darvishi et al., 2023](#); [Sabati et al., 2018](#); [Doostmohammadi et al., 2018](#); [AleAhmad et al., 2016, 2009](#); [Shojaei, 2024](#)). However, many are limited in scale, are not openly accessible, or focus on specific content types such as news or blogs. Large multilingual web corpora ([Nguyen et al., 2024](#); [Oepen et al., 2025](#); [Sagot and Romary, 2019](#); [Xue et al., 2021](#); [Kudugunta et al., 2023](#)) provide broader Persian coverage but also include noisy documents, such as boilerplate, SEO-oriented pages, machine-translated text, and archive pages. Such documents increase corpus size while reducing its value for linguistic research and language model training.

We present HPLT3-PerRef, a 105.9-billion-token Persian corpus derived from the Persian portion of HPLT3 ([Oepen et al., 2025](#)). Starting from HPLT3’s heuristic-filtered corpus, we further refine the Persian subset using supervised document-level quality classification. Despite this additional filtering, HPLT3-PerRef remains among the largest publicly available Persian corpora (Appendix Table 4). Specifically, we (i) manually annotate a subset of HPLT3 documents to train a document quality classifier, (ii) apply the classifier to the full HPLT3 Persian part to remove low-quality documents and assign quality confidence scores to the retained ones, and (iii) publicly release the refined corpus, the classifier, the document-level quality scores, and the source code. Finally, we show that the classifier complements HPLT’s Web Document Score (WDS) by identifying additional low-quality documents.

2 Related Work

Existing Persian language resources span a wide range of domains and purposes. Large-scale web corpora include Matina ([Hosseinbeigi et al., 2025](#)),

Targoman (Ayoubi, 2022), and MirasText (Sabeti et al., 2018). Domain-specific datasets cover blogs (HmBlogs Khansari and Shamsfard 2021, irBlogs AleAhmad et al. 2016), news (PerKey Doostmohammadi et al. 2018, Hamshahri AleAhmad et al. 2009), and PDF documents (PDC Shojaei 2024). Multi-source corpora such as Hakim (Sarmadi et al., 2025) and naab (Sabouri et al., 2024) integrate web text with resources including Wikipedia, Common Crawl, and social media. Moreover, PersianQA (Ayoubi, 2021) and PQuAD (Darvishi et al., 2023) provide task-specific benchmarks for question answering. However, many existing monolingual resources are limited in scale, focus on specific domains or tasks, or face accessibility and reproducibility challenges. For example, the largest publicly released monolingual Persian corpus, Matina, contains 72.9B tokens (Appendix 4), making it smaller than HPLT3-PerRef. In addition, access to the corpus requires approval through Hugging Face, which limits ease of access and reproducibility.

Furthermore, open-source large multilingual web corpora provide broader coverage of Persian. Representative examples are CultureX (Nguyen et al., 2024), HPLT3 (Oepen et al., 2025), OSCAR (Sagot and Romary, 2019), mC4 (Xue et al., 2021), and Madlad-400 (Kudugunta et al., 2023), which include Persian alongside many other languages. Among these, HPLT3 provides one of the largest publicly available collections of Persian web text and has a document-level filtering pipeline.

Despite these existing filtering efforts, web corpora often retain document-level noise. HPLT3 employs the WDS, a heuristic-based scoring method that estimates document quality from structural and statistical properties of web pages. Other approaches, including Propella (Idahl et al., 2026) and WebOrganizer (Wettig et al., 2025), use learned document quality models for large-scale web corpus filtering; however, they have not been developed or evaluated for Persian. These observations motivate our use of HPLT3 as the foundation for HPLT3-PerRef: it contains one of the largest publicly available collections of Persian web text, yet its heuristic WDS filtering still retains some document-level noise. HPLT3-PerRef therefore further refines the Persian subset using a supervised document-level quality classifier.

3 Corpus Construction

3.1 Source Corpus

Our corpus is constructed from the Persian portion of HPLT3. Each document corresponds to the cleaned textual content extracted from a single web page. The Persian subset contains approximately 127.96 million documents and 157.76 billion tokens, collected from the Internet Archive (2012–2020) and 57 Common Crawl snapshots (2014–2025). HPLT3 applies several large-scale processing steps, including language identification, quality filtering using the WDS, and global near-duplicate removal with MinHash, resulting in broad domain coverage and reduced duplication.

Although these preprocessing steps provide a strong foundation, additional Persian-specific quality refinement can further improve the suitability of the corpus for linguistic research and language model training. Through repeated manual inspections of randomly sampled documents from the Persian portion of HPLT3, we observed that some low-quality content remains after the existing filtering process. This suggests that heuristic quality measures may not capture all types of document-level noise.

3.2 Annotation

Building on the preprocessing performed by HPLT3, we further refine its Persian subset using a Persian-specific document quality classifier trained on a manually annotated dataset. Documents were assigned one of two labels based on their overall quality: high-quality (class 1) or low-quality (class 0).

High-quality documents were defined as coherent and complete texts forming a unified document. Low-quality documents included incomplete texts (e.g., fewer than two complete sentences), incoherent content, documents dominated by boilerplate, special characters, numerical sequences, headline fragments, lists, download pages, image captions, keyword lists, repeated search terms, link collections, or SEO-oriented content. Examples of high- and low-quality documents are shown in Figure 1.

3.3 Quality Classifier

The binary document quality classifier was trained on an annotated dataset of 6,144 Persian web documents (3,998 high-quality and 2,146 low-quality) sampled from the HPLT3 Persian corpus, and divided into training, validation, and test sets using

Low-quality: SEO page
فیلم ملکه - خشک کردن سبزی - دانلود بازی تخته نرد <i>Translation:</i> -Queen's Movie -Drying vegetables - Download backgammon game.
Low-quality: Boilerplate
صفحه اصلی تماس با ما قوانین <i>Translation:</i> Home Contact Us Terms
High-quality
تیم استقلال ساری بدون حامی مالی و به عنوان تنها نماینده از مازندران وارد رقابت‌های لیگ برتر فوتسال بانوان شده و در 6 مسابقه ابتدایی خود پنج بار متحمل شکست شده است. <i>Translation:</i> Esteghlal Sari team, without any financial support and as the only representative from Mazandaran, entered the competition of the women's futsal league and lost five out of their first six matches.

Figure 1: Examples of manually annotated high- and low-quality documents.

an 80/10/10 split.

We evaluated two multilingual transformer encoders, XLM-RoBERTa-large (Conneau et al., 2020) and BGE-M3 (Chen et al., 2024), which are suitable for Persian web data that contains diverse linguistic patterns, including non-standard writing, code-switching, multilingual fragments, and other web-specific noise. XLM-RoBERTa was selected as a strong multilingual baseline, while BGE-M3 was chosen for its longer input sequence length (1,024 vs. 512 tokens), enabling more comprehensive document-level quality estimation. Both models were fine-tuned as binary sequence classifiers, with hyperparameters optimized using Bayesian optimization with Optuna on the training and validation sets. The search space, optimization procedure, and selected hyperparameter configuration are described in Appendix A.2. The best configuration was retrained from scratch and evaluated on the held-out test set.

Metric	BGE-M3	XLM-R
Accuracy	96.91%	97.56%
Macro F1	96.58%	97.31%
Weighted F1	96.90%	97.56%
Total errors	19/615	15/615

Table 1: Performance of quality classification models on the test set.

Although XLM-RoBERTa achieved slightly higher classification performance (Table 1), we selected BGE-M3 for large-scale corpus filtering because our goal was not only binary classification but also reliable document-level quality scoring across a continuous range. Its 1,024-token input capacity allows it to evaluate a larger portion of each web page, producing more representative quality estimates while maintaining comparable performance.

The classifier outputs a continuous score between 0 and 1 representing the predicted probability of a document belonging to the low-quality class, with higher scores indicating lower quality. Unlike a fixed binary decision, this probability-based output allows the filtering threshold to be adjusted according to the desired trade-off between corpus size and quality. We selected the final threshold on the validation set by evaluating precision and recall for the low-quality class across different thresholds, aiming to maximize low-quality document removal while retaining high-quality documents.

The BGE-M3 classifier contains approximately 567 million parameters, while XLM-RoBERTa-large contains approximately 559 million parameters. Model training was performed on a single NVIDIA A100 GPU with 256 GB of host memory and 8 CPU cores, and took approximately 3.5 hours for BGE-M3-RetroMAE and 10 minutes for XLM-RoBERTa-large. Applying the classifier to the HPLT3 data was performed on a single NVIDIA A100 GPU with 256 GB of host memory and 16 CPU cores. The HPLT3 data were processed using streaming inference across multiple jobs, with a cumulative elapsed job time of approximately 53 GPU hours.

3.4 Intrinsic Corpus-Quality Check

To provide an intrinsic validation beyond agreement with the human annotations, we conducted a controlled language-model pretraining experiment. We trained identical, randomly initialized GPT-2-style decoder-only Transformers on three corpora: HPLT3, HPLT3-PerRef, and a random-removal control created by randomly removing approximately 20% of documents from HPLT3.

The models have 6 Transformer layers, 12 attention heads, a hidden size of 768, and a context length of 2,048 tokens. We used the Qwen3.6-35B-A3B (Qwen Team, 2026) tokenizer for all three conditions, but we did not use the pre-

trained Qwen3.6-35B-A3B model or its pretrained weights. Each model was trained from a random initialization using the same initialization seed, architecture, optimization settings, and 100M-token training budget. This ensures that differences in validation loss are attributable to the training corpora rather than to differences in model configuration or training conditions.

After training, all models were evaluated on the same held-out Persian test set using next-token prediction loss and perplexity. The results are shown in Table 2. HPLT3-PerRef achieved lower held-out cross-entropy and perplexity than both the original HPLT3 corpus and the random-removal control. This result suggests that the documents removed by the supervised filtering procedure are not simply a random subset of HPLT3. Under a fixed token budget, the resulting HPLT3-PerRef corpus produces a more predictable training distribution according to this intrinsic language-modeling criterion.

Training corpus	Cross-entropy	Perplexity
HPLT3	13.14	511K
Random removal	13.22	553K
HPLT3-PerRef	12.93	414K

Table 2: Held-out language-modeling cross-entropy and perplexity on Persian test data.

3.5 HPLT3-PerRef Creation and WDS

Applying the trained BGE-M3 classifier to the Persian portion of HPLT3 produced HPLT3-PerRef, containing 102.37 million documents and approximately 105.9 billion tokens. Overall, 20% of the original documents were removed.

Because WDS relies solely on surface-level features, it may assign high scores to documents that our annotation guidelines classify as low quality. It may also assign lower scores to documents with unusual formatting, even when they contain useful linguistic content.

To investigate the relationship between WDS and our Persian-specific quality classifier, we compare the distribution of WDS scores for the entire Persian subset of HPLT3 with those of documents classified as low quality. This analysis assesses whether the proposed classifier identifies additional low-quality documents beyond those captured by WDS.

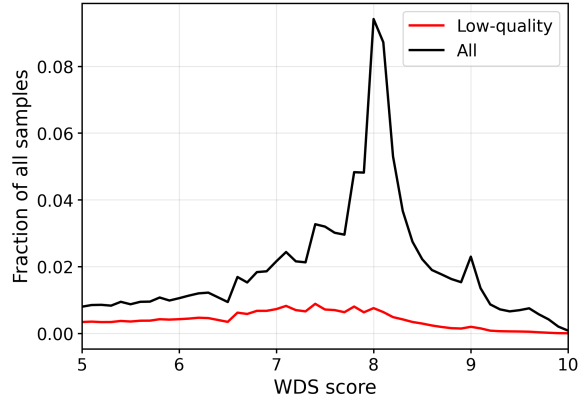


Figure 2: Distribution of WDS scores for the full HPLT3-PerRef and documents classified as low quality.

3.6 WDS analysis

Figure 2 shows the marginal distribution of WDS scores for the entire corpus and low-quality documents. Since the overall corpus is concentrated around a WDS score of approximately 8, most low-quality documents also appear in this region. However, this distribution shows the frequency of WDS scores among low-quality documents, not the probability that a document with a given WDS score is low-quality. Equivalent distributions for the remaining quality scores are in Appendix Figure 4.

Figure 3 shows the proportion of documents classified as low-quality within each WDS score bin. After normalising by the number of documents in each bin, a clear inverse relationship emerges: the proportion of low-quality documents decreases from approximately 42% at WDS = 5 to below 5% for WDS scores above 9.5. We fit a weighted four-parameter logistic (4PL) model to the observed prevalence, which provides a good fit to the empirical distribution ($R^2=0.967$)¹. The fitted curve captures the upper and lower plateaus and the transition region around WDS scores of 7.4–7.9, showing that the likelihood of a document being classified as low-quality decreases smoothly as WDS increases. However, low-quality documents remain present across the entire WDS score range, including the highest-scoring bins. WDS is therefore a useful quality indicator but cannot fully identify all low-quality documents, motivating the use of the proposed classifier as a complementary filtering approach. The corresponding prevalence plots for all quality metrics are in Figure 5.

¹The weighted 4PL model outperformed quadratic ($R^2 = 0.809$) and cubic ($R^2 = 0.904$) alternatives.

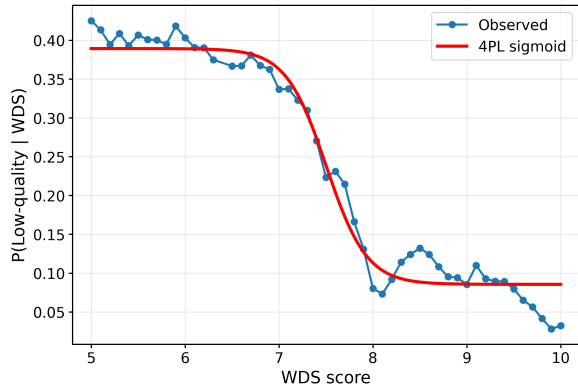


Figure 3: Proportion of documents classified as low quality across WDS score bins.

Pearson correlation analysis over the entire corpus was used to measure the linear relationship between the continuous WDS scores and the classifier’s predicted low-quality probabilities. It further indicates that the proposed classifier captures information complementary to the existing HPLT document quality scores. The overall WDS score shows a moderate negative correlation with the probability of being classified as low-quality ($r=-0.279$). Among individual WDS components, correlations remain generally weak, suggesting that the classifier captures quality-related patterns that are not fully represented by the heuristic scores.

4 Conclusion

We presented HPLT3-PerRef, a refined Persian web corpus constructed from the Persian portion of HPLT3 using a supervised document quality filtering pipeline. The filtering process removed approximately 20% of the original corpus while preserving broad web coverage. Our analysis demonstrated that the proposed classifier complements HPLT3’s Web Document Score (WDS) by identifying additional low-quality documents that are not fully captured by heuristic quality measures.

HPLT3-PerRef contains approximately 105.9 billion tokens across more than 102 million documents, making it one of the largest publicly available refined Persian corpora for general-purpose NLP and language model pretraining. HPLT3-PerRef and its associated resources are publicly released to support reproducible research, available at [GitHub](#), [Hugging Face \(Data\)](#) and [Hugging Face \(Quality classifier\)](#).

Future work includes extending the annotation framework to finer-grained quality categories, in-

corporating additional semantic quality signals, and evaluating the impact of the refined corpus on Persian language model pretraining and downstream tasks.

5 Limitations and Further Work

The quality classifier was trained on a manually annotated subset of Persian HPLT3 documents. Since the annotations were produced by a single annotator, inter-annotator agreement could not be measured, and the labels may reflect subjective quality judgments. Furthermore, the annotated dataset represents only a subset of Persian web content, and the distribution of document types and quality levels within this sample may influence the classifier’s decisions. Under-represented document types or WDS score ranges may therefore be classified less reliably.

As further work, the relationship between the heuristic WDS score and the supervised classifier could be investigated more systematically, including evaluating WDS as a baseline and exploring whether its component weights can be optimized to better reflect the annotated quality judgments. In addition, moving beyond binary quality labels toward multi-dimensional annotations could provide a more nuanced representation of the diverse quality characteristics of web documents and better capture the complexity of web data.

6 Data Disclaimer

The documents contained in HPLT3-PerRef are derived from publicly available web content and may contain opinions, claims, or other statements expressed by the original authors of those documents. Their inclusion in the corpus does not imply that the authors of this work endorse, support, or agree with any such content. The corpus is provided for research purposes, and the views and statements expressed in the included documents should not be attributed to the authors of this work.

Acknowledgments

Alireza Razzaghi received funding from the European Union’s Horizon Europe research and innovation program under the Marie Skłodowska-Curie grant agreement No 101177564—HAIF. Co-funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive

Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

Tomasz Galica received funding from OpenEuroLLM and LLMs4EU projects, under Digital Europe Programme grant agreements No 101195233 and 101198470.

This work was supported by the Research Council of Finland through several projects: FIN-CLARIAH research infrastructure (project 358720, which has also received funding from the European Union– NextGenerationEU instrument), “Mechanisms of Register Variation in Massively Multilingual Web-Scale Corpora” (project 362459), and the “Centre of Excellence Human Diversity through Contacts” (project 374223). We also wish to acknowledge CSC– IT Center for Science Ltd. for providing computational resources

References

- Abolfazl AleAhmad, Hadi Amiri, Ehsan Darrudi, Masoud Rahgozar, and Farhad Oroumchian. 2009. [Hamshahri: A standard persian text collection](#). *Knowledge-Based Systems*, 22(5):382–387.
- Abolfazl AleAhmad, MohammadSadeh Zahedi, Masoud Rahgozar, and Behzad Moshiri. 2016. [irblogs: A standard collection for studying persian bloggers](#). *Computers in Human Behavior*, 57:195–207.
- Mohammad Yasin Ayoubi, Sajjad & Davoodeh. 2021. Persianqa: a dataset for persian question answering. <https://github.com/SajjjadAyobi/PersianQA>.
- Mohammad Yasin Ayoubi, Sajjad & Davoodeh. 2022. Targoman large persian corpus. <https://huggingface.co/datasets/Targoman/TLPC>.
- Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. [M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. [Unsupervised cross-lingual representation learning at scale](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.
- Kasra Darvishi, Newsha Shahbodaghkhan, Zahra Abbasiantaeb, and Saeedeh Momtazi. 2023. [Pquad: A persian question answering dataset](#). *Computer Speech and Language*, 80:101486.
- Ehsan Doostmohammadi, Mohammad Hadi Bokaei, and Hossein Sameti. 2018. [Perkey: A persian news corpus for keyphrase extraction and generation](#). In *2018 9th International Symposium on Telecommunications (IST)*, page 460–465. IEEE.
- Sara Bourbour Hosseinbeigi, Fatemeh Taherinezhad, Hesham Faily, Hamed Baghbani, Fatemeh Nadi, and Mostafa Amiri. 2025. [Matina: A large-scale 73B token Persian text corpus](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9143–9157, Albuquerque, New Mexico. Association for Computational Linguistics.
- Maximilian Idahl, Benedikt Droste, Björn Plüster, and Jan Philipp Harries. 2026. [propella-1: Multi-property document annotation for llm data curation at scale](#). *Preprint*, arXiv:2602.12414.
- Hamzeh Motahari Khansari and Mehrnosh Shamsfard. 2021. [Hmblogs: A big general persian corpus](#). *Preprint*, arXiv:2111.02362.
- Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. [Madlad-400: A multilingual and document-level large audited dataset](#). *Preprint*, arXiv:2309.04662.
- Thuat Nguyen, Chien Van Nguyen, Viet Dac Lai, Hieu Man, Nghia Trung Ngo, Franck Dernoncourt, Ryan A. Rossi, and Thien Huu Nguyen. 2024. [CulturaX: A cleaned, enormous, and multilingual dataset for large language models in 167 languages](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4226–4237, Torino, Italia. ELRA and ICCL.
- Stephan Oepen, Nikolay Arefev, Mikko Aulamo, Marta Bañón, Maja Buljan, Laurie Burchell, Lucas Georges Gabriel Charpentier, Pinzhen Chen, Mariya Fedorova, Ona de Gibert, Barry Haddow, Jan Hajic, Jindrich Helcl, Andrey Kutuzov, Veronika Laippala, Zihao Li, Risto Luukkainen, Bhavitvya Malik, Vladislav Mikhailov, and 13 others. 2025. [HPLT 3.0: Very large-scale multilingual resources for LLM and MT. mono- and bi-lingual data, multilingual evaluation, and pre-trained models](#). *CoRR*, abs/2511.01066.
- Qwen Team. 2026. [Qwen3.6-35B-A3B: Agentic coding power, now open to all](#).
- Behnam Sabeti, Hossein Abedi Firouzjaee, Ali Janalizadeh Choobbasti, S.H.E. Mortazavi Najafabadi,

and Amir Vaheb. 2018. *MirasText: An automatically generated text corpus for Persian*. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Sadra Sabouri, Elnaz Rahmati, Soroush Gooran, and Hossein Sameti. 2024. *naab: A ready-to-use plug-and-play corpus for farsi*. *Journal of Artificial Intelligence Applications and Innovations*, 1(2).

Benoît Sagot and Laurent Romary. 2019. *Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures*. page 8.

Mehran Sarmadi, Morteza Alikhani, Erfan Zinvandi, and Zahra Pourbahman. 2025. *Hakim: Farsi text embedding model*. *Preprint*, arXiv:2505.08435.

Mohammad Shojaei. 2024. Persian document corpus. <https://huggingface.co/datasets/mshojaei77/persian-document-corpus>.

Alexander Wettig, Kyle Lo, Sewon Min, Hannaneh Hajishirzi, Danqi Chen, and Luca Soldaini. 2025. *Organize the web: Constructing domains enhances pre-training data curation*. *Preprint*, arXiv:2502.10341.

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*. *Preprint*, arXiv:2010.11934.

A Appendix

A.1 WDS Score

The WDS quality score is composed of several heuristic subscores. `language_score` measures the proportion of alphabetic characters belonging to the target language, `urls_score` the proportion of URLs, `punctuation_score` the proportion of punctuation characters, `singular_chars_score` the proportion of dominant single characters, `numbers_score` the proportion of numeric characters, `repeated_score` the amount of repeated text segments, `n_long_segments_score` the amount of long text segments, `great_segment_score` the length of the longest text segment, and `informativeness_score` the document’s compression-based informativeness estimated using Zstandard.

Figures 4 and 5 extend the analyses presented in Figures 2 and 3 to the individual WDS heuristic subscores. Figure 4 shows the marginal distribution of each score for the full corpus and for documents classified as low quality. These distributions describe the frequency of low-quality docu-

Score	Pearson’s r	N
WDS score	-0.2790	127,957,098
lang_score	-0.0853	127,957,098
url_score	0.0024	127,957,098
punctuation_score	-0.4186	127,957,098
singular_chars_score	0.0143	127,957,098
numbers_score	-0.0444	127,957,098
repeated_score	-0.2299	127,957,098
informativeness_score	0.1424	127,957,098
n_long_segment_score	0.1002	127,957,098
great_segment_score	-0.1405	127,957,098

Table 3: Pearson correlation between individual quality scores and the target variable.

Corpora	n-token/word	n-document/pages
HPLT3 Per-	157.76B (t)	127.96M (d)
sian		
HPLT3-PerRef	105.9B (t)	102.37M (d)
Matina	72.9B (t)	-
Hakim	58.8B (t)	-
mC4 Persian	52B (t)	54M (p)
Targoman	41B (t)	75M (d)
Madlad-400 Persian	29.2B (t)	58.1M (d)
naab	15B (w)	-
HmBlogs	6.8B (t)	20M (d)
OSCAR Per-	4.3B (t)	-
sian		
ParsBERT	-	4M (d)
MirasText	1.4B (t)	2.8M (d)

Table 4: Persian Corpora size comparison

ments across the score range and should not be interpreted as the probability of a document being low quality given its score. Figure 5, in contrast, shows the proportion of documents classified as low quality within each score bin, obtained by normalising the number of low-quality documents by the total number of documents in the corresponding bin.

A.2 Optuna Hyperparameters Optimization

Hyperparameter optimization was performed using Optuna. The optimization searched over the learning rate, weight decay, warmup ratio, per-device training batch size, and gradient accumulation steps. The learning rate was sampled between 5×10^{-6} and 3×10^{-5} on a logarithmic scale. Weight decay was sampled between 0 and

0.1, while the warmup ratio was sampled between 0 and 0.15. The per-device training batch size was selected from 4 or 8, and the gradient accumulation steps were selected from 4 or 8.

For each trial, the model was initialized from the pretrained model and trained for up to 10 epochs. Early stopping was applied with a patience of 3 epochs. The per-device evaluation batch size was fixed at 16. Evaluation and logging were performed after each epoch, and the best model checkpoint was retained based on recall for class 0 on the development set (eval_recall_0). The optimization objective was therefore to maximize class-0 recall on the development set.

The remaining training settings were kept fixed across trials. These included a random seed of 42 and bfloat16 (BF16) mixed-precision training. The complete hyperparameter configuration selected by Optuna for the final model was: learning rate = 7.79×10^{-6} , weight decay = 0.0508, warmup ratio = 0.00184, per-device training batch size = 8, and gradient accumulation steps = 8. The configuration reported above corresponds to the trial achieving the highest class-0 recall on the development set.

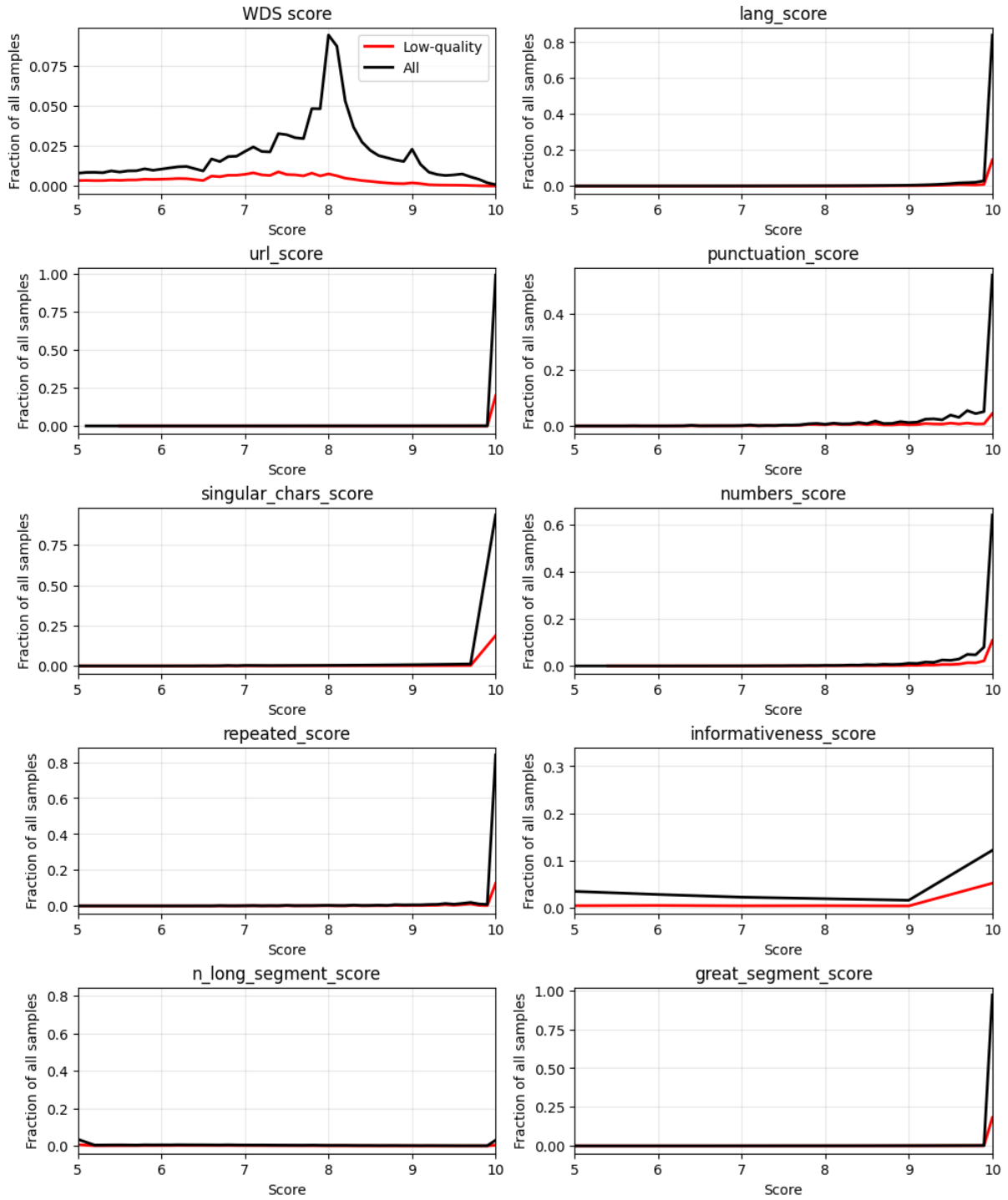


Figure 4: Marginal distributions of the WDS score and its heuristic subscores for the full HPLT3-PerRef corpus and documents classified as low quality. The WDS distribution corresponds to Figure 2 in the main text; the additional panels show the corresponding distributions for the individual heuristic subscores.

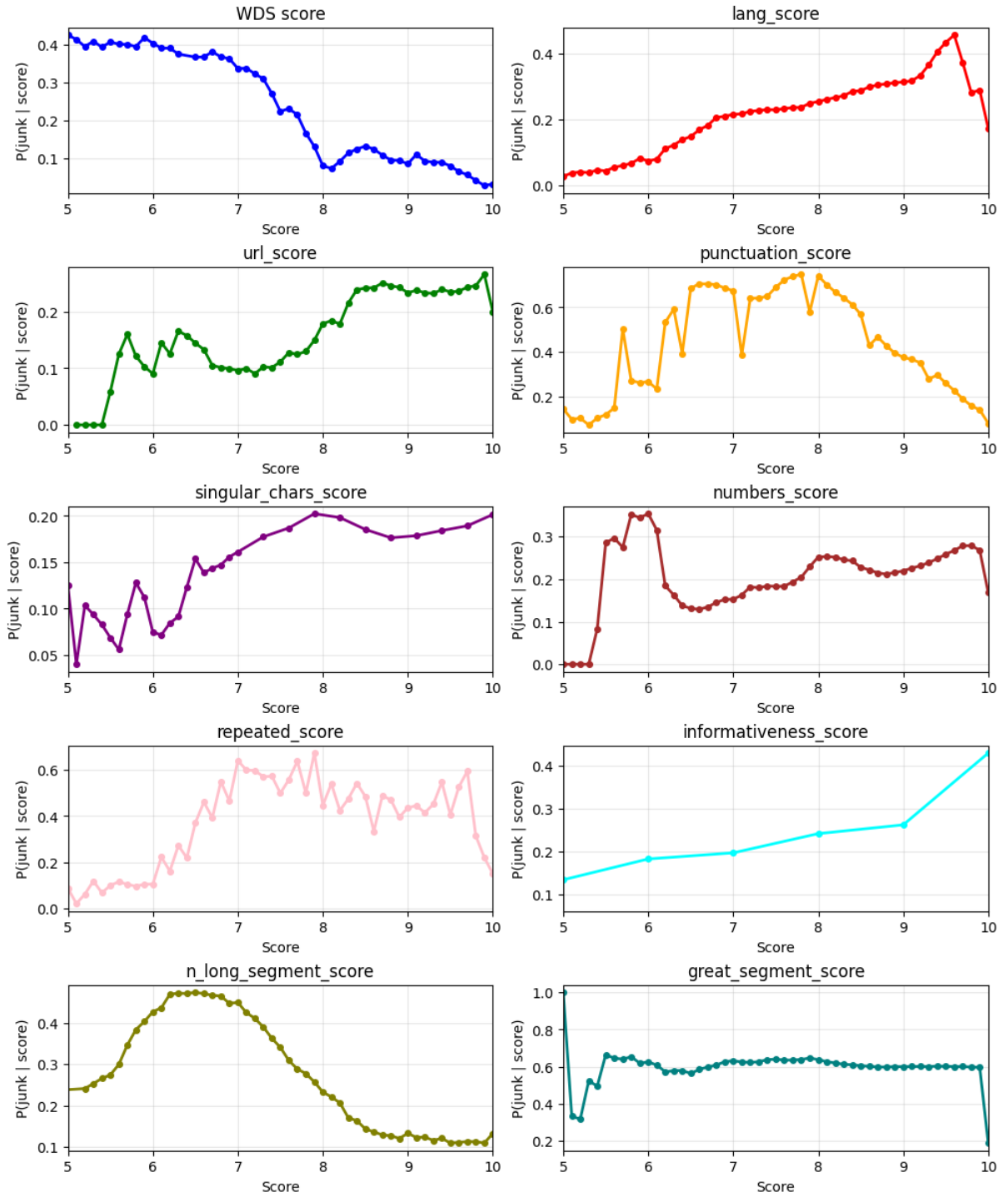


Figure 5: Proportion of documents classified as low quality within bins of the WDS score and its heuristic subscores. The WDS relationship corresponds to Figure 3 in the main text; the additional panels show the corresponding relationships for the individual heuristic subscores.