Multilingual Benchmarking of Main Content Extractors

Anonymous Author(s)

ABSTRACT

Tools designed for extracting main content from web pages require thorough evaluation. However, existing benchmarks disproportionately represent English-language datasets. As a consequence, previous studies have demonstrated that while these extractors are well-optimized for English, their effectiveness diminishes partially or entirely in other languages. This study reproduces and extends recent benchmarks for main content extractors by incorporating multilingual datasets as a key consideration. We analyze extractor performance across five languages - Greek, English, Polish, Russian, and Chinese - highlighting the need to adapt extraction models to linguistic variations to enhance their overall effectiveness.

CCS CONCEPTS

• **Information systems** → **Content analysis and feature selection**; **Test collections**; **Document filtering**.

KEYWORDS

Boilerplate removal, Multilingual settings, Web content extraction

ACM Reference Format:

Anonymous Author(s). 2025. Multilingual Benchmarking of Main Content Extractors. In *Proceedings of The 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2025).* ACM, New York, NY, USA, 5 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

With over 1.93 billion websites in 2023 and nearly 5 billion Internet users worldwide, the web has evolved into a central repository of humanity's accumulated knowledge. Search engines and indexing tools have emerged to organize this vast content by relevance. To enhance user experience, some browsers now incorporate reading modules – both visual and audio – to filter out superfluous elements, even offering accessibility features for visually impaired users.

Isolating the *main content* of web pages allows to extract meaningful content, making the web a valuable resource for high-quality data. However, in addition to the primary content, there is a certain amount of *boilerplate*, including navigation menus, footers, and advertisements. According to Bevendorff et al. [1], main content refers to the central article content (if present) or any non-redundant material on a website. Comments are considered external to this definition as they are neither central content nor site-wide redundant structures such as navigation menus. Alternatively, main content can be defined as the user's expected information when visiting a web page, with everything else categorized as *boilerplate*.

Main content extractors use either heuristic-based or machine learning-based approaches. Heuristic methods rely on predefined rules, analyzing HTML structures like tag and content density, making them efficient but dependent on human expertise, as seen in tools like Readability and Trafilatura [2]. In contrast, machine learning-based extractors use classification models to identify content, leveraging text patterns and structural features, as seen in Boilerpipe [3] and BoilerNet [4]. While these models can be more accurate, they require labeled training data and are computationally expensive. Each approach has trade-offs between efficiency, accuracy, and adaptability.

When evaluating content extractors, studies have found that tool performance varies significantly based on language and HTML structure [5]. Indeed, most extractors are optimized for English, as there exists a wide range of English-language datasets for evaluation, and thus perform well in that language but struggle with multilingual or non-English pages.

In order to consider the multilingual aspect of main content extraction and confirm previous findings in the field, we reproduced the experiments from the web Content Extraction Benchmark (WCEB) [1] in a multilingual setting. First, we discovered inconsistencies in the benchmark datasets, regarding the presence of comments in the gold standard, that can bias current conclusions. Second, we added the multilingual dataset DAnIEL[1] [6] in order to show the impact of language on this task, specifically focusing on majority and under-represented languages such as Greek, English, Polish, Russian and Chinese.

Our findings highlight the need to adapt extraction models to linguistic variations. By incorporating four additional languages to English, we observed structural differences, such as sentence length, which had a notable impact on extractor performance. While these variations occasionally improved results (e.g., in Greek), they more often led to performance declines. Our study also found that web page complexity is not a strong indicator of extraction effectiveness.

2 RELATED WORK

2.1 Heuristic and Machine Learning Extractors

Main content extractors use two kinds of approaches, based on *heuristics* or *machine learning*. Heuristic-based extractors rely on predefined assumptions about the structure of web pages. In order to identify the main content, these approaches often analyze HTML structure, such as tag density, content density, and link presence. They are computationally efficient and require no training data but depend heavily on human expertise to craft these rules. Examples include Readability[2][3], which uses handcrafted rules to optimize extraction for article-like pages, Trafilatura [2], which combines XPath queries with fallbacks and tools such as jusText [7].

[1] https://github.com/rundimeco/waddle/tree/master/corpora/Corpus_daniel_v2.1

[2] https://github.com/masukomi/arc90-readability

[3] https://github.com/mozilla/readability

SIGIR 2025, July 13–18, 2025, Padua, Italia

ACM ISBN 978-1-4503-XXXX-X/2018/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

In contrast, machine learning-based extractors use classification models to locate the main content or the boilerplate (binary classification). These models can leverage features such as text density, tag patterns, or word frequencies. For example, BOILERPIPE [3] uses decision trees on shallow text and structural features, while BOILERNET [4] employs sequence-labeling models based on LSTMs. However, these models are computationally expensive and require labeled datasets for training, which are scarce.

2.2 Web Content Extraction Benchmarks

In Bevendorff et al. [1], 8 datasets were evaluated through 14 main content extractors and 5 HTML-to-text extractors. The datasets were: CETD, CleanEval, CleanPortalEval, Dragnet[4], Google-Trends-2017, L3S-GN1, Readability[5] and Scrapinghub. The main content extractors did their best to ignore the boilerplate. The HTML-to-text conversion tools were used as baselines, their role were to extract all the content from the page, including boilerplate.

The metrics used for evaluation were ROUGE-L [8] and the Levenshtein distance [9]. ROUGE-L evaluates textual similarity by capturing the longest common subsequence between two texts, while respecting word order. The Levenshtein distance measures the minimum number of operations (insertions, deletions and substitutions) required to transform one character string into another, regardless of linguistic structures.

Following their study, a page is considered *complex* if it contains a high proportion of boilerplate content. As the ground truth locates the main content, the complexity c of a web page is related to the expert annotation and it is defined in Equation 1 where T is a multiset of DOM text tokens and $truth(t)$ returns 1 if the token t belongs to the ground truth, otherwise 0.

$$c = 1 - \frac{|\{t \in T : truth(t) = 1\}|}{|T|} \tag{1}$$

According to the results of Bevendorff et al., no single method outperforms all others for extracting the main content of web pages. Heuristic approaches, based on hand-written rules, are generally more robust and efficient than machine learning approaches for complex page extraction tasks. Combining several tools improves overall performance by reducing the variance of results. But the authors mention that the field remains limited by a lack of recent large datasets.

Although the study is effective in terms of evaluation, we will add a few limitations. Firstly, there is indeed a lack of datasets specifically tuned for this task, and there is an over-representation of datasets targeting blogs and press articles genre, which obviously guides the creation and benchmarking of extractors. Similarly, the most widely used datasets are concentrated on the English language.

In our experiments, we propose to select three main content extractors: TRAFILATURA and READABILITY, due to their strong heuristic performance, and BOILERPIPE as the top machine learning extractor. We also included HTML_TEXT[6] for its high recall as an HTML-to-text tool, using it as the sole baseline.

[4]https://github.com/dragnet-org/dragnet
[5]https://github.com/mozilla/readability
[6]https://github.com/TeamHG-Memex/html-text

2.3 Multilingual Evaluation of Extraction Tools

The multilingual study of the main content extractors by Barbaresi and Lejeune [5] compared their performance on the DAnIEL dataset, which comprises web pages in Greek, Chinese, Polish, Russian and English. In particular, they used 6 evaluation metrics. The metrics were *CleanEval*, evaluating similarity by the longest common subsequence, regardless of token order; *VocEval*, comparing vocabularies, while *OccEval* integrated occurrences; *Cosine* and *Euclidean* distances were calculated on the Bag of Words [10] of extracted text and ground truth; *KL-Divergence* measured the divergence between token distributions.

This study concluded that the performance of the web extraction tools varies considerably according to the languages and HTML structures of the web pages. Tools designed primarily for English showed superior results in that language, but inferior performance on multilingual pages or pages written entirely in another language. Although this is highly language-dependent, heuristic approaches such as JUSTEXT and READABILITY were the most stable.

Although the findings of Barbaresi and Lejeune are instructive, they lack the same strong comparison with the widely used existing benchmarks as proposed by Bevendorff et al. [1]. As a consequence, in this paper, we propose to combine the best of each studies to perform a robust evaluation of web content extractors in a multilingual setting, so that conclusive findings can be drawn.

3 BENCHMARK DATASETS

In this section, we detail the modifications we have undertaken to improve experimental conditions, i.e. corpus cleaning by comments removal and introduction of a multilingual dataset.

3.1 Comments Removal

Upon reviewing web pages with poor performance, we observed that comments in blogs and news articles were still present in the ground truth of certain datasets. By default, extractors are generally configured not to extract comments, which can introduce a bias in the actual performance of the extractors. Since the ground truth in the Dragnet dataset was labeled, we were able to quickly remove the comments and retest the extractors. All extractors performed better on the cleaned dataset, with READABILITY benefiting the most. Finally, we compared the size of Dragnet before and after cleaning, based on the number of characters. It turned out that 48% of Dragnet content were comments. The CETD and CleanEval datasets suffered from a similar issue. As the comments were not marked in the ground truth, we preferred to exclude them from the study. Table 1 shows extraction results with ROUGE-L Precision, Recall and F_1 scores evidencing the positive impact of comment removal. The F_1 score clearly improves when the cleaned version of Dragnet is used, exclusively obtained by the increase in Recall. Indeed, while Precision is steadily higher with the original Dragnet, Recall is boosted when comments are removed, thus showing the importance of curated benchmarks.

3.2 Adding DAnIEL dataset

To improve the experimental conditions, we added the DAnIEL dataset – Diverse And Novel Information Extraction from Languages, introduced by Lejeune et al. [6]. DAnIEL was designed for

Table 1: ROUGE-L average comparison between cleaned Dragnet and the original Dragnet.

Model	Original Dragnet			Cleaned Dragnet		
	Precision	Recall	F_1	Precision	Recall	F_1
HTML_TEXT	**0.474**	0.995	**0.604**	0.368	**0.997**	0.501
BOILERPIPE	**0.888**	0.750	0.773	0.852	**0.858**	**0.838**
READABILITY	**0.929**	0.771	0.806	**0.916**	**0.899**	**0.896**
TRAFILATURA	**0.916**	0.834	0.839	0.858	**0.925**	**0.861**

news extraction and monitoring in a multilingual context. The data originates from multilingual news feeds collected from various online sources, covering several languages: Greek, English, Polish, Russian, and Chinese. Table 2 shows the number of web pages for each language in the version of the dataset we accessed[7]. The corpus was filtered using a language detector[8] followed by a secondary manual review to separate it into the five languages.

Table 2: DAnIEL web page distribution grouped by language.

Languages	Number of occurences
Greek	273
English	476
Polish	274
Russian	266
Chinese	400

Figure 1 shows the complexity levels of pages from the DAnIEL dataset and the 8 other datasets included in the web Content Extraction Benchmark (WCEB). Interestingly, every sub-corpus of DAnIEL has a significantly higher level of complexity than all the other datasets included in WCEB proposed by Bevendorff et al. [1].

4 RESULTS

The evaluation of main content extractors was conducted using the ROUGE-L metric, with its Precision, Recall anf F_1 scores. This metric allows for a robust comparison of performance across different models and datasets.

4.1 Language-independent Performances

The language-independent performances of the content extraction models, evaluated on both macro- and micro-average, is summarized in Table 3. Macro-average stands for the mean of each dataset performance, ignoring size, whereas micro-average weights each dataset performance by its size. To ensure a balanced evaluation across linguistic diversity, macro-averaging is employed by grouping classes equally based on their respective languages.

Models such as READABILITY and BOILERPIPE consistently demonstrated high performance across Precision, Recall, and F_1 scores, with READABILITY achieving a macroF_1 score of 0.838 and BOILERPIPE achieving 0.805. TRAFILATURA performs much worse than expected with a macro$F1_1$ score achieving only 0.77, which means

[7] https://github.com/rundimeco/waddle/tree/master/corpora/Corpus_daniel_v2.1
[8] https://pypi.org/project/langdetect/

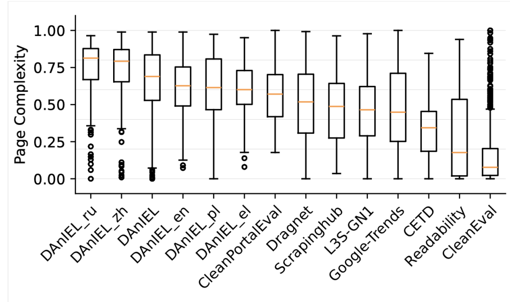

Figure 1: Web page complexity scores per dataset. Boxplots show maximum, minimum, median and standard deviation values. *ru* stands for Russian, *zh* for Chinese, *en* for English, *pl* for Polish and *el* for Greek.

it is less stable on non-English languages. The baseline extractor, HTML_TEXT, had of course very good Recall but poor Precision.Nevertheless, note that micro-averaging less penalizes TRAFILATURA, although globally same tendencies are observed between macro- and micro-averaging.

Table 3: ROUGE-L metrics, Precision, Recall and F_1 over all 13 datasets by macro- and micro-averaging per language.

Macro-averaging per language						
Model	Mean			Median		
	Precision	Recall	F_1	Precision	Recall	F_1
HTML_TEXT	0.300	**0.970**	0.414	0.280	**1.000**	0.412
BOILERPIPE	0.826	0.830	0.805	0.909	0.928	0.898
READABILITY	**0.862**	0.849	**0.838**	**0.935**	0.927	**0.914**
TRAFILATURA	0.749	0.880	0.770	0.828	0.955	0.852
Micro-averaging per language						
Model	Mean			Median		
	Precision	Recall	F_1	Precision	Recall	F_1
HTML_TEXT	0.407	**0.983**	0.531	0.393	**1.000**	0.562
BOILERPIPE	0.869	0.827	0.820	0.965	0.973	0.945
READABILITY	**0.894**	0.839	**0.841**	**0.988**	0.954	**0.953**
TRAFILATURA	0.835	0.883	0.823	0.968	0.964	0.934

4.2 Language-dependent Performances

As detailed in Table 4, the performance of three different extractors varies significantly across languages. In particular, we focused on READABILITY, BOILERPIPE and TRAFILATURA leaving out HTML_TEXT due to lack of space. READABILITY outperforms most models for the vast majority of languages, achieving an F_1 score of 0.962 (Greek), 0.862 (Polish), 0.840 (Russian) and 0.672 (Chinese). The only exception is evidenced by TRAFILATURA for English, with a maximum F_1 score value of 0.883. All models show similar behavior with respect to different languages, showing that Greek and English are the best performing settings, while Chinese poses the greatest challenge.

Table 4: ROUGE-L metrics, Precision, Recall and F_1 over DAnIEL sub-corpora for Readability, Boilerpipe and Trafilatura.

Readability						
Model	Mean			Median		
	Precision	Recall	F_1	Precision	Recall	F_1
Greek	0.969	0.964	**0.962**	0.992	0.986	**0.983**
English	0.912	0.865	0.862	0.987	0.974	**0.972**
Polish	0.875	0.888	**0.862**	0.970	0.981	0.937
Russian	0.855	0.849	**0.840**	0.960	0.945	**0.937**
Chinese	0.688	0.702	**0.672**	0.759	0.762	**0.750**

Boilerpipe						
Model	Mean			Median		
	Precision	Recall	F_1	Precision	Recall	F_1
Greek	0.954	0.975	0.961	0.975	1.000	**0.983**
English	0.893	0.854	0.847	0.968	0.987	0.959
Polish	0.870	0.891	0.861	0.963	0.991	**0.966**
Russian	0.740	0.825	0.750	0.926	0.983	0.925
Chinese	0.664	0.622	0.611	0.711	0.688	0.667

Trafilatura						
Model	Mean			Median		
	Precision	Recall	F_1	Precision	Recall	F_1
Greek	0.833	0.933	0.868	0.934	0.984	0.938
English	0.899	0.911	**0.883**	0.984	0.968	0.956
Polish	0.784	0.886	0.800	0.899	0.989	0.922
Russian	0.729	0.856	0.759	0.841	0.951	0.857
Chinese	0.501	0.836	0.555	0.479	0.889	0.588

4.3 Low ROUGE-L F_1 Web Pages

To highlight performance differences across languages, we measured the proportion of *problematic* web pages in the DAnIEL dataset. A web page is considered problematic when its handling by one or several extractors leads to a ROUGE-L F_1 score below a given threshold. In particular, we tested thresholds from 0.1 to 0.5 with a step of 0.1. This analysis excluded html_text, as this extractor serves only as a baseline. Table 5 presents the proportions of problematic pages below every threshold for the different languages. It reveals that the languages of a web pages have a considerable impact of the extraction. For the worst performance, almost all pages are written in Chinese. The Greek proportion is in line with its global performances.

5 DISCUSSION

In order to strengthen our evaluation, we propose to analyze the correlation between web page complexity and performance scores as well as we draw concluding remarks about feature language agnosticity.

5.1 Extractors Stability

Low performance is not necessarily related to high page complexity. We can verify this by examining the Pearson and Spearman rank correlation measures presented in Table 6. The Pearson correlation coefficient [11] quantifies the linear relationship between two variables, assuming normally distributed data, while the Spearman rank correlation [12] measures the monotonic relationship between ranked variables, making it more robust to non-linear dependencies.

Results when comparing complexity c and ROUGE-L F_1 scores reveal a relatively weak to moderate negative correlation for extractors and a strong negative correlation for html_text. This analysis indicates that Readability is the most stable extractor across all levels of complexity compared to the other extractors studied. Boilerpipe and Trafilatura exhibit equivalent levels of stability. Non surprisingly, html_text decreases in ROUGE-L F_1 score when complexity c increases.

Table 5: Proportion (in percentage) of *problematic* web pages for one or several extractors by language.

$F_1 \leq$ n	n = 0.1	n = 0.2	n = 0.3	n = 0.4	n = 0.5
Greek	4	9	16	27	44
English	8	13	16	28	40
Polish	8	18	27	35	49
Russian	15	30	51	66	78
Chinese	67	87	95	98	99

Table 6: Correlation between Complexity and ROUGE-L F_1.

Extractor	Pearson	Spearman
Boilerpipe	-0.265	-0.357
Readability	**-0.184**	**-0.286**
Trafilatura	-0.294	-0.313
html_text	-0.946	-0.959

5.2 Language Agnosticity of Features

Main content extractors score the analyzed web page blocks to decide if they are part of the main content. For example, Readability uses features such as character count, comma count, and link density. Significant differences in characters or punctuation, such as between English and Chinese, highlight that these features are not optimized for multilingual extraction. Features heavily influence performance in languages that differ greatly from those of the extractor creators. Handcrafted features may be accidentally more effective for certain languages, such as Greek.

6 CONCLUSION

In this article, we provided a robust analysis of the multilingual behavior of main web content extractors. Our study underscores the importance of adapting extraction models to linguistic nuances. With adding four languages, i.e. Greek, Polish, Russian and Chinese, to the dominant English datasets, structural differences appeared. The multilingual settings significantly impacted the extractors, occasionally enhancing performance (e.g., in Greek) but more often reducing it. In addition, we showed that the complexity of a web page does not determine the performance of the extraction task.

REFERENCES

[1] Janek Bevendorff et al. "An Empirical Comparison of Web Content Extraction Algorithms". In: *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '23: The 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. Taipei Taiwan: ACM, July 19, 2023, pp. 2594–2603. ISBN: 978-1-4503-9408-6. DOI: 10.1145/3539618.3591920. URL: https://dl.acm.org/doi/10.1145/3539618.3591920 (visited on 12/20/2024).

[2] Adrien Barbaresi. "Trafilatura: A web scraping library and command-line tool for text discovery and extraction". In: *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations*. 2021, pp. 122–131. URL: https://aclanthology.org/2021.acl-demo.15/ (visited on 01/22/2025).

[3] Christian Kohlschütter, Peter Fankhauser, and Wolfgang Nejdl. "Boilerplate detection using shallow text features". In: *Proceedings of the third ACM international conference on Web search and data mining*. WSDM'10: Third ACM International Conference on Web Search and Data Mining. New York New York USA: ACM, Feb. 4, 2010, pp. 441–450. ISBN: 978-1-60558-889-6. DOI: 10.1145/1718487.1718542. URL: https://dl.acm.org/doi/10.1145/1718487.1718542 (visited on 01/22/2025).

[4] *Boilerplate Removal using a Neural Sequence Labeling Model. - Google Scholar*. URL: https://scholar.google.fr/scholar?hl=en&as_sdt=0%2C5&q=Boilerplate+Removal+using+a+Neural+Sequence+Labeling+Model.&btnG= (visited on 01/22/2025).

[5] Adrien Barbaresi and Gaël Lejeune. "Out-of-the-box and into the ditch? multilingual evaluation of generic text extraction tools". In: *Language Resources and Evaluation Conference (LREC 2020)*. 2020, pp. 5–13. URL: https://hal.science/hal-02732851/ (visited on 01/06/2025).

[6] Gaël Lejeune et al. "DAnIEL: Language Independent Character-Based News Surveillance". In: *Advances in Natural Language Processing*. Ed. by Hitoshi Isahara and Kyoko Kanzaki. Red. by David Hutchison et al. Vol. 7614. Series Title: Lecture Notes in Computer Science. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 64–75. ISBN: 978-3-642-33982-0 978-3-642-33983-7. DOI: 10.1007/978-3-642-33983-7_7. URL: http://link.springer.com/10.1007/978-3-642-33983-7_7 (visited on 01/15/2025).

[7] Jan Pomikálek. "Removing boilerplate and duplicate content from web corpora". In: *Disertacnı práce, Masarykova univerzita, Fakulta informatiky* (2011). URL: https://is.muni.cz/th/45523/fi_d/Dr.Baroni.pdf (visited on 01/22/2025).

[8] Chin-Yew Lin. "Rouge: A package for automatic evaluation of summaries". In: *Text summarization branches out*. 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013.pdf (visited on 01/22/2025).

[9] V. I. Lcvenshtcin. "Binary coors capable or 'correcting deletions, insertions, and reversals". In: *Soviet Physics-Doklady*. Vol. 10. Issue: 8. 1966. URL: http://profs.sci.univr.it/~liptak/ALBioinfo/2012_2013/2011_2012/files/levenshtein66.pdf (visited on 01/22/2025).

[10] Wisam A. Qader, Musa M. Ameen, and Bilal I. Ahmed. "An overview of bag of words; importance, implementation, applications, and challenges". In: *2019 international engineering conference (IEC)*. IEEE, 2019, pp. 200–204. URL: https://ieeexplore.ieee.org/abstract/document/8950616/ (visited on 01/22/2025).

[11] "VII. Note on regression and inheritance in the case of two parents". In: *Proceedings of the Royal Society of London* 58.347 (Dec. 31, 1895), pp. 240–242. ISSN: 0370-1662, 2053-9126. DOI: 10.1098/rspl.1895.0041. URL: https://royalsocietypublishing.org/doi/10.1098/rspl.1895.0041 (visited on 01/22/2025).

[12] Charles Spearman. "The proof and measurement of association between two things". In: *International journal of epidemiology* 39.5 (2010). Publisher: Oxford University Press, pp. 1137–1150. URL: https://academic.oup.com/ije/article-abstract/39/5/1137/806468 (visited on 01/22/2025).