DOI: 10.15514/ISPRAS-2026-38(2)-10

NewsXLM: A Multilingual Dataset and Model for Information Extraction from News Web Pages

[1, 2] P.A. Bedrin, ORCID: 0009-0008-7523-8298 <pbedrin@ispras.ru>
[1] M.I. Varlamov, ORCID: 0000-0002-1083-6210 <varlamov@ispras.ru>
[1, 2] A.K. Yatskov, ORCID: 0000-0002-1312-1675 <yatskov@ispras.ru>

[1] Institute for System Programming of the Russian Academy of Sciences,
25, Alexander Solzhenitsyn st., Moscow, 109004, Russia.

[2] Lomonosov Moscow State University,
GSP-1, Leninskie Gory, Moscow, 119991, Russia.

Abstract. This paper addresses the challenge of automatically extracting attributes from news article web pages across multiple languages. Automatic extraction of structured information from news web pages is crucial for multilingual web mining, aggregation, and analytics applications. Recent neural approaches, while effective on web page extraction datasets in English, are pre-trained on English data, limiting their applicability to other languages. We present the first large-scale multilingual dataset for news web page attribute extraction, containing 29,081 annotated pages from 759 websites across 56 languages. Each page includes DOM-node-linked annotations for up to five key attributes (title, publication date, text, authors, and tags), together with HTML and MHTML sources, English-translated versions, screenshots, and node-level render metadata. We evaluate a variety of open-source extraction methods, including heuristic tools and modern transformer-based models. Specifically, we fine-tune the English pre-trained MarkupLM on both original and English-translated pages, and pre-train a multilingual DOM-LM-based model from scratch on a multilingual news web corpus before fine-tuning it on our dataset. Experimental results show that the multilingual DOM-LM achieves the best overall performance across most attributes and languages without relying on translation, while MarkupLM benefits from translation but remains less consistent across languages. The collected dataset and all trained models are publicly available to support practical use and future research in multilingual web information extraction and downstream applications in the news domain.

Keywords: web data extraction; information extraction; web page dataset; news; multilingual dataset; multilingual model; neural networks.

For citation: Bedrin P.A., Varlamov M.I., Yatskov A.K. NewsXLM: A Multilingual Dataset and Model for Information Extraction from News Web Pages. Trudy ISP RAN/Proc. ISP RAS, vol. 38, issue 2, 2026, pp. 149-164. DOI: 10.15514/ISPRAS-2026-38(2)-10.

NewsXLM: Мультиязычный набор данных и модель для извлечения информации из новостных веб-страниц

[1, 2] П.А. Бедрин, ORCID: 0009-0008-7523-8298 <pbedrin@ispras.ru>
[1] М.И. Варламов, ORCID: 0000-0002-1083-6210 <varlamov@ispras.ru>
[1, 2] А.К. Яцков, ORCID: 0000-0002-1312-1675 <yatskov@ispras.ru>

[1] Институт системного программирования РАН,
Россия, 109004, г. Москва, ул. А. Солженицына, д. 25.

[2] Московский государственный университет имени М.В. Ломоносова,
Россия, 119991, Москва, Ленинские горы, д. 1.

Аннотация. В данной работе рассматривается задача автоматического извлечения атрибутов из новостных веб-страниц на разных языках. Автоматическое извлечение структурированной информации из новостных веб-страниц имеет ключевое значение для мультиязычного веб-майнинга, агрегаторов данных и аналитических приложений. Актуальные нейросетевые подходы, хотя и демонстрируют высокую эффективность на англоязычных наборах данных для задачи извлечения из веб-страниц, предварительно обучены на англоязычных данных, что ограничивает их применимость к другим языкам. Мы представляем первый крупномасштабный мультиязычный набор данных для извлечения атрибутов новостных веб-страниц, включающий 29 081 аннотированные веб-страницы с 759 веб-сайтов на 56 языках. Каждая страница содержит привязанные к DOM-узлам аннотации до пяти ключевых атрибутов новости (заголовок, дата публикации, текст, авторы и теги), а также исходные HTML- и MHTML- файлы, их переведённые на английский язык версии, скриншоты и метаданные рендеринга на уровне узлов. Мы оцениваем ряд открытых методов извлечения данных, включая эвристические инструменты и современные трансформерные модели. В частности, мы дообучаем предобученную англоязычную модель MarkupLM как на оригинальных, так и на переведённых на английский страницах, а также с нуля предобучаем мультиязычную модель на основе DOM-LM на мультиязычном корпусе новостных веб-страниц с последующим дообучением на нашем наборе данных. Экспериментальная оценка показывает, что мультиязычный DOM-LM демонстрирует лучшее общее качество по большинству атрибутов и языков без использования машинного перевода, тогда как MarkupLM выигрывает от перевода, но является менее стабильным в мультиязычном сценарии. Представленный набор данных и все обученные модели опубликованы для практического использования и будущих исследований в области мультиязычного извлечения информации из Интернета и связанных приложений в новостном домене.

Ключевые слова: извлечение веб-данных; извлечение информации; набор данных веб-страниц; новости; мультиязычный набор данных; мультиязычная модель; нейронные сети

Для цитирования: Бедрин П.А., Варламов М.И., Яцков А.К. NewsXLM: Мультиязычный набор данных и модель для извлечения информации из новостных веб-страниц. Труды ИСП РАН, том 38, вып. 2, 2026 г., стр. 149–164 (на английском языке). DOI: 10.15514/ISPRAS–2026–38(2)–10.

1. Introduction

The rapid expansion of information on the Internet, which encompasses a vast amount of useful knowledge, underscores the importance of automatically extracting structured content from web documents to support a wide range of downstream applications. However, most web content is presented in a semi-structured HTML format that integrates natural language text with markup elements. Automatically transforming this heterogeneous content into structured knowledge presents a fundamental challenge in the Information Extraction (IE) task [1].

Our work focuses on information extraction from news web pages, each representing a single news article. The objective is to automatically identify and extract text spans corresponding to a predefined set of attributes from each page. Automatic attribute extraction from news web pages is crucial for various real-world applications, including large-scale news aggregation and analytics, real-time media monitoring and reputation management, knowledge base construction, and more.

Despite their semi-structured appearance, HTML documents are difficult to process automatically at scale. Websites may visually present the same semantic attributes in widely diverse layouts under distinct DOM paths. Manual or wrapper-based approaches require developing site-specific extractors that are expensive to maintain and quickly become obsolete as page layouts change [2]. Therefore, there is a strong need for scalable and robust methods that can generalize across languages and sites.

Recent studies have demonstrated that transformer-based models pre-trained on large unlabeled web corpora achieve state-of-the-art performance on the downstream task of automatic web page attribute extraction. Architectures such as MarkupLM [3] and DOM-LM [4] jointly encode textual and structural information from HTML and have shown strong results on standard benchmarks.

However, several limitations remain. First, most SOTA models are initialized from English LM (e.g., RoBERTa [5]) and further pre-trained and evaluated on English-only datasets such as CommonCrawl [6] and SWDE [7]. Second, even self-supervised pre-trained models still require a substantial number of labeled pages in the fine-tuning corpus to adapt to a specific domain [8]. Third, existing benchmark datasets for news attribute extraction are typically small-scale and confined to a single language. This reliance on monolingual data significantly hinders model performance and application in real-world, diverse multilingual news domain.

To address the above problems, we introduce the first large-scale multilingual dataset for news web page attribute extraction, comprising nearly 30,000 annotated pages collected from diverse news sites across multiple languages (Fig. 1). As shown in Fig. 2, each page in the dataset encompasses:

- *Attribute annotations.* Ground-truth containing the extracted text and the corresponding node's XPath for five key news attributes: title, publication date, text, authors, and tags.
- *HTML sources.* The original HTML along with its English-translated version.
- *Visual and state features.* MHTML snapshots preserving the rendered state of the page, and visual features, including full-page screenshots and DOM elements metadata, such as style and coordinates.

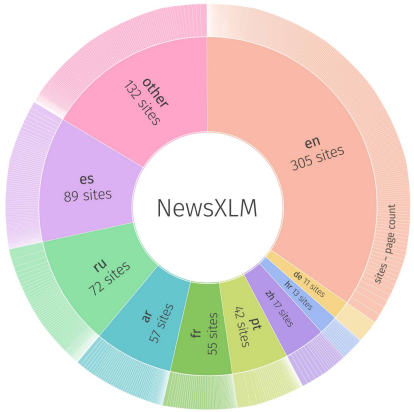

Fig. 1. Data distribution across languages in the NewsXLM dataset.

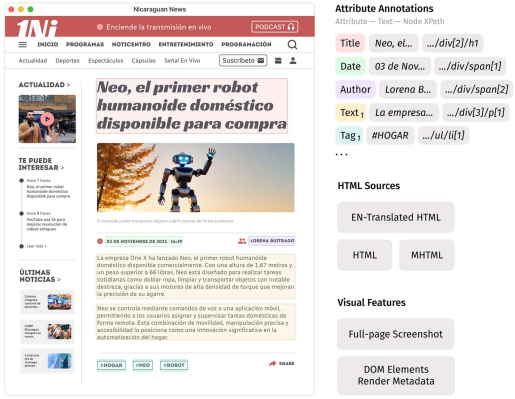

Fig. 2. The NewsXLM dataset components.

Based on this resource, we:

- conduct an evaluation of publicly available modern transformer-based models and open-source tools on both existing news datasets and the newly introduced corpus;

- develop a multilingual model for news attribute extraction built upon the DOM-LM architecture. It achieves the highest extraction performance on most attributes across languages in our dataset, demonstrating its effectiveness in a multilingual scenario.

We refer to our model and dataset as **NewsXLM**.

2. Related work

In this section, we review existing publicly available datasets for structured web data extraction, along with modern solutions for this task, with a particular focus on BERT-like transformer models.

2.1 Datasets

The Structured Web Data Extraction (SWDE) dataset [7] (2011) is the standard benchmark for training and evaluating models in the structured web data extraction task. It consists of over 124,000 web pages collected from 80 websites across 8 domains (e.g., *book* and *restaurant*). For each domain (or *vertical*), the task is to extract 3 to 5 predefined attributes. While SWDE is widely used, it suffers from two limitations: it is monolingual (English-only), and the attribute values are stored separately as plain text, which lacks the DOM-node linkage required by structural IE methods.

The limitations of SWDE have prompted an augmented version [9] (2019), which extended the annotations for 3/8 verticals: *movie, university, NBAPlayer*. The researchers selected 21 sites and significantly expanded the number of annotations. While the original SWDE provided only 4,480 annotations for an average of three attributes on those sites, the extension extracted an average of 41,000 items for 36 attributes.

Another dataset addressing web data extraction is the Klarna Product Page [10] (2021). It comprises 51,701 e-commerce product pages collected from 8,175 websites. The corpus is relevant for multilingual tasks, as it is distributed across eight European languages. For each page, the authors collected a MHTML and screenshot, alongside five labeled attributes relevant to e-commerce.

The CoVA dataset (2022) [11] specifically designed for approaches relying on visual page analysis rather than HTML code. The dataset contains 7,740 pages from 408 e-commerce websites in various languages. Each page in the dataset includes screenshot coupled with detailed node characteristics, such as boundaries and text length. The dataset aims at extracting three e-commerce attributes.

Another e-commerce-focused dataset is the Zyte Product Extraction Benchmark [12] (2021). It is considerably smaller than previous ones, containing only 140 pages from various websites with five labeled attributes. It is primarily designed for quality assessment.

The task of Main Content Extraction, also known as Boilerplate Removal, is closely related to the news domain. It involves extracting the core article text while removing surrounding template elements or advertisements. However, only a limited number of datasets exist for this task [14]. For example, the Zyte benchmark [13] (2020) for article text extraction includes 181 pages, and the collection is approximately 90% English.

Finally, only a limited number of datasets exist in the news domain. They are monolingual and are an order of magnitude smaller than the collections mentioned above.

A dataset of Russian-language news web pages [15] (2022) consists of 722 news pages from 112 websites, with attributes manually annotated per page using Label Studio [16]. Nine attributes are provided (*title, subtitle, publication and modification dates, text, tags, category, author, and source*). The annotations include the labeled text and the XPath of the corresponding nodes.

Yet another benchmark dataset [17] was presented in 2024 by the authors of Newspaper4k [18], an open-source library for news attribute extraction. This collection consists of 424 pages (two pages per site) sampled from 212 websites. It features four labeled news attributes: *title, text, publication date, and author*. The corpus is English, except for three pages.

2.2 Extraction Methods

Automated methods allow to extract attributes from any website within a specific domain without having to manually create a rule-based wrapper for each site. These methods can be broadly classified into two groups: heuristic tools and approaches based on neural network models.

Focusing on the first group, there exist several open-source heuristic-based libraries that extract attributes from a given HTML page by relying on typical XPaths, meta tags, and other structural heuristics. Notable tools that cover multiple news attributes and exhibit strong performance on the Zyte benchmark are Newspaper4k [18], Newsplease [19], and Trafilatura [20].

Among the neural network models, we focus on BERT-based transformer models, which are currently among the most promising approaches in our task [4].

MarkupLM [3] (2022) is a pre-trained model specifically designed for document understanding tasks that utilize markup languages, such as HTML, where text and markup information are jointly pre-trained. Text is encoded by a RoBERTa [5] model, while the relationships between document elements are defined by the XPath expressions of DOM nodes. MarkupLM was pre-trained on 24M English web pages from CommonCrawl dataset with three strategies: Masked Markup Language Modeling, Node Relation Prediction and Title Page Matching. The pre-trained MarkupLM model has been released for fine-tuning by the developers, with a focus on two specific downstream tasks: Information Extraction and Reading Comprehension. The information extraction task is formulated as token classification into $n+1$ classes (where n is the number of extracted attributes, additional class is *None*).

DOM-LM [4] (2022) is also RoBERTa-based pre-trained model, but unlike MarkupLM, DOM-LM considers various structural features of nodes instead of relying on XPath expressions. The authors propose the DOM Tree Processor algorithm for splitting DOM tree into subtrees that preserve local context. The textual representation of each DOM node is defined by the concatenation of its HTML tag, the tag's HTML attributes, and the node's textual content. Node depth and index, parent node index and some other characteristics are used as DOM position features. DOM-LM was pre-trained on over 120,000 English web pages from SWDE using an adapted Masked Language Modeling strategy. An implementation for the pre-training stage is available on GitHub [21].

Structor [22] (2023) is based on MarkupLM and incorporates site-level information and attribute patterns by regular expressions for each DOM node. The authors retrive a node in the same position from another DOM tree on the same website and then incorporate it into the input sequence, approximating character patterns of DOM nodes using regular expressions and integrating them into the neural networks logits.

WebLM [23] (2024) is a multimodal pre-trained network designed to address the limitations of solely modeling text and structure modalities of HTML. It integrates the hierarchical structure of document screenshots to enhance the understanding of markup-language-based documents. Specifically, the bounding box and spatial location on the page are encoded for each node, linking the structural and visual information.

Table 1 summarizes comparative results of existing models on the SWDE, reporting *page-level* F1 scores as cited in the corresponding papers. Extraction for a page is considered successful if at least one of the model's predictions matches the ground truth. For each SWDE vertical, models were fine-tuned on $k=2$ and $k=5$ random sites and then evaluated on the remaining $10 - k$. The ten training splits were constructed by cyclically shifting the list of sites.

Table 1. Extraction performance of existing models on SWDE (F1-score).

	MarkupLM	DOM-LM	Structor	WebLM
$k = 2$	91.29	91.70	92.12	93.17
$k = 5$	95.89	95.70	96.08	96.78

2.3 Conclusion

Despite the high practical relevance of news web page attribute extraction, only a few small, predominantly monolingual datasets currently exist. To enable the training and comprehensive evaluation of modern extraction models and tools in real-world, large-scale multilingual scenarios, we decided to construct a new dataset. It encompasses a diverse collection of websites spanning multiple languages, ensuring broad coverage of layout patterns and page structures. The dataset is designed to provide a rich set of features compatible with various model architectures.

Regarding the extraction methods, we observe that state-of-the-art approaches achieve strong performance on existing benchmarks. In this work, we focus on MarkupLM and DOM-LM, as they combine competitive extraction quality with publicly available pre-trained weights or implementations, making them suitable for reproducible research. Specifically, we evaluate *(1)* the English pre-trained MarkupLM, which we fine-tune both on original and English-translated pages, and *(2)* a DOM-LM based model that we pre-train and fine-tune from scratch for the multilingual news attribute extraction task. We will also train a multilingual text-based baseline, XLM-RoBERTa, and evaluate the performance of the mentioned heuristic tools.

3. Dataset Construction

Manual page-by-page annotation, as adopted in some prior works [10][15], requires an extremely long time and human labor for constructing large-scale datasets such as ours. Therefore, we adopted a site-level markup strategy, like that used in SWDE. We utilized the Web Scraper GUI tool [24] to interactively create website wrappers, referred to as sitemaps.

Each sitemap defines a set of CSS selectors or XPaths that specify *(1)* the paths to news article links from multi-record pages and *(2)* the paths to the target attributes to be extracted from article pages. This site-level approach significantly improved the efficiency of the annotation process, enabling the collection of a large-scale multilingual dataset within a reasonable timeframe.

We curated a list of world news publishers for data collection, encompassing 784 unique websites.

Our team of human annotators manually created a sitemap for each website. They were instructed to label the five most common and important news attributes, namely *title, publication date, article text, authors,* and *tags*, by clicking them directly on the rendered page within Web Scraper. All annotators followed a comprehensive guideline that included attribute definitions and labeling examples. The key rules were as follows:

- Avoid unnecessary text: attribute selectors should locate only the relevant content.
- Use semantic and generalizable selectors: avoid positional or index-based selectors (e.g., :nth-child(3), div[1]), use stable features such as class names or other semantic attributes.
- If an attribute appears in inconsistent locations across pages of a site and a universal selector cannot be defined, the attribute is excluded from labeling on that site.
- No overlapping selectors: each attribute must have a distinct selector.
- Exclude duplicates: if an attribute occurs multiple times, only the first primary visible instance is selected.

To collect target news links and download corresponding pages, we developed a rendering pipeline based on the DrissionPage [25] library. It is a modern Chromium automation tool capable of bypassing certain click-based captchas. During the rendering process, both HTML and MHTML versions of each page were saved. We limited the collection to up to 50 news articles per website.

Attribute texts were then extracted from the downloaded HTML using the sitemap selectors, followed by a multi-step data validation and error correction procedure. We identified problematic cases where selectors returned no nodes or produced outputs violating heuristic constraints, and each of these cases was manually inspected and fixed.

Firstly, we resolved the issues related to page loading:

- If a page failed to load completely, it was re-downloaded.
- If access was restricted by region, we used a set of proxies to download it.
- Sites with complex captchas, unavailable pages, or redirects were excluded.

Next, we fixed problems with the correctness of selectors:

- If a selector returned no text, it was corrected.
- If multiple nodes were extracted for attributes that should be unique (*title* or *publication date*), the selector was refined.

For each page, we stored its URL, HTML, MHTML, and per-node annotations, including the full node XPath, extracted text, and corresponding label.

To identify the language of each page, we used the Langdetect library [26]. We also applied Googletrans [27], which implements the Google Translate API, as a fallback when Langdetect returned "unknown". The detector was executed for every page, using the extracted *text* attribute or, when unavailable, the entire textual content of the page. Although some websites contained pages in multiple languages, occasional detection errors occurred. To reduce noise, we filtered out pages written in languages that represented less than 10% of all website pages.

Translation was performed using the Googletrans. Texts from page nodes were grouped into batches, translated asynchronously, and then inserted back into the corresponding nodes. As a result, each non-English page in our dataset has an English-translated counterpart.

We rendered all collected MHTML files to capture full-page screenshots and utilized Klarna's Web Traversal Library [28] to extract metadata of rendered elements. For each visible DOM node, we collected its coordinates, dimensions, style properties, and other characteristics. This rich metadata can be aligned with the screenshots, providing a valuable resource for visual and multimodal IE methods.

In total, the final dataset comprises 29,081 news article pages collected from 759 websites across 56 languages, featuring 626,460 labeled nodes.

The distribution of sites by languages is illustrated in Fig. 1. The compositional components of the dataset are detailed in Fig. 2. The percentage of pages containing each attribute, averaged across languages, is shown in Fig. 3. The full per-language statistics will be published in the dataset repository.

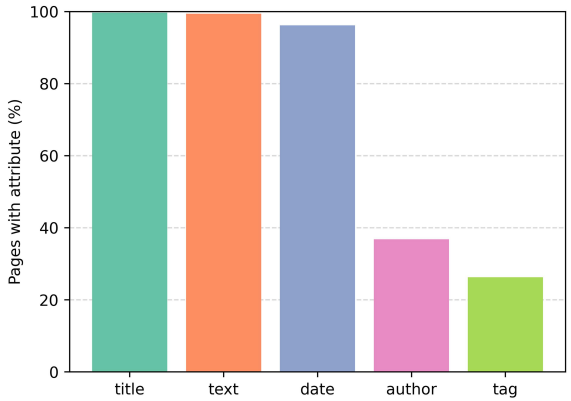

Fig. 3. Attribute coverage (%) averaged across languages in the NewsXLM dataset.

4. Models Training

4.1 MarkupLM as a Monolingual Baseline

As noted earlier, MarkupLM is initialized from the English RoBERTa model, pre-trained on 24M English web pages, and uses the RoBERTa tokenizer for text encoding. It is the only publicly available model for our web extraction task with publicly released pre-trained weights suitable for direct fine-tuning.

MarkupLM uses a byte-level Byte-Pair Encoding (BBPE) [29] tokenizer, which processes text as byte sequences rather than characters. Therefore, despite the small size of the base dictionary (256 bytes), this design allows the model to represent any character sequence without out-of-vocabulary tokens.

We hypothesized that this ability to encode arbitrary characters, combined with awareness of HTML markup, would enable MarkupLM to perform effectively on multilingual data despite its English-only pre-training. Therefore, in our experiments we fine-tune MarkupLM in two configurations: on the original multilingual pages (denoted *MarkupLM*) and on the English-translated pages (denoted *MarkupLM-Translated*) to compare its performance under multilingual conditions.

4.1.1 Data Preprocessing

We prepared data for MarkupLM following the methodology described by its authors. A DOM tree was built from each HTML page after cleaning unnecessary tags (e.g., <*script*>), extra whitespace, and control characters. In training mode, nodes were labeled according to the dataset annotations. Each tree was serialized using preorder traversal into a list of tuples containing {node text, node XPath, node label}.

Texts were tokenized, and each token was linked to its label and to the sequence of tokenized XPath tags and subscripts corresponding to token's node. Tokenized pages were then split into intersecting chunks of length *max_seq_length*, with a fixed stride *doc_stride*. We set max_seq_length = 512, which is the maximum input length of MarkupLM, and doc_stride = 170 (approximately 1/3 of the sequence length, following the authors).

The model's objective is token classification. In MarkupLM, the first token of each node is used for prediction, and the final node-level prediction is obtained by averaging the model's predictions for this token across all chunks containing that node.

4.2 DOM-LM-Based Multilingual Model

Modern BERT-like models can be extended to multilingual settings by initializing from a multilingual language model, using a multilingual tokenizer, and training on multilingual data.

We selected DOM-LM as the base architecture because it achieves performance comparable to MarkupLM on SWDE while requiring significantly less pre-training data and have publicly available pre-training implementation that facilitates reproducibility.

We adapted the implementation of DOM-LM pre-training to our multilingual news domain and subsequently fine-tuned the resulting model on the downstream news attribute extraction task. The resulting multilingual model we term *NewsXLM* further in the paper.

4.2.1 Pre-Training

To obtain a multilingual DOM-LM, we initialized its weights from XLM-RoBERTa [30] instead of the English RoBERTa model as in the original paper. XLM-RoBERTa was trained on 2.5 TB of text covering 100 languages from the CommonCrawl dataset.

DOM-LM utilizes a self-supervised pre-training strategy that extends the traditional Masked Language Model (MLM) objective by incorporating the masking of entire nodes and their attributes,

compelling the model to learn representations that capture both textual semantics and tree-level structural context.

DOM-LM pre-training was performed on both our multilingual news dataset and a one-day sample of 37,473 news pages of different languages from the CommonCrawl-News dataset [31].

4.2.2 Data Preprocessing

The preprocessing pipeline for DOM-LM followed the same cleaning procedure as for MarkupLM. Each cleaned DOM tree was split into subtrees designed to preserve local context, using the DOM-LM algorithm. Subtrees were serialized in preorder traversal.

For every node, textual features (a concatenation of tag, attributes, and text) and structural features were generated. In the algorithm, the chunk (subtree) size *M* was set to 512 tokens and the stride *S* to 128.

Each tokenized node begins with a special BOS token. During fine-tuning, this token is used for both training and prediction. We assigned an ignore index to the remaining tokens. Final node-level predictions are obtained by averaging the BOS token predictions across all subtrees containing the corresponding node, analogous to the MarkupLM.

4.3 Non-Structural Multilingual Baseline

For comparison, we also trained a multilingual non-structural baseline based on XLM-RoBERTa to evaluate extraction quality without structural markup features. Page chunking followed the same subtree-based approach as DOM-LM. Only node text was used, without tags and attributes, allowing us to assess the performance of a purely text-based multilingual model.

5. Experimental Setup

In this section, we outline the evaluation metric for the subsequent experiments and the parameters used for pre-training and fine-tuning the models.

5.1 Evaluation Metric

Since our evaluation covers both node classification models and heuristic text extractors, we assess performance at the level of extracted texts and adopt the evaluation metrics proposed in [15], which are based on Zyte's benchmarks for article and product extraction.

For each news page, the attribute values extracted by automatic methods are compared with the ground truth values. Based on this, True Positives (TP), False Positives (FP) and False Negatives (FN) are computed, and the $F1$-score is then calculated at the language-level for each attribute.

Different matching strategies were applied depending on the attributes:

- *Titles* and *texts* were compared as bags of 4-grams over all alphanumeric sequences, with whitespace and punctuation removed. Compared to using single words, this approach reduces false matches with irrelevant text and penalizes repeated extractions by taking n-gram frequencies into account [13].
- *Authors* and *tags* were compared as sets of individual values, with lowercase and strip normalization applied.
- For *dates*, a parsing pipeline was employed that combines Python's standard *datetime* parsers with Zyte's *dateparser* [32] library for complex cases. If the ground truth date lacked a particular component (e.g., time), the same component was omitted from the extracted date to ensure fair evaluation. When a date could be parsed ambiguously (e.g., DMY/MDY/YMD), a match was considered valid if any of the interpretations aligned with the ground truth. If the parsers failed, dates were compared as raw strings.

For the *title* and *date* attributes, we retained only the single highest-scoring node predicted by the models, as these attributes are expected to have a single value.

5.2 Settings

We pre-trained the NewsXLM model almost according to the DOM-LM paper. The model was initialized with pre-trained XLM-RoBERTa-Base weights using Transformers library [33]. Then it was pre-trained on raw HTML documents without labels for 5 epochs using a batch size of 24, a linear learning rate scheduler with a maximum learning rate of $1e^{-4}$ and warm up for the first half epoch.

For fine-tuning, all models were trained for one epoch using a batch size of 32, the AdamW optimizer with a linear scheduler. The scheduler had a learning rate of $3e^{-5}$, a weight decay of 0, and an ϵ value of $1e^{-8}$.

The models were trained on one NVIDIA A100 80GB GPU. It took approximately 5 days to pre-train NewsXLM. The results were obtained using the equipment of the Shared Research Facility «Shared Research Center of the Ivannikov Institute for System Programming of the Russian Academy of Sciences (SRC ISP RAS).

We filtered datasets for our experiments by removing pages where the ground truth contained more than one node for the *title* and *date* attributes. Then we limited the number of pages to a maximum of ten pages per website to keep the training time manageable. As a result, we use 6,890 pages from the original NewsXLM dataset.

6. Experiments

In this section, we report the experiments conducted with the described models and open-source tools. We performed two main sets of experiments: multilingual evaluation on the NewsXLM dataset, and cross-dataset evaluation on external news domain datasets we mentioned earlier.

6.1 Multilingual Evaluation

This primary experiment evaluates the extraction quality of all described methods on our collected multilingual dataset, demonstrating real-world performance when the page language is arbitrary and the model is trained on all available multilingual data.

The models were assessed using 3-fold cross-validation with stratification by languages. The list of websites was divided into three parts: two used for training and one for testing. Metrics were computed per language, and final results were average across three folds. This CV setup ensures uniform data utilization for reliable model assessment.

We adhere to a zero-shot scenario by ensuring that the sites used for training and evaluation do not intersect. This is done to assess the models' ability to learn site-agnostic knowledge which can be transferred to unseen websites in domain. The same splits were used for all methods.

Results averaged across languages are presented in Table 2. Additionally, we report the *Win Count* for each method: the number of languages in which the method achieves the best performance on a majority of evaluated attributes (Table 3).

6.1.1 Discussion

Across most attributes, NewsXLM achieves performance comparable or better than other methods. It also has the highest Win Count.

MarkupLM-Translated delivers comparable or superior F1-scores for several attributes, but requires the computational overhead to page pre-translation. This performance is also likely attributed to its significantly larger pre-training corpus.

Table 2. Results for NewsXLM dataset, averaged across languages. F1-score.

Method	*Title*	*Date*	*Text*	*Author*	*Tag*	*Macro-Avg*
NewsXLM	0.91	**0.94**	**0.92**	**0.55**	0.47	**0.76**
XLM-RoBERTa	0.80	0.84	0.90	0.37	0.44	0.67
MarkupLM-Translated	**0.96**	0.91	**0.92**	0.40	**0.50**	0.74
MarkupLM	0.94	0.86	0.88	0.31	0.40	0.68
Trafilatura	0.83	0.17	0.88	0.27	0.10	0.45
Newsplease	0.86	0.24	0.66	0.19	-	0.49
Newspaper4k	0.83	0.16	0.73	0.19	-	0.48

Table 3. Win counts on NewsXLM dataset.

Method	*Win Count*
NewsXLM	**33**
XLM-RoBERTa	4
MarkupLM-Translated	21
MarkupLM	10
Trafilatura	4
Newsplease	0
Newspaper4k	0

For all attributes, translation improves the extraction quality of MarkupLM. However, on the original multilingual pages the results remain competitive, indicating that MarkupLM's byte-level tokenizer and structural awareness provide a degree of cross-lingual transfer despite monolingual pre-training.

On average, the XLM-RoBERTa baseline exhibits the lowest averaged score among the evaluated transformers. However, its score on the *text* and *tag* is closely comparable to that of the structural NewsXLM. This suggests the importance of textual features for these attributes.

The heuristic tools exhibit the lowest overall quality. Newsplease and Newspaper4k do not support tags extraction. Date matching here remains particularly challenging: tools often extract dates from meta-tags, causing mismatches with ground truth labeled from visible nodes. This indicates a need for future research, including exploration of language models for the date parsing task.

6.2 External Datasets Evaluation

To assess cross-dataset generalization, we fine-tuned all models exclusively on the NewsXLM dataset and evaluated them, along with heuristic tools, on three external datasets:

- a Russian-language dataset [15],
- the Newspaper dataset [17],
- the Zyte Article Extraction benchmark [13].

Results for the first two datasets with multiple news article attributes are shown in Table 4 and Table 5, respectively.

We evaluated methods on the Zyte benchmark to compare them with commercial solutions. The metrics for proprietary services, Zyte Automatic Extraction (Zyte AE) and Diffbot, are taken from Zyte's benchmark report. These results are presented in Table 6.

Table 4. Results for Russian-language News Dataset. F1-score.

Method	*Title*	*Date*	*Text*	*Author*	*Tag*	*Macro-Avg*
NewsXLM	**0.98**	0.91	**0.83**	0.66	**0.65**	**0.81**
XLM-RoBERTa	0.89	0.88	0.80	**0.71**	0.60	0.78
MarkupLM-Translated	**0.98**	**0.94**	**0.83**	0.56	0.62	0.79
MarkupLM	0.97	0.91	0.79	0.23	0.58	0.70
Trafilatura	0.97	0.23	0.81	0.44	0.22	0.53
Newsplease	0.96	0.42	0.71	0.34	-	0.61
Newspaper4k	0.90	0.30	0.79	0.34	-	0.58

Table 5. Results for Newspaper News Dataset. F1-score.

Method	*Title*	*Date*	*Text*	*Author*	*Macro-Avg*
NewsXLM	0.97	0.67	0.91	0.66	0.80
XLM-RoBERTa	0.89	0.66	0.91	0.66	0.78
MarkupLM-Translated	**0.98**	**0.70**	**0.92**	0.71	**0.83**
MarkupLM	0.97	**0.70**	0.90	**0.74**	**0.83**
Trafilatura	0.97	0.76	0.93	0.53	0.80
Newsplease	0.83	0.63	0.91	0.38	0.69
Newspaper4k	0.83	0.68	0.91	0.38	0.70

Table 6. Results for Zyte AE Dataset.

Method	*F1*	*Precision*	*Recall*
Zyte AE	**0.97**	**0.98**	0.96
Diffbot	0.95	0.96	0.94
NewsXLM	0.94	0.95	0.93
XLM-RoBERTa	0.94	0.96	0.93
MarkupLM-Translated	0.96	0.95	0.96
MarkupLM	0.95	0.96	0.94
Trafilatura	0.96	0.94	**0.98**
Newsplease	0.95	0.97	0.94
Newspaper4k	0.93	0.96	0.89

6.2.1 Discussion

On both the Russian and Newspaper datasets, all models achieve comparable high score. On the Russian dataset, NewsXLM performs comparably or better than other methods across most attributes.

On the predominantly English Newspaper dataset, the MarkupLM models demonstrate higher scores, consistent with the model's English-only massive pre-training and RoBERTa tokenization.

Notably, across both datasets, XLM-RoBERTa shows its largest performance drop on the *title*, underscoring the importance of structural features for accurate extraction of this attribute.

In Zyte benchmark, all methods exhibit a high F1-score. Most model errors occur on non-news articles (e.g., blogs, forums). These pages often contain long comments that the model treats as text, atypically many subheadings, unfamiliar elements (e.g., tables) that are labeled as text in ground truth.

7. Conclusion

In this paper, we addressed the task of automatic attribute extraction from multilingual news web pages, introducing the first large-scale dataset and model designed specifically for this purpose.

We developed and released a multilingual dataset comprising 29,081 annotated pages from 759 websites across 56 languages, labeled with five main news attributes: *title*, *publication date*, *text*, *authors*, and *tags*. The dataset also includes page sources, English-translated versions, and visual features such as screenshots and page render metadata.

We adapted and fine-tuned modern open-source transformer-based models, including the English MarkupLM and a newly pre-trained multilingual DOM-LM model, referred to as *NewsXLM*. Experimental results show that translating pages into English improves the performance of MarkupLM, yet NewsXLM achieves higher scores across most languages and attributes, confirming the effectiveness of structure-aware multilingual pre-training.

The NewsXLM model can be further improved by expanding the volume, linguistic diversity, and balance of the pre-training corpus, and increasing the number of labeled sites per language. Future work may also explore implementing new or existing models that incorporate visual features of web pages, a direction for which our dataset is well suited.

All dataset resources and trained models are publicly available [34] to support practical use and foster further research in web information extraction and related downstream tasks within the news domain.

References

[1]. Chang, C. H., Kayed, M., Girgis, M. R., & Shaalan, K. F. (2006). A survey of web information extraction systems. IEEE transactions on knowledge and data engineering, 18(10), 1411-1428.

[2]. Azir, M. A. B. M., & Ahmad, K. B. (2017, November). Wrapper approaches for web data extraction: A review. In 2017 6th International Conference on Electrical Engineering and Informatics (ICEEI) (pp. 1-6). IEEE.

[3]. Li, J., Xu, Y., Cui, L., & Wei, F. (2022, May). MarkupLM: Pre-training of text and markup language for visually rich document understanding. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 6078-6087).

[4]. Deng, X., Shiralkar, P., Lockard, C., Huang, B., & Sun, H. (2022). DOM-LM: Learning Generalizable Representations for HTML Documents. arXiv e-prints, arXiv-2201.

[5]. Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.

[6]. CommonCrawl. Available at: https://commoncrawl.org/, accessed 25.09.2025

[7]. Hao, Q., Cai, R., Pang, Y., & Zhang, L. (2011, July). From one tree to a forest: a unified solution for structured web data extraction. In Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval (pp. 775-784).

[8]. Sarkhel, R., Huang, B., Lockard, C., & Shiralkar, P. (2023). Self-training for label-efficient information extraction from semi-structured web-pages. Proceedings of the VLDB Endowment, 16(11), 3098-3110.

[9]. Lockard, C., Shiralkar, P., & Dong, X. L. (2019, June). Openceres: When open information extraction meets the semi-structured web. 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) (pp. 3047-3056).

[10]. Hotti, A., Risuleo, R. S., Magureanu, S., Moradi, A., & Lagergren, J. (2021). The Klarna Product Page Dataset: Web Element Nomination with Graph Neural Networks and Large Language Models. arXiv preprint arXiv:2111.02168.

[11]. Kumar, A., Morabia, K., Wang, W., Chang, K., & Schwing, A. (2022, May). CoVA: context-aware visual attention for webpage information extraction. In Proceedings of the Fifth Workshop on e-Commerce and NLP (ECNLP 5) (pp. 80-90).

[12]. Zyte product extraction benchmark. Available at: https://github.com/scrapinghub/product-extraction-benchmark, accessed 26.09.2025.

[13]. Zyte article extraction benchmark. Available at: https://github.com/scrapinghub/article-extraction-benchmark, accessed 26.09.2025.

[14]. Bevendorff, J., Gupta, S., Kiesel, J., & Stein, B. (2023, July). An empirical comparison of web content extraction algorithms. In Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 2594-2603).

[15]. Varlamov, M., Galanin, D., Bedrin, P., Duda, S., Lazarev, V., & Yatskov, A. (2022, December). A dataset for information extraction from news web pages. In 2022 Ivannikov Ispras Open Conference (ISPRAS) (pp. 100-106). IEEE.

[16]. Tkachenko M, Malyuk M, Holmanyuk A, et al. Label Studio: Data labeling software. Available at: https://github.com/HumanSignal/label-studio, accessed 21.10.2025.

[17]. Newspaper article extraction dataset. Available at: https://github.com/AndyTheFactory/article-extraction-dataset/, accessed 26.09.2025.

[18]. Newspaper4k. Available at: https://github.com/AndyTheFactory/newspaper4k/, accessed 10.10.2025.

[19]. Hamborg F., Meuschke N., Breitinger C., Gipp B. news-please: A Generic News Crawler and Extractor. Proceedings of the 15th International Symposium of Information Science, 2017, pp. 218-223. DOI: 10.5281/zenodo.4120316.

[20]. Barbaresi A. Trafilatura: Discover and Extract Text Data on the Web. Available at: https://github.com/adbar/trafilatura/, accessed 21.09.2025.

[21]. DOM-LM implementation. Available at: https://github.com/ilyalasy/DOM-LM, accessed 21.09.2025.

[22]. Zhang, Z., Yu, B., Liu, T., Liu, T., Wang, Y., & Guo, L. (2023, April). Learning structural co-occurrences for structured web data extraction in low-resource settings. In Proceedings of the ACM Web Conference 2023 (pp. 1683-1692).

[23]. Xu, H., Chen, L., Zhao, Z., Ma, D., Cao, R., Zhu, Z., & Yu, K. (2024, March). Hierarchical multimodal pre-training for visually rich webpage understanding. In Proceedings of the 17th ACM International Conference on Web Search and Data Mining (pp. 864-872).

[24]. Web Scraper. Available at: https://webscraper.io/, accessed 21.10.2025.

[25]. DrissionPage. Available at: https://github.com/g1879/DrissionPage, accessed 21.10.2025.

[26]. Langdetect. Available at: https://pypi.org/project/langdetect/, accessed 21.10.2025.

[27]. Googletrans. Available at: https://github.com/ssut/py-googletrans, accessed 21.10.2025.

[28]. Webtraversallibrary. Available at: https://github.com/klarna-incubator/webtraversallibrary, accessed 21.10.2025.

[29]. Wang, C., Cho, K., & Gu, J. (2020, April). Neural machine translation with byte-level subwords. In Proceedings of the AAAI conference on artificial intelligence (Vol. 34, No. 05, pp. 9154-9160).

[30]. A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzman, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, “Unsupervised Cross-lingual Representation Learning at Scale”, in Annual Meeting of the Association for Computational Linguistics, 2019.

[31]. CommonCrawl-News Dataset. Available at: https://data.commoncrawl.org/crawl-data/CC-NEWS/index.html, accessed 21.10.2025.

[32]. Dateparser. Available at: https://github.com/scrapinghub/dateparser, accessed 21.10.2025.

[33]. FacebookAI/xlm-roberta-base. Available at: https://huggingface.co/FacebookAI/xlm-roberta-base, accessed 21.10.2025.

[34]. ISPRAS-CRAWLERS's Collections. NewsXLM. Available at: https://huggingface.co/collections/ispras-crawlers/newsxlm, accessed 16.02.2026.

Информация об авторах / Information about authors

Павел Александрович БЕДРИН – старший лаборант Института системного программирования, магистрант ВМК МГУ. Сфера научных интересов: сбор данных из веб-ресурсов, автоматизация процесса сбора данных, извлечение информации, машинное обучение.

Pavel Alexandrovich BEDRIN – senior laboratory assistant at the Institute of System Programming of the RAS, master student of the CMC faculty of Lomonosov Moscow State University. Research interests: data collection from web resources, automation of the data collection process, information extraction, machine learning.

Максим Игоревич ВАРЛАМОВ – научный сотрудник Института системного программирования. Сфера научных интересов: сбор данных из веб-ресурсов, автоматизация процесса сбора данных, извлечение информации, машинное обучение.

Maxim Igorevich VARLAMOV – researcher at the Institute of System Programming of the RAS. Research interests: data collection from web resources, automation of the data collection process, information extraction, machine learning.

Александр Константинович ЯЦКОВ – младший научный сотрудник Института системного программирования, ведущий программист ВМК МГУ. Сфера научных интересов: сбор данных из веб-ресурсов, автоматизация процесса сбора данных, извлечение информации, машинное обучение.

Alexander Konstantinovich YATSKOV – junior researcher at the Institute of System Programming of the RAS, leading programmer at the CMC faculty of Lomonosov Moscow State University. Research interests: data collection from web resources, automation of the data collection process, information extraction, machine learning.