e-JURNAL PENYELIDIKAN DAN INOVASI
ISLAMIC INNOVATION CENTER FOR TECHNOLOGY, UIS
https://ejpi.uis.edu.my/
e-ISSN 2289-7909

VOL. 12 NO. 4 (DECEMBER, 2025): page 200-212
DOI: 10.53840/ejpi.v12i4.291

Received: 07 Oct 2025 | Accepted: 01 Dec 2025 | Published: 20 Dec 2025

AI-Driven Web Content Scraping Focusing on Transformer-Based: A Systematic Literature Review

Pengikisan Kandungan Web Didorong Kepintaran Buatan (AI) dengan Tumpuan kepada Model Berasaskan Transformer – Suatu Ulasan Literatur Sistematik

Khirulnizam Abd Rahman, Syed Arbaz Ahmed, Che Wan Shamsul Bahri Che Wan Ahmad, Syarbaini Ahmad, Sazanah Md Ali, Siti Azrehan Aziz, Nurul Ibtisam Yaacob & Rafiza Kasbun.
Universiti Islam Selangor, Bandar Seri Putra, 43000 Kajang, Selangor, Malaysia
Email: khirulnizam@uis.edu.my, syed_arbaz@hotmail.com

ABSTRACT

Web scraping has become a key technique for harvesting large volumes of data across varied online platforms, underpinning applications in business intelligence, market research, academic studies, and social media oversight. Traditional scraping strategies, which are usually rule-based and depend on fixed pattern-matching techniques—such as XPath, CSS selectors, and regular expressions—tend to fail when web pages undergo structural modification or when content is rendered dynamically via JavaScript. Such brittleness compromises the longevity and flexibility of standard solutions, especially when confronted with diverse and frequently refreshed sites. Recently, breakthroughs in artificial intelligence, particularly the advent of transformer architectures like Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformers (GPT), have paved the way for smarter, context-sensitive content harvesting. These models leverage their profound grasp of contextual semantics and their ability to decipher intricate language constructs, yielding higher precision in information retrieval and greater resilience to changes in page layout. This systematic literature review (SLR) aggregates the prevailing research on embedding transformer-driven models within web scraping pipelines. It surveys current system designs, fusion methods, metric frameworks, and comparative performance records while illuminating outstanding research voids and pathways for continued exploration. The results are designed to give researchers and practitioners a thorough understanding that can help them improve scraping systems by embedding transformer-based AI technologies.

Keywords: Web scraping; AI-driven; transformer-based; information retrieval; systematic literature review

ABSTRAK

Pengikisan web telah menjadi teknik penting untuk mengumpul data dalam jumlah besar merentasi pelbagai platform dalam talian, sekaligus menyokong aplikasi dalam kecerdasan perniagaan, penyelidikan pasaran, kajian akademik, dan pemantauan media sosial. Strategi pengikisan tradisional, yang biasanya berasaskan peraturan dan bergantung pada teknik padanan corak tetap—seperti XPath, pemilih CSS, dan ungkapan biasa—sering gagal apabila halaman web mengalami pengubahsuaian struktur atau apabila kandungan dipaparkan secara dinamik melalui JavaScript. Kerapuhan ini menjejaskan ketahanan dan fleksibiliti penyelesaian standard, terutamanya apabila berdepan dengan laman web yang pelbagai serta kerap dikemas kini. Baru-baru ini, kemajuan dalam kecerdasan buatan, khususnya kemunculan seni bina transformer seperti Bidirectional Encoder Representations from Transformers (BERT) dan Generative Pre-trained Transformers (GPT), telah membuka jalan kepada

pengikisan kandungan yang lebih pintar serta peka terhadap konteks. Model-model ini memanfaatkan pemahaman mendalam mereka terhadap semantik kontekstual serta keupayaan untuk mentafsir struktur bahasa yang rumit, sekali gus menghasilkan ketepatan lebih tinggi dalam pengambilan maklumat serta ketahanan yang lebih baik terhadap perubahan susun atur halaman. Kajian literatur sistematik (SLR) ini mengumpulkan penyelidikan semasa tentang penggabungan model berasaskan transformer dalam saluran pengikisan web. Ia meninjau reka bentuk sistem terkini, kaedah gabungan, rangka kerja metrik, serta rekod perbandingan prestasi, sambil mendedahkan kekosongan penyelidikan yang wujud serta hala tuju untuk penerokaan lanjut. Hasilnya direka untuk memberikan penyelidik dan pengamal pemahaman menyeluruh yang dapat membantu mereka menambah baik sistem pengikisan dengan mengintegrasikan teknologi AI berasaskan transformer.

Kata kunci: Pengikisan web; berasaskan AI; berasaskan transformer; pengambilan maklumat; tinjauan literatur sistematik

INTRODUCTION

In recent years, the exponential growth of online data has led to web scraping becoming a vital technique for large-scale data acquisition, enabling applications in business intelligence, market analysis, academic research, and social media monitoring (Khder, 2021). Old-school web scraping techniques mostly use rules like XPath, CSS selectors, and regex to dig out data. While they were once popular, they struggle as web pages change layout, load parts of their content through scripts, or hide information to block automated gathering (Xu & Zheng, 2021). When sites start to use more advanced front-end frameworks and add layers of anti-scraping defenses, scrapers based solely on fixed rules become brittle and fall behind, leading to missing or unreliable information (Ferrara et al., 2014).

Advances in artificial intelligence, especially the transformer approaches, have opened fresh pathways for smart web-content harvesting. Models like Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformers (GPT) grasp language in rich, contextual layers and adjust easily, which lets them dig into ever-shifting combinations of web layouts (Yaman et al., 2025). Studies such as DocFormer (Appalaraju et al., 2021) and WebFormer (Wang et al., 2022) show how transformers can learn to decode document layouts and tag HTML elements, beating older rule-based and conventional machine-learning methods by a wide margin. Likewise, retrieval-augmented generation (RAG) techniques enhance the semantic classification and chunking of web pages, reinforcing transformers' value in the scraping pipeline (Ahluwalia & Wani, 2024).

In addition, large language models (LLMs) have started to craft scraper code on the fly and to automatically tune extraction rules, which softens the brittleness that plagues many traditional approaches (Ahluwalia & Wani, 2024). These trends imply that transformers could move web scraping beyond rigid, brittle scripts to flexible, context-sensitive systems. At the same time, the ecosystem lacks cohesion; researchers employ varying methods, datasets, and metrics, making cross-comparison and collective progress difficult.

Therefore, this research will perform a systematic literature review (SLR) to aggregate previous works on transformer-based AI models in web contents scraping. The purposes of this review are (i) to identify how transformer-based architectures have been applied in web data extraction tasks,(ii) identify the advantages and limitations compared to traditional scraping approaches, and (iii) which type of datasets and metrics are general in empirical works. Mapping the existing knowledge base, SLRs aim to shed light on the current state and possible outlook of the phenomena that develops around the use of transformers for web scraping, for both researchers and practitioners.

METHODOLOGY

This section outlines the research methodology used to conduct systematic literature review (SLR). The process follows the guidelines of Kitchenham et al. (2009) and Petersen et al. (2008), ensuring a

structured and reproducible review. The methodology includes the formulation of research questions, definition of inclusion and exclusion criteria, search strategy, quality assessment, data collection, and analysis approach.

Research Questions

The primary aim of this SLR is to synthesize research on the use of transformer-based artificial intelligence models for web content scraping. The objectives were translated into the following research questions (RQs):

- **RQ1:** How have transformer-based models been applied in web content scraping and data extraction tasks?
- **RQ2:** What advantages do transformer-based approaches offer compared to traditional rule-based or earlier machine learning methods?
- **RQ3:** What datasets, tasks, and evaluation metrics have been employed in studies on transformer-based web scraping?
- **RQ4:** What are the reported limitations and challenges of applying transformer-based models to web scraping, and what research gaps remain?

Search Process

We conducted a comprehensive search through peer-reviewed journals and conference proceedings from a variety of digital databases. Databases search included Semantic Scholar, Scopus and Google Scholar. The search captures the period following the introduction of transformer architectures in natural language processing and lasts between January 2015 and August 2025.
The steps forward were as follows:

i. At first, the process of conducting a systematic review started with a search for studies using the pre-established keywords: ("web scraping" OR "web content scraping") AND ("transformer-based model").
ii. Secondly the edited keywords: ("web scraping" OR "web content scraping") AND ("transformer-based model") AND ("BERT" OR "GPT" OR "AI-driven" OR "AI-powered" OR "AI-based" OR "artificial intelligence") AND ("information extraction" OR "data extraction").
iii. Third: relevant studies were included and unrelated studies were excluded.
iv. Duplicated were removed.

The search terms were utilized for reviewing the paper's title, abstract, and keywords. Table 1 displays the journals and conferences that have been identified.

Table 1: Database Library

Resource Type	Resource Name
Search Engine and Online Database	Google Scholar, Semantic Scholar, and Scopus

Inclusion and Exclusion Criteria

There were articles on the following subjects:

i. Journal, conference.
ii. Papers composed in the English language.
iii. Published in or 2015 and later.
iv. Papers come up from the search string.
v. Research that complements studies on Web Content Scraping.

Articles on the following topics were excluded:

i. Articles that don’t concentrate on Web Content Scraping.
ii. Other literature reviews.

iii. PowerPoint and presentation Slides.
iv. Logs or web pages.

Quality Assessment

There are six questions for quality assessment (QA) as follows:

i. **QA1:** Does the paper clearly describe the application of transformer-based models for web content scraping, extraction, or parsing?
ii. **QA2:** Does the paper provide sufficient technical details (architecture, training, datasets, preprocessing) for reproducibility or understanding?
iii. **QA3:** Does the paper include empirical evaluation with performance metrics?
iv. **QA4:** Does the paper discuss challenges or limitations of applying transformer-based models to web scraping?
v. **QA5:** Does the paper clearly state the research objectives or problem statement?
vi. **QA6:** Does the study adequately answer its research objectives with supporting?

The questions in the quality assessment are scored as shown in Table 2:

Table 2: Quality Assessment Scoring Procedure

QA1. Does the paper clearly describe the application of transformer-based models for web content scraping, extraction, or parsing?
Y (Yes): The paper details what transformer-based models for web scraping are required for, including the types of work and aims.
P (Partly): The paper discusses web scraping with transformer-based models, yet it falls short on specifics.
N (No): The paper does not explain at all how to employ transformer-based techniques for web scraping tasks.
QA2. Does the paper provide sufficient technical details (architecture, training, datasets, preprocessing) for reproducibility or understanding?
Y (Yes): The paper thoroughly details the authors´ technical methods with which the high-level model itself, pre-training, data collection and experiment configuration are rendered and left open for others to use.
P (Partly): The paper supplies some technical details, but on important parts there is some ambiguity or loss of understanding.
N (No): The paper gives very little technical detail, which makes it impossible for others to actually recreate or understand.
QA3. Does the paper include empirical evaluation with performance metrics?
Y (Yes): The paper reports result on large datasets or with widely used methods. Appropriate benchmarks and metrics are included to show how these results compare across different work environments.
P (Partly): The paper reports some results, but metrics are incomplete or poorly explained.
N (No): The paper does not provide quantitative evaluations or performance metrics at all
QA4. Does the paper discuss challenges or limitations of applying transformer-based models to web scraping?
Y (Yes): The paper explicitly discusses challenges and limitations (scalability, cost, ethical issues, generalization).
P (Partly): The paper briefly mentions some limitations but does not discuss them in detail.
N (No): The paper does not mention any challenges or limitations
QA5. Does the paper clearly state the research objectives or problem statement?
Y (Yes): The objectives or problem statement are explicitly stated and well-defined.
P (Partly): The objectives are mentioned but not clearly defined or explained.
N (No): The paper does not state any research objectives or problem statement**.**
QA6. Does the study adequately answer its research objectives with supporting?

Y (Yes): The study addresses its objectives thoroughly with evidence and analysis.
P (Partly): The study addresses some objectives but leaves gaps or lacks strong evidence.
N (No): The study does not provide answers or evidence related to its stated objectives.
The scoring procedure are Y = 1, P = 0.5, N = 0

Data Collection

The data that is retrieved out of all papers includes:
i. The complete reference and the source (journal or conference).
ii. Main topic area / issues / limitations.
iii. Methodology applied.
iv. Main finding / Discussion.

Analysis of Data

The data was shown in a form of table to display:
i. An overview of the review steps 1-3's outcomes.
ii. Each SLR's quality score.
iii. Category distribution of transformer-based approaches applied in web content scraping.

RESULT

Search Result
Quantitative outcomes attained after every step can be observed in Table 3 below. Search result included from digital databases.

Table 3: Number of Papers According to The Search

Result after inclusion and exclusion	Google Scholar	Semantic Scholar	Scopus
STEP 1	465	356	69
STEP 2	102	24	4
STEP 3	6	1	2

A collection of 9 relevant, related works has been assembled after searching with the strings mentioned and browsing online databases. Every paper has been organized into a table, which (as shown in Table 4) has the following details in its contents: (a) Title of study; (b) Author or authors/ citation; and (c) Publication year.

Table 4: Data Collection of Related Study

Paper ID.	Title	Author/citation	Publication year
1.	A Framework for the Unsupervised Modelling and Extraction of Polarization Knowledge from News Media	(Paschalides et al., 2025)	2025
2.	A Hybrid Approach for Key-Value Extraction from Technical Specification Documents	(Lee, 2024)	2024
3.	Bringing order into the realm of Transformer-based language models for artificial intelligence and law	(Greco & Tagarelli, 2024)	2024
4.	DeB3RTa: A Transformer-Based Model for the Portuguese Financial Domain	(Pires et al., 2025)	2024
5.	Enhancing Web Scraping with Artificial Intelligence: A Review	(Weerasinghe et al., 2024)	2024
6.	Large language model-based framework for automated extraction of genetic interactions from unstructured data	(Gill et al., 2024)	2024

7.	AI-Powered Web Scraping and Parsing: A Browser Extension Using LLMs for Adaptive Data Extraction	(Nevgi et al., 2025)	2025
8.	A Novel Approach to Web Article Summarization	(Suresh et al., 2025)	2025
9.	Metadata Extraction from Scholarly Document Using Deep Learning	(Raval & Bhaidasna, 2025)	2025

Quality Assessment of Related Papers

The quality assessment questions that were previously discussed were used to analyse and score the 9 papers that were chosen for this study. The outcomes are displayed in Table 5 below:

Table 5: Systematic Review Studies

Paper ID.	Article Type	QA1	QA2	QA3	QA4	QA5	QA6	Total score
1.	Journal	1	1	1	0.5	1	1	5.5
2.	Master's thesis	1	0.5	1	0.5	1	1	5
3.	Review Article	0.5	0.5	0	1	1	0.5	3.5
4.	Journal	0.5	1	1	0.5	1	1	5
5.	Review Article	1	0.5	0.5	1	1	0.5	4.5
6.	Research paper	0.5	1	1	1	1	1	5.5
7.	Journal	1	0.5	0.5	1	0.5	1	4.5
8.	Research Paper	0.5	1	1	0.5	1	1	5
9.	Conference	0.5	1	1	1	1	1	5.5

Performance Comparison of Transformer-Based Methods Included Studies

This table 6 summarizes the transformer-based models used in the studies, the tasks they were applied to, and their reported best performance scores.

Table 6: Performance comparison in included studies

Paper ID.	Author/citation	Transformer Model	Task	Accuracy score
1.	(Paschalides et al., 2025)	Polarization Data Model (PDM)	Capturing entities' attitudes toward various topics, aligning politically cohesive fellowships with their respective party manifestos, and identifying domain-specific topics along with their degree of polarization	Accuracy score = 0.9100
2.	(Lee, 2024)	CNN-RNN encoder-decoder, RCNN, and LLM	Extract key-value pairs from technical specification documents with a minimum accuracy of 85%	Accuracy score by MPLs = 97.5%
3.	(Greco & Tagarelli, 2024)	GPT-3, mT5, and XLM-RoBERTa	Sequence labeling for personal data extraction, and case law retrieval,	Average accuracy score of 71%
4.	(Pires et al., 2025)	DeB3RTa	Assign labels to input text, specifically in the context of financial natural language	Accuracy score = 0.9953

			processing (NLP) in Portuguese-speaking markets	
5.	(Weerasinghe et al., 2024)	Not addressed	Scrape job advertisements and extract structured data from unstructured sources, such as images and web pages	Not addressed
6.	(Gill et al., 2024)	LLM, BioBERT, and BERN2	Automated extraction of gene interactions from literature, optimizing sentence selection and providing confidence factors for the extracted relations	Precision score = 86.30%
7.	(Nevgi et al., 2025)	LLM and Integrations (BeautifulSoup and Selenium)	Web scrapping and browser automation	Not addressed
8.	(Suresh et al., 2025)	BERT, T5, and GPT-3	web scraping, text preprocessing, and summarization	ROUGE-1 score = 0.40 ROUGE-2 score = 0.45 ROUGE-L score = 0.21 Precision = 0.44 Recall = 0.46 F1-Score = 0.45
9.	(Raval & Bhaidasna, 2025)	BERT, GPT and NLP	Metadata extraction	Accuracy score = 21.5

Category Distribution

It has been observed after Step 3, that the remaining studies are accumulated around two categories as shown in Figure 1.

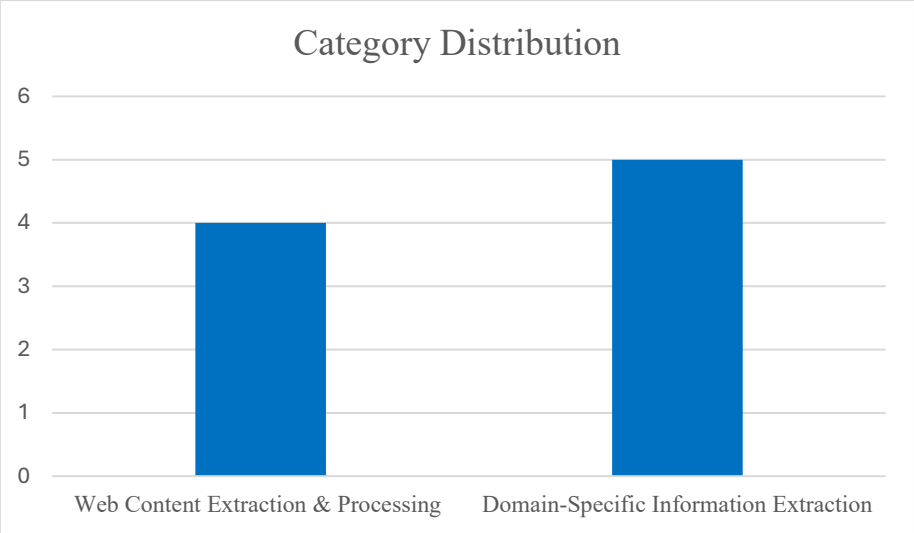
I. Web Content Extraction & Processing (4)

Focuses on using transformer models for web scraping, content summarization, and sentiment/polarization analysis. Papers in this category (Nevgi et al., 2025; Paschalides et al., 2025; Suresh et al., 2025; Weerasinghe et al., 2024) use transformers to extract, summarize, and analyze web content, making information more accessible and understandable.

II. Domain-Specific Information Extraction (5)

This category focuses on transformer models that are intended for specific domains such as law, finance, biology or technology. Using domain-specific transformer models, these papers (Gill et al., 2024; Greco & Tagarelli, 2024; Lee, 2024; Pires et al., 2025; Raval & Bhaidasna, 2025) automate the extraction of structured data and insights, enhancing productivity and decision-making in specialized sectors.

Figure 1: Category Distribution



Subcategories within Web Content Extraction & Processing

Subcategories distribution is shown in Figure 2. Several subcategories within web content extraction & processing includes:

I. **Web Scraping & Data Extraction**
Transforming models for information extraction from the web pages have become increasingly popular. Now even unstructured, dynamic or constantly changing contents can be extracted with ease by them. However, transformer models mean less work for you. Compared with conventional scrapers that must restructure themselves for any new website structure, a transformer model will automatically identify the new layout and adapt itself accordingly. This category also covers techniques used to extract data for web pages with complicated structures and return only those points that matter.

II. **Content Summarization**
This subcategory focuses about what kind of transformer model can be used for long text conversion into a shorter form, such as abstracts or summaries. Summarization can be either extractive (selecting key sentences or phrases) or abstractive (generating new sentences that summarize the content). It is particularly useful for websites, articles, and news platforms where users need quick access to the most relevant information.

III. **Sentiment & Polarization Analysis**
Transformer models in this subcategory are used for sentiment analysis and polarization detection in web content. Sentiment analysis classifies text by the emotional tone (positive, negative, neutral) and polarization analysis detects ideological bias in the content. These models can recognize the public opinion, trends on social media, as well as media bias in the various channels.

Subcategories within Domain-Specific Information Extraction

Several subcategories within domain-specific information extraction includes:

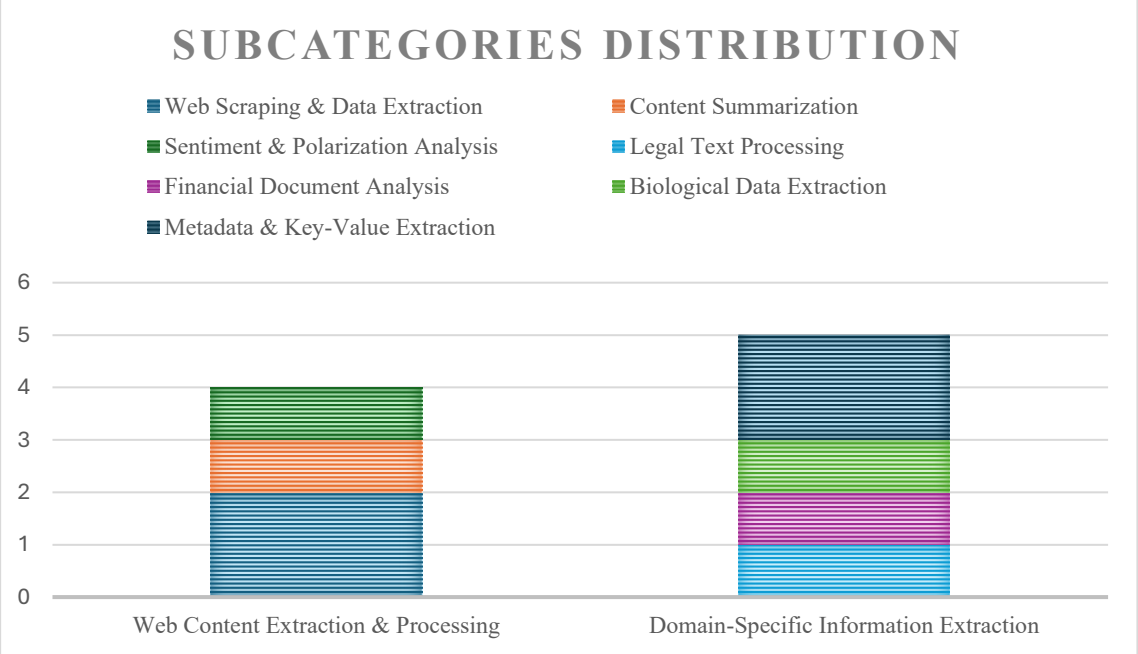
I. **Legal Text Processing**
Transformer models are employed to deal with complex legal language, which helps in tasks such as retrieval of case law, analysis of legal documents, and forecasting legal outcomes. However, legal texts often involve more difficult terminologies than spoken usages and specific patterns in form which make it necessary for highly complex techniques capable of handling which type of text to be developed.

II. **Financial Document Analysis**
By means of transformer models the technical term used in financial documents applies to financial documents such as annual reports, market analyses, and financial statements. These models make it possible to automate data extraction and precision of financial decision-making, and to adapt the specialized terminology found in financial documents.

III. **Biological Data Extraction**
Increasingly, transformer models are used for automatic extraction of genetic interactions and biological data from scientific research papers. These models extract relationships between genes, proteins, or other biological entities crucial for speeding up scientific research in fields like genomics.

IV. **Metadata & Key-Value Extraction**
This subcategory applies transformer models to extract structured data from documents, such as key-value pairs, metadata, and other technical information. These methods are important for technical specification documents, academic articles, and scholarly papers were extracting specific data points (author names, publication details, citations) is crucial.

Figure 2: Subcategories Distribution



FINDINGS AND DISCUSSIONS

RQ1: How have transformer-based models been applied in web content scraping and data extraction tasks?

Transformer-based models have transformed the realm of web data harvesting and data extraction via providing the ability for models to react to the dynamic web. These models, for example, BERT, T5 and LLaMA, can be effectively used to extract information in structure (product specifications, reviews) out from the web unstructured content. They have been used for tasks such as summarization and sentiment analysis over web articles. For instance, Nevgi et al. (2025) created a browser plugin that utilizes LLMs to scrape information from dynamic web pages and Weerasinghe et al. (2024) describe how AI, and particularly transformer models, can improve web scraping by learning from complex website structures.

RQ2: What advantages do transformer-based approaches offer compared to traditional rule-based or earlier machine learning methods?

Transformer-based models offer several advantages over traditional rule-based and earlier machine learning methods:

1. **Adaptability** Unlike rule-based models, that need predefined rules for each individual page structure, transformers can generalize to any changes of web pages and layouts. This capacity significantly reduces the necessity of manual corrections (Nevgi et al., 2025).
2. **Contextual Understanding**: Transformers model relationship between words in context and hence are much suited for tasks such as summarization and sentiment analysis. Previous models were often built on the use of manually extracted features, which may fail to capture subtleties in the text (Suresh et al., 2025).
3. **Handling Unstructured Data**: While conventional systems need input in structured form, or at least require extensive preprocessing, transformers can process raw, unstructured data directly which leads to increased accuracy and efficiency when extracting data (Weerasinghe et al., 2024).

RQ3: What datasets, tasks, and evaluation metrics have been employed in studies on transformer-based web scraping?

1. **Datasets**: Typical applied datasets for transformer-based web scraping are web contents, news articles and academic papers. Datasets such as CNN/Daily Mail and XSum are commonly used for summarization (Suresh et al., 2025). Paschalides et al. (2025) used political news media datasets for sentiment and polarization analysis.
2. **Tasks**: Key tasks in transformer-based web scraping include:
    - **Web Scraping**: Extracting structured data from unstructured web content (product descriptions, user reviews).
    - **Summarization**: Translating long web articles to short readable abstracts for the consumption (Suresh et al., 2025).
    - **Sentiment and Polarization Analysis**: Studying the sentiment of context or detecting ideological bias, especially in news articles (Paschalides et al., 2025).
3. **Evaluation Metrics**: The performance of these tasks is commonly measured using metrics like:
    - **Precision, Recall, F1-score**: To evaluate the accuracy of extracted data and predictions (Suresh et al., 2025).
    - **BLEU**: Used for evaluating the quality of generated summaries (Suresh et al., 2025).
    - **Accuracy**: Often used in tasks like sentiment and polarization analysis to measure classification performance (Paschalides et al., 2025).

RQ4: What are the reported limitations and challenges of applying transformer-based models to web scraping, and what research gaps remain?

Limitations and Challenges:

1. **High Computational Cost**: Transformer-based models, particularly the larger models like BERT, GPT, are computational costly and resources-demanding if used for real-time web scraping. Weerasinghe et al. (2024) refer to the high resource needs that occur when employing a transformer at scale.
2. **Handling Dynamic Content**: Although transformers are model-agnostic, web pages with dynamically dynamic content (heavy JavaScript dependence) still present issues. This means that the extraction process can fail if the web page needs some extra processing that cannot be done by the transformers alone (Nevgi et al., 2025).
3. **Ethical Issues and Data Privacy**: The secondary use of data is a concern in ethics, particularly when considering that scraping could be performed without consent. There's also the risk that particularly invasive or personal information is scraped, something transformers might not necessarily be trained to silence (Ford et al., 2021).

CONCLUSION AND FUTURE WORK

Scraping and Extracting Data from the Web. The use of Transformer-based models has greatly contributed to the success of web content scraping and data extraction. These models, includes BERT, T5 and LLaMA, have the strengths of being able to adapt to content that has been changing over time, and the ability to handle text, and extract meaningful information from complex web pages. Compared to rule-based systems, the transformer can automatically adapt to pages of various atom structure with few manual inputs and therefore can be scaled up in web scraping. For example, Nevgi et al. (2025) and Weerasinghe et al. (2024) showcased how transformers improve data retrieval in dynamic, unstructured content

Aside from scraping, transformers have seen much success in other tasks such as summarization, sentiment analysis and polarization detection. This makes them particularly useful in extracting structured information from unstructured web sources, as demonstrated by Suresh et al. (2025) and Paschalides et al. (2025), who utilized transformers for news articles and pretrained transformers for sentiment analysis and content summarization. But even with this progress, there are still multiple obstacles to overcome. Computation burden is an important consideration since large transformer models tend to be computationally expensive for training and inference, which may hinder their scalability when empowering real-time web scraping (Weerasinghe et al., 2024). Additionally, for a large variety of web scraping cases transformers are not so good at handling dynamic content and JavaScript-laden sites we've got heaps of nowadays on the web. Ultimately, ethical considerations (such as the issue of data privacy and biased scraped content) have to be taken into account, so transformer-based web scraping is responsibly used.

Here are a few things to work on moving forward. An important next step is to extend the transformer models in such a way that they can be applied at the scale of large-scale web scraping without exhausting computational resources. Techniques like model pruning, quantization or smaller models like DistilBERT(Sanh et al., 2019) could be useful in order to lower the computational cost without sacrificing performance. Also, transformers' must be made better so that they can handle dynamic content. A lot of webpages are JavaScript heavy for content that current transformer-based models struggle with. One option could be to incorporate transformers with other web scraping methods, such as Selenium or BeatutifulSoup, which are more robust for dynamic content (Nevgi et al., 2025).

Furthermore, ethical concerns concerning bias detection and privacy must also be tackled. In future work, we aim to learn and correct for the biases present in the training data and models. Formulating ethical frameworks for web scraping, particularly with respect to regulations, such as GDPR, is also necessary to guarantee responsible practices of AI in web scraping applications (Binns, 2017). With more and more transformer-based models in use, openness and ethical considerations become crucial.

Finally, the utilization of transformer models in novel domains like healthcare and smart cities could provide major growth in terms of automated content extraction for these domains. Moreover, how to incorporate transformers to multi-model (images, video and text) data would extend a broad prospect for comprehensive data extraction bolstered by their versatility in different businesses.

In conclusion, transformer-based have demonstrated its extraordinary ability of enhancing the ordinary methods of extracting and scraping the web content for the data extraction. Yet there are issues such as computational efficiency, dynamic content management, and ethical issues that need to tackle. Project future studies on developing these models as lean, adaptable, and ethical tools to harness not only web scraping but its power across many areas.

Table 7: Research Gaps and Priority

Gap	Priority	Explanation
High computational burden of transformer models (Weerasinghe et al., 2024) (Raval & Bhaidasna, 2025)	High	Large transformer models are expensive to train and run, making them difficult to use for real-time or large-scale web scraping.
Ethical concerns: data privacy, bias in scraped content (Binns, 2017) (Weerasinghe et al., 2024)	High	Scraping may expose private user data or amplify biased content. Ethical frameworks and compliance with regulations like GDPR are necessary to prevent misuse.
Lack of transformer optimization for large-scale web scraping (Pires et al., 2025)	Medium	Models need to be made smaller/faster so they can run efficiently at scale without high resource consumption.
Limited integration of transformers with existing scraping pipelines (Selenium, BeautifulSoup) (Nevgi et al., 2025)	Medium	Transformers are powerful for text understanding but weak at navigating web structures. Integrating them with traditional scraping tools could improve performance on dynamic content.
Need for bias detection and correction in training data (Greco & Tagarelli, 2024)	Medium	Transformer models reflect the biases of their training data. Identifying and correcting these biases is important for fair and accurate scraped outputs.
Insufficient ethical standards for AI-based web scraping practices (Weerasinghe et al., 2024)	Medium	With more AI scrapers being deployed, clear guidelines and industry standards are required to ensure responsible use.
Limited exploration of transformers in emerging domains (healthcare, smart cities) **(**Greco & Tagarelli, 2024)	Low	These domains have high potential, but current research is limited. Expanding transformer-based scraping into these areas could bring new opportunities.
Need for improved multimodal capabilities (images, videos, text) (Lee, 2024; Weerasinghe et al., 2024)	Low	Many websites contain more than just text. Enhancing transformers to process multimodal data would improve comprehensive web content extraction.

REFERENCES

Ahluwalia, A., & Wani, S. (2024). *Leveraging Large Language Models for Web Scraping* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2406.08246

Appalaraju, S., Jasani, B., Kota, B. U., Xie, Y., & Manmatha, R. (2021). *DocFormer: End-to-End Transformer for Document Understanding* (Version 2). arXiv. https://doi.org/10.48550/ARXIV.2106.11539

Binns, R. (2017). *Fairness in Machine Learning: Lessons from Political Philosophy*. https://doi.org/10.48550/ARXIV.1712.03586

Ferrara, E., De Meo, P., Fiumara, G., & Baumgartner, R. (2014). Web data extraction, applications and techniques: A survey. *Knowledge-Based Systems*, *70*, 301–323. https://doi.org/10.1016/j.knosys.2014.07.007

Ford, E., Shepherd, S., Jones, K., & Hassan, L. (2021). Toward an Ethical Framework for the Text Mining of Social Media for Health Research: A Systematic Review. *Frontiers in Digital Health*, *2*, 592237. https://doi.org/10.3389/fdgth.2020.592237

Gill, J. K., Chetty, M., Lim, S., & Hallinan, J. (2024). Large language model based framework for automated extraction of genetic interactions from unstructured data. *PLOS ONE*, *19*(5), e0303231. https://doi.org/10.1371/journal.pone.0303231

Greco, C. M., & Tagarelli, A. (2024). Bringing order into the realm of Transformer-based language models for artificial intelligence and law. *Artificial Intelligence and Law*, *32*(4), 863–1010. https://doi.org/10.1007/s10506-023-09374-7

Khder, M. (2021). Web Scraping or Web Crawling: State of Art, Techniques, Approaches and Application. *International Journal of Advances in Soft Computing and Its Applications*, *13*(3), 145–168. https://doi.org/10.15849/IJASCA.211128.11

Lee, S. S. (2024). *A Hybrid Approach for Key-Value Extraction from Technical Specification Documents*.

Nevgi, S., Kadam, S., Haldankar, S., Jadhav, S., & Rashmi More, Prof. (2025). AI-Powered Web Scraping and Parsing: A Browser Extension Using LLMs for Adaptive Data Extraction. *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, *09*(04), 1–9. https://doi.org/10.55041/IJSREM44113

Paschalides, D., Pallis, G., & Dikaiakos, M. (2025). A Framework for the Unsupervised Modeling and Extraction of Polarization Knowledge from News Media. *ACM Transactions on Social Computing*, *8*(1–2), 1–38. https://doi.org/10.1145/3703594

Pires, H., Paucar, L., & Carvalho, J. P. (2025). DeB3RTa: A Transformer-Based Model for the Portuguese Financial Domain. *Big Data and Cognitive Computing*, *9*(3), 51. https://doi.org/10.3390/bdcc9030051

Raval, P., & Bhaidasna, H. (2025). Metadata Extraction From Scholarly Document Using Deep Learning. *2025 3rd International Conference on Inventive Computing and Informatics (ICICI)*, 1–5. https://doi.org/10.1109/ICICI65870.2025.11069873

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter* (Version 4). arXiv. https://doi.org/10.48550/ARXIV.1910.01108

Suresh, P., Kavya, V., S, V. G., Harshitha, B., & Ashritha, D. (2025). A Novel Approach to Web Article Summarization. *2025 3rd International Conference on Inventive Computing and Informatics (ICICI)*, 37–42. https://doi.org/10.1109/ICICI65870.2025.11069807

Wang, Q., Fang, Y., Ravula, A., Feng, F., Quan, X., & Liu, D. (2022). WebFormer: The Web-page Transformer for Structure Information Extraction. *Proceedings of the ACM Web Conference 2022*, 3124–3133. https://doi.org/10.1145/3485447.3512032

Weerasinghe, K., Maduranga, D. M., & Kawya, M. (2024). *Enhancing Web Scraping with Artificial Intelligence: A Review*.

Xu, X., & Zheng, X. (2021). Hybrid Model for Network Anomaly Detection with Gradient Boosting Decision Trees and Tabtransformer. *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 8538–8542. https://doi.org/10.1109/ICASSP39728.2021.9414766

Yaman, A., Schwab, J., Nitsche, C., Sinha, A., & Huber, M. (2025). *Comparison of Large Language Models for Deployment Requirements* (Version 1). arXiv. https://doi.org/10.48550/ARXIV.2508.00185