Article

Enhancing News Articles: Automatic SEO Linked Data Injection for Semantic Web Integration

Hamza Salem * , Hadi Salloum * , Kamil Sabbagh , Manuel Mazzara

Posted Date: 31 October 2024

doi: 10.20944/preprints202410.2489.v1

Keywords: Semantic Web; Pattern Mining; OpenAI GPT-3; News Web Pages; Readability; Structured Data.)

Article

Enhancing News Articles: Automatic SEO Linked Data Injection for Semantic Web Integration

Hamza Salem [1], **Hadi Salloum** [1,2,*], **Kamil Sabbagh** [1,2], **Manuel Mazzara** [1]

[1] Innopolis University, Innopolis, Republic of Tatrstan, Russian Federation
[2] QDeep, Innopolis, Republic of Tatrstan, Russian Federation
[*] Correspondence: h.salloum@innopolis.university

Abstract: This paper presents a novel solution aimed at enhancing news web pages for seamless integration into the Semantic Web. By utilizing advanced pattern mining techniques alongside OpenAI's GPT-3, we rewrite news articles to improve their readability and accessibility for Google News aggregators. Our approach is characterized by its methodological rigor and is evaluated through quantitative metrics, validated using Google's Rich Results Test, which confirms the effectiveness of our generated structured data. The impact of our work is threefold: it advances the technological integration of a substantial segment of the web into the Semantic Web, promotes the adoption of Semantic Web technologies within the news sector, and significantly enhances the discoverability of news articles in aggregator platforms. Furthermore, our solution facilitates the broader dissemination of news content to diverse audiences. This submission introduces an innovative solution substantiated by empirical evidence of its impact and methodological soundness, thereby making a significant contribution to the field of Semantic Web research, particularly in the context of news and media articles.

Keywords: semantic web; pattern mining; OpenAI GPT-3; news web pages; readability; structured data

1. Introduction

The Semantic Web is a visionary extension of the traditional World Wide Web, aimed at making data more accessible and interpretable by machines. It focuses on embedding structured, machine-readable information into web content to enhance how search engines and other services understand, process, and display that content. In the domain of news web portals, the Semantic Web has transformed how articles and other information are indexed and presented, particularly through platforms like Google News, which leverages structured data to improve the discoverability and relevance of content [1,2]. This structured data, commonly referred to as SEO-linked data, is embedded directly into HTML using formats such as JSON-LD, providing search engines with additional context to deliver more accurate search results.

Rich Results, which are enhanced search results displaying additional visual and contextual information, are made possible through structured data. These results can include images, star ratings, and product details, offering users a more informative experience in search engine results pages (SERPs). Despite the availability of tools that facilitate the automatic generation of structured data—such as plugins integrated with popular content management systems (CMS)—a large number of legacy websites still lack appropriate structured data annotations. This gap presents challenges for both search engine optimization (SEO) and user engagement, as unstructured websites struggle to be effectively indexed by search engines [3].

This paper addresses the problem of insufficient structured data annotations by introducing a robust solution for the automatic generation of JSON-LD scripts for news websites. Building upon our previously published algorithm based on pattern mining [4] and state-of-the-art natural language models like OpenAI GPT-3, we propose an advanced method that extracts relevant article features and uses these features to create suitable JSON-LD annotations. By applying this method to a dataset of news articles, we demonstrate the automatic insertion of structured data into existing HTML, ensuring compliance with modern search engine standards.

To validate the generated structured data, we propose a framework that incorporates the removal of pre-existing JSON-LD scripts, automatic regeneration via our algorithm, and subsequent verification through a set of mathematical models. This process is formalized using a matrix representation of article features. Specifically, let the article data be represented as a matrix $A \in \mathbb{R}^{n \times m}$, where n is the number of articles, and m represents the extracted features (e.g., headline, author, date, and content). The goal is to map this feature matrix to a structured data matrix $S \in \mathbb{R}^{n \times k}$, where k denotes the set of JSON-LD elements required for structured data compliance.

This approach can be formalized as an optimization problem where the objective is to minimize the difference between the ground truth JSON-LD data and the generated JSON-LD data. The aim is to ensure that the generated annotations are as close as possible to what would be manually created. The objective is thus to minimize:

$$\min_{\mathbf{S}} |\text{Ground Truth JSON-LD} - \text{Generated JSON-LD}|_2^2 \tag{1}$$

This formulation ensures that the generated structured data aligns accurately with the ground truth. The difference is measured using a suitable comparison method, which could include either a rich search output comparison or a word-by-word evaluation of the structured data fields. This ensures that the structured data annotations are correct and optimized for search engines.

Furthermore, we define the accuracy of the structured data generation process in terms of a matching function $M(A, S)$, which evaluates how well the structured data aligns with the original article features. This matching function is computed as:

$$M(A, S) = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} \mathbf{1}(A_{ij} = S_{ij})}{n \times m} \tag{2}$$

where $\mathbf{1}(A_{ij} = S_{ij})$ is an indicator function that returns 1 if the element in the i-th article and j-th feature matches between the original and structured data matrices, and 0 otherwise. This metric provides a quantitative assessment of the automated annotation process.

Our research tackles the critical issue of structured data scarcity across the web, with particular emphasis on legacy news websites that do not benefit from modern CMS tools. By leveraging advanced pattern recognition techniques and language models, we present an innovative solution for generating JSON-LD annotations that enhance search engine visibility and compliance with current standards. Our validation process incorporates the use of tools like the Rich Text Verifier, which allows us to check the correctness and compliance of our generated structured data against established search engine requirements. This ensures that the JSON-LD scripts meet the necessary standards for displaying rich results, thereby increasing the visibility of the annotated web pages. This work contributes to the broader Semantic Web ecosystem, improving content discoverability, indexing, and interoperability. It also has far-reaching implications for the web, as billions of existing pages that lack structured data annotations can now be effectively indexed, representing a significant step forward in automated web data structuring.

Structure of the Paper

This paper is organized into several sections that collectively present our approach to enhancing news web pages for integration into the Semantic Web. Section 2, titled Literature Review. This review highlights both historical developments and contemporary approaches, situating our contribution within the broader context of Semantic Web research. Section 3, Methodology, outlines our novel approach, illustrated by a flowchart. Here, we detail how our methodology infers linked data from web page content without prior knowledge of the HTML structure, utilizing articles from Google News as the primary data source. Section 4 presents the Implementation, where we elaborate on the practical application of our methodology. In Section 5, titled Comments and Results, we discuss the outcomes of our approach and provide critical insights based on our findings. Following this, Section 6 is dedicated to Impact Analysis, assessing the technological, social, and business implications of our solution. Finally, Section 7 concludes the paper by summarizing our findings and emphasizing the significance of our contributions to the field of Semantic Web research.

2. Literature Review

This section provides a comprehensive review of the existing literature, organized into two main threads: (i) the generation of metadata and linked data from raw text, and (ii) methods for the automated extraction and organization of data to be used in metadata generation. The review underscores both historical developments and current approaches, contextualizing our contribution within this body of knowledge.

2.1. Metadata Generation and Linked Data Representation

The vision of the Semantic Web, first articulated over two decades ago by Tim Berners-Lee and his team, has been fundamental in reshaping how web content is perceived by both humans and machines. Berners-Lee's goal was to create a web where information is not only accessible by humans but also structured in a way that machines can understand and process autonomously. This vision laid the groundwork for the development of standards and technologies aimed at transforming the web into a more intelligent system.

To this end, the World Wide Web Consortium (W3C) has defined key standards for representing structured data on the web. Two of the most significant standards include the *Resource Description Framework (RDF)* and the *Web Ontology Language (OWL)* [6]. RDF provides a syntax for encoding relationships between resources, while OWL allows for the creation of complex ontologies—formal vocabularies that define concepts within specific domains. These frameworks have seen increasing adoption across academia and industry [7]. However, for these technologies to reach their full potential, it is essential to develop domain-specific vocabularies (ontologies). Many of these vocabularies have been crafted manually for fields such as healthcare, law, and geography, but advancements in machine learning are enabling the automatic or semi-automatic generation of these ontologies from large text corpora [8]. This progress reduces the time and expertise required to build and maintain these knowledge representations, making the Semantic Web more accessible and scalable.

2.2. Automated Data Acquisition and Content Extraction

Extracting the main content from unstructured web pages has been a long-standing challenge in web data mining. Early approaches focused on parsing the *Document Object Model (DOM)* structure of web pages. For instance, researchers at Columbia University proposed a method to detect the largest text body within a web page, classifying it by counting the words in each HTML element [9]. This approach, which is the basis of our method for identifying the body of an article, proves effective for detecting main content sections but struggles with title identification due to the variability in webpage layouts.

More sophisticated content extraction methods have since been proposed. For example, [10] introduced an approach combining *structural* and *contextual analysis* to improve the accuracy of content extraction. While this method has shown promise, it has yet to be fully implemented, suggesting that further refinement and experimentation are needed.

The literature reveals two dominant approaches in the field of web data extraction:

2.2.1. DOM-Based Approaches

DOM-based approaches leverage the structural properties of HTML documents to identify and extract relevant content. These methods rely on traversing the DOM tree to locate elements such as text bodies and headings based on properties such as size, position, and relative relationships between elements [11–13]. While effective in well-structured environments, the major limitation of DOM-based methods is their sensitivity to changes in webpage layouts. As web pages evolve, their structural composition can change, rendering static DOM-based techniques less effective over time.

2.2.2. AI-Driven Approaches

Artificial Intelligence (AI)-based methods represent a more dynamic approach to content extraction. These techniques typically involve training machine learning models on large, labeled datasets of web pages. The models learn to recognize patterns and features within the page that correspond to different types of content, such as article titles, bodies, and advertisements [14–16]. AI-based methods have demonstrated strong performance, particularly in handling diverse web layouts. However, they require significant computational resources and large amounts of labeled data for training. Moreover, the generalization of these models to new, unseen web layouts can be challenging, especially when the underlying page structure differs significantly from the training data.

2.3. Challenges and Opportunities

Both DOM-based and AI-based approaches offer distinct advantages and limitations. *AI-based methods*, though flexible and adaptive, require significant infrastructure and resources, and their performance is often contingent upon the availability of extensive training data. They also face difficulties in adapting to new or evolving web layouts unless retrained periodically. In contrast, *DOM-based methods* are lightweight and efficient, particularly for structured data extraction, but are sensitive to changes in webpage architecture. These methods must be updated regularly to remain effective. In this research, we adopt a *DOM-based approach* to automatically extract article data from web pages. This choice is motivated by the simplicity and computational efficiency of DOM-based techniques, which align well with the structured nature of news articles and their presentation on modern web pages. Furthermore, we extend the traditional DOM-based methodology by integrating pattern mining techniques to enhance content accuracy and robustness. Our method allows for the automated extraction and re-injection of structured data (e.g., article titles and bodies) back into web pages as linked data, eliminating the need for manual intervention [17–22].

The review of existing literature reveals a clear division between DOM-based and AI-driven approaches to web content extraction. While *AI-driven approaches* are more flexible, they require extensive datasets and computational power to function effectively. *DOM-based approaches*, as employed in this research, offer a more scalable and efficient solution for content extraction, particularly for domains with consistent web structures such as news articles. By leveraging the strengths of DOM-based methods and enhancing them with pattern recognition and linked data injection, this study contributes a novel methodology that addresses the limitations of both traditional DOM-based and AI-driven approaches.

3. Methodology

Our methodology, depicted in the flowchart in Figure 1, enables the inference of linked data from the entire web page content without requiring prior knowledge of the structure of the source HTML code. The results produced are comparable to those generated by web page authors.

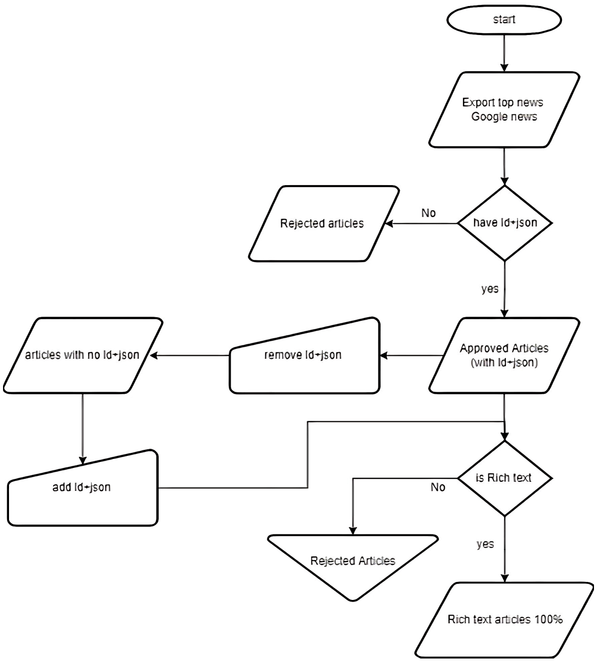
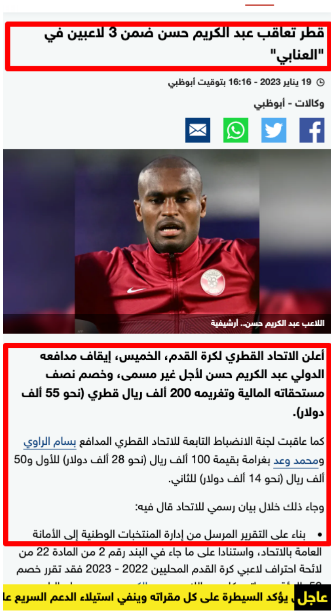


Figure 1. Flowchart

To perform the analysis, we utilized Google News as our source of articles, which already possesses linked data JSON objects (JSON-LD) attached to each page, hereafter referred to as the original JSON-LD. Our approach employs a general data extraction method that applies pattern mining to news sites, as published in [4]. The pattern miner scrapes the title and body of the news article to extract the plain text content. The text with the largest font size on the page is identified as the title, while the largest *div* element (by character length) is designated as the body, as illustrated in Figure 2. Furthermore, we extended our method from [4] to clean the title and article body after scraping, removing unrelated words, links, images, or advertisements using the OpenAI API GPT-3 [23].



قطر تعاقب عبد الكريم حسن ضمن 3 لاعبين في
"العنابي"

19 يناير 2023 - 16:16 بتوقيت أبوظبي

وكالات - أبوظبي

اللاعب عبد الكريم حسن.. أرشيفية

أعلن الاتحاد القطري لكرة القدم، الخميس، إيقاف مدافعه الدولي عبد الكريم حسن لأجل غير مسمى، وخصم نصف مستحقاته المالية وتغريمه 200 ألف ريال قطري (نحو 55 ألف دولار).

كما عاقبت لجنة الانضباط التابعة للاتحاد القطري المدافع بسام الراوي ومحمد وعد بغرامة بقيمة 100 ألف ريال (نحو 28 ألف دولار) للأول و50 ألف ريال (نحو 14 ألف دولار) للثاني.

وجاء ذلك خلال بيان رسمي للاتحاد قال فيه:

- بناء على التقرير المرسل من إدارة المنتخبات الوطنية إلى الأمانة العامة بالاتحاد، واستنادا على ما جاء في البند رقم 2 من المادة 22 من لائحة احتراف لاعبي كرة القدم المحليين 2022 - 2023 فقد تقرر خصم

عاجل يؤكد السيطرة على كل مقراته وينفي استيلاء الدعم السريع ع

Figure 2. News Title and Body Detection

Utilizing this methodology, we extract the most important properties for the JSON-LD object, such as the title, article body, images, and URLs, thus creating a new JSON object—the generated

doi:10.20944/preprints202410.2489.v1

JSON-LD. To verify our results, we compare the generated JSON-LD with the original using two distinct methods:

1. Validate the results on the Google Rich Results Checker [24].
2. Perform a word-by-word comparison of the original and generated texts.

While the Google Rich Results Checker serves as the primary validation tool, we also employ a word-by-word verification method to ensure content accuracy, as the rich results validator only checks the structural integrity of the object.

To quantify the content similarity, we utilize **Jensen-Shannon Divergence** [25], which measures the similarity between two probability distributions P and Q. This divergence is defined as:

$$D_{JS}(P||Q) = \frac{1}{2}D_{KL}(P||M) + \frac{1}{2}D_{KL}(Q||M) \tag{3}$$

where $M = \frac{1}{2}(P+Q)$ and D_{KL} is the Kullback-Leibler divergence, defined as:

$$D_{KL}(P||Q) = \sum_x P(x)\log\left(\frac{P(x)}{Q(x)}\right) \tag{4}$$

This allows us to quantify the differences in content distribution between the original JSON-LD and the generated JSON-LD.

Additionally, we can measure the entropy H of the generated content using the equation:

$$H(X) = -\sum_i p(x_i)\log(p(x_i)) \tag{5}$$

where $p(x_i)$ represents the probability of the occurrence of each unique word x_i in the content.

The methodology can be succinctly represented in five steps:

1. **Extraction of News Articles:** Gather news web pages containing the original JSON-LD object.
2. **Analyze/Remove JSON-LD:** Remove the original JSON-LD object and save it externally for comparison.
3. **Generate & Inject JSON-LD:** Use pattern mining to generate a new JSON-LD object and inject it into the original page to replace the original.
4. **Check JSON-LD:** Validate the new page with the injected JSON-LD using the Rich Results Checker.
5. **Check the JSON-LD Content:** Compare the content word by word to compute a similarity score between properties.

The following pseudocode outlines the algorithms utilized in our methodology for extracting and generating JSON-LD annotations.

3.1. Algorithm 1: Extraction of News Articles

This algorithm describes the process of extracting news articles from web pages.

Algorithm 1 ExtractNewsArticles

```
procedure ExtractNewsArticles(URL)
    Fetch the web page content from the URL
    Parse the HTML content
    Identify the title:
        title ← FindElementByLargestFontSize()
    Identify the body:
        body ← FindElementByLargestDiv()
    Identify additional properties:
        img ← FindImageElements()
        url ← FindPageURL()
    Return (title, body, img, url)
end procedure
```

3.2. *Algorithm 2: Cleaning the Article Body*

This algorithm cleans the extracted article body by removing unrelated content, such as stop words and images.

Algorithm 2 CleanArticleBody

```
procedure CleanArticleBody(RawArticleBody)
    Initialize cleanedBody as an empty string
    for each word in RawArticleBody do
        if word is not in the stop words list then
            cleanedBody ← cleanedBody + word
        else if word is a link or image then
            continue
        end if
    end for
    Return cleanedBody
end procedure
```

3.3. *Algorithm 3: Generation of JSON-LD*

This algorithm generates the new JSON-LD object based on the cleaned title and body, including other relevant properties.

Algorithm 3 GenerateJSONLD

```
Input: Title, Cleaned Body, Image, URL
Output: Generated JSON-LD object
Initialize an empty JSON object jsonLD as {}
Set jsonLD["title"] ← Title
Set jsonLD["body"] ← Cleaned Body
Set jsonLD["image"] ← Image
Set jsonLD["url"] ← URL
Add any additional metadata to jsonLD
return jsonLD
```

3.4. *Algorithm 4: Validation of JSON-LD*

This algorithm validates the generated JSON-LD using the Google Rich Results Checker and performs a content comparison.

Algorithm 4 ValidateJSONLD

```
1: **Input:** Original JSON-LD, Generated JSON-LD
2: **Output:** Validation result
3: Validate JSON structure using Google Rich Results Checker
4: validationResult ← CheckRichResults(Generated JSON-LD)
5: Perform word-by-word comparison between Original and Generated JSON-LD
6: similarityScore ← CompareWordByWord(Original JSON-LD, Generated JSON-LD)
7: **return** (validationResult, similarityScore)
```

4. Implementation

This section presents a detailed review of the implemented solution, fulfilling all the requirements outlined earlier. We begin with an overview of the framework architecture, followed by an in-depth description of the individual components, as depicted in Figure 3.

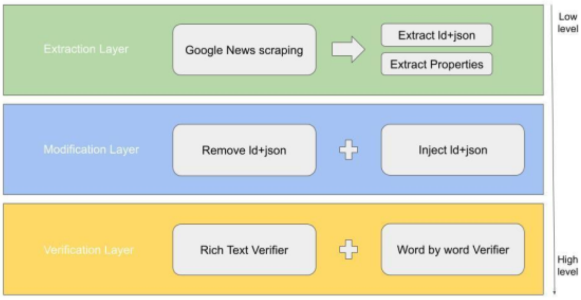
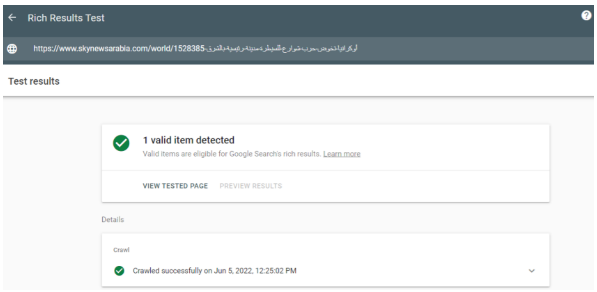


Figure 3. Architecture Overview



https://www.skynewsarabia.com/world/1528385-أوكرانيا-خوض-حرب-شوارع-للسيطرة-مدينة-رئيسية-بالشرق

Test results

1 valid item detected

Valid items are eligible for Google Search's rich results. Learn more

Details

Crawl

Crawled successfully on Jun 5, 2022, 12:25:02 PM

Figure 4. Data Flow

4.1. *Data Extraction Layer*

The Data Extraction Layer consists of:

- **Web Scraping Module:** This module employs the Beautiful Soup library for HTML parsing and the requests library to fetch the web content from Google News articles.
- **Data Storage:** The extracted data is stored in a NoSQL database (MongoDB) to facilitate easy access and manipulation during subsequent analysis stages.

4.2. *Transformation Layer*

The Transformation Layer encompasses:

- **Pattern Mining Algorithm:** This algorithm utilizes a set of predefined rules to identify and extract the title and body of the news articles based on font sizes and element types.
- **Data Cleaning Module:** Implemented using the OpenAI API, this module cleans the extracted article body by removing irrelevant content, enhancing the quality of the information retrieved.

4.3. Loading Layer

The Loading Layer involves:

- **JSON-LD Generation:** The cleaned and structured data is converted into a JSON-LD format, ensuring compliance with schema.org standards for linked data representation.
- **Database Integration:** The generated JSON-LD is stored back into the database for future retrieval and validation purposes.

4.4. Validation Layer

Finally, the Validation Layer includes:

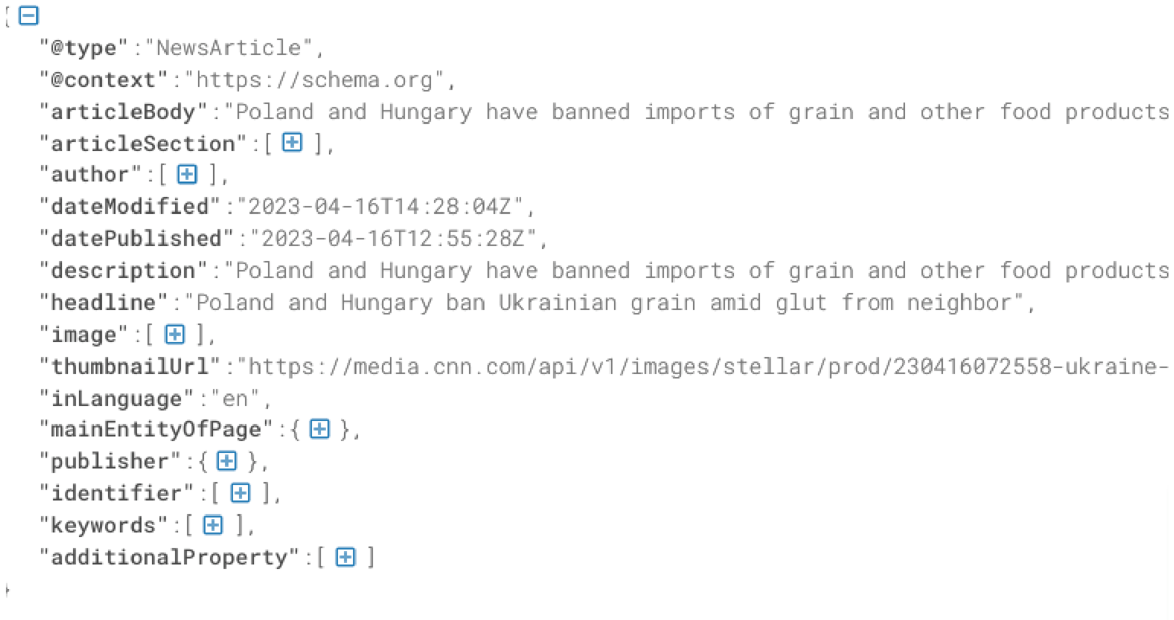
- **Google Rich Results Checker Integration:** This component automates the validation of the generated JSON-LD, verifying its structural integrity against Google's standards.
- **Similarity Comparison Module:** This module employs the Jensen-Shannon Divergence metric to quantitatively assess the similarity between the original and generated JSON-LD content.

The implementation described above provides a comprehensive framework for the extraction, transformation, generation, and validation of linked data from news articles, facilitating the enhancement of data accessibility and interoperability on the web.

5. Experiment and Results

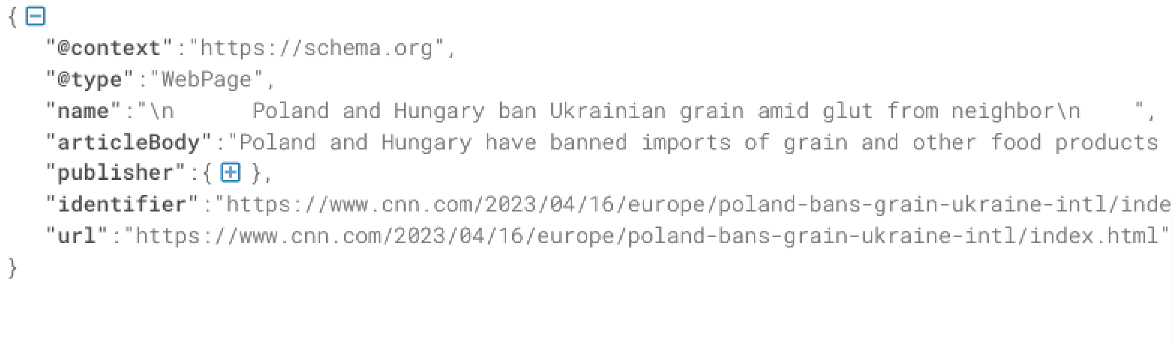
The dataset used for the experiment consists of 1100 web pages from 18 reputable news sources, written in both English and Arabic. The original and generated JSON-LD objects for each article are compared using word-by-word matching, and validated using Google Rich Results.

Figures 5 and 6 illustrate examples of the original and generated JSON-LD objects, respectively. A quantitative comparison of the performance across various websites is presented in Table 1.



```
{⊟
  "@type":"NewsArticle",
  "@context":"https://schema.org",
  "articleBody":"Poland and Hungary have banned imports of grain and other food products
  "articleSection":[⊞],
  "author":[⊞],
  "dateModified":"2023-04-16T14:28:04Z",
  "datePublished":"2023-04-16T12:55:28Z",
  "description":"Poland and Hungary have banned imports of grain and other food products
  "headline":"Poland and Hungary ban Ukrainian grain amid glut from neighbor",
  "image":[⊞],
  "thumbnailUrl":"https://media.cnn.com/api/v1/images/stellar/prod/230416072558-ukraine-
  "inLanguage":"en",
  "mainEntityOfPage":{⊞},
  "publisher":{⊞},
  "identifier":[⊞],
  "keywords":[⊞],
  "additionalProperty":[⊞]
}
```

Figure 5. Original JSON-LD data



```
{⊟
   "@context":"https://schema.org",
   "@type":"WebPage",
   "name":"\n      Poland and Hungary ban Ukrainian grain amid glut from neighbor\n    ",
   "articleBody":"Poland and Hungary have banned imports of grain and other food products
   "publisher":{⊞},
   "identifier":"https://www.cnn.com/2023/04/16/europe/poland-bans-grain-ukraine-intl/inde
   "url":"https://www.cnn.com/2023/04/16/europe/poland-bans-grain-ukraine-intl/index.html"
}
```

Figure 6. Generated JSON-LD data

The dataset evaluation is summarized in Table 1, indicating the pass rates and similarity scores for the title and article body, broken down by website.

Table 1. Dataset Evaluation and Results

Website	Rich Results	Title Similarity	Article Body Similarity	Lang.
Skynewsarabia.com	Pass	>93%	>90%	AR
arabic.cnn.com	Pass	>93%	>90%	AR
youm7.com	Pass	>93%	>90%	AR
bbc.com	Pass	>93%	>90%	AR
cnn.com	Pass	>95%	>90%	EN
reuters.com	Pass	>95%	>90%	EN

To quantify the similarities between the original and generated JSON-LD objects, we used a word-by-word comparison tool. Figures 7 and 8 display histograms of the similarity percentages for the title and article body, respectively.

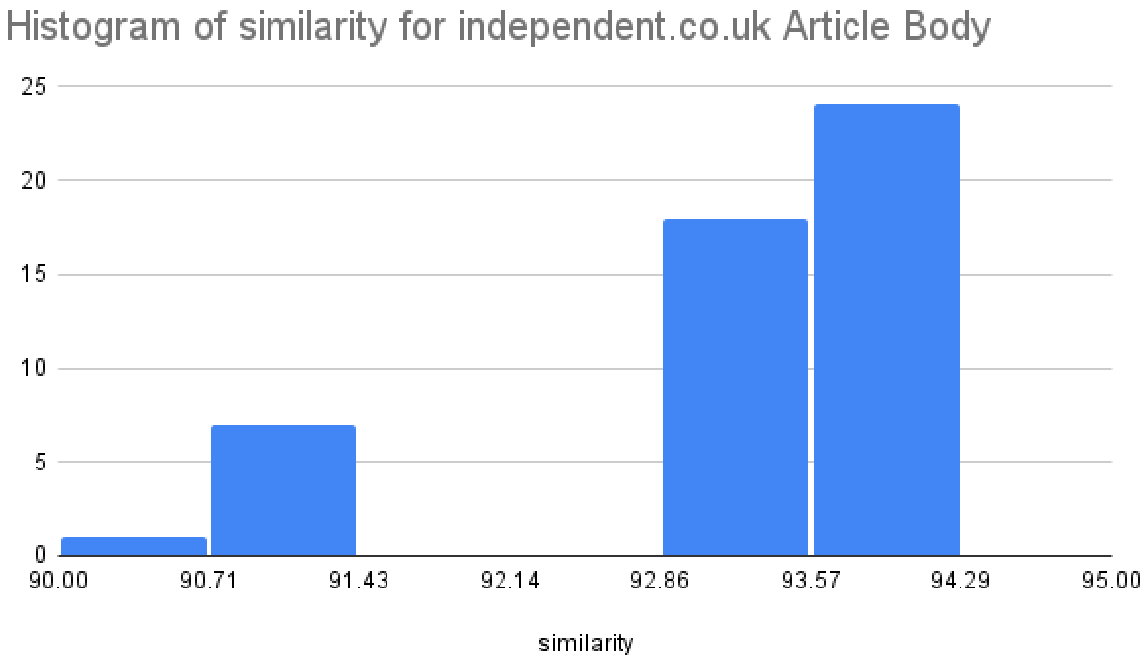


Figure 7. Similarity of Article Body (independent.co.uk)

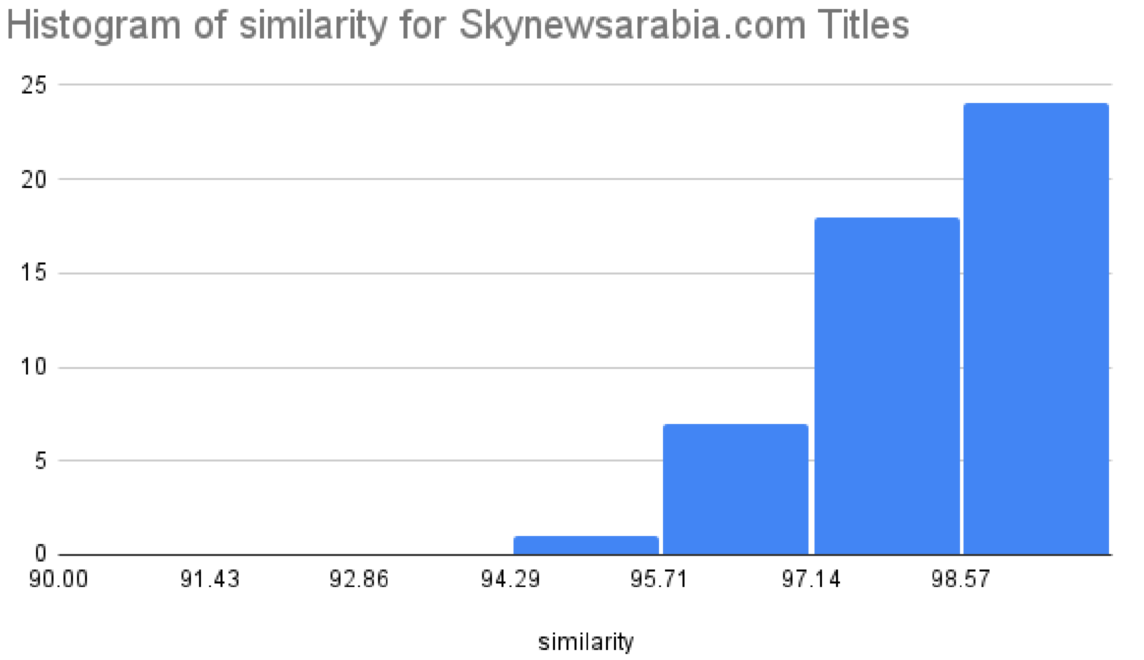


Figure 8. Similarity of Article Title (Skynewsarabia.com)

The implemented framework successfully scraped and processed news articles from diverse sources, generating accurate JSON-LD objects with high similarity to the original data. Validation through Google Rich Results indicated high pass rates, and content similarity consistently exceeded 90%. Future work could involve enhancing the data cleaning process and expanding the dataset to include more websites in multiple languages.

6. Impact Analysis

The implications of our proposed solution for automatically generating JSON-LD annotations extend across several dimensions, including technological advancements, business benefits, and social impacts. This multifaceted impact underscores our work's significance in the Semantic Web context and broader data accessibility initiatives.

6.1. Technological Impact

From a technological standpoint, our work significantly contributes to the evolution of the Semantic Web by enhancing the machine-readability of web content. The automated generation of structured data through advanced algorithms and language models, such as GPT-3, facilitates the integration of artificial intelligence into web development processes. One of the key benefits of this approach is its scalability; it allows for the rapid scaling of structured data across millions of web pages, accommodating the growing demand for data-driven insights and services. Furthermore, by adhering to JSON-LD standards, the generated structured data enhances interoperability between different systems and platforms, enabling better data exchange and collaboration among various stakeholders in the digital ecosystem. This also leads to enhanced SEO, as the implementation of structured data improves visibility in search engine results, providing a competitive advantage to websites that adopt our solution [29].

6.2. Business Impact

The business ramifications of our solution are equally substantial. The ability to automate structured data generation presents several commercial opportunities. Notably, automating the JSON-LD generation process minimizes the need for extensive manual labor, significantly reducing operational costs associated with website maintenance and SEO. This cost reduction enables organizations to allocate resources more efficiently. Additionally, businesses that implement our structured data generation solution can distinguish themselves from competitors by offering enhanced user experiences, personalized content recommendations, and improved engagement metrics. Furthermore, better access to structured data allows organizations to leverage analytics and insights to inform strategic decision-making, optimizing their content strategies and improving audience targeting.

6.3. Social Impact

On a societal level, our work addresses critical issues related to information accessibility and digital equity. By improving the indexing of news content, our solution ensures that users have better access to timely and relevant information, fostering a more informed society. Moreover, smaller news organizations that may lack the resources for sophisticated SEO strategies can leverage our automated solution to enhance their online presence, thus leveling the playing field in the digital landscape. This increased visibility not only empowers these organizations but also promotes a more diverse media landscape by ensuring that a wider range of perspectives and narratives are accessible to the public. Consequently, our work contributes to a more democratic society by facilitating the dissemination of varied viewpoints.

In conclusion, the technological, business, and social impacts of our work are profound and far-reaching. By enabling automated structured data generation, we advance the capabilities of web technologies while empowering organizations and individuals to harness the full potential of information in the digital age. The comprehensive nature of these impacts highlights the relevance and necessity of our research in fostering an inclusive and innovative information ecosystem.

7. Conclusions

In this research, we have proposed a comprehensive solution for embedding linked data objects into news articles sourced from Google News. Our dataset comprised 1,100 web pages representing news articles from 18 distinct news websites in both English and Arabic. The results demonstrated a substantial similarity rate exceeding 93% for news titles and over 90% for the article body within the generated linked data JSON objects. Notably, we utilized ChatGPT, a sophisticated language model developed by OpenAI, to refine the titles and article content, ensuring high-quality output. The algorithm we developed holds promise for seamless integration as a plugin within various content management systems (CMS), facilitating the automatic injection of linked data annotations into web pages. This innovation has the potential to significantly enhance the discoverability and interoperability of news articles on the web by incorporating structured data.

Moreover, the applicability of our approach extends beyond news articles, suggesting broader utility for a wide range of web page types. This versatility can greatly enhance the functionality and user experience of the web for both consumers and developers alike. Looking ahead, we aim to expand our dataset further and enhance the accuracy and efficiency of our algorithm. We believe our research has significant implications for advancing the adoption of linked data practices across the web, fostering improved information accessibility and integration.

Funding: Not applicable.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

References

1. Yu, Liyang. Introduction to the semantic web and semantic web services. Chapman and Hall/CRC, 2007.
2. Wang, Qun. Normalization and differentiation in Google News: a multi-method analysis of the world's largest news aggregator. Diss. Rutgers The State University of New Jersey, School of Graduate Studies, 2020.
3. Bizer, C.,Meusel, R., Primpeli, A., Brinkmann, A.: Web Data Commons - RDFa, Microdata Microformat Data Sets - Section 3.2 Extraction Results from the October 2022 Common Crawl Corpus. 2023. accessed on 2023-01-29. Retrieved from: `http://webdatacommons.org/structureddata/index.html#toc4`
4. Salem, Hamza, and Manuel Mazzara. "Pattern Matching-based scraping of news websites." Journal of Physics: Conference Series. Vol. 1694. No. 1. IOP Publishing, 2020.
5. Sporny, Manu, et al. "JSON-LD 1.0." W3C recommendation 16 (2014): 41.
6. Berners-Lee, Tim, James Hendler, and Ora Lassila. "The semantic web." Scientific american 284.5 (2001): 34-43.
7. Adida, Ben, et al. "RDFa in XHTML: Syntax and processing." Recommendation, W3C 7.41 (2008): 14.
8. Chandrasekaran, Balakrishnan, John R. Josephson, and V. Richard Benjamins. "What are ontologies, and why do we need them?." IEEE Intelligent Systems and their applications 14.1 (1999): 20-26.
9. McKeown, Kathleen, et al. "Columbia multi-document summarization: Approach and evaluation." (2001).
10. A. F. R. Rahman, H. Alam and R. Hartono. "Content Extraction from HTML Documents". In 1st Int. Workshop on Web Document Analysis (WDA2001), 2001.
11. Fumarola, F.; Weninger, T.; Barber, R.; et al.: Extracting General Lists from Web Documents: A Hybrid Approach. In Proceedings of the 24th International Conference on Industrial Engineering and Other Applications of Applied Intelligent Systems Conference on Modern Approaches in Applied Intelligence - Volume Part I. IEA/AIE'11. Berlin, Heidelberg: Springer-Verlag. 2011. ISBN 978-3-642-21821-7. pp. 285–294.
12. Hong, J. L.; Siew, E.-G.; Egerton, S.: Information Extraction for Search Engines Using Fast Heuristic Techniques. Data Knowl. Eng.. vol. 69, no. 2. February 2010: pp. 169–196. ISSN 0169-023X. doi:10.1016/j.datak.2009.10.002.
13. Safi, W.; Maurel, F.; Routoure, J.-M.; et al.: A Hybrid Segmentation of Web Pages for Vibro-Tactile Access on Touch-Screen Devices. In 3rd Workshop on Vision and Language (VL 2014) associated to 25th International Conference on Computational Linguistics (COLING 2014). dublin, Ireland. Aug 2014. pp. 95 – 102.

14. Lima, R.; Espinasse, B.; Oliveira, H.; et al.: Information Extraction from the Web: An Ontology-Based Method Using Inductive Logic Programming. In 2013 IEEE 25th International Conference on Tools with Artificial Intelligence. Nov 2013. ISSN 2375-0197. pp. 951–958. doi:10.1109/ICTAI.2013.114.
15. Zheng, S.; Song, R.; Wen, J.-R.: Template-Independent News Extraction Based on Visual Consistency. In Proceedings of the 22nd National Conference on Artificial Intelligence - Volume 2. AAAI'07. AAAI Press. 2007. ISBN 9781579953232. pp. 1507–1512.
16. Zhu, W.; Dai, S.; Song, Y.; et al.: Extracting news content with visual unit of web pages. In Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing (SNPD), 2015 16th IEEE/ACIS International Conference on. June 2015. pp. 1–5
17. Suhit Gupta, Gail Kaiser, David Neistadt, and Peter Grimm. *DOM-based content extraction of HTML documents*. In *Proceedings of the 12th International Conference on World Wide Web*, pages 207–214, 2003.
18. Mehdi Mirzaaghaei and Ali Mesbah. *DOM-based test adequacy criteria for web applications*. In *Proceedings of the 2014 International Symposium on Software Testing and Analysis*, pages 71–81, 2014.
19. Johannes Behr, Peter Eschler, Yvonne Jung, and Michael Zöllner. *X3DOM: a DOM-based HTML5/X3D integration model*. In *Proceedings of the 14th International Conference on 3D Web Technology*, pages 127–135, 2009.
20. A. Milani Fard, M. Mirzaaghaei, and A. Mesbah, “Leveraging existing tests in automated test generation for web applications,” in *Proceedings of the 29th ACM/IEEE International Conference on Automated Software Engineering*, pp. 67–78, 2014.
21. J. Xia, F. Xie, Y. Zhang, and C. Caulfield, “Artificial intelligence and data mining: algorithms and applications,” *Abstract and Applied Analysis*, vol. 2013, no. 1, pp. 524720, 2013. Hindawi Publishing Corporation.
22. D. Menaga and S. Saravanan, “Application of artificial intelligence in the perspective of data mining,” in *Artificial Intelligence in Data Mining*, pp. 133–154, Elsevier, 2021.
23. OpenAI, T. B. "Chatgpt: Optimizing language models for dialogue." OpenAI (2022).
24. “Rich Results Test.” Google Search Console, Google, https://search.Google.com/test/rich-results.
25. M. L. Menéndez, J. A. Pardo, L. Pardo, and M. C. Pardo, “The Jensen-Shannon divergence,” *Journal of the Franklin Institute*, vol. 334, no. 2, pp. 307–318, 1997. Elsevier.
26. Mitchell, Ryan. Web scraping with Python: Collecting more data from the modern web. " O'Reilly Media, Inc.", 2018.
27. Patel, Jay M. "Web Scraping in Python Using Beautiful Soup Library." Getting Structured Data from the Internet. Apress, Berkeley, CA, 2020. 31-84.
28. “Plagiarism Detector Software: Anti-Plagiarism Tools.” Copyleaks, 4 Oct. 2022, https://copyleaks.com/.
29. N. Shadbolt, T. Berners-Lee, and W. Hall, “The Semantic Web Revisited,” *IEEE Intelligent Systems*, vol. 21, no. 3, pp. 96–101, 2006.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.