# ADAPTATION OF WEB CONTENT INTO VIDEO FORMAT FOR USERS WITH SPECIAL NEEDS

Shatylo I., Chala L., Bezkorovainyi V.

Department of Systems Engineering, Department of Artificial Intelligence
Kharkiv National University of Radio Electronics,
Ukraine

E-mail: ihor.shatylo@nure.ua,
larysa.chala@nure.ua,
vladimir.beskorovainyi@nure.ua

**Abstract**

The report presents a system designed to transform unstructured web articles into accessible video formats and is designed to overcome significant barriers faced by users with visual impairments or cognitive disabilities. The architecture is based on a hybrid content extraction method that combines dynamic page rendering and heuristic analysis, ensuring the separation of the main text from the "noise" of the website. Further processing is performed by a microservice pipeline based on artificial intelligence that supports distributed and scalable work. The experimental results show that abstract generalization models, combined with high-quality neural speech synthesis and generative models for music creation, are capable of forming full-fledged, accessible video content. Experimental exploitation confirms the effectiveness of the chosen approach and its potential in increasing the level of information accessibility. Based on the results obtained, it is advisable to develop recommendations for large-scale automated adaptation of digital content, which will help ensure equal access to information for all users.

The modern digital space is characterised by the dominance of visual-textual forms of information presentation. Websites, news portals, blogs, and educational resources are predominantly text-centric, which creates a significant barrier for certain categories of users. Specifically, users with visual impairments, dyslexia, or other cognitive particularities that affect information processing often cannot fully perceive data presented as continuous text. Traditional accessibility tools, such as screen readers, while useful, are sometimes ineffective on modern websites. They linearly read all DOM content, including advertising banners, navigation links, and meta-tags, which creates noise interference and significantly complicates the extraction of the useful signal. This deepens the social isolation of users with special needs and limits their access to knowledge and services.

One of the most effective ways to solve this problem is to adapt web content into synesthetic, multimedia formats, particularly video [1]. The video format, based on dual-coding theory, combines an auditory track (synthesised speech), a visual sequence (relevant videos and images), and text captions. This allows engaging multiple perception channels simultaneously, which significantly reduces cognitive load and promotes information accessibility. However, manually creating such video adaptations for the daily flow of web content is almost impossible due to high costs in time and resources.

The fundamental problem, on which the quality of all further adaptation depends, is obtaining relevant data from web sources. Modern pages are no longer static HTML documents; they are dynamic web applications built on frameworks (React, Vue, Angular, etc.), where content is loaded asynchronously using JavaScript requests. Under these conditions, the main information (e.g., the text of an article or news messages) is often semantically and visually hidden among numerous auxiliary elements: navigation panels, advertising blocks, comments, footers, and widgets. This information redundancy creates an effect of «content entropy», where the signal (main content) and noise (decorative or auxiliary elements) are practically indistinguishable at the surface markup level.

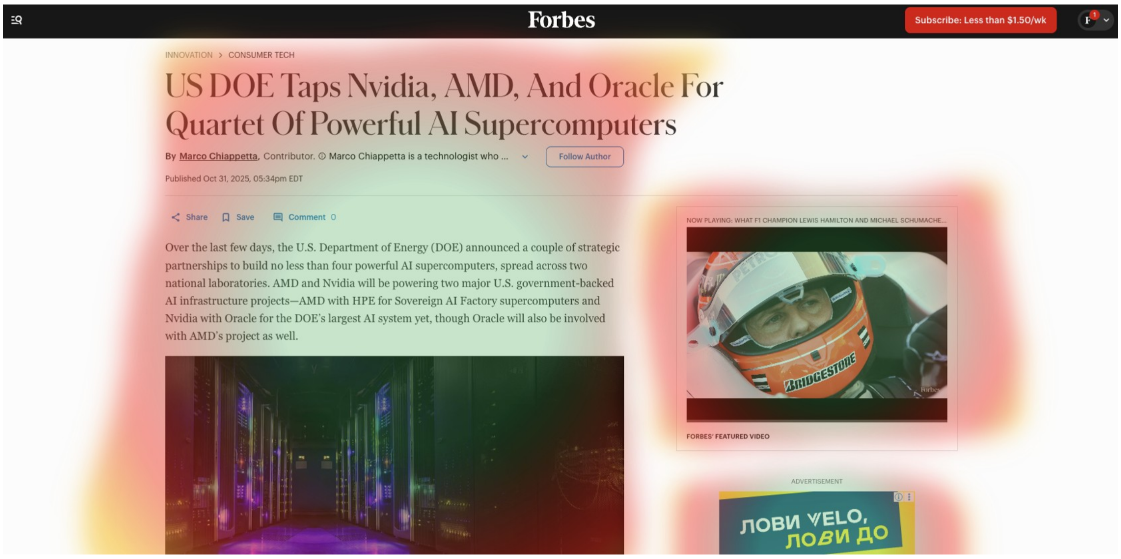
Traditional parsing methods based on static DOM analysis (e.g., searching by <p> or <article> tags) prove to be ineffective because they do not account for dynamic content loading and cannot distinguish the main text from decorative elements. Furthermore, the structural heterogeneity of web pages means that even minor changes in layout or design make extraction algorithms completely unusable.

Alternative approaches, particularly wrapper induction (creating templates for specific sites), provide some accuracy but are extremely fragile and require constant manual updating and scaling in a dynamic web environment [2]. More recently, the emergence of Large Language Models (LLMs) has introduced a new paradigm. These models can leverage their deep semantic understanding to analyse either the raw HTML or a text-only representation of the DOM, allowing them to «read» the page and identify the main content, navigation, or footers based on context rather than fixed structural rules.

However, using general-purpose LLMs for this «zero-stage» extraction presents significant challenges. It often introduces substantial computational overhead and latency, making it inefficient for rapid or high-volume adaptation [3]. Furthermore, the inherent «black-box» nature and potential for inconsistent outputs or «hallucinations» can be a critical failure point; an LLM might mistakenly summarise a prominent advertisement or a user comment section instead of the main article, completely compromising the accessibility goal.

This landscape – with fragile, high-maintenance wrappers on one end and costly, non-deterministic LLMs on the other – reinforces the critical need for hybrid content extraction methods. Such methods must be robust and efficient, ideally combining lightweight heuristics, statistical analysis, and more focused, explainable semantic models capable of adapting to changes in page structure and ensuring stable, high-quality data.

The reliability of the entire adaptation system for users with special needs depends on the system's ability to accurately and stably isolate the «signal» (main content) from the «noise» of interface elements (Fig. 1). An error at this «zero» stage (e.g., capturing advertising text instead of the article) leads to error propagation and the generation of a completely irrelevant video, which completely negates the goal of accessibility. That is why not only accuracy but also explainability is critically important for such systems. We need not only to obtain content but also to understand why the model considers this particular block to be the main one. This will allow us to reliably verify the result before the «expensive» operation of video generation.


Forbes
Subscribe: Less than $1.50/wk
INNOVATION > CONSUMER TECH
Follow Author
Share
Save
Comment 0
NOW PLAYING: WHAT F1 CHAMPION LEWIS HAMILTON AND MICHAEL SCHUMACHE...
FORBES' FEATURED VIDEO
ADVERTISEMENT
ЛОВИ VELO,
ЛОВИ ДО

US DOE Taps Nvidia, AMD, And Oracle For Quartet Of Powerful AI Supercomputers

By Marco Chiappetta, Contributor. ⓘ Marco Chiappetta is a technologist who ...

Published Oct 31, 2025, 05:34pm EDT

Over the last few days, the U.S. Department of Energy (DOE) announced a couple of strategic partnerships to build no less than four powerful AI supercomputers, spread across two national laboratories. AMD and Nvidia will be powering two major U.S. government-backed AI infrastructure projects—AMD with HPE for Sovereign AI Factory supercomputers and Nvidia with Oracle for the DOE's largest AI system yet, though Oracle will also be involved with AMD's project as well.

**Fig. 1. A schematic representation of a webpage highlighting «signal» (main text) and «noise» (navigation, advertising) blocks**

The proposed web content adaptation system, at the first stage, uses a dynamic rendering tool – Playwright [4]. It launches a full instance of a headless browser (Chromium), navigates to the specified URL, waits for all network requests and JavaScript execution to complete, and only then captures the final DOM structure of the page, identical to what the user sees. In the second stage, heuristic and statistical analysers (based on the Newspaper4k library) come into play, which do not rely on specific tags but analyse the downloaded DOM structure by recursively traversing the node tree [5]. For each node, the «text density» is calculated – the ratio of text characters to the total number of HTML tags within the node. Nodes with low density (e.g., navigation menus with many <a> links) are pruned. The system then identifies the node with the maximum density and returns it as the main content. This combination of programmatic browser control and intelligent structure analysis allows for the extraction of the main text block with fairly high accuracy.

However, such a heuristic approach, although more effective than static analysis, is still vulnerable. It may mistakenly identify blocks of comments or long legal disclaimers as main content, as they also have high «text density». This proves once again that neither simple rules nor statistical heuristics are capable of reliably distinguishing «signal» from «noise» on the modern web. This creates an urgent need for hybrid models that combine structural analysis (DOM), visual rendering (screenshots) and semantic context.

Further research should be aimed at improving this stage, particularly by implementing hybrid models and methods for web content extraction, as well as Explainable AI (XAI) models.

After extraction, the main text is passed to a processing pipeline, where it is first analysed by a language detector (Fig. 2). If the language is not English (e.g., Ukrainian), the text is automatically translated. For this, optimised transformer models from the Helsinki-NLP repository, trained for specific languages, are used. This is a necessary step, as most modern high-quality models for summarisation (BART) and sentiment analysis (RoBERTa) are trained primarily on large English-language corpora and demonstrate the best quality in English. Next, the cleaned and normalised text proceeds to the key stage – abstractive summarisation.

The system deliberately avoids extractive methods (like TextRank) because they simply pull existing sentences, resulting in a text that sounds unnatural and has abrupt transitions. Instead, the system uses the *facebook/bart-large-cnn* model, which is based on the BART (Bidirectional and Auto-Regressive Transformers) architecture. This model was trained on the CNN/DailyMail dataset and allows for analysing an article and creating a short summary for it. Thanks to its architecture (a bidirectional encoder, like in BERT, and an auto-regressive decoder, like in GPT), BART does not copy but paraphrases and synthesises new sentences, creating an entirely new, concise, and grammatically coherent text. This generated short summary becomes the script for the future video. This is convenient for users with cognitive particularities, as instead of a 10-minute article, they receive a 1-minute summary.

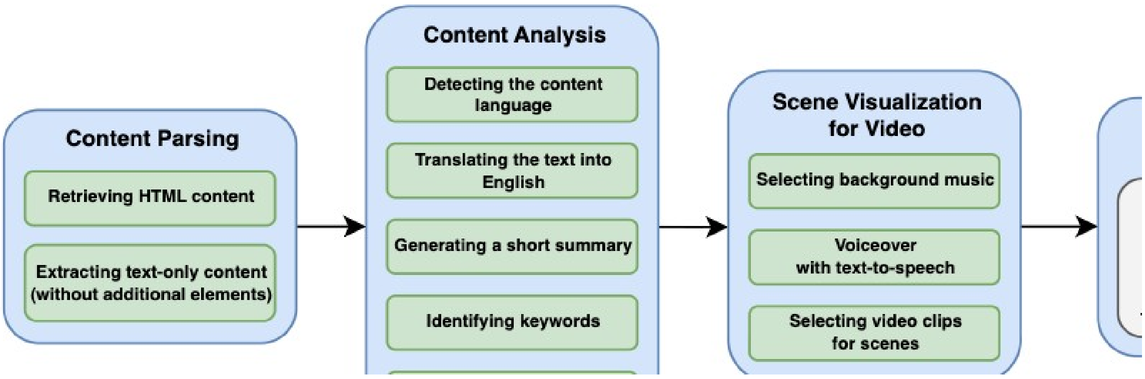

Content Parsing
Retrieving HTML content
Extracting text-only content (without additional elements)
Content Analysis
Detecting the content language
Translating the text into English
Generating a short summary
Identifying keywords
Scene Visualization for Video
Selecting background music
Voiceover with text-to-speech
Selecting video clips for scenes

**Fig. 2. Pipeline for text processing and final video generation**

The resulting summary-script undergoes the next layer of semantic analysis to enrich it with the context needed for video generation. Each component of this layer performs a clear function:

*1. Keyword Extraction:* the KeyBERT model, based on BERT embeddings, extracts keywords and phrases. Unlike classical TF-IDF, KeyBERT finds semantically significant terms, even if they are used infrequently. These keywords will become search queries for selecting stock video.

*2. Sentiment Analysis:* the *siebert/sentiment-roberta-large-english* model is used. It determines the overall emotional tone of the text (positive, negative, neutral). This metadata is critically important and is fed directly into the music generator.

*3. Classification:* a model based on the GRU architecture determines the content category (e.g., «Technology», «Politics», «Sports»).

*4. Named Entity Recognition (NER):* multilingual models based on *XLM-RoBERTa* are used. The model finds mentions of people (PER), organisations (ORG), and locations (LOC) in the text. This is one of the key accessibility features.

The formed set of metadata is critically important for the next stage – multimedia synthesis. The short summary is broken down into individual sentences, and each sentence becomes a separate scene in the future video.

For each scene, the system generates visual and auditory accompaniment. The visual sequence is formed by querying a stock video API, using the keywords obtained from KeyBERT for that scene. If nothing is found for the specific keywords of the scene, the system uses the «global» keywords extracted for the entire article, thus providing a reliable fallback mechanism and avoiding black screens. To improve accessibility, if named entities (NER) were detected in the scene, the system overlays the video with a corresponding image (e.g., a company logo or a person's photo), which the user can upload via the interface. This provides important visual context for users who may not know what the mentioned person or organisation looks like.

The auditory accompaniment consists of two layers. The first is background music, created using the generative model MusicGen [6]. To automate this creative process, the prompt for MusicGen is formed using another LLM (TinyLLaMA), which receives the category and sentiment as input and generates a description (e.g., «a somber, tense, atmospheric track suitable for news»). This description is fed into the MusicGen model. The second layer, which is a crucial accessibility enhancement, is synthesised speech. The text of each scene is voiced using the high-quality diffusion synthesis model StyleTTS2 [7]. Specific, trained models are used for this: *patriotyk/styletts2_ukrainian_single* for Ukrainian and *hexgrad/Kokoro-82M* for English. Unlike older concatenative or Tacotron-like models, StyleTTS2 generates natural speech with realistic intonations, which is critical for maintaining attention and preventing listening fatigue.

In the final stage, the final video file is assembled (Fig. 3). This process is implemented using the MoviePy library [8], which serves as a programmatic interface to *ffmpeg*. After forming the list of all scenes, they are sequentially joined. The generated background music from MusicGen is overlaid on the entire final video track as a separate audio track with reduced volume, so as not to overpower the narrator's voice. The final result is a complete, short, informative video clip that accurately conveys the essence of the original web article in a format accessible to a wide range of users.

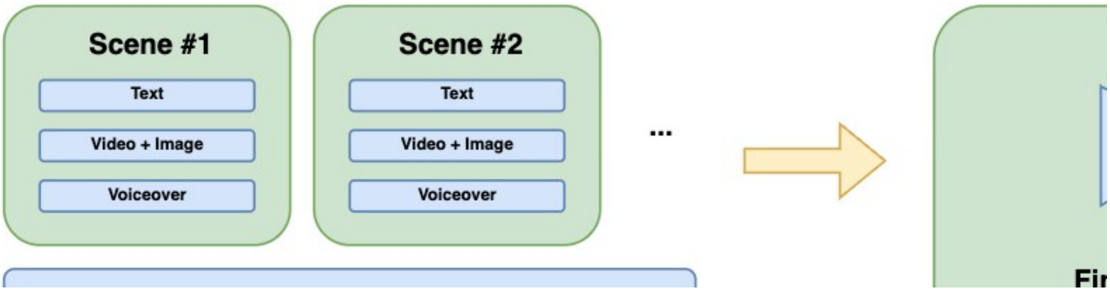


**Fig. 3. Diagram of combining all parts for the scenes**

Thus, the presented system is not just a converter of text materials into video, but a comprehensive digital accessibility bridge that combines technologies of artificial intelligence, cognitive psychology, and multimodal data analysis. Unlike traditional means of voice-over or subtitling, this system provides a deep transformation of content – not just a formal reproduction of text by voice, but the creation of a new semantic layer where meaning, intonation, rhythm, and visual images work in sync. It overcomes several barriers at once: cognitive – through abstractive summarisation, which compresses and structures information into an easy-to-perceive format; visual – through natural synthesised voice-over, which allows content to be received without visual contact; and contextual – through the use of named entities and images, which provide associative reinforcement and emotional understanding of the content.

In fact, such a system can be seen as a new level of user interaction with the information environment, in which the user ceases to be a passive consumer of text and becomes an active participant in a multimedia experience. The intelligent modules at its core are capable not only of reproducing data but also of interpreting it, determining significance, sentiment, and emotional context. This brings the technology closer to understanding natural information perception, where content is not just voiced but experienced through the integration of sensory channels.

Further development of the system should be linked to the improvement of hybrid data extraction methods that combine rules, statistical models, and deep learning. This will increase the accuracy of relevant content extraction even in complex, dynamic structures of constantly updated web applications. At the same time, a promising direction is the implementation of Explainable Artificial Intelligence (XAI) principles, which will make the system's operation more transparent to users. Thanks to this, the user will understand why a particular text fragment was chosen for voice-over or how the emotional accompaniment for a scene was formed.

In a broader context, such technologies become an element of the social infrastructure for digital inclusion. Their application can change approaches to online education, e-government, and media communications, ensuring equal access to information for all categories of users. In conditions of rapid societal digitalisation, such integration of artificial intelligence not only solves technical problems but also shapes the humanistic dimension of modern information and communication technologies, which become a tool for ensuring social equality.

## References

1. Shatylo I., Chala L. Using artificial intelligence technologies to create text-based videos // 29th International Forum of Young Scientists Radio Electronics And Youth In The Xxi Century: materials from the 29th International Youth Forum, Kharkiv, 16–19 April 2025. Kharkiv, 2025. Vol. 6. P. 87–89.

2. Lotfi C. et al. Web Scraping Techniques and Applications: A Literature Review / / Scrs Conference Proceedings On Intelligent Systems. 2021. P. 381–394.

3. Peykani P. et al. Large Language Models: A Structured Taxonomy and Review of Challenges, Limitations, Solutions, and Future Directions // Applied Sciences. 2025. Vol. 15, No. 14. P. 8103.

4. Zanini A. Web Scraping With Playwright and Node. 2025. Bright Data. [Електронний ресурс] // Режим доступу: https://brightdata.com/blog/how-tos/playwright-web-scraping.

5. Web article scraping, analysis & processing. Newspaper4k [Електронний ресурс] // Режим доступу: https://newspaper4k.readthedocs.io/en/latest/.

6. Wei L. et al. From Tools to Creators: A Review on the Development and Application of Artificial Intelligence Music Generation // Information. 2025. Vol. 16, No. 8. P. 656.

7. Aaron Y.Li et al. StyleTTS 2: Towards Human-Level Text-to-Speech through Style Diffusion and Adversarial Training with Large Speech Language Models // New Orleans, Red Hook, NY, USA, 2023. P. 19594–19621.

8. Automate Video Editing with MoviePy in Python. Best AI Tools Directory & AI Tools List – Toolify [Електронний ресурс] // Режим доступу: https://www.toolify.ai/ai-news/automate-video-editing-with-moviepy-in-python-1200309.