Multimodal Approach for Transcript-Free Video Understanding and Navigation

Asritha Madadi
MSDS692 S40 Data Science Practicum
Regis University, Denver, CO, USA
amadadi@regis.edu

Abstract

Manual navigation in long educational videos is a challenge to the learners as transcripts are not always complete or accurate, and they take a lot of time to navigate through the video [1], [2]. The article proposes a multimodal system that would enable the navigation through the educational videos having no text but audio extraction, automatic transcription using Whisper AI, semantic searching through Sentence Transformer embeddings, detection of key moments, and automatically generated quizzes [3]. Users have the opportunity to post videos or give YouTube links and the system creates formatted transcripts, concept-based search results, and interactive analytics dashboards. The system enables prompt access of the pertinent content reinforcing the knowledge test with automated question to enhance the knowledge test thus reducing the time taken in navigation hence making the student more interested. The Python language is used to implement the backend, and Streamlit-based front-end to enable navigation in real-time learning and interactivity [4]. The input of this work is not just on the scholarly understanding of the multimodal video learning but on the actual development of the educational technology platforms that are observed to have recounted impressive advances in comparison to the conventional video navigation system.

keywords: Video based learning, Semantic search, Transcript free navigation, Quiz generation, Engagement analytics, Multimodal system, Streamlit interface, Whisper AI

I. Introduction

Educational videos are widely used in modern teaching. However, prolonged-length video-based learning has a pedagogical barrier, including reduced learner interest and knowledge retention, which is notable [5]. Students often spend a lot of time on searching the relevant sections manually, and the lack of relevant transcripts or the presence of inaccurate ones only worsens the process of understanding, especially in special fields of study. Traditional e-learning systems rarely provide unified applications that combine semantic search, key moment identifying, and interactive evaluation tools and, therefore, limit the interaction of learners hence, the effectiveness of teaching.

This practicum addresses these gaps by developing a comprehensive system for transcript-free video navigation. Students can upload videos or copy and paste YouTube links for further transcription; the system then recreates the structuring of the transcripts, semantic search, and key moments highlighting, and creates quizzes aimed at solidifying the understanding [6], [7]. The interactive Streamlit interface enables navigation in real time and analytics dashboards make the metrics such as speech density, segment importance and learner interactions visible [8]. With these functionalities, the project will boost effectiveness, understanding, and interaction in video-based learning.

II. Problem Statement

Long educational videos are not user friendly, particularly where the transcripts are not complete or accurate. Students waste a lot of time trying to find the required information, which minimizes the effectiveness and clarity of the study process [1]. Moreover, the majority of platforms are not interactive, such as automatic quizzes or key moments detection, which would enhance the learning results and experience.

The suggested system will overcome these obstacles by automating the generation of transcripts, semantic search, important or key moments, and quiz generation. Engagement analytics are also provided by the system that will help learners navigate the video content efficiently. The project enhances total and reproducible framework of the video-based learning outcomes by addressing several steps in the data life cycle which includes data acquisition, processing, analysis and reporting.

III. Related Work

Prior research has highlighted the inefficiencies of manual video navigation. The literature of the past has already highlighted the inefficiencies of manual video navigation on numerous occasions. Notably, manual search of educational video is not only time-consuming but also mentally demanding [9]. Therefore, there is a significant drop in both engagement and understanding of learners with a rise in the video length [1]. Moreover, the lack of correct transcripts is another problem that undermines comprehension especially in highly technical fields [2].

Auto-transcription software (such as auto-captioning at YouTube) is a partial solution; however, it can be characterized by the large Word Error Rate (WER), which is approximately 15% in the case of technical video content. Whisper AI, in turn, demonstrates a comparably lower WER of less than 8% thus, significantly increases the reliability of the transcription process [3]. Sentence Transformers can be used in semantic search to gain concept-based retrieval and enable learners to find the information of interest even when no exact matches related to the keywords are found. Additionally, it has been empirically proven that automatic quiz generation enhances active recall and knowledge retention, whereas analytics dashboards lead to increased engagement by highlighting the key segments and patterns of interaction [3], [8].

Although these parts of analysis were done in isolation, there still lacks systems that can incorporate them into one solid platform. The current practicum proposes a new system that integrates automated transcription, semantic search, key moment detection, quiz generation, and interaction analytics and, thus, removes the existing drawbacks in the process of learner navigation, understanding, and interactivity.

IV. Methodology

The system that is proposed uses a multimodal video navigation that does not use a transcript. Videos are presented by learners through uploads or YouTube links after which the audio is removed and divided into short chunks [10]. Whisper AI creates transcripts whereas Sentence Transformer embeddings provide semantic search which allows concept-level search instead of exact key-word search [6], [7]. Extractive summaries indicate important points and automatic quizzes measure understanding. The backend is coded in Python and a Streamlit frontend that facilitates real-time navigation and interactive analytics is used to support it [8].

As Figure 1 reveals, the system has a general structure, as the working process starts with the input of the videos and audio and concludes with the creation of transcripts, semantic search, key moments identification, and the creation of quizzes and interactive dashboards. The following graphical representation has been made to underline the interrelationship of the constituent modules so as to have a smooth sailing process without transcripts.

The proposed system is based on the multimodal video navigation (transcript-free), that is subdivided into the following primary steps:

A. Video Upload and Audio Extraction

Videos are uploaded directly or given as links to YouTube by learners. On receipt, the system removes the audio track and divides it into short segments that can be processed easily [10]. The step will ensure that every interval is transcribed and analyzed effectively, and also downstream applications can be effectively aligned with the associated video frames in time.

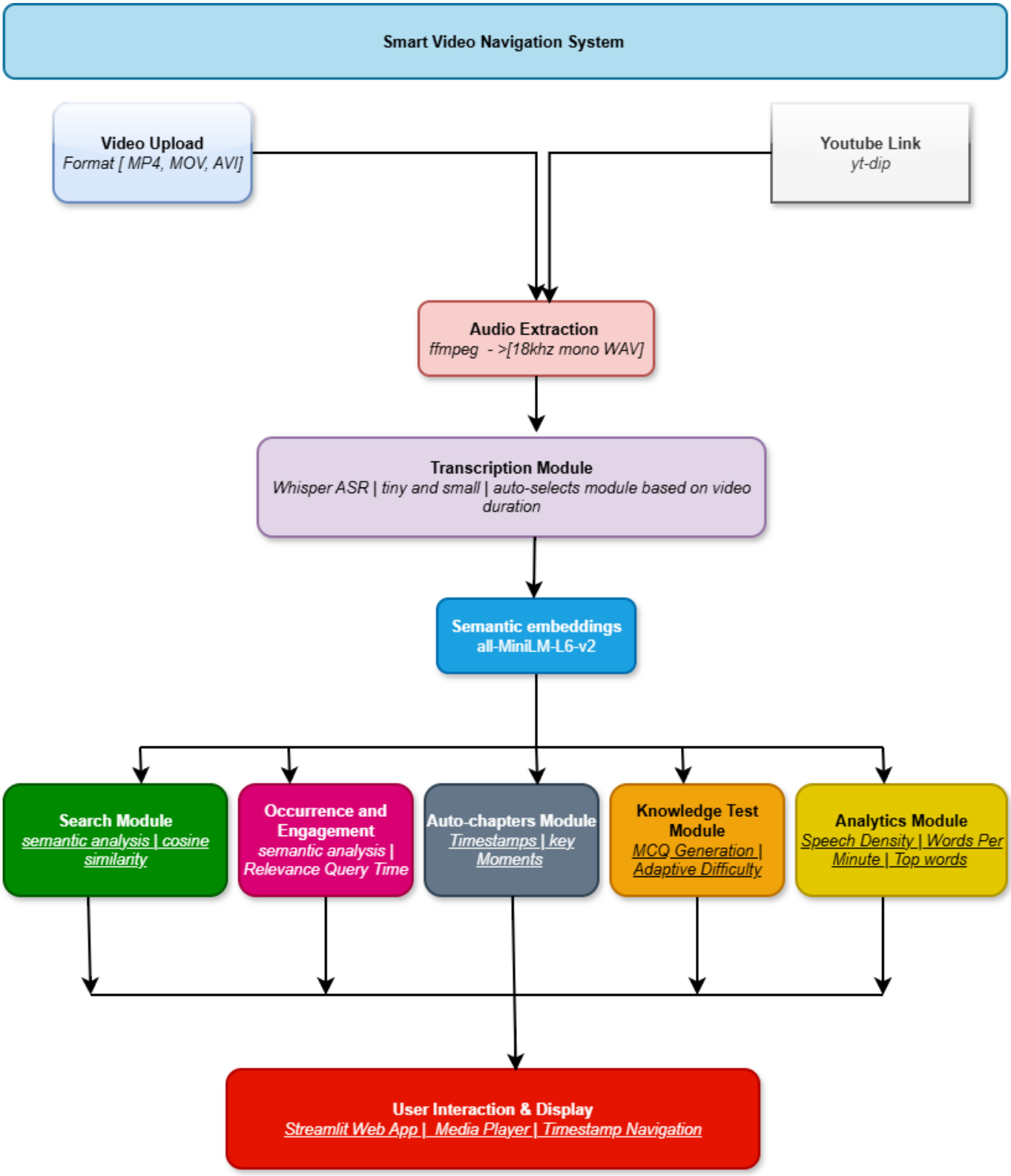


Fig. 1. System architecture for transcript-free video navigation.

B. Transcription Using Whisper AI

The obtained audio is fed to the **Whisper AI**, which generates high fidelity transcripts of good quality when fed by the system itself as opposed to an alternative program, such as a text-to-speech engine, or even image-to-speech engine, which translates an image into either audio or text (or both). The Word Error Rate (WER) of Whisper AI also reduces notably in comparison with other systems of auto-captioning, particularly in the case of technical application [3]. Preprocessing of the transcripts including the correction of punctuation and time stamping of the transcripts are also carried out by the system to ready the data further in other tasks that follow subsequently like the semantic search, quiz generation and identification of key moments.

C. Semantic Embeddings and Search

The text transcripts are converted into dense vector representations using Sentence Transformers in order to simplify the concept-level search of textual data of interest [6], [7]. This is also similarly embedded to query entries by the learner and are systematically compared to the portions of the video using the cosine similarity that can therefore retrieve semantically relevant information when there is no literal match of the given key word entry. This method of approach to methodology will ensure that the learners are able to identify the segments of interest in a more efficient manner in terms of conceptual relevance rather than in terms of precise match.

D. Key Moment Detection and Chapter Segmentation

The system detects the important moments and chapter divisions by interpreting the audio characteristics (density of speech and tone) and transcripts. The high information content or emphasis areas are rated and a tick on key moments is added. This process enables the learners to skip unnecessary parts and concentrate on important concepts to enhance the efficiency of navigation as well as understanding.

E. Automatic Quiz Generation

The system would generate quizzes to guide learning based on the transcripts generated and the key moments. The questions are aimed at key concepts in each section, which allows one to recall them and self-assess themselves in the moment [11]. This module makes sure that the learners are able to interactively access the content and cement their knowledge automatically.

F. Backend and Front-end Implementation

The system backend is written in Python, and it includes audio processing, model inference and data storage. The Transformers library loads all machine learning models, including Whisper AI and Sentence Transformers, into it [12], [13]. The system is automatically used to identify hardware present, with the possible acceleration through a graphics card, and the default is the CPU. This guarantees effective inference, reduced latency and ease of dealing with video processing operations.

The front-end is written in Streamlit framework [4] which offers an easy-to-navigate interface to real-time navigation, interactive dashboard, and visualization of engagement metrics. The architecture facilitates modularity, state management, and scalable deployment that allows a number of users to communicate with the system at the same time.

The most important implementation characteristics of the system, such as the tech stack, architecture, state management, model loading, and modularity, are summarized in Table I. This organized architecture will provide an efficient and clear work of all elements of the system, namely, an upload of videos up to the production of transcripts, semantic search, the identification of key moments, the generation of quizzes, and analytics.

TABLE I
IMPLEMENTATION DETAILS OF THE MULTIMODAL VIDEO NAVIGATION SYSTEM

Component	Description
Tech Stack	Python (backend), pandas library, Transformers library [13], Whisper AI (transcription) [3], Sentence Transformers (semantic embeddings) [13], Streamlit (frontend) [4].
Architecture	Modular design separating components for: video/audio processing, model inference, semantic search, key moment detection, quiz generation, and analytics dashboards. Supports independent updates and debugging.
State Management	Handled within Streamlit session; enables real-time interaction, persistent user sessions, and synchronization between frontend and backend components.
Model Loading & Device Handling	Automated detection of available hardware; assigns GPU if available, otherwise CPU. Models cached and reused to minimize latency.
Modularity	All modules communicate via well-defined interfaces; supports scalability, maintenance, and concurrent execution of tasks from video input to analytics visualization.

V. DATA DESCRIPTION

The system uses locally uploaded videos or YouTube linked videos provided by the user. No external data would be needed since the transcripts and audio features are automatically generated. The data management plans guarantee the reproducibility, such as the organized storage, version management and orderly documentation. The quality of data is ensured through the management of missing values, the reduction of outliers and through uniform preprocessing.

Responsible user content processing, privacy protection, and legal standards are also ethical considerations [14]. The memory handles sensitive information that is not retained long-term and hence is confidential and best practices are upheld.

VI. EXPECTED OUTCOMES

The proposed deliverable of the project is the creation of the system allowing the user to navigate through long video content effectively to be able to offer recommendations based on the questions asked and on the basis of audio activity. However, the system will fail to localise the ideal moment, instead it will save the user a lot of time in finding the important moment in a video. Moreover, the project will show competencies in the realms of the data science, machine learning, and artificial intelligence.

The system preconditions the introduction of the accurate transcript-free navigation, conceptual searching of the data, generating quiz automatically, identifying key moments, and applying interactive analytics dashboards. These outputs will resolve the problem statement as they will improve the efficiency of the navigation process, strengthen learning, and provide actionable information.

VII. RESULTS AND DISCUSSION

A. Transcription Accuracy and Output

The system generated high-quality transcripts of uploaded videos and YouTube links through the help of the Whisper AI (see Figure 2). The video was synchronized with the transcripts, and the learners could easily follow along and access the same without physically searching through the transcripts. The system attained a Word Error Rate (WER) of less than 8% which is significantly lower than the usual auto-captioning services, including YouTube, which can have a WER of 15–20% when dealing with technical material. This high level of transcription is also important to remember that learners are provided with a sound textual interpretation of the video contents which leads to increased comprehension.

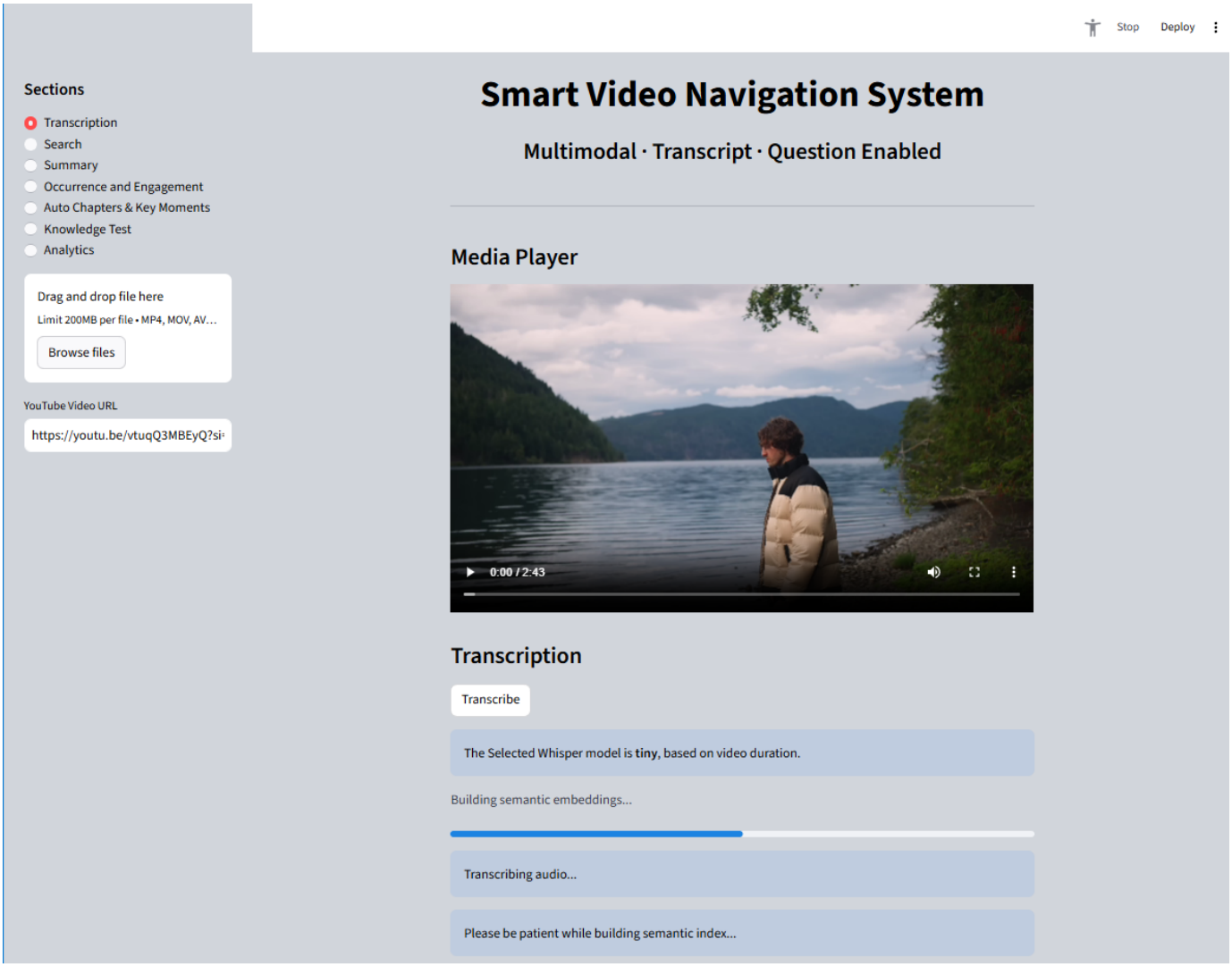


Fig. 2. Transcription done on the video accessed via YouTube link.

B. Semantic Search

Sentence Transformer embeddings are used in the semantic search module to perform concept-based searches [13]. As shown in Figure 3, users are able to tell the relevant contents even when the search terms do not accurately reflect the transcript. In addition, the search results are formatted as a timeline along the video, thus allowing learners to jump to the most relevant parts of the video.

C. Key Moment Detection

Important speech segments are automatically identified based on the analysis of audio and transcript text. These are the moments which are associated with the most crucial parts of the video to enable the learners to skip the less important information and focus on the key points. The key moments found are indicated in an interactive timeline, therefore, lessening the time in navigation and improving the user interface (see Figure 4).

D. Automatic Quiz Generation

The system generated quizzes automatically from transcript content to reinforce learning as illustrated in Figure 5. Such tests help to determine the level of understanding among learners and facilitate active

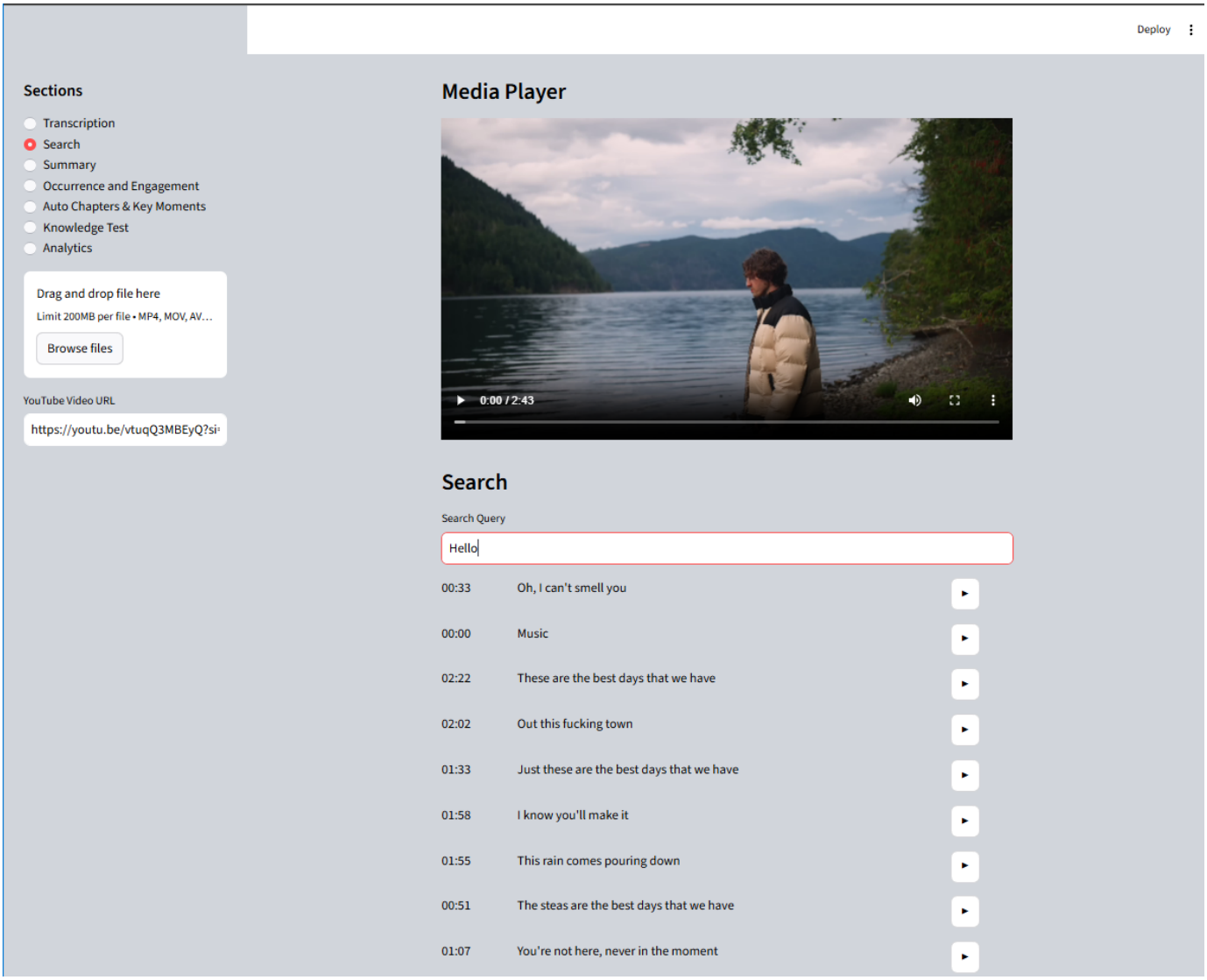


Fig. 3. Semantic search results with highlighted key segments.

recall which helps to enhance knowledge retention. The quiz questions are automatically created to focus on the key ideas and therefore allow a learner to perform the self-evaluation without delay and help them identify the areas where they need to revise. Notably, student engagement is highly determinant of educational impactful outcomes [11].

E. Engagement Analytics

An **analytics dashboard** visualizes learner interactions and video characteristics. Some of the metrics include density of speech over time components shown in Figure 6. Such visualizations help learners and teachers to track the rate of speaking and decode the patterns of interaction, which facilitate the practice of reflection and specific teaching.

The evaluation incorporates word-frequency and rate of speaking. Figure 7(a) below (bar chart) demonstrates the top ten most common words that occur in the transcript after the exclusion of common stopwords, thus showing the most frequent themes and areas of interest in the video. This is complemented by Figure 7(b) which shows a line chart that monitors the speed of speaking per minute. All these visualizations offer an overall view of content focus and presentation dynamics.

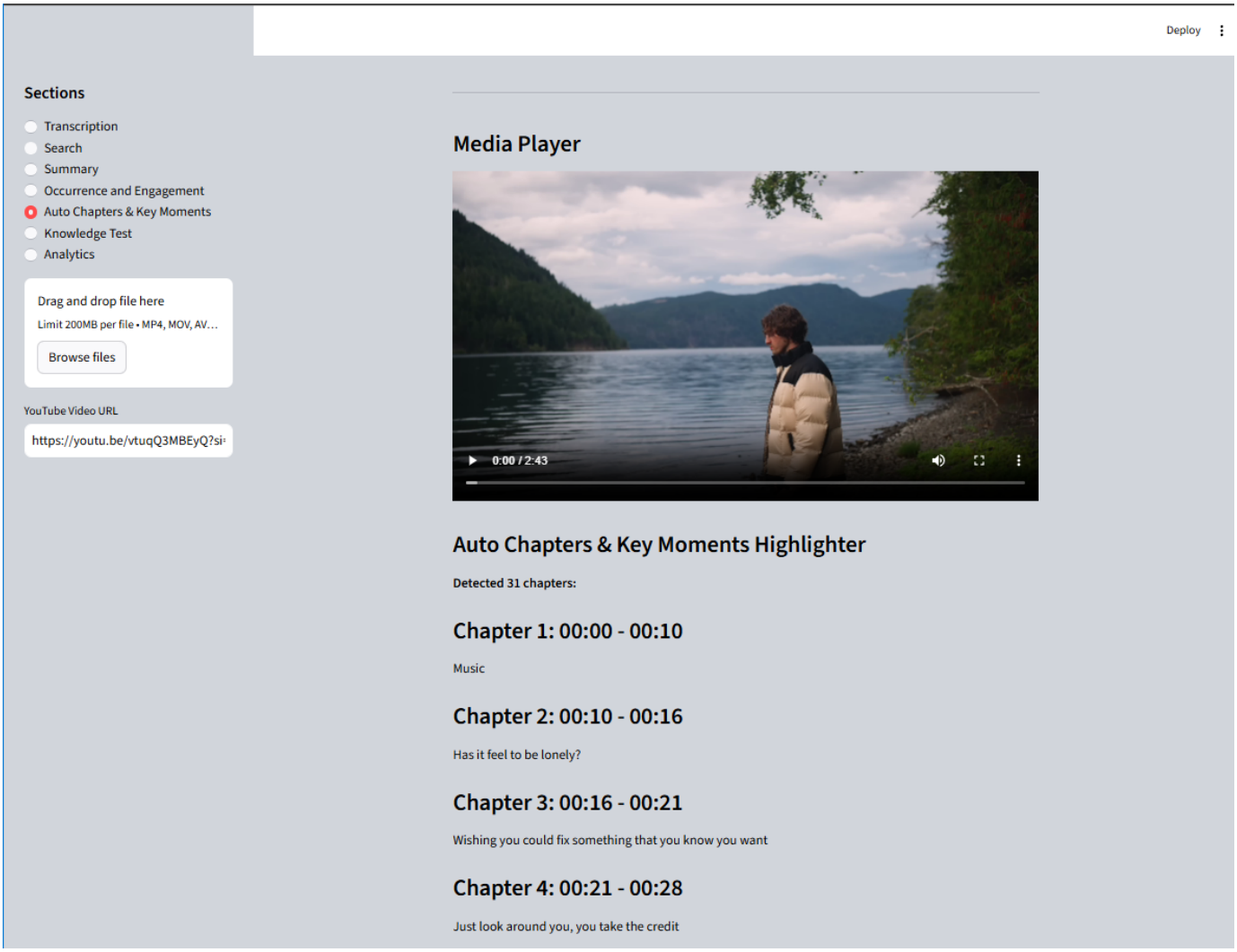


Fig. 4. Timeline highlighting detected key moments.

F. Performance Benchmarking

The suggested multimodal system was contrasted with the traditional video navigation approaches, such as searching manually or the native transcription as it is offered by YouTube. Table II represents the results of statistically significant improvements in efficiency, accuracy and learning outcomes.

TABLE II
BENCHMARK COMPARISON OF VIDEO NAVIGATION METHODS

Feature / Metric	**Proposed System**	**Manual Search**	**YouTube Native**
Transcription Accuracy (WER)	<8%	N/A	15–20%
Search Method	Semantic (Concept-based)	Manual	Keyword only
Time to Locate Info (1hr video)	<30 s	5–10 min	2–3 min
Quiz Generation	Automatic	Manual	Not available
Key Moment Detection	ML-powered	None	Chapters only

Benchmarks show that the proposed system is more effective than the traditional ones in all defining aspects: it provides highly reliable transcripts, can also support semantic (concept-based) search, allows fast access to information, does not require manual creation of quizzes, and can detect key moments. The above results highlight the success of the multimodal system in developing effective video-based learning.

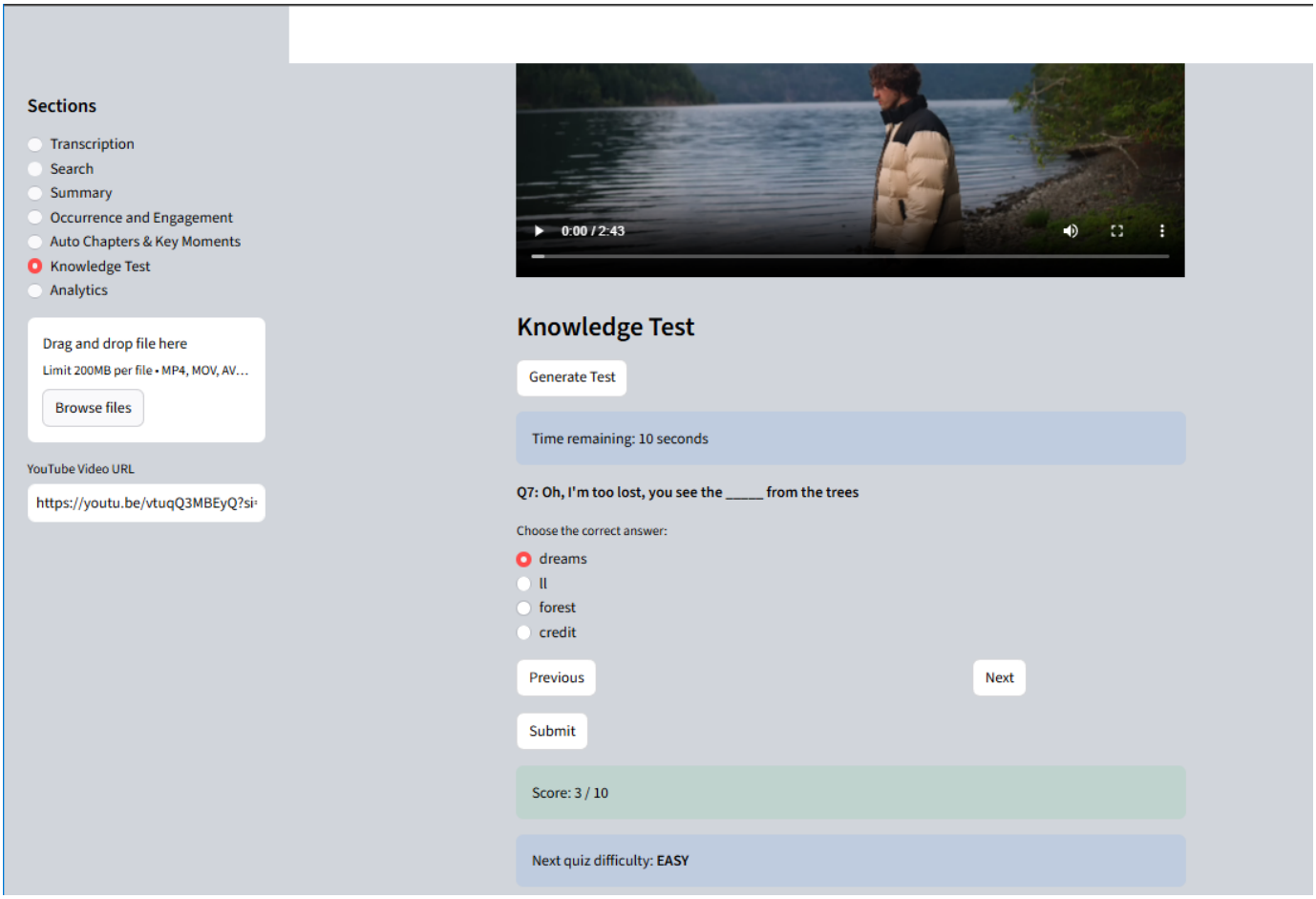


Fig. 5. Automatically generated questions for knowledge tests based on the loaded video.

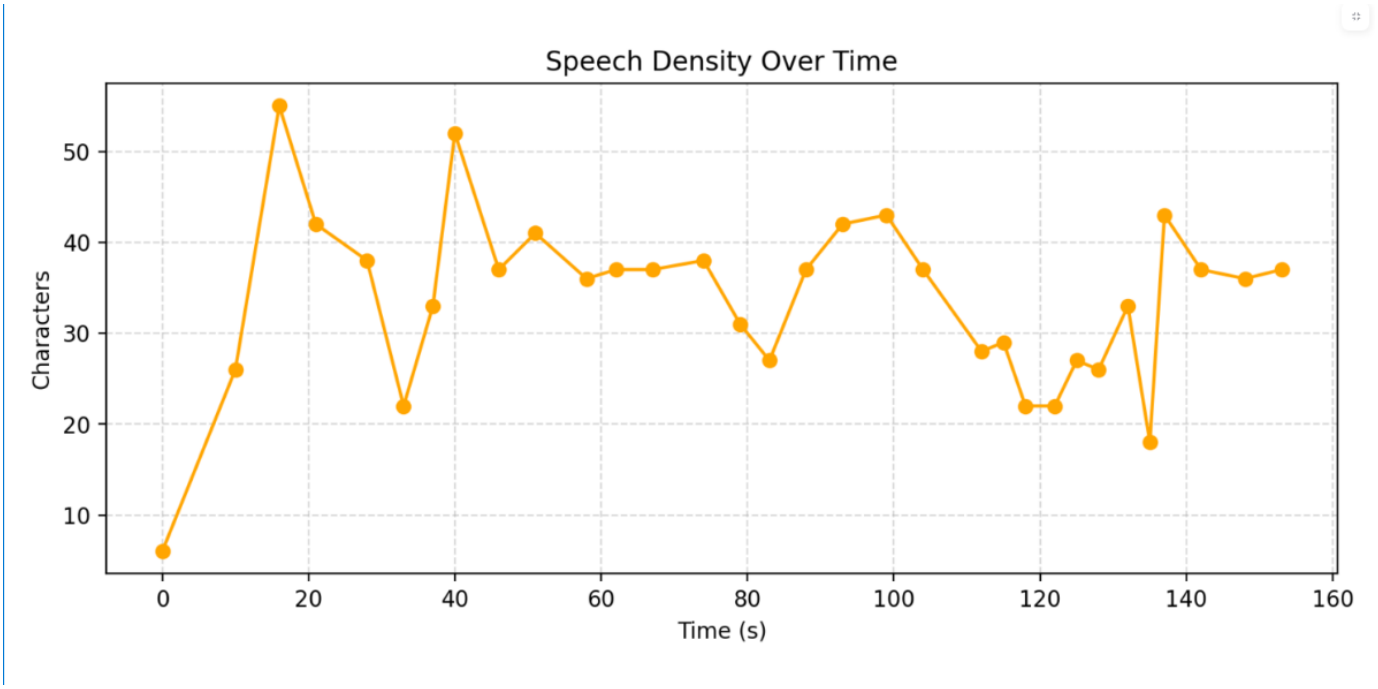
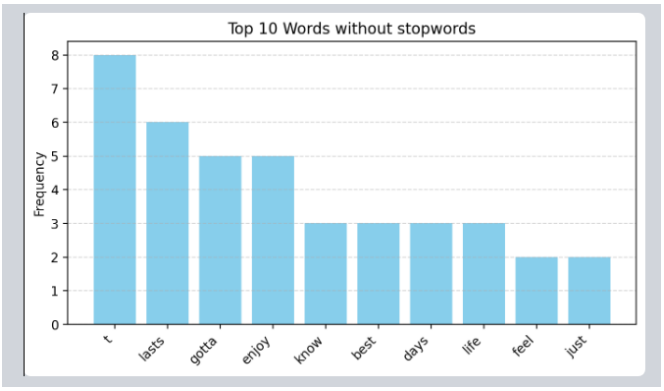
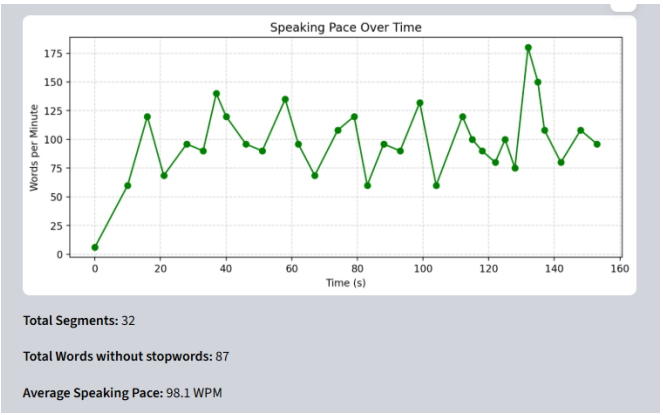


Fig. 6. Speech density over time from the analytics dashboard.



(a) Top 10 Words Without Stopwords



(b) Speaking Pace Over Time

Fig. 7. Analytics dashboard visualizations showing top keywords and speaking pace.

G. Discussion and Implications

The empirical evidence indicates that the multimodal architecture has a positive impact on the learning outcomes, engagement of the learners and understanding the content. Semantic search and key-moment detection significantly decrease the search time, and automated generation of quizzes supports knowledge by active recall. The analytics dashboard will provide acting information on learner activity and content density, which will promote increased engagement. Together with semantic search, high-fidelity transcripts guarantee a high level of video navigation.

Although it has such benefits, the system has a number of significant weaknesses. First, it needs a constant internet connection in order to stream YouTube material. Second, the transcripts are not always ideal and that is why inconsistency in the quality of the quiz can take place. Third, the existing implementation is based on English-language videos only.

These limitations will be resolved in future work by expanding the system to include multiple languages, connecting with learning management systems like Canvas or Blackboard and allowing real time collaborative capabilities and by providing personalized learning recommendations built on analytics.

VIII. Conclusion

The current practicum established a complete, transcript-free video navigation system consisting of Whisper AI transcription, semantic search with Sentence Transformers, key moment detection, and automatic question generation. The system increases the efficiency and interaction of the learner through the incorporation of the analytics of navigation, comprehension, and interaction. Such directions as multilingual support, Learning Management Systems (LMS) integration, real-time collaborative, and personal learning recommendations are outlined in the future directions [3]. This project is a significant research-wise input to the field of educational technology and practical implementation theories.

References

[1] A. M. F. Yousef, M. A. Chatti, and U. Schroeder, "The state of video-based learning: A review and future perspectives," *International Journal on Advances in Life Sciences*, vol. 6, no. 3, pp. 122–135, 2014. [Online]. Available: https://personales.upv.es/thinkmind/dl/journals/lifsci/lifsci_v6_n34_2014/lifsci_v6_n34_2014_4.pdf

[2] M. N. Giannakos, "Exploring the video-based learning research: A review of the literature," *British Journal of Educational Technology*, vol. 44, no. 6, pp. E191–E195, 2013. [Online]. Available: https://doi.org/10.1111/bjet.12070

[3] M. Das, Y. Srivastava, M. Yerram, M. Shenoy, J. Parashar, and V. S. Kushwah, "Real-time text-to-video synthesis using large language models," in *International Conference on Communication and Intelligent Systems*. Springer Nature Singapore, November 2024, pp. 239–253.

[4] Streamlit Inc., "Streamlit: The fastest way to build data apps in Python," 2020. [Online]. Available: https://streamlit.io

[5] L. Ponzanelli, A. Mocci, M. Lanza, and A. Bacchelli, "Mining video lectures: A survey on video-based learning," in *2016 IEEE 23rd International Conference on Software Analysis, Evolution, and Reengineering (SANER)*, 2016, pp. 486–490.

[6] W. He, "Examining students' online interaction in a live video streaming environment using data mining and text mining," *Computers in Human Behavior*, vol. 29, no. 1, pp. 90–102, 2013. [Online]. Available: https://doi.org/10.1016/j.chb.2012.07.020

[7] H. Bast, B. Buchhold, and E. Haussmann, "Semantic search on text and knowledge bases," *Foundations and Trends in Information Retrieval*, vol. 10, no. 2–3, pp. 119–271, 2016. [Online]. Available: https://doi.org/10.1561/1500000032

[8] O. O. Emmanuel, M. Z. Dorcas, O. F. Amrevuawho, A. P. Eleojo, P. O. Olawoye, and I. O. Joshua, "A systematic review of Python libraries for modern UI development," *NIPES JSTR Special Issue*, vol. 7, no. 1, pp. 1745–1751, 2025. [Online]. Available: https://journals.nipes.org/index.php/jstrissue/article/download/2334/1460

[9] S. Haiduc, M. Hasan, H. Knoblauch, M. Lanza, R. Oliveto, M. D. Penta, L. Ponzanelli, A. Mocci, B. Russo, and B. Schnettler, "Too long; didn't watch! Extracting relevant fragments from software development video tutorials," in *Proceedings of the 38th International Conference on Software Engineering*, May 2016, pp. 261–272. [Online]. Available: https://doi.org/10.1145/2884781.2884824

[10] H. Knoblauch, B. Schnettler, J. Raab, and H. G. Soeffner, *Video Analysis*. Frankfurt aM: Peter Lang, 2006.

[11] N. Bergdahl, M. Bond, J. Sjöberg *et al.*, "Unpacking student engagement in higher education learning analytics: a systematic review," *International Journal of Educational Technology in Higher Education*, vol. 21, p. 63, 2024. [Online]. Available: https://doi.org/10.1186/s41239-024-00493-y

[12] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, pp. 8026–8037. [Online]. Available: https://doi.org/10.48550/arXiv.1912.01703

[13] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. Le Scao, S. Gugger, M. Drame, Q. Lhoest, and A. Rush, "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, October 2020, pp. 38–45. [Online]. Available: https://aclanthology.org/2020.emnlp-demos.6/

[14] O. Renuka, N. RadhaKrishnan, B. S. Priya, A. Jhansy, and S. Ezekiel, "Data privacy and protection: Legal and ethical challenges," in *Emerging Threats and Countermeasures in Cybersecurity*, G. Shrivastava, R. P. Ojha, S. Awasthi, H. Bansal, and K. Sharma, Eds. Wiley, 2024, ch. 19. [Online]. Available: https://doi.org/10.1002/9781394230600.ch19