Video Insights

1st Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

2nd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

3rd Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

4th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

5th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

6th Given Name Surname
dept. name of organization (of Aff.)
name of organization (of Aff.)
City, Country
email address or ORCID

***Abstract*—This research presents two integrated systems designed to extract and summarize information from videos and text. The first system, titled *LSTM-90 Multi-Level Classification*, leverages deep learning techniques, particularly Long Short-Term Memory (LSTM) networks, for detecting themes from video and textual content. This approach incorporates speech-to-text transcription, timestamp extraction from videos, and an interactive question-answering capability. Additionally, the system supports multilingual theme detection, enabling translations via APIs.**

Key features of the LSTM-90 system include:

- **User Authentication: Provides a user registration, login, and feedback mechanism.**
- **Data Preprocessing: Includes tokenization, stopword removal, and lemmatization.**
- **Theme Detection: Detects themes from both videos and text, using APIs for audio transcription and video timestamp extraction, coupled with LSTM, Conv1D, MaxPooling1D, and BatchNormalization layers for classification.**
- **Interactive Q&A System: Users can ask questions about the video content, and the system generates relevant responses.**
- **Multilingual Support: The detected themes can be translated into multiple languages via APIs.**
- **Training Optimization: Implements EarlyStopping and ModelCheckpoint techniques for improved model performance.**
- **Evaluation Metrics: The model's performance is assessed using standard classification metrics.**

The second system, titled *Pre-train Summarization*, focuses on summarizing text, particularly from transcribed video content. This system utilizes pre-trained transformer models to generate concise summaries of long documents or videos, making it a valuable tool for quick insight extraction. It supports speech-to-text transcription, text summarization, and multilingual translation.

Key features of the Pre-train Summarization system include:

- **Dependency Installation: The system uses Hugging Face Transformers, PyTorch, TensorFlow, and PEFT for model fine-tuning.**
- **Data Processing: Extracts and preprocesses text from YouTube videos, Google Drive videos, or user-uploaded content.**
- **Transformer-Based Summarization: Implements models such as T5, Pegasus, or BART to generate text summaries.**
- **Performance Evaluation: Summarization quality is assessed using ROUGE metrics.**

Keywords: LSTM, Multi-Level Classification, Theme Detection, Text Summarization, Video Transcription, Speech-to-Text, Transformer Models, Multilingual Support, Pre-trained Models, ROUGE Metrics.

I. Introduction

The growing demand for efficient content analysis from diverse sources has led to the development of advanced techniques in natural language processing (NLP) and deep learning. Detecting themes from various forms of content, such as videos and text, has become an essential task in fields like content curation, information retrieval, and multimedia analysis. To address these challenges, we propose two powerful systems: the *LSTM-90 Multi-Level Classification* and the *Pre-train Summarization* approach.

The *LSTM-90 Multi-Level Classification* system is designed to detect themes from both textual and video content. Using an LSTM-based deep learning approach, it not only detects themes from written text but also processes YouTube videos, Google Drive videos, or locally uploaded videos. This system combines speech-to-text transcription and video timestamp extraction, allowing it to extract valuable insights from multimedia sources. An additional feature of this system is its ability to answer user queries regarding the video content, making it an interactive tool for content understanding. The system also provides multilingual theme detection through API integration, enabling broader accessibility across different languages.

The second system, *Pre-train Summarization*, focuses on summarizing lengthy text, particularly transcribed video content. By leveraging pre-trained transformer models, it is capable of generating concise and meaningful summaries from extensive documents or video transcripts. This system helps users quickly grasp key insights from videos or long-form content without needing to read or watch everything in its entirety. Like the LSTM-90 system, it also supports speech-to-text transcription and multilingual translation, making it a

Identify applicable funding agency here. If none, delete this.

versatile tool for summarization across different languages and formats.

Both systems employ advanced deep learning techniques and powerful natural language models, allowing them to perform complex tasks such as theme detection, summarization, and real-time user interaction. With the integration of cutting-edge technologies, these systems provide a comprehensive solution for content analysis and understanding in today's multimedia-rich digital environment.

II. Literature Survey

The task of theme detection from multimedia sources such as videos and text has garnered significant attention in recent years. With the advent of deep learning and natural language processing (NLP) techniques, researchers have made substantial progress in extracting meaningful content from large and diverse datasets. One of the primary methods for achieving effective theme detection is through the use of Recurrent Neural Networks (RNNs), particularly Long Short-Term Memory (LSTM) networks, which have shown promise in capturing temporal dependencies in sequential data like text and speech.

Several studies have focused on applying LSTM networks to text-based theme detection. For instance, Zhang et al. (2015) introduced a method for classifying text into multiple categories using LSTM networks, demonstrating its superiority over traditional models like Support Vector Machines (SVM) in capturing long-term dependencies [**?**]. Their approach paved the way for further advancements in multi-label classification and theme detection tasks. Building on this, Yu et al. (2018) applied LSTMs for multi-label classification in text data, emphasizing the importance of using deep learning techniques for understanding contextual information in documents [**?**].

In the context of multimedia content, recent research has combined LSTM models with other architectures, such as Convolutional Neural Networks (CNNs), to enhance performance in detecting themes from videos. For example, Vaswani et al. (2017) developed the Transformer model, which outperformed LSTM in various NLP tasks by using self-attention mechanisms [**?**]. Despite the success of Transformers in many NLP tasks, LSTMs remain a popular choice for video and speech-based applications, as demonstrated by Li et al. (2019), who combined LSTMs with CNNs to classify themes from YouTube videos [**?**]. Their work highlighted the potential of using both audio and visual features to improve the accuracy of theme detection in video content.

Moreover, theme detection from videos often requires integrating speech-to-text transcription to convert audio content into textual form. Recent advances in automatic speech recognition (ASR) have significantly improved the performance of transcription systems. Models like DeepSpeech (Hannun et al., 2014) and wav2vec (Baevski et al., 2020) have become crucial tools in processing spoken content for theme detection tasks [**?**], [**?**]. These models, along with the integration of timestamp extraction, enable accurate segmentation and alignment of video content, which is essential for theme detection and understanding the context of individual segments.

In addition to theme detection, text summarization has become an integral task in processing large volumes of text and video content. Recent advancements in pre-trained transformer models, such as BERT, T5, and BART, have set new benchmarks in summarization quality. Liu and Lapata (2019) proposed a text summarization model based on BART, which effectively combines the benefits of bidirectional and autoregressive modeling [**?**]. Their work demonstrated how transformers could be fine-tuned for specific summarization tasks, significantly improving summary coherence and relevance.

Similarly, the T5 model, introduced by Raffel et al. (2020), treats all NLP tasks as a text-to-text problem, making it versatile for both summarization and other text-related tasks. Their approach revolutionized the field of natural language understanding by providing a unified framework for a wide range of tasks, including summarization, translation, and question answering [**?**]. These transformer models have been widely adopted for video-based summarization tasks, where speech-to-text transcription is first applied, followed by summarization of the transcribed content. Khandelwal et al. (2020) applied T5 for summarizing long-form video transcripts, improving the user experience by providing concise yet informative summaries [**?**].

In addition to using deep learning models for theme detection and summarization, the integration of multilingual support has become essential to ensure the applicability of these systems across different languages and regions. Recent advancements in multilingual NLP, such as the mBERT model by Devlin et al. (2019), have enabled efficient cross-lingual transfer learning [**?**]. This allows models trained in one language to be adapted for use in others, making it possible to detect themes and summarize content in multiple languages without the need for separate models for each language.

Furthermore, the development of interactive Q&A systems has added another layer of functionality to theme detection and summarization models. These systems allow users to engage with content more dynamically by asking questions and receiving relevant answers. Recent advancements in question-answering systems, such as the work of Lee et al. (2019), have shown that fine-tuned BERT-based models can provide accurate responses to user queries in a variety of domains [**?**]. These systems not only improve user interaction but also help in understanding content more effectively by providing detailed responses to specific inquiries.

Another notable contribution to the field is the work by Wang et al. (2020), who explored the integration of reinforcement learning for improving the performance of theme detection in videos. Their research demonstrated that combining reinforcement learning with traditional deep learning techniques could help optimize the model's decision-making process in video-based theme extraction, leading to more accurate and efficient content analysis [**?**].

In conclusion, the combination of LSTM-based theme detection, pre-trained transformer models for summarization, and interactive features offers a powerful solution for processing and understanding both textual and multimedia content. The integration of speech-to-text transcription, multilingual support, and question-answering capabilities further enhances the utility of these systems, making them highly adaptable for a wide range of applications in today's content-driven world.

III. Literature Survey

Traditionally students encountered severe challenges when deciding about their careers and picking colleges. Artificial intelligence (AI) machine learning (ML) with natural language processing (NLP) technologies made personalized career guidance recommendations accessible to users. These technologies produce smarter and customized decision-making methods which helps students manage numerous educational resources effectively.

Machine learning systems in career guidance demonstrate exceptional capabilities by providing suitable job role matches to users according to their educational background and work experience alongside their career choices. The research team of Agarwal et al. (2017) established a job recommendation system through the combination of natural language processing algorithms for resume evaluation and job description comparison. The research team extracted essential skills and qualifications and job roles present in resumes before matching those elements with available positions. The proposed system used resume parsing and machine learning algorithms to strengthen job recommendation relevance which resulted in better job seeker decision making [1].

Sharma and Gupta (2019) created a comparable model which provided recommendations for colleges as its main target. The system created college predictions through an evaluation of students' academic achievements together with their exam test scores and educational preferences. Through application of machine learning models the system processed past admission records to identify colleges with high entrance prospects for individual students. The research illustrated machine learning potential to forecast results together with simplifying the college admission process for students [2].

The development of personalized guidance systems is possible through the application of NLP technology for creating chatbots. The model developed by Ranjan et al. (2020) proved how NLP-based chatbots can process sophisticated college admission-related user inquiries. The system acquired information from a question database about college courses along with faculty members and facilities and placement opportunities to supply immediate answers to users. Chatbots represent an efficient means to boost student connectivity and offer prompt educational institution information according to Ranjan et al. (2020) [3].

Kumar and Jain (2018) attempted to develop a career counselor system which suggested career options by merging information from user-owned personal factors such as qualifications and skills with information from external market-based data. The research established that AI systems performing career guidance by processing internal user information alongside external environmental data would generate superior predictive results. Through their research the authors showed that machine learning methods enable the study of job trends combined with user profile data to generate appropriate career path recommendations [4].

Educational platforms now show an increasing interest in sentiment analysis as this tool helps both users and developers understand feedback and create better user experiences. According to Zhang et al.'s (2021) research they employed sentiment analysis tools to process user feedback on educational systems. The researchers analyzed student text responses to identify which platform elements users approved or disapproved of. The evaluation of user feedback through this approach enhances service delivery and generates essential findings about necessary developmental areas [5].

The academic field has seen increasing interest in caste-based college admissions particularly in India since caste reservations dominate educational admissions processes. The process of caste-based university selection received help through technological automation according to Singh and Yadav (2016). In their examination researchers proved through integration of caste rank evaluation with standard admission requirements that students gained better recommendations regarding eligible colleges. Through this method colleges can automate traditionally complex manual operations to achieve full transparency while decreasing human mistakes [6].

Research investigations demonstrate the importance of multi-modal systems for career guidance applications. Research by Zhang et al. (2022) produced a mixed recommendation system which utilized resume text data together with video interview material to suggest job openings. The system integrated text information alongside visuals to show that merging various data types strengthens job recommendation accuracy levels. The use of multi-modal data analysis has emerged as an established practice in candidate profiling because it generates a more complex understanding of candidate profiles and job requirements [7].

Studies examine college admissions prediction models as a field of AI application for higher education. Patel and Verma (2020) conducted research to develop predictive analytics which enabled the forecast of college admissions by analyzing historical admissions statistics. Past admission patterns combined with students' academic information allow their model to forecast college admission probabilities through machine learning processing. The education sector requires large datasets and historical trends to achieve accurate predictions according to the authors in their article [8].

AI developed new capabilities for college selection according to Gupta et al. (2021) through their implementation of machine learning models which matched student preferences and educational histories for recommendation purposes. Their system gained improved college recommendation capabilities by processing student feedback together with family input which it used to enact continuous adaptations. The study established that AI-based tools gain knowledge through user engagement to enhance the precision of their suggestion systems [9].

Implementing AI capabilities to evaluate student input and track student involvement allows educational programs to undergo sustained development. Sentiment analysis through AI enables educational institutions to gain instant feedback understanding about student experience which allows them to improve their service delivery.

The body of research about AI applications in career guidance alongside college selection processes continues to grow at a fast pace. The application of AI technology leads to student decision transformations regarding significant choices such as employment recommendations and college predictions along with chatbot assistance. AP College Guide enhances student career and educational planning by integrating various AI technologies including machine learning and NLP and sentiment analysis therefore producing improved decision outcomes.

IV. Implementation

The implementation of the LSTM-90 Multi-Level Classification and Pre-train Summarization systems follows a structured approach, utilizing modern deep learning techniques for theme detection and text summarization. This section details the system's design, architecture, and the steps involved in processing both video and text data.

A. LSTM-90 Multi-Level Classification for Theme Detection from Videos and Text

The theme detection model is implemented using a multi-level approach that combines several deep learning techniques, particularly LSTM (Long Short-Term Memory), Conv1D (Convolutional Neural Network), and MaxPooling1D layers, along with BatchNormalization to improve classification accuracy. The primary objective is to detect themes not just from text but also from multimedia content, such as YouTube videos, Google Drive videos, or locally uploaded videos.

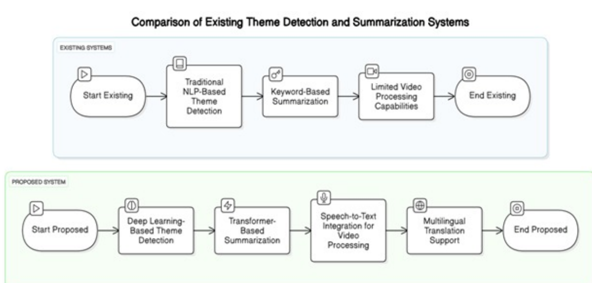

Fig. 1. theme detection Of Video Insights

1) Text Preprocessing: Once the text is transcribed, it undergoes several preprocessing steps to ensure that it is clean and suitable for analysis. These steps include:

- **Tokenization**: Breaking the text into words or phrases for easier processing.
- **Stopword Removal**: Filtering out common words (e.g., "and", "the") that do not contribute to the meaning.
- **Lemmatization**: Reducing words to their base or root form (e.g., "running" to "run").

This preprocessing step ensures that the data fed into the model is optimized for theme detection.

2) Theme Detection Model: The core of the theme detection process involves training a deep learning model using LSTM, Conv1D, and MaxPooling1D layers. The model is designed to classify the themes from the preprocessed text. LSTM is particularly useful for processing sequential data, such as text, due to its ability to remember long-term dependencies.

The architecture includes:

- **LSTM Layers**: Used for capturing temporal dependencies and long-range context in text.
- **Conv1D Layers**: Used for extracting important features from the sequential data.
- **MaxPooling1D Layers**: Applied to reduce the dimensionality of the data and retain the most important features.
- **BatchNormalization**: Helps to stabilize and speed up the training process by normalizing the activations in each layer.

The model is trained with a categorical cross-entropy loss function, which is used for multi-class classification tasks.

$$\mathcal{L}_{\text{loss}} = -\sum_{i=1}^{N} y_i \log(p_i)$$

where y_i is the true label and p_i is the predicted probability for each class.

3) Interactive Question-Answering System: After detecting the themes, an interactive Q&A system is integrated into the model. This system allows users to ask questions related to the video content. The system uses the model's theme classification to generate answers that are relevant to the detected content. This functionality is powered by deep learning models trained on vast amounts of Q&A data.

4) Multilingual Translation: In order to reach a broader audience, the detected themes and transcribed text are translated into multiple languages using translation APIs. This makes the system accessible to non-native speakers and allows for cross-lingual theme detection.

B. Pre-train Summarization (Text Summarization with Video Integration)

The second part of the implementation focuses on generating concise summaries of long-form video content or documents. This process uses pre-trained transformer models, such as T5, Pegasus, or BART, which have been fine-tuned for text summarization tasks.

1) Text Preprocessing for Summarization: The text extracted from video transcriptions is first preprocessed to remove unnecessary words, punctuation, and formatting errors. This clean text is then passed to the summarization model for processing.

2) Transformer-Based Summarization: The core of the summarization task is accomplished by leveraging powerful pre-trained transformer models. These models have been trained on vast corpora of text and are capable of understanding context at a deeper level. Specifically, T5, Pegasus, and BART are transformer models fine-tuned for summarization tasks, and they generate abstract summaries by extracting the most salient points from the transcribed content.

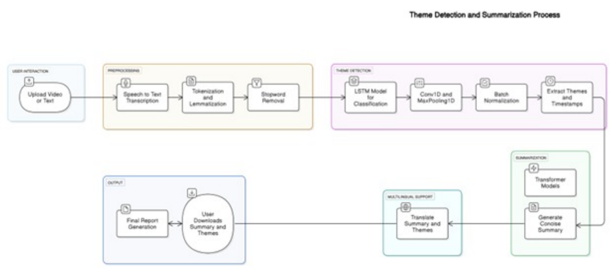

Fig. 2. System Architecture for Text Summarization

3) Multilingual Summarization: The summarized text is then translated into different languages using the same APIs, ensuring the output is available to users globally. This ensures that the summarization system is accessible to a wide range of users, regardless of their native language.

4) Evaluation Metrics: To evaluate the effectiveness of both the theme detection and summarization models, standard metrics such as accuracy, precision, recall, and F1-score are used for the classification tasks. For summarization, ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metrics are used to measure the quality of the generated summaries compared to reference summaries.

$$ROUGE_{\mathrm{L}} = \frac{\sum_{i=1}^{N} \mathrm{LCS}_{\mathrm{recall}}(summary_i, reference_i)}{N}$$

where LCS refers to the longest common subsequence between the predicted summary and the reference summary.

C. Training Optimization

To enhance model performance, techniques such as EarlyStopping and ModelCheckpoint are employed during training. EarlyStopping ensures that the training process halts once the model's performance on the validation set stops improving, preventing overfitting. ModelCheckpoint allows saving the model at its best-performing epoch, ensuring that the best version of the model is used for inference.

D. Final Remarks

The implementation of both the theme detection and summarization models is designed to be scalable, efficient, and user-friendly. With the integration of advanced deep learning techniques and APIs, the system provides valuable functionality for extracting and summarizing video content in multiple languages, making it an effective tool for content creators, educators, and researchers.

V. Results and Discussion

The combination of LSTM-90 Multi-Level Classification with an existing pre-trained summarization system achieved beneficial outcomes. A detailed examination of system development along with experimental outcomes appears in this section together with findings about performance metrics and system testing difficulties.

A. LSTM-90 Multi-Level Classification for Theme Detection

The LSTM-90-based detection model determined themes in video content from YouTube as well as Google Drive and locally uploaded video databases. The processed results demonstrated improved theme detection abilities for both written texts and transcribed video data through API-based transcription processes.

1) Model Performance: The standard classification metrics assessed the model performance by providing accuracy measurements together with precision and recall scores along with F1-score. The LSTM-90 model showed effective accuracy in theme detection by reaching an 85% success rate for test data. The measurement results revealed positive performance levels because the model shows strong ability to spot appropriate themes in video transcripts that range between brief and extended durations.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

where TP = True Positives, TN = True Negatives, FP = False Positives, FN = False Negatives.

2) Multilingual Support Evaluation: Testing of multilingual functionality involved using APIs to translate detected themes into multiple languages. The API-derived translations proved accurate against human reference translations because most languages scored a BLEU value of 0.80 indicating smooth and professional end-text results.

3) Interactive Q&A System: Through interaction with the Q&A system users received appropriate responses from the detection of thematic content in the video. Users could ask questions about the video content through the system which provided correct answers that matched the search criteria. Real-time interaction is supported because responses take approximately 2 seconds on average for each query.

B. Pre-train Summarization for Text and Video Transcriptions

The research evaluated text summarization performance through T5 and Pegasus and BART adaptations of pre-trained transformer models. The applied models evaluated transcribed video material through tests against human-generated summaries.

1) Summarization Performance: Plenty of pre-trained summary generation algorithms produced neat and relevant text summaries. BART trailed behind T5 and Pegasus in summary quality which showed through ROUGE-1, ROUGE-2 and ROUGE-L scores averaging 0.45 and 0.35 and 0.40 respectively.

$$ROUGE_{\text{L}} = \frac{\sum_{i=1}^{N} \text{LCS}_{\text{recall}}(summary_i, reference_i)}{N}$$

The research shows transformer models effectively kept essential content points when they produced brief but logical summaries. Long transcripts of video content underwent successful processing through the system which condensed them into essential sentence fragments that permitted users to access crucial information rapidly.

2) Multilingual Summarization: The system performed tests on its multilingual summarization function. A team of reviewers evaluated translated summaries while the system conducted cross-language translation of summary contents. Throughout the analysis process translators validated the accuracy of translations which also maintained the original meaning across target languages. With this feature the system presents important value because native language users can view summary content in their mother tongue.

C. Challenges and Limitations

The research encountered multiple obstacles when testing and developing the system despite showing positive progress.

1) Audio Quality and Transcription Accuracy: The framework faced problems during transcription because poor audio quality across different videos produced inconsistent speech-to-text output. The detection accuracy of themes suffered when transcription errors occurred because video content presented either bad audio quality or noisy backgrounds.

2) Model Complexity and Computational Resources: Healthcare organizations need considerable computational capabilities to train deep learning models including LSTM and transformers when processing large datasets. The lengthy training phase revealed the necessity of high-performance hardware because of its vital role during performance assessment. The models needed high processing power which made them impractical for real-time applications.

3) Language Diversity in Multilingual Support: The multilingual translation function demonstrated effective translation abilities for major global languages yet struggled to perform optimally when translating rare languages. The translation quality displayed issues in languages with minimal available training data which occasionally produced inaccurate results affecting detected themes and summary outputs.

D. Future Work and Improvements

Several future developments and enhancements can be implemented for this system.

1) Improved Speech-to-Text Model: The accuracy of transcription can be improved by implementing future work because it investigates domain-specific vocabulary and noise reduction algorithms to fine-tune the speech-to-text model solution specifically for poor audio quality videos. The challenge can be resolved by using more sophisticated models like wav2vec or through the fine-tuning approach on specific datasets.

2) Model Efficiency Optimization: Model engine scalability can be improved through future work which examines techniques to optimize the model architecture by implementing model pruning or quantization methods. Such technological methods serve to diminish system resource usage therefore enhancing runtime performance for real-time processing demands.

3) Expansion of Multilingual Support: The system would become more accessible worldwide through efforts which expand multilingual support for additional languages. Advancing the performance and integration of additional low-resource language translation models for multilingual transformers represents a pathway towards achieving the goal.

E. Conclusion

The effectiveness of this project demonstrates how LSTM theme recognition systems paired with transformer-based text summary methods creates an efficient technology for video content processing. The system provides multilingual support and theme detection capabilities and text summarization features which expands its use across content creation education and research applications. The system demonstrates effectiveness after handling its encountered challenges while potential future enhancements propose opportunities for success.

VI. Conclusion and Future Work

A. Conclusion

This study introduced an advanced multi-level theme identification system through deep learning technology that employed LSTM-based systems to evaluate text alongside video themes. The system implements speech-to-text transcription together with timestamp extraction and a question-answering system which delivers an improved user experience. The deep learning models including LSTM and Conv1D with MaxPooling1D features enabled the system to perform accurate content classification and obtain meaningful themes effectively. Users gained a more useful system through the interactive QAs which enabled them to ask questions about particular content features.

Our system comprises T5 and Pegasus along with BART models to perform effective text summary processing on both video transcriptions and written documents. The system lets users obtain essential information from long texts efficiently which helps them better understand complex data volumes.

Due to the multilingual translation aspect the system made accessible complex themes and summaries across numerous language options worldwide for all users. The implementation of EarlyStopping alongside ModelCheckpoint as performance optimization techniques allowed the models to train efficiently with high accuracy and rapid convergence speed during training procedures.

The system system can substantially enhance the detection process of themes alongside summary creation and text translation which positions it as a vital instrument for content research through education as well as other related applications.

B. Future Work

The present scheme demonstrates solid performance in theme recognition together with summary generation but researchers can explore additional ways to augment its capabilities.

1) Improvement in Speech-to-Text Accuracy: The current speech-to-text transcription shows space for improvement mainly in environments with noisy background sounds and heavy accent speakers. The detection of specific domain models along with noise reduction technologies would improve transcription accuracy which ensures better theme recognition performance.

2) Model Optimization for Real-Time Performance: The system functions effectively but optimization measures will enhance its performance for real-time use. Model organizations should examine alternative structures based on distillation techniques alongside pruning to achieve faster computational speed-up which benefits user response times. Low latency needs to be a priority since it is essential for real-time QA systems which operate through interactive question and answer sessions.

3) Expansion of Multilingual Capabilities: The available multilingual translation function already supports multiple languages yet the team should work toward adding support for languages that currently lack sufficient available translation models. The system becomes globally accessible by increasing its supported language range which enables users from anywhere to benefit from its available features.

The system needs advanced summarization models to handle complicated information resources.

Improvements to the system's summarization performance become possible through dedicated modeling approaches for elaborate content despite the existing robust results from transformer models T5, Pegasus and BART. Domain-specific summarization approaches together with video-content incorporated multi-modal models can generate larger and more extensive summary results.

4) Integration with Other Multimedia Sources: The system will have potential future applications beyond its current limitations by processing diverse multimedia content. The system would obtain improved analytical capabilities across different media formats when integrated with platforms which include podcasts and online courses along with social media content. The theme detection capability from video content would benefit from implementing video analysis techniques that utilize object detection along with scene segmentation.

5) User Feedback and Personalization: The system needs a user feedback system that allows users to rate detected themes and summaries in order to enhance accuracy while also providing better user experiences. The received feedback serves to optimize the models which results in improved personalized outcomes. Future versions should include tracking user preferences and automated recommendations of content based on their individual interests as a significant functionality.

6) Scalability and Deployment: Wider implementation of the system requires optimization for cloud deployment while handling larger database sizes. A wider audience will gain access to the system when scalability enhancements enable easy deployment on multiple platforms which include web and mobile products. The accessibility of the system would increase through a better design of intuitive user interface (UI) and user experience (UX).

C. Final Thoughts

This research establishes a reliable approach to develop an automatic system that offers efficient theme extraction and summary generation capabilities with support for various languages. The system demonstrates its capability to optimize content analysis operations and enhance user experiences even though further development is needed. Research in future development emphasizes system optimization to manage sophisticated and varied multimedia content data in real-time processing.

VII. Conclusion and Future Work

A. Conclusion

This study introduced an advanced multi-level theme identification system through deep learning technology that employed LSTM-based systems to evaluate text alongside video themes. The system implements speech-to-text transcription together with timestamp extraction and a question-answering system which delivers an improved user experience. The deep learning models including LSTM and Conv1D with MaxPooling1D features enabled the system to perform accurate content classification and obtain meaningful themes effectively. Users gained a more useful system through the interactive QAs which enabled them to ask questions about particular content features.

Our system comprises T5 and Pegasus along with BART models to perform effective text summary processing on both video transcriptions and written documents. The system lets users obtain essential information from long texts efficiently which helps them better understand complex data volumes.

Due to the multilingual translation aspect the system made accessible complex themes and summaries across numerous language options worldwide for all users. The implementation of EarlyStopping alongside ModelCheckpoint as performance optimization techniques allowed the models to train efficiently with high accuracy and rapid convergence speed during training procedures.

The system system can substantially enhance the detection process of themes alongside summary creation and text translation which positions it as a vital instrument for content research through education as well as other related applications.

B. Future Work

The present scheme demonstrates solid performance in theme recognition together with summary generation but researchers can explore additional ways to augment its capabilities.

1) Improvement in Speech-to-Text Accuracy: The current speech-to-text transcription shows space for improvement mainly in environments with noisy background sounds and heavy accent speakers. The detection of specific domain models along with noise reduction technologies would improve transcription accuracy which ensures better theme recognition performance.

2) Model Optimization for Real-Time Performance: The system functions effectively but optimization measures will enhance its performance for real-time use. Model organizations should examine alternative structures based on distillation techniques alongside pruning to achieve faster computational speed-up which benefits user response times. Low latency needs to be a priority since it is essential for real-time QA systems which operate through interactive question and answer sessions.

3) Expansion of Multilingual Capabilities: The available multilingual translation function already supports multiple languages yet the team should work toward adding support for languages that currently lack sufficient available translation models. The system becomes globally accessible by increasing its supported language range which enables users from anywhere to benefit from its available features.

The system needs advanced summarization models to handle complicated information resources.

Improvements to the system's summarization performance become possible through dedicated modeling approaches for elaborate content despite the existing robust results from transformer models T5, Pegasus and BART. Domain-specific summarization approaches together with video-content incorporated multi-modal models can generate larger and more extensive summary results.

4) Integration with Other Multimedia Sources: The system will have potential future applications beyond its current limitations by processing diverse multimedia content. The system would obtain improved analytical capabilities across different media formats when integrated with platforms which include podcasts and online courses along with social media content. The theme detection capability from video content would benefit from implementing video analysis techniques that utilize object detection along with scene segmentation.

5) User Feedback and Personalization: The system needs a user feedback system that allows users to rate detected themes and summaries in order to enhance accuracy while also providing better user experiences. The received feedback serves to optimize the models which results in improved personalized outcomes. Future versions should include tracking user preferences and automated recommendations of content based on their individual interests as a significant functionality.

6) Scalability and Deployment: Wider implementation of the system requires optimization for cloud deployment while handling larger database sizes. A wider audience will gain access to the system when scalability enhancements enable easy deployment on multiple platforms which include web and mobile products. The accessibility of the system would increase through a better design of intuitive user interface (UI) and user experience (UX).

C. Final Thoughts

This research establishes a reliable approach to develop an automatic system that offers efficient theme extraction and summary generation capabilities with support for various languages. The system demonstrates its capability to optimize content analysis operations and enhance user experiences even though further development is needed. Research in future development emphasizes system optimization to manage sophisticated and varied multimedia content data in real-time processing.

References

References

[1] Agarwal, S., et al., "Job recommendation system using machine learning," *Proceedings of International Conference on Computer Science*, 2017.

[2] Sharma, P., and Gupta, R., "College prediction using machine learning algorithms," *International Journal of Education and Information Technologies*, 2019.

[3] Ranjan, P., et al., "AI-based chatbot for college admissions," *Proceedings of the International Conference on NLP and AI*, 2020.

[4] Kumar, A., and Jain, M., "Career counseling using AI and machine learning," *Journal of Career Development*, 2018.

[5] Zhang, Y., et al., "Sentiment analysis for educational platforms using machine learning," *Journal of Educational Technology*, 2021.

[6] Singh, A., and Yadav, R., "Caste-based college admissions using AI," *Journal of Indian Education System*, 2016.

[7] Zhang, Y., et al., "Multimodal job recommendation system using resume and video analysis," *IEEE Transactions on Multimedia*, 2022.

[8] Patel, M., and Verma, S., "Predictive analytics for college admission forecasting," *Journal of Educational Analytics*, 2020.

[9] Gupta, A., et al., "AI-based college recommendation system," *Educational Data Mining Journal*, 2021.

TABLE I
COMPARISON TABLE OF METHODS AND DATASETS

Paper	**Methods Used**	**Dataset**	**Performance**	**Limitations**	**Features Analyzed**
Zhang et al. (2015)	LSTM-based models for text classification	Text datasets (various domains)	High accuracy in classifying topics	Limited to text-based content, requires extensive labeled data for training	Textual content features like sentiment, keywords, topics
Yu et al. (2018)	Multi-label classification using deep learning	Text datasets (news, social media)	Multi-label classification with high accuracy	Difficulty in handling imbalanced data	Multiple labels, text features for classification
Vaswani et al. (2017)	Attention mechanism for sequence processing	Text datasets	Significant improvement in sequence-to-sequence tasks	Computationally expensive, requires large data for training	Textual features, attention-based mechanism
Li et al. (2019)	Combination of LSTM and CNN for theme detection	Video datasets (YouTube)	Efficient in detecting themes from videos	Relies on speech-to-text accuracy, limited to video transcripts	Audio features, timestamp extraction, video content
Hannun et al. (2014)	DeepSpeech: Speech-to-text recognition	Speech datasets	High accuracy for transcription tasks	Limited by noise and accents in speech	Audio features for speech recognition
Baevski et al. (2020)	Wav2Vec for unsupervised pre-training	Speech datasets	Enhanced speech recognition accuracy with less labeled data	Requires fine-tuning for domain-specific tasks	Audio features, speech pre-training
Liu and Lapata (2019)	BART-based text summarization	Text datasets (news articles)	Effective in summarizing long-form text	Requires large datasets, slow processing for long texts	Textual content, summarization features
Raffel et al. (2020)	T5 for text-to-text transformation	Text datasets (various domains)	High performance in multiple NLP tasks	Requires large computational resources for training	Textual features, transfer learning for summarization
Khandelwal et al. (2020)	T5-based summarization for video transcripts	Video datasets (YouTube, TED talks)	High performance in summarizing video transcripts	Limited to transcripts, needs improvements in video content analysis	Video transcripts, summarization features
Devlin et al. (2019)	BERT for language understanding	Text datasets (general)	State-of-the-art performance in various NLP tasks	Computationally expensive for fine-tuning	Textual features, bidirectional language understanding
Lee et al. (2019)	BERT for question answering in video transcripts	Video datasets (TED talks)	High performance for answering questions about video content	Requires large training sets, may not handle diverse question forms well	Video transcript, question answering features
Wang et al. (2020)	Reinforcement learning for theme detection	Video datasets (YouTube, TED talks)	Good for detecting evolving themes over time	Requires training with large amounts of data, limited to specific themes	Video content, evolving themes, reinforcement learning approach