PROJECT PHASE II REPORT

ON

DIALOGUE ANALYZER (DiAna)

Submitted by

GOURI S GOVIND (SJC20AD034)

GEORGE JOYAL VINCENT (SJC20AD032)

JUDIN AUGUSTIN (SJC20AD045)

NOYAL JOSEPH (SJC20AD049)

to

the APJ Abdul Kalam Technological University

in partial fulfillment of the requirements for the award of the degree

of

Bachelor of Technology

in

Artificial Intelligence and Data Science

Department of Artificial Intelligence and Data Science

St. Joseph's College of Engineering and Technology, Palai

MAY : 2024

Declaration

We declare that the Project Phase II report on **DIALOGUE ANALYZER (DiAna)**, submitted for partial fulfillment of the requirements for the award of degree of Bachelor of Technology of the APJ Abdul Kalam Technological University, Kerala, is a bonafide work done by us under supervision of **Ms. Aswathy James**. This submission represents our ideas in our own words and where ideas or words of others have been included. We have adequately and accurately cited and referenced the original sources. We also declare that we have adhered to ethics of academic honesty and integrity and have not misrepresented or fabricated any data or idea or fact or source in our submission. We understand that any violation of the above will be a cause for disciplinary action by the institute and/or the University and can also evoke penal action from the sources which have thus not been properly cited or from whom proper permission has not been obtained. This report has not been previously formed the basis for the award of any degree, diploma or similar title of any other University.

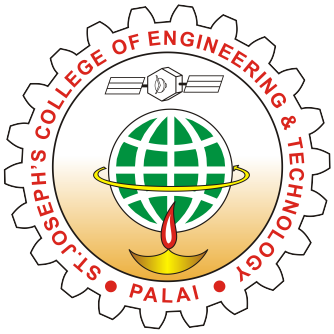
Name and Signature of Students

Gouri S Govind (SJC20AD034)
George Joyal Vincent (SJC20AD032)
Judin Augustin (SJC20AD045)
Noyal Joseph (SJC20AD049)

Place: Choondacherry
Date: 03-05-2024

ST. JOSEPH'S COLLEGE OF ENGINEERING AND TECHNOLOGY, PALAI

DEPARTMENT OF ARTIFICIAL INTELLIGENCE AND DATA SCIENCE



CERTIFICATE

This is to certify that the report entitled "**DIALOGUE ANALYZER (DiAna)**" submitted by **Gouri S Govind (SJC20AD034), George Joyal Vincent (SJC20AD032), Judin Augustin (SJC20AD045) and Noyal Joseph (SJC20AD049)** to the APJ Abdul Kalam Technological University in partial fulfillment of the requirements for the award of the Degree of Bachelor of Technology in Artificial Intelligence and Data Science is a bonafide record of the project work carried out by them under my guidance and supervision.

Project Guide
Ms.Aswathy James
Assistant Professor
Department of AD

Project Coordinator
Mr.Jacob Thomas
Assistant Professor
Department of AD

Head of Department
Dr.Renjith Thomas
Head of Department
Department of AD

Place : Choondacherry
Date : 03-05-2024

Acknowledgement

The success and final outcome of this project phase II required a lot of guidance and assistance from many people, and we are extremely privileged to have received their support throughout the completion of this project. All that we have accomplished is only possible due to their supervision and assistance, and we are sincerely grateful to them. We would like to express our respect and gratitude to the management of St. Joseph's College of Engineering and Technology for providing us with the opportunity and platform to work on this project. A special word of thanks goes to our beloved Principal, **Dr. V. P. Devassia**, for providing invaluable support and necessary facilities to carry out this project. We are extremely indebted to **Dr. Deepa V**, School of AI & Robotics MG University Kottayam, Kerala and **Dr. Renjith Thomas**, Head of the Department of Artificial Intelligence and Data Science, for their valuable suggestions and encouragement throughout the course of this project work. We would also like to express our gratitude to our project coordinator, **Mr. Jacob Thomas**, Assistant Professor in the Department of Artificial Intelligence and Data Science, for his valuable suggestions and guidelines during the entire duration of this project. We truly appreciate his contributions and technical support in preparing this report. Our heartfelt thanks go to our project guide, **Ms.Aswathy James**, Assistant Professor in the Department of Artificial Intelligence and Data Science, displayed a keen interest in this project and provided guidance and all the necessary information for developing a robust system. Our deepest gratitude goes to our family, whose unwavering support, understanding, and love have been our pillars of strength. Lastly, we would like to extend our thanks to all those individuals who have directly or indirectly influenced our work as a whole. Their contributions have been indispensable in making this project a reality.

Gouri S Govind
George Joyal Vincent
Judin Augustin
Noyal Joseph

Abstract

Multimedia content consumption has become increasingly prevalent in today's digital landscape, necessitating innovative solutions to enhance user experiences. In this study, we present a comprehensive framework for the analysis of podcast and interview content, integrating advanced technologies such as diarization, sentiment analysis, and summarization. Our approach aims to provide users with personalized and emotionally informed experiences by deciphering conversational dynamics and emotional nuances within multimedia narratives. Leveraging state-of-the-art models including DialogLED, Whisperx large v2, and RoBERTa, we delve into tasks such as speaker segmentation, sentiment classification, and content summarization. Through detailed analysis and visualization, we offer valuable insights into audience engagement, emotional resonance, and content reception, with potential applications in podcast analytics, content recommendations, and market research. By showcasing the potential of our innovative approach, we aim to pave the way for enhanced multimedia experiences and deeper connections between creators and audiences in the digital age.

Contents

List of Abbreviations

ASR Automatic Speech Recognition

EEND End-to-End Neural Diarization

LED Longformer-Encoder-Decoder

LM Language Model

LSA Latent Semantic Analysis

ML Machine Learning

NLP Natural Language Processing

RSAN Recurrent Selective Attention Network

List of Figures

List of Tables

Chapter 1

Introduction

In today's digital era, multimedia content has become an integral part of the lives, offering a diverse range of content types across various platforms. From podcasts to webinars, these mediums cater to different interests and preferences, fueled by advancements in technology and the widespread use of social media. As people increasingly rely on digital platforms for entertainment, education, and social interaction, the demand for innovative solutions to enhance the multimedia experience continues to grow.

The analysis of multimedia content has emerged as a crucial area of research, drawing significant attention from content creators, marketers, researchers, and consumers alike. Understanding the subtle nuances of emotions embedded within multimedia narratives, extracting actionable insights from vast volumes of audiovisual data, and deciphering complex content structures are now essential tasks. However, the endeavor is not without the challenges, including managing the sheer volume of data, deciphering the intricacies of human language and emotions, and developing scalable analysis techniques.

In response to these challenges, the study aims to contribute to multimedia analysis by presenting a comprehensive framework for podcast and interview content. Leveraging advanced technologies like diarization, sentiment analysis, and summarization, the framework offers a holistic approach to understanding multimedia narratives [1] [2] [3]. Central is the recognition of emotions' pivotal role in shaping audience engagement and content reception, providing users with personalized, emotionally informed experiences.

Moreover, the research is motivated by the growing significance of multimedia content across various domains, including entertainment, education, marketing, and journalism. Podcasts, in particular, have witnessed a surge in popularity, emerging as a powerful medium for storytelling, knowledge dissemination, and audience engagement. Similarly, interviews serve as valuable platforms for dialogue, debate, and knowledge exchange, offering insights into diverse perspectives and experiences. Understanding the dynamics of podcast and interview content not only holds scholarly relevance but also has practical implications for content creators, advertisers, and policymakers seeking to harness the power of multimedia for various purposes.

By bridging the gap between theory and practice in the analysis of multimedia content, the study seeks to empower stakeholders with the tools and techniques needed to navigate the complex landscape of multimedia communication effectively [4] [5]. Through the research, is to the advancement of knowledge in the field and pave the way for enhanced multimedia experiences that enrich and inspire audiences worldwide.

1.1 Background

Multimedia content, encompassing a diverse array of audio, video, and interactive elements, has become an integral part of modern communication and entertainment [6]. With the proliferation of digital platforms and the advent of high-speed internet, multimedia consumption has surged, reshaping how individuals engage with information and entertainment. From streaming services to social media platforms, multimedia content permeates various aspects of daily life, offering unparalleled opportunities for communication, learning, and expression [7].

The evolution of multimedia technology has been driven by several key factors, including advances in computing power, data storage, and internet connectivity. These advancements have facilitated the creation and distribution of multimedia content on a scale never

before imagined, democratizing access to information and empowering individuals to become content creators in their own right. Additionally, the rise of artificial intelligence and machine learning has revolutionized how multimedia content is processed, analyzed, and personalized, opening up new avenues for content discovery and consumption. As a result, multimedia has become an integral part of modern communication, shaping the way interact with information and each other in the digital age.

Against the backdrop of rapid technological innovation and shifting consumer preferences, the study seeks to address the growing need for advanced multimedia analysis techniques [8] [10]. By leveraging cutting-edge technologies such as natural language processing, speech recognition, and sentiment analysis, to enhance the understanding and interpretation of multimedia content, enabling more personalized and engaging user experiences. The research is to contribute to the ongoing evolution of multimedia technology, driving innovation and fostering deeper connections between content creators and consumers in an increasingly digital world [9].

1.2 Motivation

The study stems from the growing significance of multimedia content in today's digital landscape and the inherent challenges associated with analyzing and extracting meaningful insights from such diverse and voluminous data [13]. With the proliferation of podcasts, interviews, and other multimedia formats, there arises a pressing need for innovative approaches to enhance user experiences, optimize content delivery, and derive actionable insights. The advent of advanced deep learning models and NLP techniques offers unprecedented opportunities to tackle these challenges and revolutionize the way multimedia content is processed, summarized, and understood [11] [15].

By delving into the intricacies of multimedia analysis, the study seeks to address several key motivations. Firstly, the study aims to bridge the gap between raw multimedia data and actionable insights by developing comprehensive frameworks for tasks such as di-

arization, sentiment analysis, and summarization [16] [17]. Secondly, the study endeavors to empower content creators, analysts, and consumers alike with tools and methodologies to navigate the vast landscape of multimedia content more efficiently and effectively. Thirdly, to contribute to the advancement of research in the field of multimedia analysis by exploring novel techniques and methodologies and evaluating their efficacy in real-world scenarios.

Moreover, the motivation behind the study extends beyond academic curiosity to practical applications in various domains such as podcast analytics, content recommendation systems, market research, and sentiment analysis [18] [19]. By harnessing the power of advanced technologies and cutting-edge research, the study aims to address real-world challenges and unlock new possibilities for understanding, analyzing, and leveraging multimedia content in an increasingly digital world. Ultimately, the motivation driving the study is to empower stakeholders across industries with the tools, insights, and methodologies needed to extract maximum value from multimedia content and enhance user experiences in the digital age.

1.3 Objective and Scope

The primary objective of the study is to develop and implement a comprehensive framework for the analysis of multimedia content, encompassing tasks such as audio transcription, speaker diarization, sentiment analysis, and content summarization. By leveraging state-of-the-art technologies and methodologies, including machine learning models and natural language processing algorithms, to extract meaningful insights from multimedia data and enhance the overall user experience.

Furthermore, the research endeavors to address several key challenges and limitations inherent in existing multimedia analysis techniques. These challenges include the need for more accurate and efficient methods of speaker identification, sentiment analysis, and content summarization, as well as the need to adapt these techniques to diverse multi-

media formats and genres. Through rigorous experimentation and validation, the study aims to develop robust and scalable solutions that can be applied across a wide range of multimedia content types and platforms. By doing so, the study seeks to advance the field of multimedia analysis and contribute to the development of more effective tools and technologies for understanding and leveraging multimedia content in various contexts.

The scope of the study encompasses both theoretical research and practical implementation, with a focus on real-world applications and use cases. To demonstrate the efficacy and utility of the proposed framework through empirical testing and evaluation, using benchmark datasets and performance metrics to assess the effectiveness of the methods. Additionally, the aim to explore the potential implications of the research for various industries and domains, including media and entertainment, education, healthcare, and marketing.

Overall, the main objectives are twofold: to advance the state of the art in multimedia analysis through innovative research and to provide practical tools and insights that can be leveraged by content creators, developers, and end-users alike. By addressing the objectives, meaningful contributions to the field of multimedia technology and pave the way for future advancements in the rapidly evolving domain.

Chapter 2

Literature Review

In an effort to situat research in the context of the wider landscape of podcast consumption, conducted a comprehensive review of the existing literature on podcast summarization and diarization. The goal is to reveal insights from previous studies, applied methodologies, and technological advances to provide a basis for the development of an innovative system. The review seeks to identify gaps and trends and guides approach to optimizing the podcast listening experience using intelligent speaker summarization and segmentation techniques.

2.1 Existing Solutions

1. **Extractive Text-Image Summarization Using Multi-Modal RNN**

   Jingqiang Chen and Hai Zhuge

   In the field of multimodal content augmentation, the paper presents an innovative approach of multimodal summarization [1]. Using a multimodal RNN, deftly tackles challenges such as sentence-image alignment, assessing the impact of an image on text summarization, and optimizing image selection. With hierarchical RNN, VGGNet, and logistic classifier, the model outperforms existing methods and shows the effectiveness through dataset expansion and hierarchical RNN application, which represents significant progress in multimodal document summarization.

2. **A Combined Extractive With Abstractive Model for Summarization**

Wenfeng Liu, Yaling Gao, Jinming Li, Yuzhen Yang

The surge in web text data has heightened the demand for efficient thematic extraction, known as text summarization or semantic extraction. Two prevalent paradigms, extractive and abstractive, tackle the task [2]. Extractive methods utilize features like sentence position and word frequency, while abstractive models such as CopyNet and Ptr-Net leverage encoder-decoder architectures but may introduce novel words. To address limitations in readability and semantics, a hybrid abstractive-extractive approach is proposed. The unsupervised model combines an initial extractive phase with a subsequent abstractive phase, demonstrating improved performance on duc2004 and Chinese datasets. Future work aims to tackle challenges in multi-document and cross-document summarization.

3. **Encoder-Decoder Based Attractors for End-to-End Neural Diarization**

Shota Horiguchi et al.

Traditional speaker diarization methods treat the task as a partition problem, relying on cascaded approaches that struggle with speaker overlaps. In contrast, end-to-end neural methods such as Recurrent Selective Attention Network (RSAN) and End-to-End Neural Diarization (EEND) naturally handle speaker overlaps with a single network [3]. The EEND-EDA model, introduced in previous work, addresses unknown numbers of speakers by employing an encoder-decoder-based attractor calculation module. The paper revisits EEND-EDA, proposing training refinements and discussing the relationship with the original EEND. To facilitate fair comparisons with cascaded methods, SAD post-processing is introduced. An iterative inference approach is also proposed to overcome empirical limitations in the number of outputs. Comprehensive evaluations on various datasets demonstrate the efficacy and versatility of EEND-EDA in speaker diarization.

4. **Extractive Document Summarization Based on Dynamic Feature Space Mapping**

   Samira Ghodratnama et al.

   In the era of information overload driven by the internet and smartphones, automated document summarization has become a critical tool for distilling essential insights from vast amounts of textual data [4]. The ExDoS approach presented in the article stands out by seamlessly integrating supervised and unsupervised algorithms, offering a novel solution to the summarization challenge. With dynamic feature weighting, coherent sentence selection, and interpretability, ExDoS not only achieves state-of-the-art performance but also addresses the limitations of redundancy in summarization. The model's ability to produce informative and less redundant summaries is validated through human evaluation experiments, showcasing effectiveness in navigating the complexities of large-scale, unstructured information. The aforementioned innovative approach marks a significant advancement in the field of automated document summarization, providing researchers and practitioners with a powerful tool for efficiently extracting key insights from textual data.

5. **Extractive Summarization of Call Transcripts**

   Pratik K. Biswas and Aleksandr Iakubovich

   The paper introduces an innovative extractive summarization technique tailored for call transcripts [5]. Overcoming challenges like small talk and punctuation issues, the study combines channel separation, topic modeling, and sentence selection with punctuation restoration. Unique elements include evaluating various topic models and using a BERT transformer-based model for punctuation restoration. Extensive evaluations confirm the effectiveness in providing readable and accurate summaries for call transcripts, distinguishing in the domain-specific summarization landscape. The mentioned approach not only addresses the specific challenges inherent in call transcripts but also showcases the potential of advanced machine learning techniques in enhancing summarization accuracy and readability for specialized domains.

6. **System fusion and speaker linking for longitudinal diarization of TV shows**

   Marc Ferràs et al.

   Creating a system for speaker diarization in TV show datasets involves addressing challenges in speech expressiveness and environmental noise. The study approach fuses diarization outputs before speaker linking, reducing error rates [6]. A linear prediction technique handles noisy clusters, achieving comparable performance for linked and non-linked tasks. The system proves effective for longitudinal diarization in the Multi-Genre Broadcast Challenge, showcasing adaptability to expressive speech and variable speaker turn structures in TV shows.

7. **Neural Attention Model for Abstractive Text Summarization Using Linguistic Feature Space**

   Aniqa Dilawari et al.

   Automatic summarization condenses text efficiently using a feature-rich model that combines extractive and abstractive methods. Achieving a significant ROUGE score of 37.76% on CNN/DailyMail, the model employs diverse features for enhanced summarization [7]. The unified approach, leveraging both extractive and abstractive techniques, proves valuable across applications, outperforming prior methods and ensuring human-evaluated significance.

8. **Online Neural Diarization of Unlimited Numbers of Speakers Using Global and Local Attractors**

   Shota Horiguchi et al.

   The article introduces EEND-GLA, an end-to-end neural diarization model integrating global and local attractors for unrestricted speaker count [8]. Utilizing global and block-wise attractors, the model employs clustering for speaker correspondence between blocks, overcoming limitations on the number of speakers. Enhanced with features like speaker-balanced sampling probabilities and variable chunk-size training, EEND-GLA facilitates online inference while maintaining consistent performance across offline and online scenarios.

9. **Self-attention Recurrent Summarization Network with Reinforcement Learning for Video Summarization task**
Aniwat Phaphuangwittayakul et al.

With the surge in video data, efficient summarization is vital. The paper introduces an innovative unsupervised video summarization approach, the Deep Self-attention Recurrent Summarization Network with Reinforcement Learning (DSR-RL) [9]. Integrating self-attention, BRNN, and reinforcement learning, DSR-RL outperforms state-of-the-art methods on SumMe and TVSum datasets. Learns from videos without human annotations, enhancing importance scores and summary diversity. The mentioned novel framework combines attention network principles with reinforcement learning, offering a comprehensive and accurate video summarization solution.

10. **Singer Diarization for Polyphonic Music With Unison Singing**
Hitoshi Suda et al.

The paper presents a novel framework for singer diarization, addressing the challenge of identifying multiple singers in songs [10]. Introduces the Cosacorr score, a unique acoustic feature for accurate overlap detection in unison singing. The framework employs a singing voice separation technique and leverages the ArcFace architecture to extract discriminative singer representations. Evaluation on a dataset of unison singing voices demonstrates the framework's superiority in diarization accuracy, highlighting potential for music analysis and information retrieval.

11. **Abstractive Summarization of Meeting Conversations**
Daksha Singhal et al.

The paper explores abstractive summarization for encapsulating meeting dialogues, emphasizing the benefits in handling spontaneous utterances. The study discusses the challenges posed by the increasing volume of communication data and presents a transformer-based model for supervised abstractive summarization on the Switchboard Dataset [11]. The study suggests future work involving model training on diverse datasets and exploring advanced summarization techniques.

12. **Adaptive and Online Speaker Diarization for Meeting Data**
Giovanni Soldi et al.

Speaker diarization determines 'who spoke when' in an audio stream, crucial for various applications. Most systems are offline, but the increasing demand for online diarization, especially in challenging meeting data, lacks robust solutions [12]. The paper introduces an adaptive and online diarization system for meetings, addressing high speaker overlap and spontaneity. Experiments reveal challenges, with diarization error rates remaining high, suggesting the need for further research to enhance performance and convergence rates.

13. **A Two-Stage Transformer-Based Approach for Variable-Length Abstractive Summarization**
Ming-Hsiang Su et al.

The study presents a two-stage variable-length abstractive summarization model, integrating a text segmentation module, BERTSUM-based extractive model, and Transformer-based abstractive module [14]. Users can specify the desired summary length. The two-stage approach efficiently utilizes single-sentence headline summaries for training. Evaluation on the LCSTS dataset achieved a maximum of 70.0% accuracy in human subjective assessment, showcasing competitive performance in generating fluent and variable-length abstractive summaries.

14. **Speaker Diarization with LSTM**
Quan Wang et al.

The paper introduces a novel approach to speaker diarization by leveraging text-independent d-vector embeddings obtained from LSTM-based audio embeddings [15]. The study combine these embeddings with a non-parametric spectral clustering algorithm, achieving state-of-the-art results. Experimental evaluations on standard datasets, including NIST SRE 2000 CALLHOME, demonstrate the superiority of d-vector-based systems over traditional i-vector approaches, yielding a 12.0% diarization error rate. The proposed system showcases promising advancements in speaker diarization using deep learning techniques.

15. **Automatic Multi-Document Summarization for Indonesian Documents Using Hybrid Abstractive Extractive Summarization Technique**

Glorian Yapinus et al.

The paper introduces a hybrid abstractive-extractive summarization approach for Indonesian news documents [16]. By combining WordNet-based abstractive and title word-based extractive methods, the proposed technique achieves fast, well-compressed, and readable summaries. Compared to Latent Semantic Analysis (LSA), demonstrates efficient processing. Future work may involve applying discourse analysis and similarity techniques for further enhancements.

2.2 Summary

The literature review conducted for the study explores the landscape of multimedia content analysis, summarization, and speaker diarization, drawing from a diverse range of research articles. A prevalent trend observed is the increasing reliance on deep learning models for these tasks, with studies showcasing the effectiveness of models like multimodal RNNs and combined extractive-abstractive models in improving summarization accuracy and readability. Additionally, approaches such as EEND-GLA and RSAN demonstrate the capabilities of end-to-end neural diarization models in handling complex speaker overlap scenarios.

Hybrid approaches combining extractive and abstractive techniques emerge as another significant theme in the literature. Frameworks like ExDoS seamlessly integrate supervised and unsupervised algorithms to produce informative and less redundant document summaries, offering promising solutions for generating coherent and semantically rich summaries from large textual datasets. Furthermore, domain-specific approaches tailored to specialized types of multimedia content, such as extractive summarization techniques optimized for call transcripts and music diarization methods, address unique challenges inherent in their respective domains.

Overall, the literature review provides valuable insights into the state-of-the-art methodologies and advancements in multimedia analysis techniques. By leveraging these insights, the study aims to contribute to the advancement of multimedia analysis frameworks, ultimately enhancing the accessibility and usability of multimedia content across various domains and applications. Building upon the foundation laid by existing research, the study approach seeks to address current challenges in multimedia analysis, such as emotion recognition and narrative understanding, paving the way for more robust and efficient multimedia processing systems. Through empirical validation and practical implementation the work bridge the gap between theoretical advancements and real-world applications, fostering innovation and progress in the field of multimedia analysis.

Chapter 3

Proposed Methodology

In the digital age, the need for efficient and accurate transcription of audio and video recordings has become increasingly important. Manual transcription can be a tedious and time-consuming process, prone to errors and inaccuracies. Fortunately, advancements in artificial intelligence and natural language processing have paved the way for automated transcription systems that offer a more efficient and reliable solution. The figure 3.1 illustrates the key steps involved in the automated video/audio transcription process, highlighting the valuable role of language models and summarization techniques.

3.1 Overview of the Proposed System

The proposed system integrates cutting-edge technologies and innovative methodologies to address multifaceted challenges associated with multimedia content analysis and processing. Comprising modules dedicated to diarization, sentiment analysis, summarization, and model architecture, each tailored to extract valuable insights from diverse multimedia sources, the system leverages state-of-the-art deep learning models such as Whisperx large v2, RoBERTa, and DialogLED.

The diarization module utilizes the Whisperx large v2 model to accurately segment audio recordings by speaker identity, providing a dynamic visualization of speaker engagement and conversation dynamics [24] [25]. Meanwhile, the sentiment analysis module, powered by RoBERTa, offers sophisticated methods for deciphering emotional nuances within multimedia content, enabling comprehensive insights into audience engagement and content reception. Through the integration of advanced models, the system aims to enhance the understanding of multimedia interactions and facilitate more informed content creation and delivery strategies.

In parallel, the summarization module employs advanced techniques to condense multimedia content into concise and informative summaries, facilitating efficient content navigation and comprehension. Additionally, the model architecture component outlines the underlying frameworks and methodologies driving the system's functionality, showcasing the integration of diverse deep learning models and algorithms to optimize performance and accuracy. Through the comprehensive approach, the system aims to streamline the process of multimedia content analysis and empower users with actionable insights for informed decision-making and content creation strategies.

Overall, the proposed system represents a holistic approach to multimedia content analysis, leveraging state-of-the-art technologies and methodologies to unlock valuable insights and enhance user experiences across various domains. By combining the strengths of diarization, sentiment analysis, summarization, and model architecture, the system offers a comprehensive solution to the challenges of analyzing and processing multimedia content in the digital age.

3.2 Block Diagram

The proposed system aims to tackle the challenges associated with analyzing and processing low-quality multimedia content, such as audio and video recordings. In the digital age, where multimedia consumption is ubiquitous, ensuring the quality and understanding of such content is paramount. The system comprises several key steps, each designed to enhance the user experience and extract valuable insights from the input data. By leveraging advanced technologies and methodologies, including diarization, sentiment analysis, and summarization, the system offers a comprehensive solution to navigate and comprehend multimedia content effectively. The introduction provides an overview of the pipeline, detailing each step's significance and contribution to the overall system's functionality.

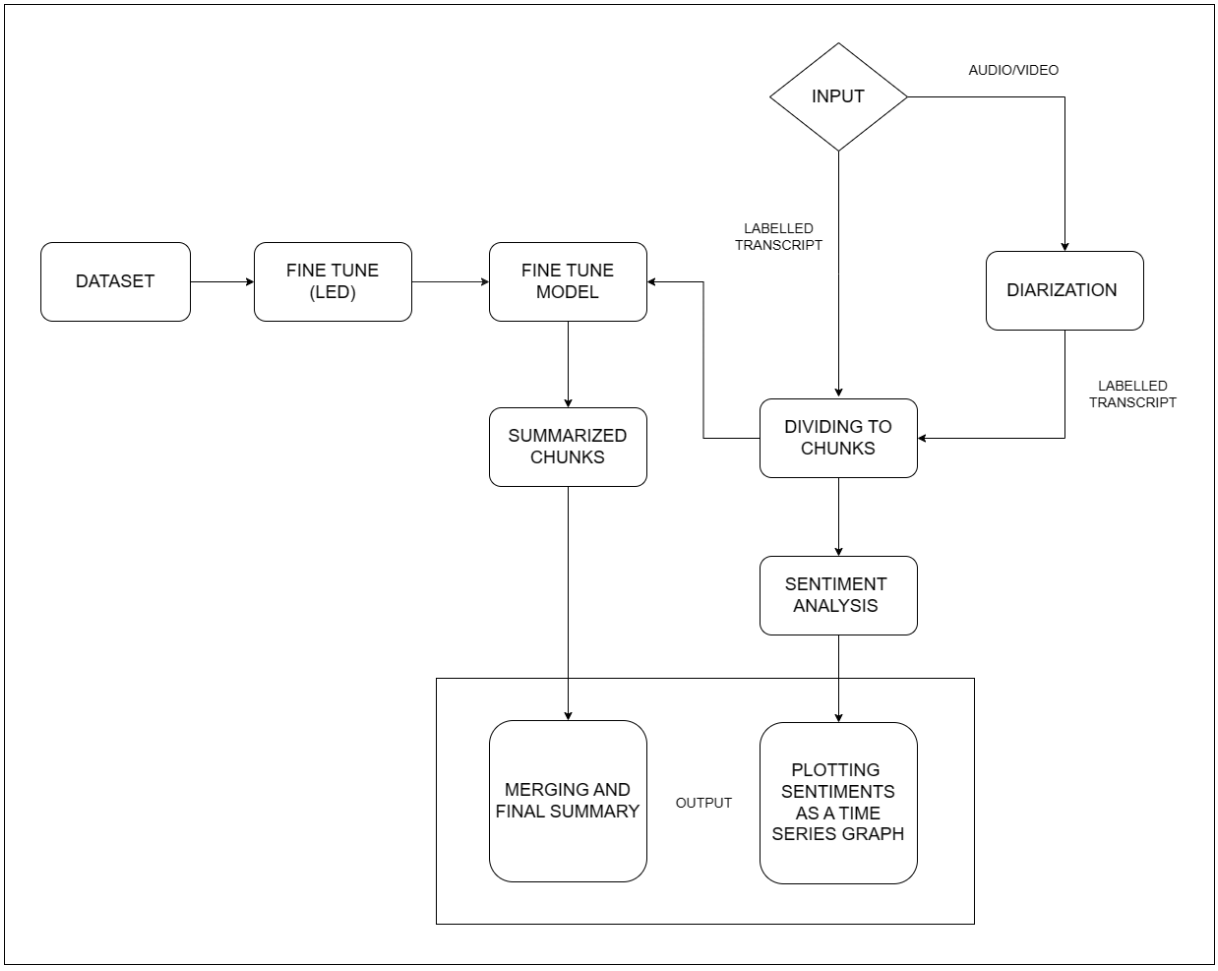


Figure 3.1: Block Diagram of the Proposed System

Step 1: Input

The initial step represents the source of the input data, which typically consists of low-quality audio or video recordings. These recordings serve as the foundation for subsequent processing steps and may originate from various sources such as interviews, podcasts, or surveillance footage. Ensuring the quality and integrity of the raw data is paramount, as the study directly impacts the accuracy and reliability of the analysis and insights derived from the multimedia content.

Step 2: Fine-tuning

In the proposed system, fine-tuning particularly using the Longformer-Encoder-Decoder (LED) model, is essential for optimizing transcription accuracy. The aforementioned phase involves adapting pre-trained language models (LMs) like Whisperx large v2 to the specifics of the target domain, refining the LM's understanding of linguistic structures and enhancing the ability to transcribe spoken words accurately. By tailoring the model to the intricacies of the target domain through LED, the system ensures robust performance in transcribing diverse multimedia content, leading to improved transcription quality and reliability.

Step 3: Labelled Transcript Generation

The proposed system utilizing the Whisperx - distil-medium.en model, transcription tasks are efficiently handled, providing both textual content and corresponding timestamps. Developed by OpenAI, Whisperx stands as an automatic speech recognition (ASR) system, adept at converting spoken language into written text. Trained on extensive multilingual and multitask supervised datasets, Whisperx demonstrates versatility across transcription endeavors. The operational mechanism involves processing audio inputs to generate textual transcripts, accompanied by precise timestamps delineating the speech segments' temporal occurrence.

Step 4: Diarization

Diarization plays a pivotal role in segmenting the input audio or video data into distinct units corresponding to individual speakers or scenes [20] [21]. The process involves analyzing acoustic features such as pitch, timbre, and vocal characteristics to differentiate between different speakers or segments within the recording [22]. By accurately identifying speaker boundaries, diarization enhances the organization and clarity of the transcribed content, facilitating further analysis and interpretation [23]. Additionally, diarization enables the creation of separate transcripts for each speaker or segment, providing a structured representation of the dialogue or commentary present in the multimedia content.

Step 5: Summarization

Summarization in the proposed system involves condensing transcribed audio or video content into concise summaries by identifying key points, essential information, and significant themes. The process enhances efficiency in content consumption, facilitating effective communication of the core message without the need to read through the entire transcript. Additionally, summarization enables users to efficiently navigate large volumes of information, focusing on relevant content and extracting actionable insights with ease, thereby optimizing their overall multimedia experience.

Step 6: Merging and Final Summary

Finally, the summary from the previous step is merged with the original labelled transcript. Merging process ensures that the final output preserves the accuracy and factual richness of the labelled transcript while incorporating the conciseness and informativeness of the summary. Merged document offers a comprehensive overview of the audio or video content, catering to users seeking both detailed information and a quick grasp of key points.

The proposed system presents a systematic approach to analyzing and enhancing low-quality multimedia content. By integrating advanced technologies and methodologies, including diarization, sentiment analysis, and summarization, the system offers a comprehensive solution to address the challenges associated with multimedia processing. Through a series of carefully orchestrated steps, the system enhances the user experience, extracts valuable insights, and enables more effective navigation and comprehension of multimedia content. Moving forward, further refinement and optimization of the system's components will continue to enhance the capabilities and impact across diverse multimedia applications.

3.3 Datasets

The datasets Samsum, Dialogsum, and QMSum represent invaluable resources for delving into the intricacies of multimedia content and advancing research in various fields. Samsum offers a balanced narrative experience with moderate dialogue and summary lengths, capturing audience attention while maintaining brevity. Dialogsum, on the other hand, immerses researchers in lengthy dialogues interspersed with concise summaries, providing a deeper understanding of conversational dynamics. Despite the attempts to finetune models with additional datasets like QMSum, Samsum, and ICSI, limited resources hindered the ability to produce useful models, highlighting the challenges in adapting models to diverse datasets and underscoring the need for further exploration and innovation in the field.

In the realm of multimedia research, Samsum, Dialogsum, and QMSum emerge as indispensable assets, each offering unique perspectives on storytelling, conversation dynamics, and emotional nuances. Samsum's balanced narrative composition provides researchers with a nuanced lens through which to explore the intricacies of storytelling, while Dialogsum's blend of lengthy dialogues and concise summaries facilitates in-depth analysis of

conversational dynamics. Despite the efforts to finetune models with additional datasets like QMSum, Samsum, and ICSI, resource constraints posed significant challenges, hindering the ability to derive meaningful insights and highlighting the importance of resource allocation and optimization in future research endeavors.

As researchers navigate through the rich tapestry of multimedia datasets like Samsum, Dialogsum, and QMSum, they uncover the complexities of human communication and emotion, driving advancements in diarization, sentiment analysis, and beyond. Despite encountering obstacles in finetuning models with additional datasets like QMSum, Samsum, and ICSI, the exploration of these diverse datasets remains crucial for understanding the multifaceted nature of multimedia content and pushing the boundaries of research in the field. Moving forward, addressing resource limitations and refining methodologies will be essential for unlocking the full potential of these datasets and harnessing their insights for innovative applications in multimedia analysis and beyond.

Table 3.1 presents the dataset statistics for Samsum, Dialogsum, and QMSum, offering valuable insights into the characteristics of these datasets. The table provides mean and median values for both dialogue and summary lengths, shedding light on the distribution and central tendencies of the data. For Samsum, the mean dialogue length is 93.79, indicating a moderate duration of dialogue segments, while the mean summary length is 20.32, suggesting concise summarization. Similarly, the median dialogue length of 73.00 and summary length of 18.00 further emphasize the balanced nature of the dataset in terms of dialogue and summary lengths.

Moving on to Dialogsum, observe slightly longer dialogue and summary lengths compared to Samsum. The mean dialogue length of 130.99 and mean summary length of 22.87 indicate a higher level of verbosity in both dialogue segments and summaries. The median values of 116.00 for dialogue and 21.00 for summary lengths confirm the trend, highlighting the propensity for longer conversational exchanges and more detailed summarization.

Table 3.1: Dataset Statistics

Dataset	Mean		Median	
	Dialogue	Summary	Dialogue	Summary
Samsum	93.79	20.32	73.00	18.00
Dialogsum	130.99	22.87	116.00	21.00
QMSum	9318.66	70.59	8693.00	66.00

In contrast, the QMSum dataset stands out for significantly longer dialogue and summary lengths. With a mean dialogue length of 9318.66 and mean summary length of 70.59, QMSum presents a considerable amount of content, reflecting the richness and complexity. The median dialogue length of 8693.00 and summary length of 66.00 further underscore the extensive nature of the dataset, making QMSum a valuable resource for in-depth analysis and exploration of multimedia content.

3.4 Summary

In today's digital era, the demand for efficient and precise transcription of audio and video recordings has reached new heights. The laborious and error-prone nature of manual transcription processes has underscored the need for automated solutions, made possible by advancements in artificial intelligence and natural language processing. The insight drawn from the proposed system's overview underscores the transformative potential of these technologies, particularly in automating transcription workflows. By leveraging state-of-the-art deep learning models like Whisperx large v2, RoBERTa, and DialogLED, the system exemplifies a holistic approach to multimedia content analysis, promising enhanced accuracy and efficiency in transcription tasks.

One of the pivotal insights gleaned from the proposed system lies in comprehensive architecture, designed to address the multifaceted challenges associated with multimedia content. The integration of modules dedicated to diarization, sentiment analysis, summarization, and model architecture signifies a concerted effort to extract valuable insights from diverse multimedia sources. The choice of advanced techniques and methodologies underscores a commitment to optimizing performance, promising actionable insights for informed decision-making and content creation strategies. However, concerns may arise regarding the complexity and computational requirements of such a sophisticated system, highlighting the importance of efficient resource management and scalability.

Delving deeper into the specifics of the proposed system, insights emerge regarding the crucial role of fine-tuning, particularly using the Longformer-Encoder-Decoder (LED) model, in optimizing transcription accuracy. The phase, aimed at tailoring pre-trained language models to the specifics of the target domain, holds immense potential for enhancing transcription quality and reliability. Nevertheless, the systematic approach outlined in the proposed system offers a promising framework for addressing these concerns, paving the way for more robust and efficient multimedia content analysis in the digital age.

Chapter 4

Results and Discussions

In the realm of performance evaluation, meticulous scrutiny of sentiment analysis models ensures their efficacy in decoding nuanced emotions from textual data. The presented tables meticulously examine model performance on the Samsum dataset, offering crucial insights for researchers and practitioners to guide real-world deployments. Meanwhile, diarization outcomes, showcased in a succinct Pie Chart, unveil speaker activity fluctuations, empowering users with a clear understanding of conversational dynamics. Transitioning to sentiment analysis, visualizations like the Line Chart and Heat Map provide rich depictions of emotional evolution over time, driven by advanced models like RoBERTa. These insights not only enrich viewer experiences but also equip content creators with cues to optimize emotional resonance.

4.1 Performance Evaluation

Sentiment analysis models undergo rigorous evaluation to ensure their effectiveness in capturing nuanced emotions from textual data. The precision, recall, and F-measure metrics provide valuable insights into the models' performance across different datasets.

The tables presented here offer a comprehensive overview of sentiment analysis model performance, specifically evaluated on the Samsum dataset. These evaluations are crucial for researchers and practitioners seeking to deploy sentiment analysis solutions in real-world applications. By examining the precision, recall, and F-measure values, stakeholders can make informed decisions about model selection and optimization strategies. Overall, these evaluations contribute to the advancement of sentiment analysis technology, enhancing the applicability and reliability in various domains.

Metric	**Precision**		
	Low	**Mid**	**High**
Rouge1	0.2636	0.3255	0.3850
Rouge2	0.0795	0.1086	0.1457
RougeL	0.1839	0.2371	0.2830
RougeLsum	0.1889	0.2360	0.2853

Table 4.1: Precision Values Evaluated on the Samsum Dataset

Precision is a metric that measures the accuracy of positive predictions made by the model. Table 4.1 displays precision values for various Rouge metrics across different levels of granularity: low, mid, and high. For instance, Rouge1 precision ranges from 0.2636 to 0.3850 across the three levels, indicating the model's ability to accurately predict unigram overlaps between the reference and system summaries. Similarly, Rouge2 precision varies from 0.0795 to 0.1457, showcasing the model's performance in predicting bigram overlaps. Overall, these precision values provide insights into the model's ability to make accurate positive predictions at different levels of granularity.

Recall, on the other hand, assesses the model's ability to capture relevant information from the dataset. Table 4.2 presents recall values for Rouge metrics across different levels.The table provide insights that recall values increase as the level of granularity shifts from low to high. The study suggests that the model effectively captures more relevant information at higher levels of granularity. For instance, Rouge1 recall ranges from 0.3915

Metric	Recall		
	Low	Mid	High
Rouge1	0.3915	0.4773	0.5797
Rouge2	0.1215	0.1709	0.2372
RougeL	0.2750	0.3638	0.4676
RougeLsum	0.2769	0.3607	0.4613

Table 4.2: Recall Values Evaluated on the Samsum Dataset

to 0.5797, indicating the model's ability to retrieve unigrams effectively. Similarly, Rouge2 recall varies from 0.1215 to 0.2372, demonstrating the model's performance in retrieving bigrams from the dataset. The higher recall values at increased levels of granularity imply that the model excels in capturing more comprehensive information, which is essential for accurate sentiment analysis and decision-making in various contexts.

Metric	F-measure		
	Low	Mid	High
Rouge1	0.3167	0.3689	0.4237
Rouge2	0.0944	0.1268	0.1723
RougeL	0.2217	0.2719	0.3311
RougeLsum	0.2234	0.2725	0.3302

Table 4.3: F-measure Values Evaluated on the Samsum Dataset

F-measure, a harmonic mean of precision and recall, provides a balanced assessment of the model's performance. Table 4.3 showcases F-measure values for Rouge metrics across different levels of granularity. These values represent the overall effectiveness of the sentiment analysis model in terms of precision and recall. For instance, Rouge1 F-measure ranges from 0.3167 to 0.4237, reflecting the balance between precision and recall in capturing unigram overlaps. Similarly, Rouge2 F-measure varies from 0.0944 to 0.1723, indicating the model's effectiveness in capturing bigram overlaps.

The tables highlight the performance of sentiment analysis models on the Samsum dataset across different levels of granularity. Precision, recall, and F-measure metrics provide valuable insights into the model's ability to accurately predict positive outcomes, capture relevant information, and achieve a balance between precision and recall. These evaluations are essential for assessing the effectiveness of sentiment analysis models and guiding future improvements in natural language processing tasks.

4.2 Diarization

Figure 4.1 presents a Pie Chart of Diarization outcomes, showcasing a breakdown of speaker activity within the analyzed multimedia content. In the framework, diarization, enabled by the Whisperx large v2 model, plays a pivotal role in segmenting and clustering audio recordings based on speaker identity. The state-of-the-art neural network architecture excels in accurately identifying and distinguishing between different speakers, laying the foundation for comprehensive analysis of speaker dynamics. The pie chart illustrates that speaker 1 contributes to 31% of the dialogue, while speaker 2 dominates with 69% of the total speaking time.

The graphical representation offers nuanced insights into the fluctuations in speaker activity throughout the duration of the multimedia content. By mapping time along the x-axis and speaking intensity along the y-axis, the visualization provides a detailed depiction of speaker engagement. Peaks in the graph signify periods of heightened speech activity, indicative of significant interaction or dialogue exchanges, while troughs represent conversational lulls or instances of relative silence. The dynamic visualization empowers users to discern the rhythm and flow of the conversation, thereby enhancing their comprehension and interpretation of the multimedia content.

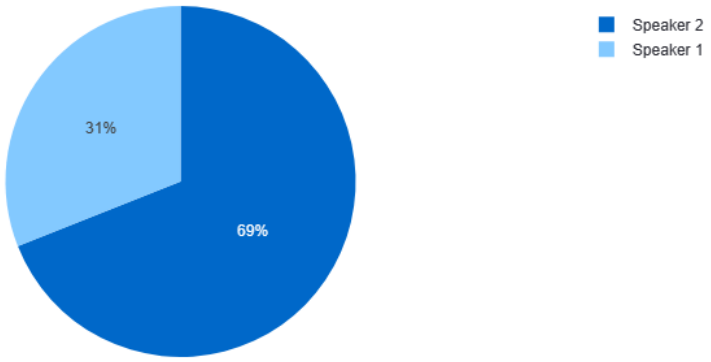


Figure 4.1: Pie Chart of Diarization

Moreover, the robust architecture of the Whisperx large v2 model, trained on diverse datasets, ensures reliable performance across various multimedia contexts. By integrating advanced diarization techniques facilitated by the model, the framework aims to provide users with deeper insights into conversational dynamics and enrich their overall multimedia consumption experience. Through such insights, users can glean valuable information about speaker participation, interaction patterns, and conversational dynamics, enabling more informed analysis and interpretation of multimedia content.

4.3 Summarization

The study utilized the DialogLED architecture, a framework built upon the LED model, which integrates innovative advancements in natural language processing. The architecture enhances dialog summarization tasks by leveraging pre-trained models fine-tuned on specific datasets. Specifically, the implementation is based on the DialogLED model, detailed in the referenced paper. Fine-tuning the MingZhong/DialogLED-base-16384 model with the Dialogsum dataset, the aim was to produce concise summaries of dialogues. Due

to resource constraints, the model was fine-tuned for 10 epochs, striking a balance between computational efficiency and model performance. Leveraging the capabilities of the DialogLED architecture, the study sought to distill complex dialogues into coherent and informative summaries, catering to diverse applications in information retrieval and conversational analysis.

The DialogLED model, with the capacity to accommodate approximately 21,000 tokens, offers a versatile solution for summarizing lengthy transcripts. To handle transcripts exceeding the token limit, devised a chunking mechanism, dividing longer transcripts into manageable segments. Each segment is then individually passed through the model, generating concise summaries. These summaries are later combined to form a comprehensive overview of the entire dialogue, ensuring no information is lost during the summarization process. The approach enables the model to effectively distill key insights from extensive conversations, facilitating efficient comprehension and analysis.

The results demonstrate the efficacy of the DialogLED architecture in producing informative and coherent summaries of dialogues. Despite the limited fine-tuning epochs, the model showcases robust performance in capturing the essence of conversations. By employing advanced natural language processing techniques, DialogLED effectively condenses lengthy transcripts into succinct summaries while maintaining semantic coherence and relevance. The capability holds significant promise for various applications, including information retrieval, conversational analysis, and content summarization, underscoring the versatility and effectiveness of the DialogLED framework.

The output generated from the DialogLED architecture reflects the proficiency in producing informative and coherent summaries of dialogues. Despite the limited fine-tuning epochs, the model adeptly captures the essence of conversations, condensing lengthy transcripts into succinct summaries while maintaining semantic coherence and relevance. By leveraging advanced natural language processing techniques, DialogLED effectively dis-

tills key insights from extensive conversations, facilitating efficient comprehension and analysis. The output holds significant promise for various applications, including information retrieval, conversational analysis, and content summarization, underscoring the versatility and effectiveness of the DialogLED framework.

4.4 Video Level Sentiment Analysis

Sentiment analysis, driven by the advanced RoBERTa model, offers a sophisticated approach to decoding emotional nuances within multimedia content. With the robust natural language processing capabilities, RoBERTa accurately categorizes emotions such as Anger, Anticipation, Disgust, Fear, Joy, Love, and Optimism, capturing subtle shifts and overarching trajectories. Integrating RoBERTa into the sentiment analysis pipeline ensures scalability and adaptability across diverse multimedia contexts, empowering stakeholders with comprehensive sentiment insights for informed decision-making and content optimization. The utilization of RoBERTa represents a significant advancement in sentiment analysis technology, allowing for more nuanced understanding and interpretation of emotional content, ultimately enhancing the user experience and driving greater engagement.

Visual representations like line charts and heat maps further enhance comprehension of emotional dynamics, facilitating the identification of key moments of resonance within the multimedia content. The Line Chart of Video Level Sentiment Analysis, depicted in Fig. 4.2 and 4.3, provides a chronological portrayal of emotional evolution over time, revealing fluctuations in emotions like Anger, Anticipation, Disgust, Fear, Joy, Love, and Optimism. For instance, Anger may peak during intense scenes before subsiding, reflecting the dynamic nature of emotional engagement. Conversely, the Heat Map, featured in Fig. 4, offers a spatial view of emotional intensity and distribution across the content's duration.

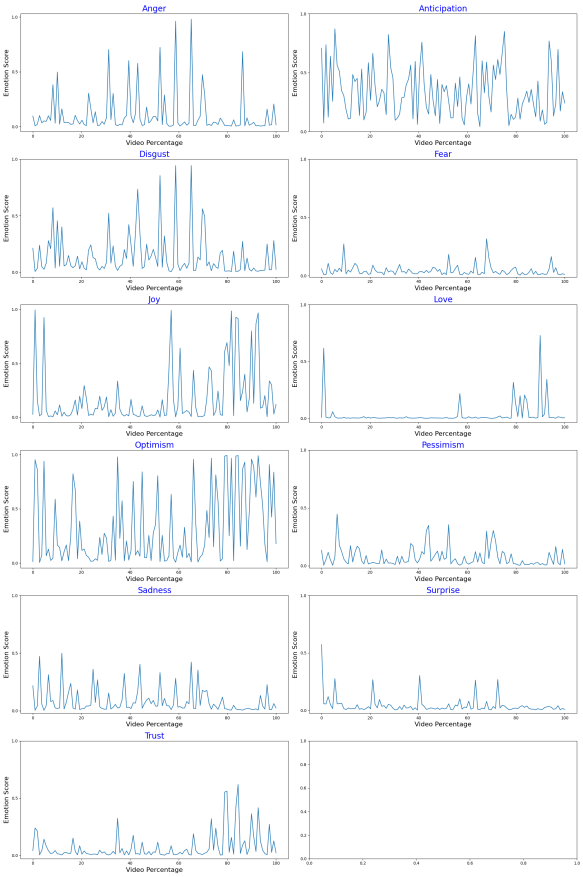


Figure 4.2: Line Map of Video Level Sentiment Analysis

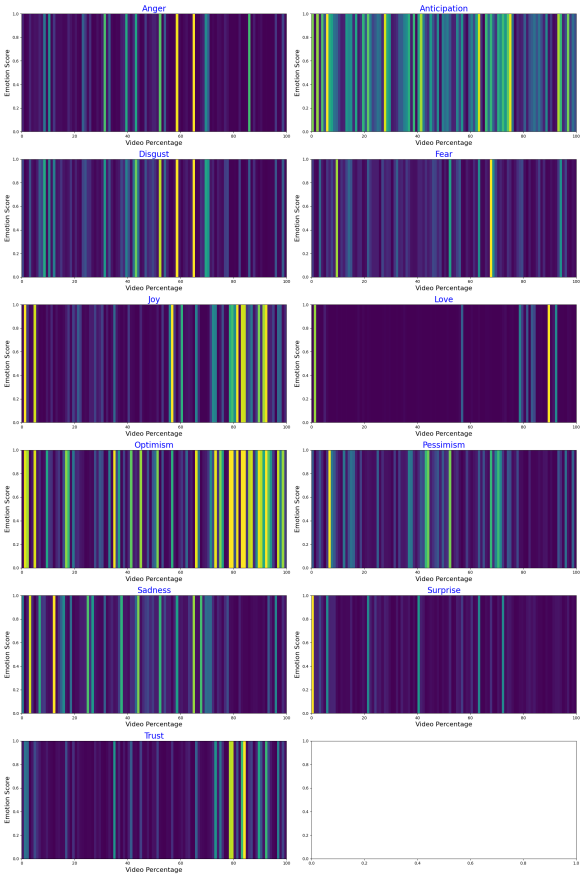


Figure 4.3: Heat Map of Video Level Sentiment Analysis

These visualizations not only serve as invaluable tools for unraveling the intricate tapestry of emotions unfolding throughout the content's duration but also offer actionable insights for content creators and analysts. Peaks in Disgust or Fear may align with scenes evoking strong negative emotions, providing cues for refining content strategies or adjusting narrative elements. Conversely, peaks in Joy signify moments of delight enhancing viewer satisfaction, offering opportunities for amplifying positive engagement. Love and Optimism sporadically spike, indicating instances of deep emotional resonance with the audience, which can be leveraged to strengthen emotional connections and foster brand loyalty. Together, these visualizations provide a comprehensive understanding of emotional dynamics, enriching the viewer's multimedia experience and informing strategic decisions in content creation and audience engagement.

The Line Map of Video Level Sentiment Analysis, as depicted in Fig. 4.2, offers a chronological depiction of emotional evolution over the duration of the multimedia content. The visualization provides insights into the fluctuation of emotions such as Anger, Anticipation, Disgust, Fear, Joy, Love, and Optimism over time. For instance, peaks in Anger may coincide with intense scenes or moments of conflict, while spikes in Joy may signify uplifting or humorous segments. By mapping these emotional trajectories, viewers can discern the emotional rhythm of the content, enhancing understanding and engagement.

On the other hand, the Heat Map of Video Level Sentiment Analysis, featured in Fig. 4.3, offers a spatial view of emotional intensity and distribution across the duration of the multimedia content. Allowing viewers to identify key moments of resonance. Peaks in emotions like Disgust or Fear may align with scenes evoking strong negative emotions, while spikes in Love and Optimism indicate moments of deep emotional connection. By visualizing emotional intensity spatially, viewers gain a deeper appreciation of the emotional landscape traversed during their multimedia journey, enriching their viewing experience and fostering deeper engagement with the content.

4.5 Speaker Level Sentiment Analysis

Sentiment analysis, driven by the advanced RoBERTa model, offers a sophisticated approach to decoding emotional nuances within multimedia content. With the robust natural language processing capabilities, RoBERTa accurately categorizes emotions such as Anger, Anticipation, Disgust, Fear, Joy, Love, and Optimism, capturing subtle shifts and overarching trajectories. Integrating RoBERTa into the pipeline ensures scalability and adaptability across diverse multimedia contexts, empowering stakeholders with comprehensive sentiment insights for informed decision-making and content optimization.

Figure 4.4 showcases the Speaker Level Sentiment Analysis of Anger and Anticipation, offering valuable insights into the emotional journey experienced by viewers engaging with multimedia content. The line chart provides a chronological portrayal of emotional evolution over time, depicting fluctuations in both Anger and Anticipation. Initially, Anger may peak during intense scenes or moments of conflict within the dialogue, reflecting heightened emotional intensity. Conversely, Anticipation exhibits peaks at the beginning of the content, indicating active viewer engagement and heightened expectations for what is to come. These fluctuations in Anger and Anticipation provide a nuanced understanding of viewer responses to different narrative elements and dialogues throughout the multimedia content.

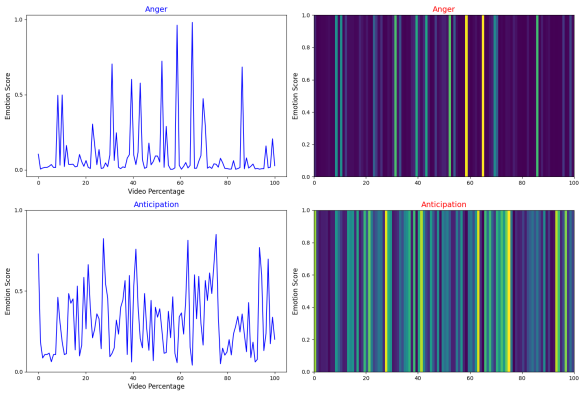


Figure 4.4: Speaker Level Sentiment Analysis of Anger and Anticipation

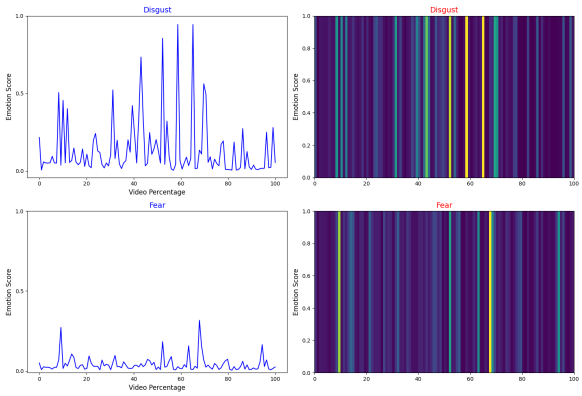


Figure 4.5: Speaker Level Sentiment Analysis of Disgust and Fear

Figure 4.5, the Speaker Level Sentiment Analysis of Disgust and Fear delves into the audience's reactions to specific content segments, revealing intriguing patterns in emotional response. The line chart depicts fluctuations in Disgust and Fear over time, highlighting moments of heightened emotional intensity or unease. Disgust may register sporadic spikes in response to particular segments, possibly indicating aversion or repulsion towards certain content elements. Similarly, Fear exhibits sporadic peaks, reflecting moments of tension or suspense within the multimedia content. These insights into Disgust and Fear contribute to a deeper understanding of audience engagement and emotional dynamics

within the dialogue. The exploration of Disgust and Fear sheds light on the emotional landscape traversed by viewers, providing valuable cues for content creators to tailor their offerings and optimize audience satisfaction.

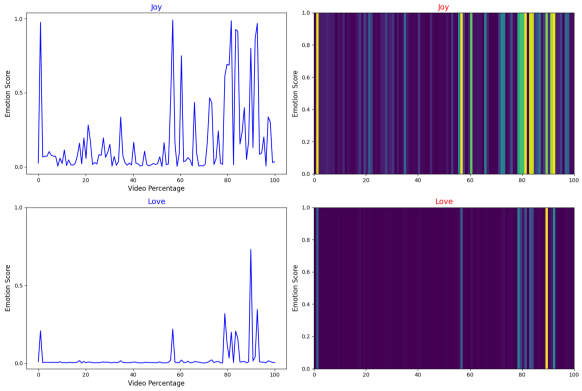


Figure 4.6: Speaker Level Sentiment Analysis of Joy and Love

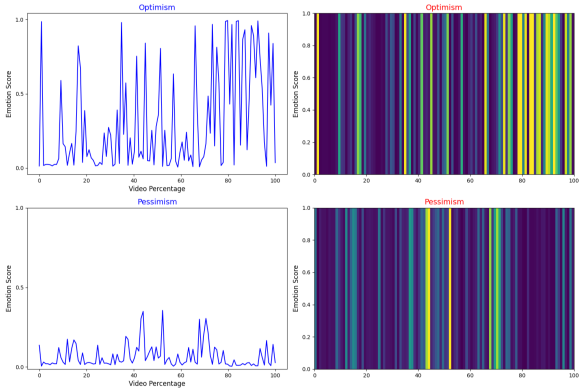


Figure 4.7: Speaker Level Sentiment Analysis of Optimism and Pessimism

Figure 4.6 shifts focus to the Speaker Level Sentiment Analysis of Joy and Love, unraveling the emotional landscape traversed by viewers throughout. The line chart presents fluctuations in Joy and Love over time, capturing moments of emotional resonance and delight within the dialogue. Joy may peak during uplifting or humorous segments, reflecting audience enjoyment and satisfaction with the content. Love shows occasional spikes, reflecting deep emotional connections with specific narrative elements. Insights into Joy and Love enrich the viewer's emotional experience and provide valuable cues for content creators to enhance audience engagement.

Exploring Figure 4.7, delve into the Sentiment Analysis of Optimism and Pessimism, uncovering nuanced emotional dynamics within the multimedia. The line chart showcases fluctuations in Optimism and Pessimism over time, revealing shifts in viewer outlook and perception throughout the dialogue. Optimism may peak during moments of hope, reflecting audience optimism towards the unfolding narrative. Conversely, Pessimism exhibits fluctuations, indicating moments of uncertainty within the dialogue. Insights into Optimism and Pessimism can guide content creators in optimizing audience satisfaction.

Figure 4.8 shifts focus to the Speaker Level Sentiment Analysis of Sadness and Surprise, shedding light on the audience's emotional responses to specific content segments. The line chart captures fluctuations in Sadness and Surprise over time, highlighting moments of emotional depth and unexpected twists within the dialogue. Sadness may register peaks during poignant or emotionally charged scenes, reflecting audience empathy and emotional investment in the narrative. Meanwhile, Surprise exhibits sporadic spikes, indicating moments of unexpected revelation or plot twists that captivate the audience's attention. These insights into Sadness and Surprise deepen the viewer's emotional engagement with the multimedia content.

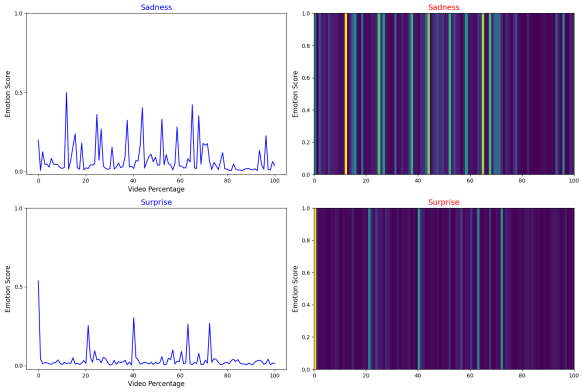


Figure 4.8: Speaker Level Sentiment Analysis of Sadness and Surprise

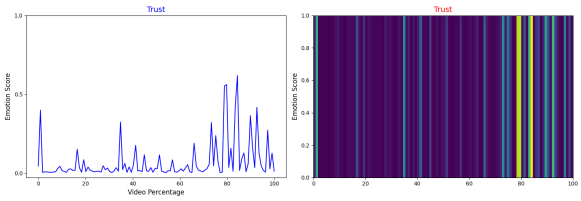


Figure 4.9: Speaker Level Sentiment Analysis of Trust

Figure 4.9 explores the Speaker Level Sentiment Analysis of Trust, revealing the audience's emotional responses to elements of trust and reliability within the dialogue. The line chart illustrates fluctuations in Trust over time, showcasing moments of trust-building or betrayal throughout the narrative. Trust may peak during instances of sincerity or honesty, reflecting audience trust in the characters or storyline, while declines may signal moments of deception or ambiguity, impacting viewer engagement. These insights offer valuable cues for content creators to foster audience trust and enhance viewer engagement.

4.6 Discussion

In the realm of diarization, as depicted in Figure 4.1, the Pie Chart of Diarization outcomes provides a succinct breakdown of speaker activity within the analyzed multimedia content. Utilizing the Whisperx large v2 model, diarization enables accurate segmentation and clustering of audio recordings based on speaker identity, allowing for comprehensive analysis of speaker dynamics. The visualization illustrates that speaker 1 contributes to 31% of the dialogue, while speaker 2 dominates with 69% of the total speaking time. The insight offers nuanced understanding into the fluctuations in speaker activity throughout the multimedia content, empowering users to discern the rhythm and flow of the conversation.

Moving to sentiment analysis, driven by the advanced RoBERTa model, the framework deciphers emotional nuances within multimedia content. The Line Chart and Heat Map of Video Level Sentiment Analysis, as shown in Figures 4.2 and 4.3, offer chronological and spatial portrayals of emotional evolution over time, capturing subtle shifts and overarching trajectories in emotions such as Anger, Anticipation, Disgust, Fear, Joy, Love, and Optimism. These visualizations serve as invaluable tools for unraveling the intricate tapestry of emotions unfolding throughout the content's duration, enriching the viewer's multimedia experience and providing valuable cues for content creators to optimize audience engagement.

Transitioning to speaker-level sentiment analysis, Figures 4.4 to 4.9 delve into the emotional journey experienced by viewers through various emotional categories. Each figure presents fluctuations in emotions over time, offering insights into audience reactions to specific content segments. These visualizations deepen the viewer's emotional engagement with the multimedia content, providing comprehensive insights into emotional dynamics. Through such insights, content creators can tailor their offerings to optimize emotional resonance and enhance overall viewer engagement.

Chapter 5

Conclusion

In conclusion, the documentation offers a comprehensive exploration of multimedia dynamics, providing insights and analysis into various aspects of content creation, consumption, and interpretation. Through advanced techniques such as diarization and sentiment analysis, powered by cutting-edge models like Whisperx and RoBERTa, the study delved into the intricate tapestry of multimedia content, unraveling the emotional nuances and conversational dynamics. From the breakdown of speaker activity to the visualization of emotional trajectories, the framework offers valuable tools for understanding and optimizing multimedia experiences.

By leveraging diarization techniques facilitated by the Whisperx large v2 model a nuanced understanding of speaker dynamics within multimedia content, empowering users to discern patterns of engagement and interaction. Meanwhile, sentiment analysis driven by RoBERTa has provided deep insights into the emotional landscape traversed by viewers, enabling stakeholders to tailor content strategies for maximum impact and resonance. Through visual representations like line charts, heat maps, and pie charts,the study enhanced comprehension and interpretation of multimedia content, fostering deeper connections between creators, distributors, and audiences.

As the study navigate the ever-evolving landscape of multimedia content, the documentation serves as a roadmap for unlocking the full potential of diarization and sentiment analysis techniques. By harnessing the power of advanced models and visualization tools, and continue to push the boundaries of multimedia exploration, driving innovation and enhancing the overall user experience. From content creators seeking to optimize engagement to analysts unraveling the complexities of human emotion, the framework offers a versatile toolkit for understanding and harnessing the power of multimedia dynamics.

5.1 Future Scope

The study opens avenues for future research and enhancements. Several key areas can be explored to advance the diarization system:

- **Advanced Diarization Techniques :** Explore advanced machine learning algorithms and neural network architectures to enhance diarization accuracy and efficiency. Integrate contextual information and multimodal data sources for improved performance in complex multimedia environments.
- **Sophisticated Sentiment Analysis Models :** Leverage state-of-the-art models like BERT and GPT to capture subtle emotional nuances within multimedia content. Fine-tune models on domain-specific datasets and explore novel training strategies for context-aware sentiment analysis.
- **Multimodal Fusion :** Develop techniques for combining audio, video, and text modalities to provide a holistic understanding of multimedia content. Explore multimodal fusion approaches to unlock new insights into conversational dynamics, emotional expression, and content semantics.
- **Real-Time Multimedia Analytics :** nvestigate lightweight algorithms and distributed computing frameworks for real-time diarization and sentiment analysis of

live audio and video streams. Enable applications such as live event monitoring, social media analysis, and interactive multimedia systems.

- **Ethical and Privacy Considerations :** Address ethical and privacy concerns related to data collection, processing, and usage in multimedia analytics systems. Collaborate with stakeholders to develop ethical guidelines and regulatory frameworks ensuring fairness, transparency, and user privacy.

The future scope encompasses a continuous pursuit of innovation and refinement, ensuring that the diarization system remains at the forefront of advancements in audio and video processing.

Chapter 6

Project Activities and Outreach

The chapter focuses on showcasing the achievements and outreach activities of the study. Including details of participation in a project competition, where the innovative idea was presented and received positive feedback from judges. Additionally, discuss the submission of a conference paper to the IEEE SPICES 2024 conference, outlining the key aspects of the research presented in the paper.

6.1 Project Competition

The team participated in a project competition conducted by Carmel College of Engineering and Technology, Alappuzha on 29th April 2024. The team presented the idea to the judges, who were quite impressed with the uniqueness of the idea and the potential.The figure 6.2 shows the brochure of the project competition.

Figure 6.1: Attending Project Competition at CCET Alappuzha

Figure 6.2: Project Competition Brochure

6.2 Conference Paper Submission

A conference paper has been submitted to the 4th IEEE International Conference on Signal Processing, Informatics, Communication, and Energy Systems 2024 (IEEE SPICES 2024), which is a flagship conference of IEEE Kerala section. The details of the conference are provided in Table 6.1

Conference Date	20 - 22 September 2024
Call for Paper Announcement	01 January 2024
Full Paper Submission Hard Deadline	10 May 2024
Notification of Acceptance	28 June 2024
Camera-Ready Paper Submission Deadline	24 July 2024
Registration Starts	01 August 2024
Registration Ends	15 August 2024
Venue	IIIT Kottayam
Mode	Offline

Table 6.1: IEEE SPICES 2024 Conference Details

The conference paper submitted was titled **"Dialogue Analyzer (DiAna)"** presents the innovative approach for efficient and accurate transcription solutions in today's multimedia landscape. With the increasing volume of audio content across various platforms, there arises a demand for tools capable of converting spoken dialogue into written text accurately and swiftly. The study aims to fulfill the need by leveraging advanced natural language processing techniques to develop a transcription model that not only ensures high accuracy but also maintains temporal context through precise timestamping. By providing a reliable transcription solution, the study seeks to enhance comprehension, analysis, and documentation of spoken content, catering to the diverse needs of users across different domains. The figure 6.2 shows the brochure of the project competition.

Figure 6.3: IEEE SPICES 2024 Brochure

References

[1] J. Chen and H. Zhuge, "Extractive Text-Image Summarization Using Multi-Modal RNN," 2018 14th International Conference on Semantics, Knowledge and Grids (SKG), Guangzhou, China, 2018, pp. 245-248, doi: 10.1109/SKG.2018.00033.

[2] W. Liu, Y. Gao, J. Li and Y. Yang, "A Combined Extractive With Abstractive Model for Summarization," in IEEE Access, vol. 9, pp. 43970-43980, 2021, doi: 10.1109/ACCESS.2021.3066484.

[3] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue and P. García, "Encoder-Decoder Based Attractors for End-to-End Neural Diarization," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 1493-1507, 2022, doi: 10.1109/TASLP.2022.3162080.

[4] S. Ghodratnama, A. Beheshti, M. Zakershahrak and F. Sobhanmanesh, "Extractive Document Summarization Based on Dynamic Feature Space Mapping," in IEEE Access, vol. 8, pp. 139084-139095, 2020, doi: 10.1109/ACCESS.2020.3012539.

[5] P. K. Biswas and A. Iakubovich, "Extractive Summarization of Call Transcripts," in IEEE Access, vol. 10, pp. 119826-119840, 2022, doi: 10.1109/ACCESS.2022.3221404.

[6] M. Ferràs, S. Madikeri, P. Motlicek and H. Bourlard, "System fusion and speaker linking for longitudinal diarization of TV shows," 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Shanghai, China, 2016, pp. 5495-5499, doi: 10.1109/ICASSP.2016.7472728.

[7] A. Dilawari, M. U. G. Khan, S. Saleem, Zahoor-Ur-Rehman and F. S. Shaikh, "Neural Attention Model for Abstractive Text Summarization Using Linguistic Feature Space," in IEEE Access, vol. 11, pp. 23557-23564, 2023, doi: 10.1109/ACCESS.2023.3249783.

[8] S. Horiguchi, S. Watanabe, P. García, Y. Takashima and Y. Kawaguchi, "Online Neural Diarization of Unlimited Numbers of Speakers Using Global and Local Attractors," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 31, pp. 706-720, 2023, doi: 10.1109/TASLP.2022.3233237.

[9] A. Phaphuangwittayakul, Y. Guo, F. Ying, W. Xu and Z. Zheng, "Self-Attention Recurrent Summarization Network with Reinforcement Learning for Video Summarization Task," 2021 IEEE International Conference on Multimedia and Expo (ICME), Shenzhen, China, 2021, pp. 1-6, doi: 10.1109/ICME51207.2021.9428142.

[10] H. Suda, D. Saito, S. Fukayama, T. Nakano and M. Goto, "Singer Diarization for Polyphonic Music With Unison Singing," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 30, pp. 1531-1545, 2022, doi: 10.1109/TASLP.2022.3166262.

[11] D. Singhal, K. Khatter, T. A and J. R, "Abstractive Summarization of Meeting Conversations," 2020 IEEE International Conference for Innovation in Technology (INOCON), Bangluru, India, 2020, pp. 1-4, doi: 10.1109/INOCON50539.2020.9298305.

[12] Ming Zhong, Pengfei Liu, Yiran Chen, Danqing Wang, Xipeng Qiu and Xuanjing Huang, "Extractive summarization as text matching", 58th Annual Meeting of the Association for Computational Linguistics, July 5-10 2020.

[13] G. Soldi, C. Beaugeant and N. Evans, "Adaptive and online speaker diarization for meeting data," 2015 23rd European Signal Processing Conference (EUSIPCO), Nice, France, 2015, pp. 2112-2116, doi: 10.1109/EUSIPCO.2015.7362757.

[14] T. Yokota, Q. Zhao and A. Cichocki, "Smooth PARAFAC Decomposition for Tensor Completion," in IEEE Transactions on Signal Processing, vol. 64, no. 20, pp. 5423-5436, 15 Oct.15, 2016, doi: 10.1109/TSP.2016.2586759.

[15] M. -H. Su, C. -H. Wu and H. -T. Cheng, "A Two-Stage Transformer-Based Approach for Variable-Length Abstractive Summarization," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 28, pp. 2061-2072, 2020, doi: 10.1109/TASLP.2020.3006731.

[16] Q. Wang, C. Downey, L. Wan, P. A. Mansfield and I. L. Moreno, "Speaker Diarization with LSTM," 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Calgary, AB, Canada, 2018, pp. 5239-5243, doi: 10.1109/ICASSP.2018.8462628.

[17] G. Yapinus, A. Erwin, M. Galinium and W. Muliady, "Automatic multi-document summarization for Indonesian documents using hybrid abstractive-extractive summarization technique," 2014 6th International Conference on Information Technology and Electrical Engineering (ICITEE), Yogyakarta, Indonesia, 2014, pp. 1-5, doi: 10.1109/ICITEED.2014.7007896.

[18] Y. Fujita, S. Watanabe, S. Horiguchi, Y. Xue and K. Nagamatsu, "End-to-end neural diarization: Reformulating speaker diarization as simple multi-label classification", ArXiv, 2020.

[19] S. Horiguchi, Y. Fujita, S. Watanabe, Y. Xue and K. Nagamatsu, "End-to-end speaker diarization for an unknown number of speakers with encoder-decoder based attractors", ArXiv, 2020.

[20] N. Ryant, K. Church, C. Cieri, A. Cristia, J. Du, S. Ganapathy, et al., "The Second DIHARD Diarization Challenge: Dataset Task and Baselines", Interspeech, pp. 978-982, 2019.

[21] S. H. Shum, N. Dehak, R. Dehak and J. R. Glass, "Unsupervised methods for speaker diarization: An integrated and iterative approach", IEEE Trans. on ASLP, vol. 21, no. 10, pp. 2015-2028, 2013.

[22] Q. Wang, C. Downey, L. Wan, P. A. Mansfield and I. L. Moreno, "Speaker diarization with LSTM", ICASSP, pp. 5239-5243, 2018.

[23] D. Snyder, D. Garcia-Romero, G. Sell, D. Povey and S. Khudanpur, "X-vectors: Robust DNN embeddings for speaker recognition", ICASSP, pp. 5329-5333, 2018.

[24] L. Wan, Q. Wang, A. Papir and I. L. Moreno, "Generalized end-to-end loss for speaker verification", ICASSP, pp. 4879-4883, 2018.

[25] Y. C. Liu, E. Han, C. Lee and A. Stolcke, "End-to-End Neural Diarization: From Transformer to Conformer", Interspeech, pp. 3081-3085, 2021.

VISION & MISSION OF THE DEPARTMENT

Vision

To achieve excellence in Artificial Intelligence and Data Science to cater to the ever-changing industrial and socio-economic needs.

Mission

- To provide high-quality and value-based technical education in the Artificial Intelligence and Data Science program.
- To establish an infrastructure fostering industry-institute interaction in order to meet global expectations and requirements.
- To empower students to become globally competent and effective problem-solvers to develop entrepreneurial skills and higher studies.