INSIGHT EXTRACTION USING MULTIMODAL ANALYSIS OF CONSUMER VIDEOS

20BCE2231| Akshat Swaminath | Dr. Uma Priya D| SCOPE

Introduction

The project aims to develop a system that can transcribe, summarize, and perform sentiment analysis on Focus Group Discussion Videos. The goal of the project is to reduce the processing time from data collection to generating and presenting reports, and to provide insights into the conversations taking place in the videos by the use of Large Language Models to make the work easier.

Motivation

Analyzing consumer videos goes beyond words, revealing hidden emotions, context, and real-time trends, leading to a deeper understanding of consumer behavior for better marketing and product development. There is a need to solve this problem as time is the most valuable asset in corporate world.

Scope of the Project

This project aims to automate the time-consuming manual process of analyzing consumer interaction videos, which typically takes almost a week to complete. By leveraging Large Language Models (LLMs) [1] like OpenAI's Whisper [2] for video transcription and Hugging Face's LLaMA 2 [3] for summarization and sentiment analysis, the project seeks to significantly reduce processing time and free up valuable human resources. The project focuses on creating user-friendly solutions that require minimal computer knowledge to operate, utilizing publicly available models that can be adapted to project needs.

Methodology

The integration of AI into problem-solving in corporate practices addresses the time-consuming nature of manual video processing, which can extend up to 2-3 weeks for a single video. The proposed solution comprises two phases: Phase 1 involves AI-assisted preparation for focus group discussions, including generating discussion questions and selecting candidates based on surveys, while Phase 2 focuses on post-recording tasks. The Whisper model is utilized for translating video transcripts and generating various file formats, while LLMs, specifically the Llama-2 model, aid in drawing insights and summarizing video content. The system provides a chat interface for human agents to query and retrieve contextualized information from the videos, significantly reducing processing time and improving efficiency.

The following are the Large Language Models that are used in this project

1. OpenAI's Whisper Model:
- Utilized for video transcription. Supports over 90 languages and is robust against noise and accents.
- Utilizes a Transformer-based architecture.
- Trained on a large dataset of audio and text, comprising 680,000 hours. Capable of translating non-English languages into English with acceptable Word Error Rates.

2. Meta AI's Llama-2-7b-chat-hf Model:
- Employed for summarization and sentiment analysis. Optimized for dialogue tasks, providing comprehensive responses.
- Trained on extensive and diverse datasets, including publicly available text and code sources.
- Fine-tuned for dialogue tasks through dedicated datasets and human-annotated examples.

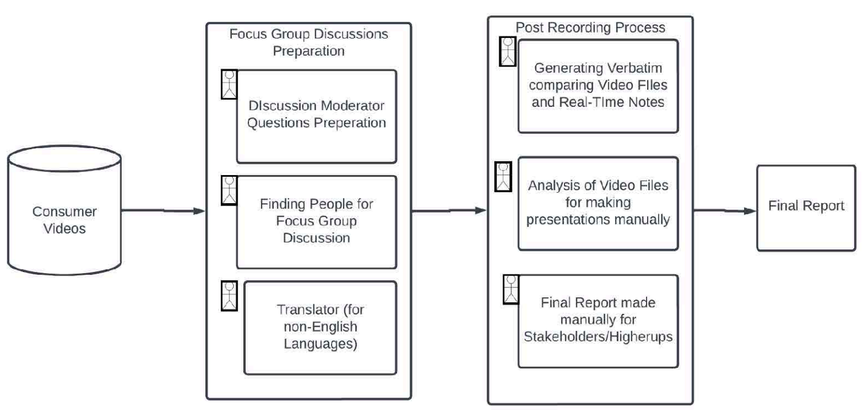

Figure 1: Conventional Workflow of Insight Extraction

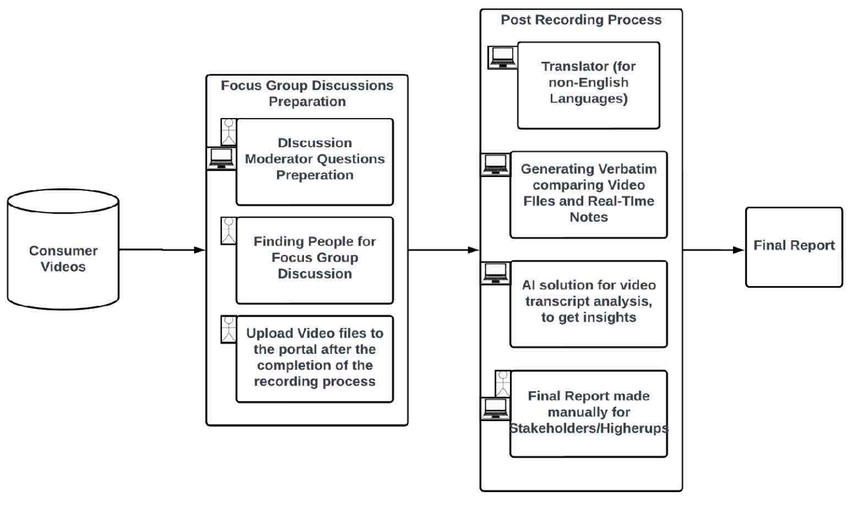

Figure 2: Proposed Workflow of Insight Extraction

Results

For the analysis of the ASR models, Deepgram Playground [4] was used as it provided a simple interface to use most of the models we wanted to compare quickly. Using the API Playground, the confidence scores for all the models were tested on the video files from the Dinner Party Corpus Dataset [5], which was used as a standard to check the scores for all 5 selected models. The following are the models used in this project:

1) Nova-2 [4]: The deepest-trained ASR model on the market, with 18% more accuracy than the previous Nova model.

2) Whisper-medium [3]: A family of encoder/decoder models trained on a large corpus of multilingual speech data by OpenAI.

3) Nova-General [4]: The deepest-trained ASR model, ideal for voice applications that require high accuracy in different contexts.

4) General-Enhanced [4]: Model useful for high accuracy timestamps, lower word error rates, and use cases that require keyword boosting.

5) Base-General [4]: Model useful for high-accuracy timestamps and large transcription volumes.

The analysis of ASR models on the Dinner Party Corpus indicates that certain models, like nova-2, whisper-medium, and nova-general, consistently produce high confidence scores, making them well-suited for precise transcription tasks. Others, such as general-enhanced and base-general, excel in tasks like timestamp accuracy and handling large transcription volumes efficiently, despite slightly lower average confidence scores. This suggests their suitability for specific use cases, such as processing extensive audio data volumes efficiently.

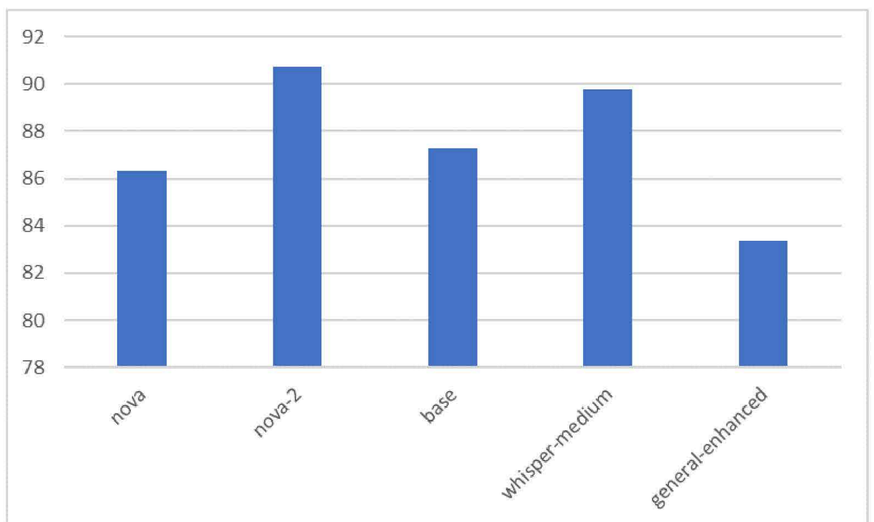

Fig 3: Average Confidence Score for all the ASR models.

Conclusion

By the experimental analysis, we were able to decide which model to use for the implementation of the project.

The confidence scores of whisper-medium and nova-2 models proved the hypothesis that ASR models are far better than conventional technologies like HMM etc, which can be used for NLP tasks.

Furthermore with the development in the area of LLMs, we can expect the technology to get better as the time passes and that would make it easier for general public to make use of these technologies.

References

[1] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell et al., “Language models are few-shot learners,” Advances in neural information processing systems, vol. 33, pp. 1877–1901, 2020.

[2] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, “Robust speech recognition via large-scale weak supervision,” 2022.

[3] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn,
S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi,
A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, “Llama 2: Open foundation and fine-tuned chat models,” 2023.

[4] D. gram, “Deepgram api playground," 2024. [Online]. Available: https://developers.deepgram.com/docs/models-languages-overview.

[5] M. V. Segbroeck, Z. Ahmed, K. Kutsenko, C. Huerta, T. Nguyen, B. Hoffmeister, J. Trmal, M. Omologo, and R. Maas, “Dipco - dinner party corpus,” in Interspeech 2020, 2019. [Online]. Available: https://www.amazon.science/publications/dipco-dinner-party-corpus.