VIOLA: Video Integration of Object Detection, language Insights and accessibility

Mr. Jeramiah T Varghese
Student pursing Engineering in Computer Science
Engineering Specialization Artificial Intelligence (AI
and Machine Learning (ML) in Dayananda Sagar University

Mr. C S Jeevan
Student pursing Engineering in Computer Science
Engineering Specialization Artificial Intelligence (AI
and Machine Learning (ML) in Dayananda Sagar University

Mr. kondanda Rama Raju kasuru
Student pursing Engineering in Computer Science
Engineering Specialization Artificial Intelligence (AI
and Machine Learning (ML) in Dayananda Sagar University

Mr. Hridanshu Ruparel
Student pursing Engineering in Computer Science
Engineering Specialization Artificial Intelligence (AI
and Machine Learning (ML) in Dayananda Sagar University

ABSTRACT

VIOLA is project proposes a cutting-edge pipeline for live video analysis by seamlessly integrating YOLO V8, a state-of-the-art object detection model, with Whisper, a leading speech-to-text AI tool developed by OpenAI. The objective is to leverage the rapid advancements in AI to create a cohesive system that enables real-time object detection and speech-to-text conversion. The integration of YOLO V8 and Whisper is anticipated to provide a comprehensive solution for live video understanding and accessibility. Furthermore, the incorporation of GPT-4 aims to enhance the system by offering contextual insights, thereby establishing a powerful trifecta for holistic live video analysis. The expected outcome is a synergistic fusion of cutting-edge technologies, paving the way for advanced applications in areas such as surveillance, accessibility, and immersive user experiences.

Key words: Yolo V8, OpenAi Whisper, GPT 4, VIOLA.

1. INTRODUCTION

This project is dedicated to advancing the accessibility and comprehension of live video streams through the seamless integration of cutting-edge AI technologies. Comprising key components such as real-time object detection, speech-to-text conversion, and languagebased insights, the initiative aims to establish an integrated system with the overarching goals of identifying objects, transcribing spoken content, and generating contextual insights. The comprehensive pipeline spans the entire process, from capturing live video through webcams to presenting enhanced insights to users. The core modules include the development of advanced object detection capabilities using YOLO v8, the integration of the Whisper Speech-to-Text system for precise speech conversion, and the connection to the powerful GPT-4 Language Model for nuanced language-based insights. A pivotal aspect of the project involves the creation of a user-friendly visual interface, facilitating the display of analyzed video content and generated insights. With the potential to revolutionize live video analysis, this initiative stands to make significant contributions to inclusivity, information dissemination, and overall accessibility in diverse application domains. In response to the growing demand for innovative solutions in live video analysis, this project embarks on a mission to enhance accessibility and comprehension by seamlessly integrating state-of-the-art AI technologies. At its core, the initiative weaves together three pivotal components: real-time object detection, speech-to-text conversion, and languagebased insights, with the overarching goal of creating a holistic system that not only identifies objects in real-time but also accurately transcribes spoken content and generates contextual insights. The scope of the project spans the entire process, commencing with the capture of live video through webcams and culminating in the presentation of enriched insights to users. The YOLO v8 Object Detection module is dedicated to the development of advanced capabilities, ensuring the system can accurately identify and track objects with precision. The Whisper Speech-to-Text module is seamlessly integrated to provide an accurate and reliable conversion of spoken content. Adding a layer of sophistication, the GPT-4 Language Model is leveraged to generate nuanced and contextually relevant insights based on the transcribed content. Recognizing the importance of user interaction, a user-friendly visual interface is crafted to facilitate the seamless presentation of analyzed video content and generated insights. The potential impact of this endeavor is profound, as it stands to revolutionize the landscape of live video analysis, contributing significantly to inclusivity, information dissemination, and overall accessibility across a myriad of applications and domains

1.1 WHY YOLO V8?

YOLOv8 is a compelling choice for its excellent performance, resource efficiency, and developer-friendly nature. If you prioritize accurate and fast object detection, especially in resource-constrained situations, YOLOv8 definitely deserves your consideration. Just keep in mind the limitations associated with its young age and evolving ecosystem Some detailed differences between the other YOLO models YOLOv5: The veteran, YOLOv5 remains a popular choice for its balance of accuracy and speed. It offers a range of model sizes from the tiny Nano to the larger X, catering to different performance needs. Its strong community and extensive documentation make it beginner-friendly. However, newer versions might outperform it in some aspects. YOLOv6: This newcomer focuses heavily on real-time inference, meaning it prioritizes speed over absolute accuracy. Its unique architecture aims for efficient calculations on mobile devices. While promising for live applications like real-time video analysis, v6 might not always match the accuracy of its peers. YOLOv7: An exciting challenger, v7 boasts improved accuracy compared to v5 for similar inference times. It introduces innovative features like "Focus-NMS" for better small object detection. While not as lightweight as v6, v7 offers a sweet spot between performance and resource efficiency

YOLOv8: The latest entrant, v8 shines in both accuracy and speed, especially with its smaller models. It utilizes novel architectural choices like the "SiLU" activation function and "Focus" modules for superior performance. However, v8 currently lacks models trained on higher resolutions like 1280, limiting its application for tasks requiring fine detail

Features	YOLO V5	YOLO V6	YOLO V7	YOLO V8
Backbone Network	CSP Darknet53	CSP Darknet53	CSP Darknet53	Efficientnet-p5
Neck	FPN	FPN	FPN	PAN
Head	Yolo V5 head	YoloV6 head	YoloV7 head	YoloV8 head
Input size	640	640	640	640
Output size	1280	1280	1280	1280
FLOPs	152.7	163.7	205.9	197.2
Params	61.3	72.8	88.0	82.9
Accuracy	0.434	0.484	0.528	0.557
Speed	62.5	70.8	54.2	61.5

Table 1.1: Differences of Yolo Models

YOLO v8 is a cutting-edge, real-time object detection and image segmentation model developed by Ultralytics. It builds upon the success of previous YOLO versions and boasts several new features and improvements, making it a powerful tool for various computer vision tasks. A state-of-the-art deep learning model for real-time object detection, classification, and segmentation. Can also handle tasks like pose estimation and tracking. Available as a Python package and command-line interface, making it accessible to a wide range of users.

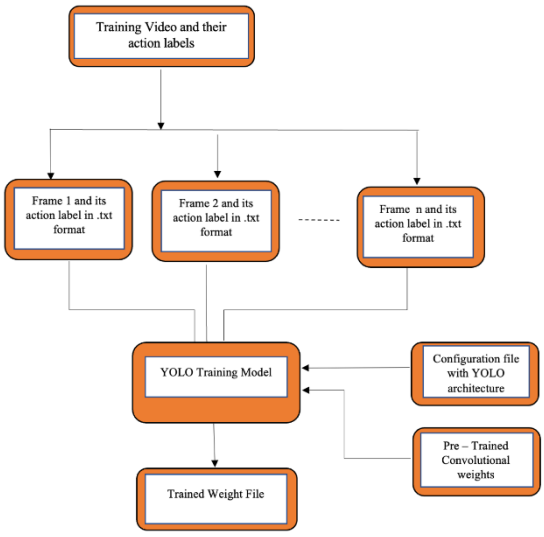


Figure 1.1: Flowchart of YoloV8

1.2 PYTTSX3

Pyttsx3, an abbreviation for Python Text-to-Speech version 3, is a versatile library designed to facilitate the conversion of text into speech within Python applications. With its cross-platform compatibility, developers can seamlessly integrate it across different operating systems, including Windows, macOS, and Linux. This broad compatibility ensures consistent functionality regardless of the platform.

One of the standout features of pyttsx3 is its support for multiple speech synthesis engines. These engines include SAPI5 on Windows, NSSpeechSynthesizer on macOS, and eSpeak on Linux. This diversity enables developers to choose the engine that best suits their needs or the requirements of their target platform.Beyond its engine flexibility, pyttsx3 offers extensive customization options. Developers can fine-tune various aspects of speech output, such as adjusting the speech rate, volume, and even selecting different voices. This level of customization allows for tailored speech generation suited to specific applications or user preferences.

Another significant advantage of pyttsx3 is its support for asynchronous operation. This means that speech generation can occur concurrently with other tasks, enhancing the responsiveness and overall user experience of applications. By enabling nonblocking speech generation, pyttsx3 ensures smooth interaction without delays or interruptions.

pyttsx3 is a powerful tool for incorporating text-to-speech functionality into Python applications. Its cross-platform compatibility, support for multiple speech engines, extensive customization options, and asynchronous operation make it a valuable asset for developers seeking to enhance the accessibility and user experience of their projects.

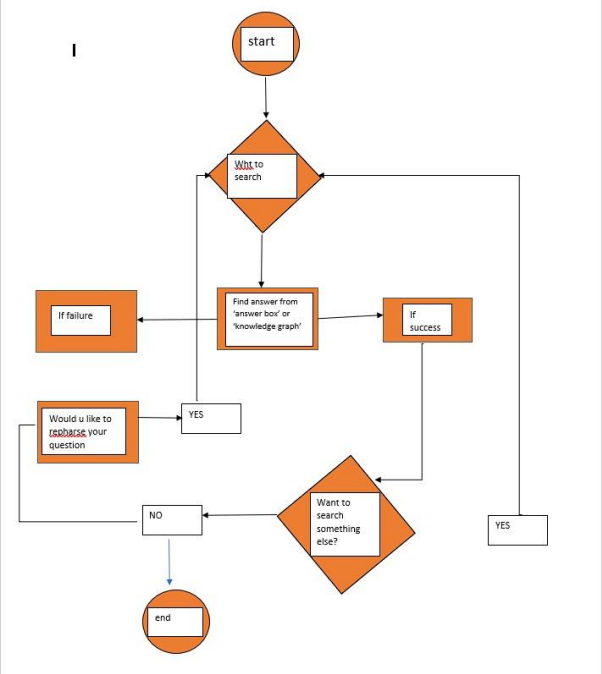


Figure 1.2: Flowchart of PYTSX3

1.3 Google Speech-To-Text

Speech Recognition & Synthesis, previously known as Speech Services, stands as a significant accessibility feature developed by Google for Android devices. This application functions as a screen reader, enabling spoken text on the screen to be read aloud, with extensive language support catering to diverse global audiences.

Google's Speech-to-Text technology, a core component of Speech Recognition & Synthesis, boasts exceptional accuracy in converting spoken language into text. This high level of accuracy ensures reliable transcription of spoken content, enhancing accessibility for users who may rely on text-based communication.

One of the key strengths of Speech Recognition & Synthesis lies in its broad language and dialect support. This inclusivity makes it suitable for a wide range of global applications, accommodating users who speak different languages or regional variations.Real-time transcription capabilities are another standout feature of Speech Recognition & Synthesis. This functionality enables live captioning and transcription services, facilitating

accessibility in various scenarios such as live events, meetings, or educational settings.Users also benefit from customization options within Speech Recognition & Synthesis. They can tailor models to recognize specific vocabularies, phrases, or domain-specific terminology, ensuring accurate transcription in specialized contexts.Moreover, Speech Recognition & Synthesis seamlessly integrates with other Google Cloud services, enhancing scalability, reliability, and security. This integration streamlines the deployment and management of speech recognition and synthesis features within cloud-based applications, ensuring a robust and efficient user experience

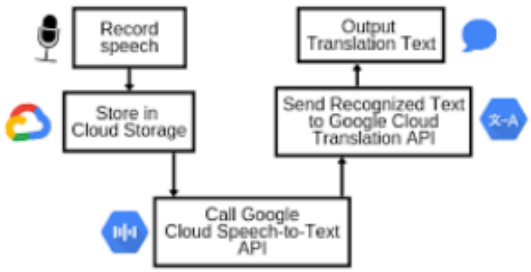


Figure 1.3: Flowchart of YoloV8

2. Literature Survey

[1] Nowadays, allowing unmanned aerial vehicles (UAVs) to accompany humans in daily life has become a hot topic. Pedestrian detection plays an important role on this application with its accuracy and real-time detection. In this paper, we design an on-board real-time pedestrian detection method for micro UAVs based on the YOLO-v8 network. More specifically, several custom network models of small scales are fine-tuned based on YOLO-v8 to detect pedestrians in real time. Besides, our detection methods are implemented and deployed on a custom micro UAV. Through simulation and real-world experiments, the results show that the custom detection network models based on YOLO-v8 can accurately detect pedestrians at up to 34 FPS on Jetson Xavier NX

[2] Accurate tobacco plant detection is crucial for effective agricultural management strategies. Due to the challenges posed by high plant variability, adverse lighting conditions, and occlusions, existing machine learning algorithms and convolutional neural networks (CNNs) have struggled to achieve satisfactory detection accuracy under real-world scenarios. This paper employs YOLOv8, a state-of-the-art object detection algorithm, to reliably detect tobacco plants. A comprehensive evaluation and benchmark of 12 state-of-the-art YOLO object detection algorithms including the YOLOv8, is established for tobacco detection. YOLOv8 demonstrates remarkable test accuracy of 96.4 compared to YOLOv5 and YOLOv7. YOLOv8's real-time object detection ability renders it an ideal solution for mobile and embedded devices, opening new possibilities for on-the-go agricultural management.

[3] Because of the detection of power line damage status in rural and remote areas, a method based on the YOLO algorithm is put forward for power line damage detection. After pre-processing the pictures taken by UAV, YOLO is used to establish the target detection model for damaged power lines, optimize the data structure through data augmentation and data classification technology, and carry out training using multi-layer convolution operation and eigenvalue pyramid structure. The test results indicate that the model can detect different power line targets in various environments, and the complete accuracy of the experiment reaches 91.51 percent. The established UAV patrol mode and back office operation interface can be competent for conducting regular inspections of electric equipment in rural and remote areas.

[4] This study explores the application of Large Language Models (LLMs) and Automatic Speech Recognition (ASR) models in the analysis of right-wing unstructured political discourse in Peru, focusing on how the concept of freedom is framed. Three types of freedom are identified: personal autonomy, economic freedom, and civil liberties. Utilizing the transcription of OpenAI's ASR Whisper and GPT-3.5 and GPT-4 models, interviews with three Peruvian right-wing political leaders are analyzed: Rafael L´opez Aliaga, Hemando de Soto, and Keiko Fujimori. The results show that GPT-4 beats GPT-3.5 in identifying dimensions of freedom, although there are

discrepancies compared to human coding. Despite challenges in classifying abstract and ambiguous concepts, the findings demonstrate GPT 4's ability to classify complexities within political discourse at comparatively small costs and easy access. The research suggests the need for additional refinement, ethical consideration, and ongoing exploration in the analysis of political speeches through Al

[5] Speech data has rich acoustic and paralinguistic information with important cues for understanding a speaker's tone, emotion, and intent, yet traditional large language models such as BERT do not incorporate this information. There has been an increased interest in multi-modal language models leveraging audio and/or visual information and text. However, current multimodal language models require both text and audio/visual data streams during inference/test time. In this work, we propose a methodology for training language models leveraging spoken language audio data but without requiring the audio stream during prediction time. We achieve this via an audio-language knowledge distillation framework, where we transfer acoustic and paralinguistic information from a pre-trained speech embedding (OpenAI Whisper) teacher model to help train a student language model on an audio-text dataset. In our experiments, the student model achieves consistent improvement over traditional language models on tasks analyzing spoken transcripts

3. Methodology

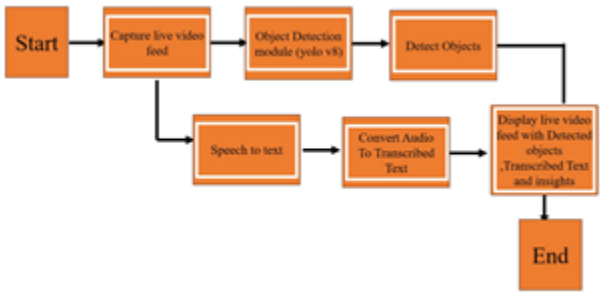


Figure 3.1: Flowchart of YoloV8

The depicted diagram illustrates the comprehensive process of converting speech to text, starting from capturing a live video feed and culminating in displaying transcribed text and insights. Here's an elaboration on each step:

1. Live Video Feed Capture: Initially, the system acquires a live video feed using a device equipped with audio and video recording capabilities, such as a webcam or microphone. This step aims to gather real-time data comprising both visual and auditory elements.

2. Object Detection Module (YOLO v8): The live video feed is then subjected to processing through an Object Detection Module, which employs the YOLO v8 deep learning model. YOLO, or You Only Look Once, is a widely-used object detection algorithm. It divides the video frames into a grid and predicts bounding boxes and class probabilities for each grid cell. This module identifies and locates various objects within the video frames, ranging from people and vehicles to animals or other pre-trained categories.

3. Speech-to-Text Module (Google Speech-to-Text and pyttsx3)**: Simultaneously, the audio segment of the live video feed is directed to a Speech-to-Text Module. This module utilizes either Google Speech-to-Text or pyttsx3, a text-to-speech conversion library. Its purpose is to transcribe spoken words from the audio track into textual representations.

4. Display Live Video Feed with Detected Objects, Transcribed Text, and Insights (User Interface): The outputs from the object detection and speech-to-text modules are integrated into a user interface. This interface overlays the live video feed with various informative elements, including detected objects marked by bounding boxes, transcribed speech content, and any additional insights derived from the data.

Examples of Insights include:

Sentiment Analysis: This involves evaluating the tone of speech to determine sentiment, whether it's positive, negative, or neutral.Person Identification: Additional features such as facial recognition may be employed if the detected object is a person, aiding in identifying individuals.Contextual Information: Relevant details related to the context, such as keywords, topics, or actions occurring in the video, can be extracted.this integrated system

provides a holistic view of the live video feed by combining visual and auditory information while offering valuable insights through object detection and speech-to-text technologies. The user interface serves as a centralized platform for understanding and interacting with the captured data, enhancing the overall comprehension and utility of the system.

4. Results & Analysis

4.1 Yolo v8 Results

The below result shows about the data getting from the camera which is performing object detection using yolo v8 model

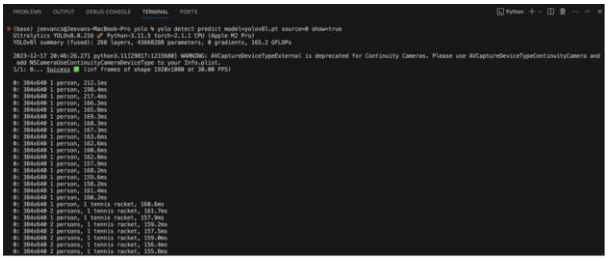
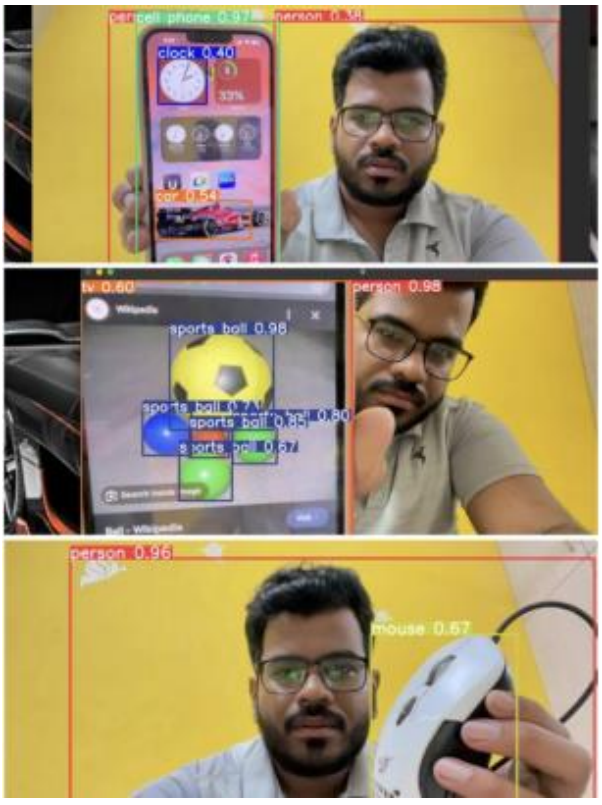


Figure 4.1.1: Result Data



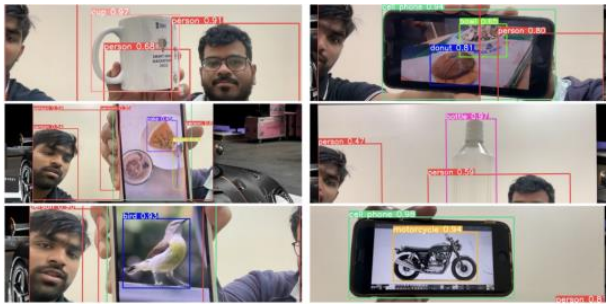


Figure 4.1.2 : output Images

4.2 Open AI- Google STT and PYTTSX3 Results

The below result shows the conversation that is taking place using AI and Whisper which is converting Speech to text and replying to it back in the form of Speech and also displaying the same

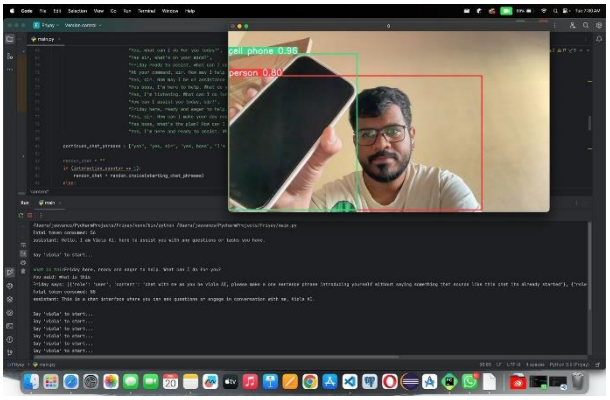


Figure 4.2.1 : output Images

5. Conclusion

In conclusion, the VIOLA project stands as a pioneering effort that integrates advanced technologies to establish a dynamic framework for real-time video analysis. By seamlessly incorporating YOLO V8 for rapid object detection, Whisper for accurate speech-to-text conversion, and GPT-4 for contextual understanding, this endeavor aims to provide a comprehensive solution applicable across various domains including surveillance, accessibility, and immersive user experiences.Positioned at the forefront of leveraging AI advancements, VIOLA commits to harmoniously blending state-of-the-art tools, poised to transform the landscape of intelligent video analysis systems. By leading innovations in live video comprehension, VIOLA aims to provide a foundational platform for sophisticated applications, capable of addressing diverse industry and domain requirements.

Through this initiative, we envision making significant contributions to technological evolution by fostering a robust infrastructure for advanced applications catering to a wide range of needs across industries and sectors.

6. Future Work

Looking forward, the future direction of VIOLA involves a comprehensive strategy aimed at enhancing its capabilities and ensuring its relevance in the evolving field of video analysis. This entails exploring advanced techniques such as multimodal fusion, continual learning, and privacy-preserving mechanisms. By integrating information from various sources and adapting dynamically to new data and scenarios, VIOLA aims to provide deeper insights while adhering to privacy regulations and protecting sensitive information.Furthermore, the focus on semantic understanding and ethical considerations highlights VIOLA's dedication to thorough analysis and responsible deployment of AI technologies. Efforts to deploy VIOLA in real-world settings and refine its user interface will improve its usability and accessibility, facilitating adoption across different domains and user demographics.

In addition, future work will prioritize optimization for edge computing environments, scalability enhancements, and support for distributed processing architectures. Leveraging edge computing infrastructure will enable VIOLA to conduct more efficient analysis closer to the data source, thereby reducing latency and bandwidth requirements. Scalability improvements and distributed processing support will equip VIOLA to handle larger volumes of video data and meet the demands of complex environments effectively.Moreover, exploring collaborative and federated learning approaches will enable VIOLA to harness collective intelligence while respecting data privacy constraints. This will facilitate knowledge sharing and collaboration across diverse organizations and domains. Through these concerted efforts, VIOLA aims to maintain its position at the forefront of intelligent video analysis, continuously evolving to address the evolving needs of industries and society as a whole.

6. References

[1] Z. Li, Z. Liu and X. Wang, ”On-Board Real-Time Pedestrian Detection for Micro Unmanned Aerial Vehicles Based on YOLO-v8,” 2023 2nd International Conference on Machine Learning, Cloud Computing and Intelligent Mining (MLCCIM), Jiuzhaigou, China, 2023, pp. 250-255, doi: 10.1109/MLCCIM60412.2023.00042.

[2] Z. Mahboob, A. Zeb and U. S. Khan, ”YOLO v5, v7 and v8: A Performance Comparison for Tobacco Detection in Field,” 2023 3rd International Conference on Digital Futures and Transformative Technologies (ICoDT2), Islamabad, Pakistan, 2023, pp. 1-6, doi: 10.1109/ICoDT259378.2023.10325705.

[3] T. Di, L. Feng and H. Guo, ”Research on Real-Time Power Line Damage Detection Method Based on YOLO Algorithm,” 2023 IEEE 3rd International Conference on Electronic Technology, Communication and Information (ICETCI), Changchun, China, 2023, pp. 671- 676, doi: 10.1109/ICETCI57876.2023.10176923.

[4] Bianchini, G. E., Zanotti, L., Mel´endez, C. (2023, August 15). Using OpenAI models as a new tool for text analysis in political leaders' unstructured discourse. https://doi.org/10.31234/osf.io/kdngb

[5] Fatema Hasan, Yulong Li, James Foulds, Shimei Pan, Bishwaranjan Bhattacharjee: Teach me with a Whisper: Enhancing Large Language Models for Analyzing Spoken Transcripts using Speech Embeddings. https://doi.org/10.48550/arXiv.2311.07014

[6] Wang, S., Yang, C.H.H., Wu, J. and Zhang, C., 2023. Can Whisper perform speech-based in-context learning. arXiv preprint arXiv:2309.07081.

7] Hasan, Fatema, Yulong Li, James Foulds, Shimei Pan, and Bishwaranjan Bhattacharjee. "Teach me with a Whisper: Enhancing Large Language Models for Analyzing Spoken Transcripts using Speech Embeddings." arXiv preprint arXiv:2311.07014 (2023).