International Journal on Science and Technology (IJSAT)
E-ISSN: 2229-7677 ● Website: www.ijsat.org ● Email: editor@ijsat.org

Sentiment Analysis And Speaker Mapping with Machine Learning

Prof. S. V. Shinde[1], Mayur Ankushrao[2], Aniket Markad[3], Sanket Pawar[4], Vishal Mule[5]

Department of Computer Engineering
PDEA's College Of Engineering, Manjari
Pune
aniket45.markad@gmail.com[1], vishalmule2004@gmail.com[2], mayurankushrao2004@gmail.com[3] , sanketpatilpawar2003@gmail.com[4]

Abstract

In an automated environment when multiple user is interacting with automated system, identification for each one and their sentiment recognition is fundamental for real-time emotion understanding, which is needed in personal and efficient response generation. This work aims at an integrated machine learning architecture that integrates speaker identification and sentiment analysis to realize precise and user-specific emotion recognition. It utilizes Mel Frequency Cepstral Coefficients (MFCC) and Dynamic Time Warping (DTW) as reliable speaker mapping method and applies Support Vector Machine (SVM), Naive Bayes, and Variance-Aware Network Decomposition and Regularization (VADER) for sentiment classification from both the acoustic feature and the text transcript. Performance evaluation has shown strong performance. It ensures that end-user will be receiving practical insights and delicate feelings about a person and that this kind of understanding will also be of service in domains of customer service analytics, education, healthcare monitoring, and recommendation systems.

Keywords—Sentiment, Speaker, User, Mapping, Video, Emotion, Audio

1. INTRODUCTION

In current time of interacting multiple users with the automated systems, distinguishing the speakers and understanding emotions in real time is also becoming important for tailoring the system responses in a personalized and efficient manner. This research tackles the problem by proposing an integrated machine learning framework that brings speaker identification together with sentiment analysis. We propose to achieve precise, speaker-sensitive emotion detection with this kind of approach by taking the profit of complementary audio and text streams. The proposed system operates with MFCC and DTW for efficient speaker mapping and uses Support Vector Machine (SVM), Naive Bayes and VADER algorithms for sentiment classification. Together, these blocks provide the real-time multi-modal fusion and accurate detection of sentiment and speaker identity. Real tests show robust performance across different cases, providing users with actionable and fine-grained emotions of each speaker in real time..

The system presented here which operates at the use level requires MFCC & DTW for reliable speaker mapping and sentiment classification to be performed by a combination of Support Vector Machine

(SVM), Naive Bayes and VADER. Furthermore, in a multi-user conversation scenario, the proposed architecture offers the multi-modal fusion between audio and text, allowing to obtain both sentiment and speaker identification in a real time. Experimental results demonstrate strong performance across all tested settings, as the system continues to provide actionable insight into feelings for the individual level at a variety of test cases, but especially on speaker level. This architecture offers valuable benefits for the fields of customer service analytics, education, healthcare monitoring, and recommender systems. The recognition of the voice and emotional state of each speaker permits us to construct more responsive human-based human computer interaction systems with conversational interfaces in multi-user settings.

This architecture offers practical benefits across domains such as customer service analytics, education, healthcare monitoring, and personalized recommender systems. By ensuring each speaker's voice and emotional state are recognized, the system paves the way for more adaptive, human-centered AI solutions in multi-user conversational contexts.

2. Literature Survey

The works on sentiment analysis and speaker mapping are very abundant in recent research and many different methods for detecting emotions and speaker identities have been proposed in audio- modality as well as text modality. From these recent advancements, classical machine learning to deep architectures have been used in many approaches.

Some recent works are like Mel Frequency Cepstral Coefficients (MFCC) features combined with the Dynamic Time Wrapping (DTW) features and SVM and Naive Bayes for speaker dependent emotional state detection from single conversational datasets which reported better performance in emotion classification. Transformer based multimodal models such as MulT outperform the state-of-the-art in sentiment analysis benchmarks by performing very efficient fusion for unaligned audio modality and text modality. Moreover, Universal Speech Representations via UniSpeech-SAT is a robust model, which is capable of integrating speaker pretraining with both sentiment and speaker recognition, as tested in a variety of domains and speakers.

Other relevant research involves spectrogram-based speaker recognition using AC-SOM neural networks for fast and efficient real-time speaker identification. Most multimodal fusion are better than unimodal sentiment analysis methods by fusing acoustic and textual data streams. Challenges such as domain adaptability, low resource languages, and sarcasm detection are yet open, which makes the need of adaptation at test time and robust fusion framework to be evident.

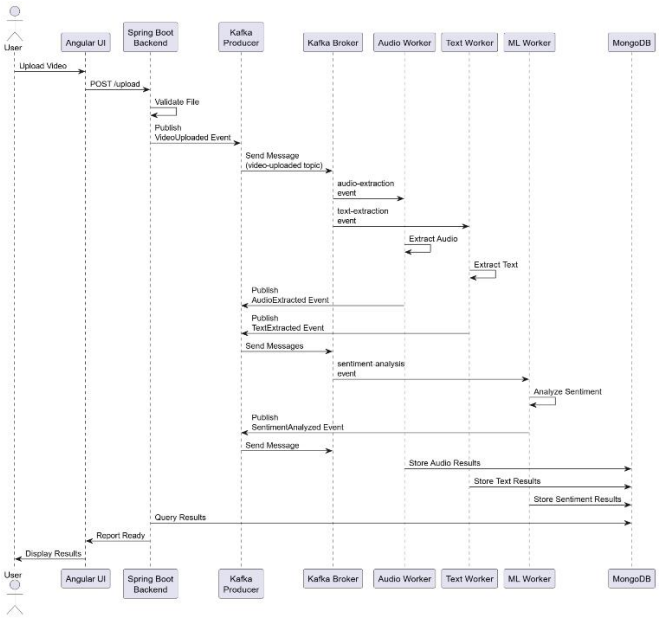
This body of knowledge provide a suitable grounding for the new work, in this paper, we combine multiple machine learning methods, as a base level SVM, NB and VADER, with some audio processing methods, i.e., MFCC and DTW, to construct a truly real time, multi modal sentiment analysis system that can be operated for the emotion state of interest for each speaker in multi-turn conversational setting.

3. Problem Statement

Develop advanced software leveraging machine learning for real-time sentiment analysis and speaker mapping in multi-speaker conversations. Integrate both audio and textual data streams to enhance the

accuracy of emotion detection, support speaker identification, and enable visualization or analytics for personalized and context-aware human-computer interaction applications

4. PROPOSED SYSTEM



This proposes that one will have an all encompassing user centered machine learning system for real-time sentiment analysis and speaker mapping of one or several speakers. It enables that each person is able to talk and is detected at the same time, therefore leading to enhanced customized interactions and analytics. The solution workflow is as follows:

- User uploads an audio or video containing multiple speakers.
- The system extracts audio and generates an accurate text transcript.
- Speaker mapping is performed using Mel Frequency Cepstral Coefficients (MFCC) and Dynamic Time Wrapping (DTW) to reliably identify and label each individual speaker.
- Sentiment analysis is conducted on both the textual transcripts and acoustic features using algorithms such as Support Vector Machine (SVM), Naive Bayes, and VADER.
- The results are fused to generate a comprehensive, speaker-specific sentiment report.
- Users can view and download detailed reports via an interactive dashboard.
- The system allows multiple simultaneous uploads and processing for scalability.
- Strong security measures ensure data privacy and protection throughout.

This streamlined process ensures users can easily upload their data, receive detailed, speaker-specific emotional insights in near real-time, and navigate the system without complexity. The modular design also allows for future extension, such as integration with additional modalities or machine learning models, and deployment on cloud platforms for robust performance and scalability.

5. PROCESS FLOW

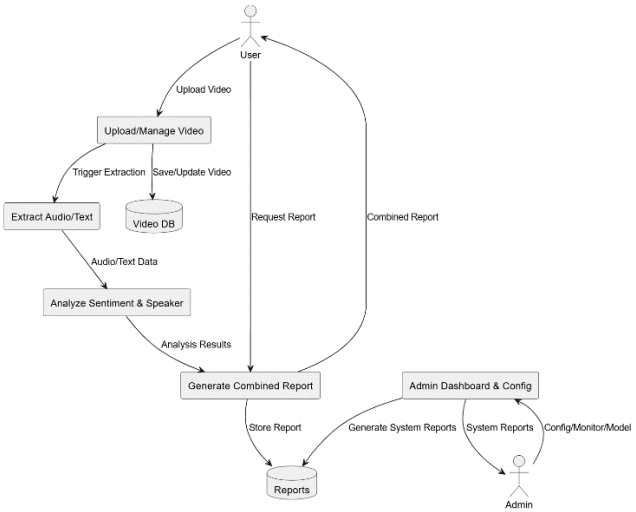


fig. Process flow diagram.

System Requirements

Here are the essential conditions necessary for a robust system that performs sentiment analysis and speaker mapping:

- Real-Time Performance: It has to analyze multi-speaker audio-video inputs at low latency (on the order of seconds) to achieve rapid output for both sentiment analysis and speaker detection, targeting at least 20 utterances per second in processing speed.
- Accuracy: The accuracy requirements (higher-than-85% classiп¬Ґcation accuracy and precision in speaker diarization) of the system should be set to prevent speaker-mapping errors and sentiment classification errors. Precise measurements are vital for dependable output.
- Deployable in edge devices with limited resources and efficient: The system should be runnable on resource-scarce hardware devices such as edge servers or mobile hardware in addition to optimized model choices and minimalistic architecture approaches to seek a reasonable compromise on both performance and power efficiency.
- Effective Fusion of Multimodal Data: It should seamlessly merge both audio and text data streams, extracting acoustic features (MFCC features for instance) and semantic cues from text to help build better emotion detectors.
- User - friendly frontend for uploading and visualization: There must be a user-friendly, web-based dashboard that permits users to upload their own data, analyze it in real time, and get detailed output reports.

6. CONCLUSION

By integrating sentiment analysis with speaker mapping we get enriched experience because we can automatically know in real time, speaker and what they are feeling about. The proposed architecture helps in reducing complex operation and extracting valuable knowledge across numerous domains. It is not only focused on innovation but scalability, usability, and data privacy to make the tech accessible and ready for

use. By developing this work we alleviate the limitations in multimodal-based sentiment analysis so that AI will be capable of 'actually' comprehending and responding to the distinct tone of every voice.

REFERENCES

1. N. Dhariwal, S. C. Akunuri, and K. Sharmila Banu, "Audio and Text Sentiment Analysis of Radio Broadcasts," IEEE Access, vol. 11, pp. 145–156, 2023.
2. Z. Guo, T. Jin, W. Xu, W. Lin, Y. Wu, "Bridging the Gap for Test-Time Multimodal Sentiment Analysis," in Proc. AAAI Conf. Artificial Intelligence, 2025, pp. 11234–11243.
3. Y. Mao, Q. Liu, Y. Zhang, "Sentiment Analysis Methods, Applications, and Challenges: A Systematic Review," Journal of King Saud University – Computer and Information Sciences, vol. 36, no. 2, pp. 1019–1039, 2024.
4. B. T. Atmaja, A. Sasou, "Sentiment Analysis and Emotion Recognition from Speech Using Universal Speech Representations," Sensors, vol. 22, no. 14, pp. 5410–5422, 2022.
5. S. Chen, Y. Wu, J. Wu, M. Zhang, X. Wu, J. Li, "UniSpeech-SAT: Universal Speech Representation Learning with Speaker-Aware Pre-Training," in Proc. IEEE ICASSP, 2022, pp. 3452–3456.
6. Y. Jia, X. Chen, J. Yu, L. Wang, Y. Xu, S. Liu, Y. Wang, "Speaker Recognition Based on Characteristic Spectrograms and AC-SOM," Complex Intelligent Systems, vol. 7, no. 4, pp. 1823–1837, 2021.
7. Y. H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L. P. Morency, R. Salakhutdinov, "Multimodal Transformer for Unaligned Multimodal Language Sequences (MulT)," EMNLP 2019.
8. S. Maghilnan, M. R. Kumar, "Sentiment Analysis on Speaker Specific Speech Data," I2C2 2017.