Audio Sentiment Analysis: Techniques, Workflow, Challenges, Applications and Literature Review

[1] Ms. Reetu Awasthi,
[2] Dr. Vinay Chavan,
[1] Research Scholar, [2] Principal
[1]Department of Electronics and Computer science
[1]Rashtrasant Tukadoji Maharaj Nagpur University, Nagpur, India

Abstract

Audio sentiment analysis is a rapidly evolving field that seeks to uncover emotional insights from audio signals, with broad applications in technology, healthcare, and user experience design. This paper presents a detailed review of recent advancements, highlighting key techniques, ongoing challenges, and real-world applications that illustrate the potential of this technology. The Literature Review provides an integrated analysis of significant findings across various studies, identifying core trends and unique contributions. Finally, the Conclusion explores promising future research directions, laying the groundwork for continued innovation in extracting sentiment from sound.

Keywords

Audio Sentiment Analysis, Sentiment Classification, Deep Learning Techniques, Feature Extraction, Transfer Learning
Ensemble Methods, Multi-Task Learning, Challenges in Sentiment Analysis, Applications of Sentiment Analysis,
Literature Review of Audio Sentiment Analysis

Introduction

Audio sentiment analysis is a field that aims to extract the emotional content from audio signals. This is an important task in many applications, including speech recognition; call centre analysis, and customer feedback analysis. Audio sentiment analysis has gained increasing attention in recent years due to the availability of large datasets and the development of advanced machine learning algorithms. In this paper, we provide a comprehensive review of the existing literature on audio sentiment analysis.This literature review on audio sentiment analysis contributes significant value over existing reviews in several key aspects:

1.1 Comprehensive Coverage: It provides a comprehensive overview of the existing literature, covering various techniques, methodologies, challenges, and applications of audio sentiment analysis. This breadth of coverage ensures that readers gain a thorough understanding of the field's current state and emerging trends.

1.2 Structured Workflow: The paper outlines a structured workflow for audio sentiment analysis, detailing the steps involved from data collection to model deployment. This workflow serves as a practical guide for researchers and practitioners, facilitating the implementation of sentiment analysis systems in real-world scenarios.

1.3 In-depth Analysis of Techniques: It delves into the details of different techniques used in audio sentiment analysis, such as feature extraction, machine learning algorithms, deep learning architectures, sentiment embeddings, and multimodal approaches. This in-depth analysis helps readers grasp the nuances of each technique and their respective strengths and weaknesses.

1.4 Discussion of Challenges: The paper highlights the challenges faced in audio sentiment analysis, such as speech variability, background noise, limited training data, lack of context, and cultural differences. By acknowledging these challenges, the review prompts further research efforts to address these issues and enhance the accuracy and robustness of sentiment analysis models.

1.5 Evaluation of Recent Research: It includes a literature review table summarizing recent research papers in the field, along with their methodologies and evaluation metrics. This evaluation provides insights into the effectiveness of different approaches and helps researchers identify promising avenues for future exploration.

1.6 Applications Across Domains: The review discusses diverse applications of audio sentiment analysis, ranging from speech recognition enhancement and call center analysis to political analysis and social media monitoring. By illustrating the broad utility of sentiment analysis, the paper underscores its relevance in various domains and industries.

Overall, this literature review consolidates and synthesizes existing knowledge on audio sentiment analysis while offering fresh insights, practical guidance, and avenues for further research, thus adding substantial value to the current understanding of the field.

2. Techniques

There are several techniques for audio sentiment analysis (figure 2.), including:

2.1. Feature extraction:

In this technique, the process of audio sentiment analysis involves the extraction of acoustic features, including pitch, energy, and spectral characteristics, from the audio signal. These features serve as essential components in training a classifier that predicts the sentiment expressed in the audio content. Features such as Mel-Frequency Cepstral Coefficients (MFCC), Chroma Features, Spectral Contrast, and Mel Spectrogram are extracted to capture important information about the audio signal. Notably, Mel-Frequency Cepstral Coefficients (MFCCs) (figure 1.) play a prominent role in this approach, serving as a widely adopted representation of the spectral features of the audio signal. MFCCs capture the frequency content of the signal by converting the power spectrum into a set of coefficients, effectively summarizing the critical frequency information for sentiment analysis. The utilization of such acoustic features, particularly MFCCs, enhances the model's ability to discern nuanced patterns and variations in the audio data, contributing to the effectiveness of sentiment prediction in diverse audio contexts.

2.1.1Temporal Feature Extraction: Delta and Delta-Delta coefficients are computed to capture temporal changes in the audio features.

2.1.2 Statistical Features: Statistical measures such as mean, standard deviation, skewness, and kurtosis may be computed to provide additional insights into the distribution of the features.

2.2. Machine Learning: Machine learning techniques encompass a diverse set of algorithms and methods designed to enable computers to learn from data and improve their performance over time without explicit programming. In supervised learning, models are trained on labelled datasets, where the algorithm learns the relationship between inputs and corresponding output labels, enabling it to make predictions on new, unseen data. Unsupervised learning, on the other hand, deals with unlabeled data, discovering inherent patterns and structures within the dataset, often used for clustering or dimensionality reduction. Semi-supervised learning combines labelled and unlabeled data to leverage the advantages of both. Reinforcement learning involves an agent learning optimal actions by interacting with an environment and receiving feedback in the form of rewards or penalties. These techniques find applications across various domains, from natural language processing and image recognition to recommendation systems and autonomous vehicles, showcasing their versatility in solving complex problems and making informed predictions.

2.2.1. Supervised Learning:

In supervised learning, the primary objective is to learn a mapping from input to output using labelled data. For instance, a spam email classifier is trained on a dataset where emails are labelled as spam or not spam. This allows the model to predict the label of new, unseen emails. Common algorithms for supervised learning include Support Vector Machines (SVM), Decision Trees, and Neural Networks.

2.2.2. Unsupervised Learning:

The objective of unsupervised learning is to discover patterns or structures within unlabeled data. An example is clustering, where customer data is grouped into segments based on purchasing behaviour without predefined categories. Algorithms such as K-Means clustering, Hierarchical clustering, and Principal Component Analysis (PCA) are frequently used in unsupervised learning tasks.

2.2.3. Semi-Supervised Learning:

Semi-supervised learning combines elements of both supervised and unsupervised learning. This approach proves valuable when obtaining a fully labelled dataset is challenging or expensive. An example could be image recognition with limited labelled data, where the model benefits from utilizing both labelled and unlabeled images. Algorithms like Self-Training and Multi-View Learning are commonly applied in semi-supervised scenarios.

2.2.4. Reinforcement Learning:

Reinforcement learning focuses on training an agent to make sequential decisions in an environment. Applications include game playing, where an agent learns optimal strategies to maximize cumulative rewards, as seen in AlphaGo. Algorithms like Q-Learning, Deep Q Network (DQN), and Policy Gradient Methods are employed for training agents in reinforcement learning settings.

2.3. Deep learning

Deep learning is a subset of machine learning that employs artificial neural networks, specifically deep neural networks, to model and solve complex problems. The term "deep" refers to the multiple layers through which the data is transformed. Deep learning has gained prominence due to its ability to automatically learn hierarchical representations from data, leading to remarkable performance in tasks such as image and speech recognition, natural language processing, and more. In deep learning, neural networks with numerous layers (deep neural networks) are utilized to learn intricate patterns and features. Common architectures include Convolutional Neural Networks (CNNs) for image-related tasks, Recurrent Neural Networks (RNNs) for sequence data, and Transformers for natural language processing. Deep learning excels in scenarios where the data is abundant and complex, allowing the model to automatically discover and represent intricate relationships within the data, resulting in state-of-the-art performance in various applications.

2.3.1. Convolutional Neural Networks (CNNs)

Application

CNNs are primarily used for image-related tasks, such as image classification, object detection, and image segmentation.

Architecture

CNNs consist of convolutional layers, pooling layers, and fully connected layers. Convolutional layers learn local patterns and features, while pooling layers down sample the spatial dimensions. These layers allow the network to recognize hierarchical visual features.

2.3.2. Recurrent Neural Networks (RNNs)

Application

RNNs are designed for sequential data, making them suitable for tasks like natural language processing, speech recognition, and time series analysis.

Architecture

RNNs have recurrent connections that allow information to persist across different time steps. This enables the network to capture dependencies and patterns in sequential data. However, traditional RNNs may struggle with long-term dependencies.

2.3.3. Transformation Models

Application

Transformers are pivotal in natural language processing tasks, including machine translation, text summarization, and language modelling.

Architecture

Transformers use self-attention mechanisms to process input data in parallel, making them highly effective for handling sequential data. They have become the backbone of state-of-the-art language models like BERT and GPT**.**

2.4. Emotion Lexicons and Databases: Emotion lexicons are dictionaries associating words with emotional categories. Databases provide labeled emotional content. Integrating these resources helps map acoustic features to emotional states.

2.5. Speech Prosody Analysis: Speech prosody focuses on the rhythm, pitch, and intonation of speech. Analyzing these features provides insights into emotional states, with variations in speech rate, pitch, and pauses indicating different sentiments.

2.6. Sentiment Embeddings: Sentiment embeddings involve representing words or audio segments in a continuous vector space. Word embeddings capture semantic relationships, and sentiment embeddings aim to represent sentiment-related nuances in the data.

2.7. Transfer learning: Transfer learning involves pre training a deep learning model on a large dataset, and then fine-tuning it on a smaller dataset for the specific task of sentiment analysis. This technique has been shown to be effective in audio sentiment analysis.

2.8. Ensemble methods: Ensemble methods involve combining the predictions of multiple models to improve performance. In audio sentiment analysis, ensemble methods can be used to combine the predictions of different models trained on different datasets or with different architectures.

2.9. Multimodal Approaches: Combining audio with other modalities like text or video can enhance sentiment analysis. Fusion techniques integrate information from different modalities, offering a more comprehensive understanding of sentiment.

2.10. Real-time Sentiment Analysis: Real-time sentiment analysis involves processing audio input on-the-fly, providing immediate insights into changing sentiments. Adaptive models can adjust to evolving sentiment patterns over time.

2.11. Multi-task learning: Multi-task learning involves training a single model to perform multiple related tasks simultaneously. In audio sentiment analysis, multi-task learning can be used to predict multiple sentiment labels (e.g., positive, negative, neutral) or to predict sentiment along with other related tasks such as speaker identification or emotion recognition.

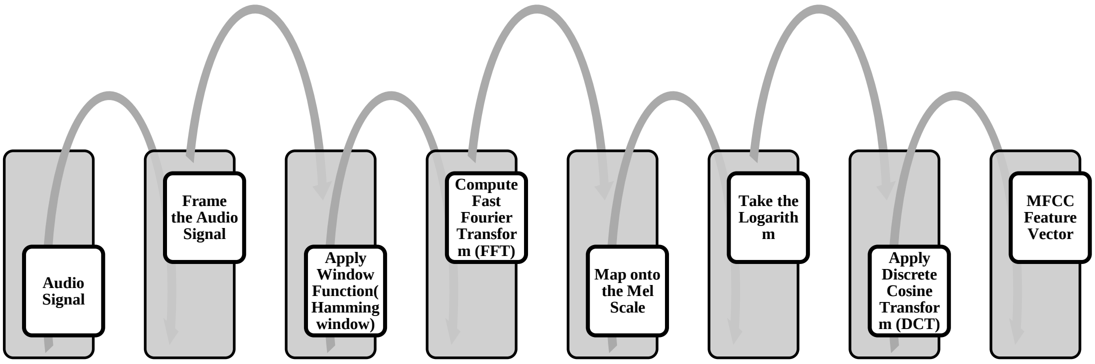
Audio Signal
Frame the Audio Signal
Apply Window Function(Hamming window)
Compute Fast Fourier Transfor m (FFT)
Map onto the Mel Scale
Take the Logarith m
Apply Discrete Cosine Transfor m (DCT)
MFCC Feature Vector

Figure1. MFCC extraction process

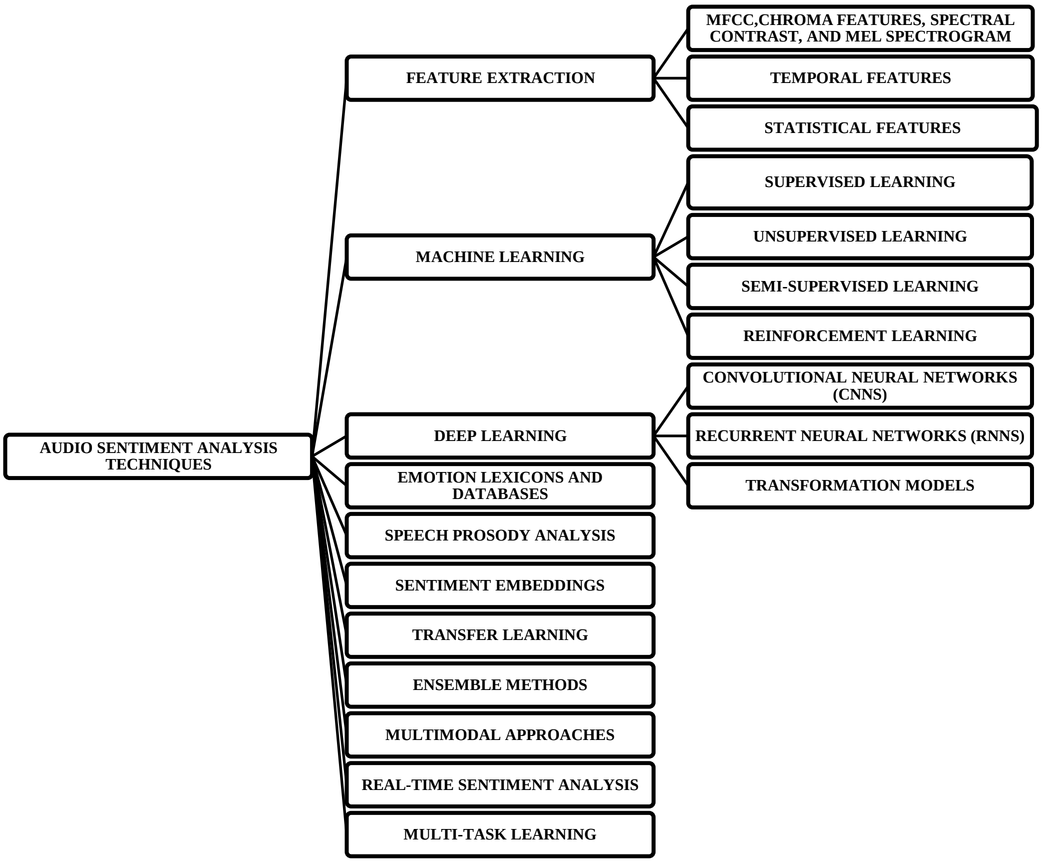

Figure 2: Techniques For Audio Sentiment Analysis

The steps involved in audio sentiment analysis can vary depending on the specific approach and techniques used, but some common steps include:

3. Workflow

3.1. Data Collection

Audio data is collected from various sources, including social media, call centre recordings, or voice assistants. This step is crucial for building a diverse and representative dataset that reflects the range of sentiments present in real-world scenarios.

Ethical considerations, privacy compliance, and obtaining proper permissions are important aspects of data collection. Ensuring a balanced representation of speakers, accents, and emotional expressions enhances the robustness of the model.

Numerous pre-designed audio databases are accessible for a range of applications, including sentiment analysis and emotion recognition. Notable examples encompass RAVDESS, which specializes in emotional speech and song analysis; IEMOCAP, designed for emotional speech and multimodal emotion recognition; SAVEE, focused on emotion recognition in speech; CREMA-D, offering audio and video clips of actors expressing various emotions; MSP-IMPROV, providing speech emotion recognition and music analysis data from improvised sessions; EmoReact, a multimodal dataset featuring audio, video, and physiological signals for emotional speech and reaction analysis; and TESS, consisting of audio clips with actors portraying different emotional expressions. These datasets contribute valuable resources for training and evaluating models, though careful consideration of licensing and dataset relevance to specific research or application needs is paramount.

3.2. Pre-processing

The collected audio data undergoes pre-processing(figure 3.) to eliminate noise or irrelevant information that may interfere with sentiment analysis. Techniques such as filtering, re-sampling, and normalization are applied to enhance the quality of the data.

Addressing noise and irrelevant information is critical for improving the accuracy of sentiment analysis models. Careful handling of noise and artefacts ensures that the extracted features represent meaningful patterns related to sentiment.

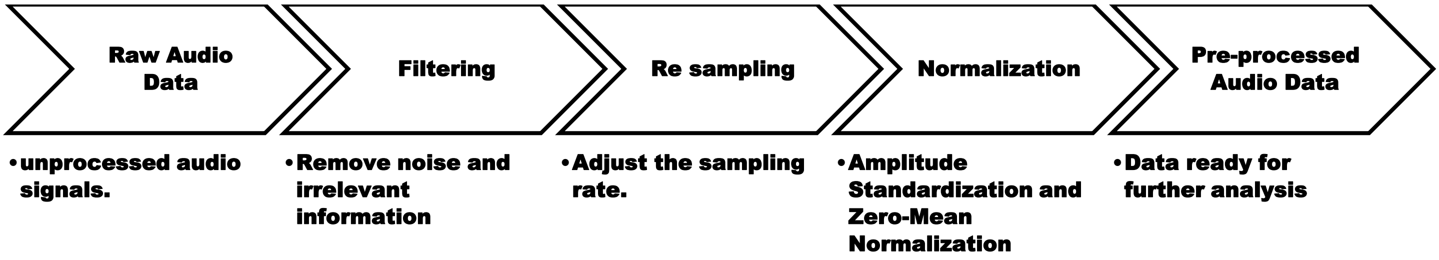
Raw Audio Data
Filtering
Re sampling
Normalization
Pre-processed Audio Data
•unprocessed audio signals.
•Remove noise and irrelevant information
•Adjust the sampling rate.
•Amplitude Standardization and Zero-Mean Normalization
•Data ready for further analysis

Figure 3: Data Pre-Processing Model

3.3. Feature Extraction

Description: Relevant features are extracted from the pre-processed audio data. Techniques may include spectral analysis, Mel-frequency cepstral coefficients (MFCCs), or deep learning-based feature extraction. These features aim to capture the acoustic characteristics that are indicative of sentiment.

The choice of feature extraction methods depends on the nature of the audio data and the specific characteristics relevant to sentiment. MFCCs, for example, are popular for their effectiveness in representing spectral information.

3.4. Sentiment Classification

A sentiment classification model is trained on the extracted features. Various machine learning or deep learning techniques can be employed, such as support vector machines (SVMs), random forests, convolutional neural networks (CNNs), or recurrent neural networks (RNNs). The model learns to associate extracted features with different sentiment classes.

Choosing the appropriate classification algorithm depends on the complexity of the task, the size of the dataset, and the specific characteristics of the audio data. Training a robust model requires careful consideration of hyper parameters and model architecture.

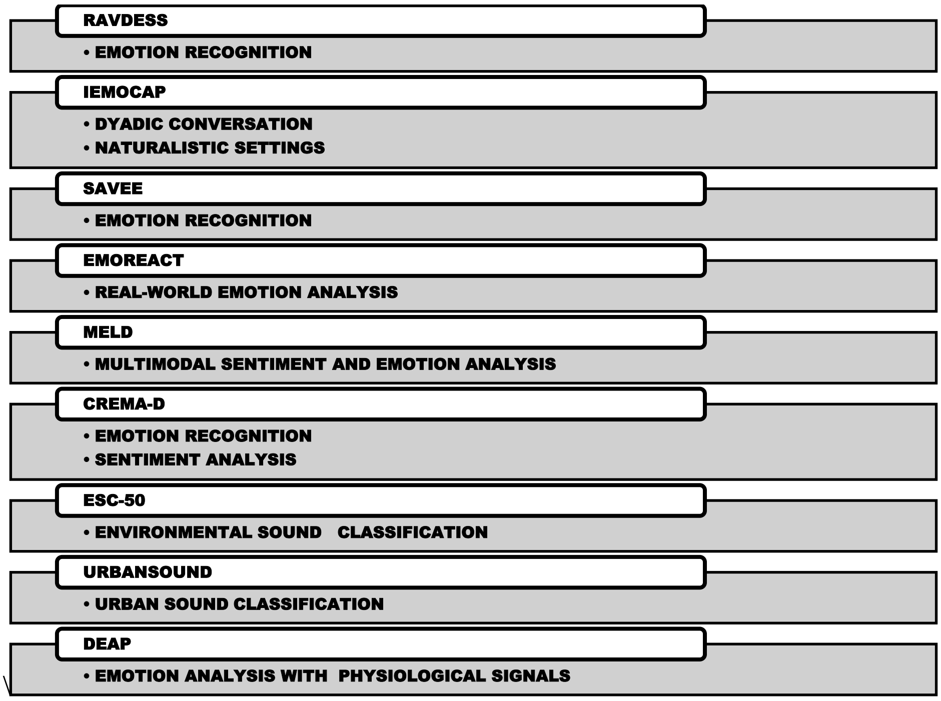

Figure4. Audio Databases And Uses

3.5. Model Evaluation

The performance of the sentiment classification model is evaluated on a separate test dataset to assess its accuracy, precision, recall, and F1-score. This step provides insights into the model's generalization to new, unseen data.

Rigorous evaluation helps identify potential over fitting or under fitting issues and ensures that the model performs well across different instances of sentiment expressions. Fine-tuning may be necessary based on the evaluation results.

3.6. Deployment

The trained model is deployed into a production environment where it can perform tasks on new, unseen data. The deployment phase involves several key steps. First, the trained model needs to be packaged in a format compatible with the deployment environment, ensuring compatibility with the production system's infrastructure. Next, the model is deployed to the target platform, whether it's on-premises servers, cloud services, or edge devices. Integration with the larger system or application, such as a customer service chatbot or a social media monitoring platform, is crucial during deployment. Real-time processing requirements, scalability considerations, and ongoing monitoring for model performance are vital aspects to address. Regular updates and maintenance are often necessary to adapt the deployed model to changes in the data distribution or shifts in the patterns of sentiment expression. Successful deployment ensures that the trained model effectively contributes to the intended application, providing valuable insights into the sentiment of new audio data in real-world scenarios. Deployment considerations include real-time processing requirements, scalability, and the integration of the sentiment analysis model into the overall system architecture. Ongoing monitoring and updates may be necessary to adapt to changes in the data distribution or sentiment expression patterns**.**

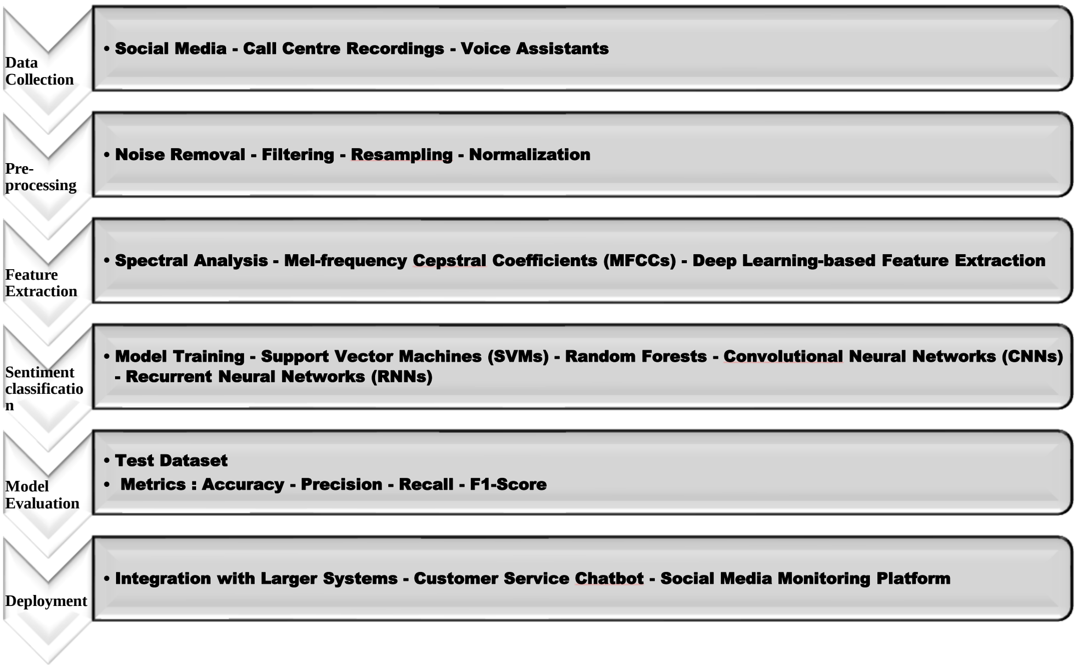
Data Collection
• Social Media - Call Centre Recordings - Voice Assistants
Pre-processing
• Noise Removal - Filtering - Resampling - Normalization
Feature Extraction
• Spectral Analysis - Mel-frequency Cepstral Coefficients (MFCCs) - Deep Learning-based Feature Extraction
Sentiment classification
• Model Training - Support Vector Machines (SVMs) - Random Forests - Convolutional Neural Networks (CNNs) - Recurrent Neural Networks (RNNs)
Model Evaluation
• Test Dataset
• Metrics : Accuracy - Precision - Recall - F1-Score
Deployment
• Integration with Larger Systems - Customer Service Chatbot - Social Media Monitoring Platform

Figure 4: Audio Sentiment Analysis Workflow

4. Challenges

Audio sentiment analysis faces several challenges that can affect the accuracy of its results. Some of these challenges include:

4.1. Variability in speech patterns

Human speech is highly variable, with differences in tone, pitch, accent, and intonation. This variability can make it difficult to accurately identify emotions and sentiments from audio recordings.

4.2. Background noise

Audio recordings frequently include background noise, presenting a substantial challenge for sentiment analysis algorithms and complicating the accurate classification of emotions. Background noise encompasses various unwanted sounds such as ambient conversations, machinery hum, or environmental sounds that coexist with the primary speech signal. This interference can adversely impact the performance of sentiment analysis algorithms, as the presence of noise introduces additional acoustic complexities. The challenge lies in distinguishing between the intended emotional content in speech and the extraneous noise, requiring sophisticated noise reduction and filtering techniques.

4.3. Limited training data: Audio sentiment analysis requires large amounts of labelled training data to accurately identify emotions and sentiments. However, obtaining labelled audio data is often more difficult and expensive than obtaining labelled text data, which can limit the development and performance of sentiment analysis algorithms.

4.4. Lack of context: Sentiment analysis algorithms may struggle to accurately classify emotions and sentiments without the context of the conversation or the speaker's background and personality.

4.5. Cultural and linguistic differences: Emotions and sentiments are expressed differently across cultures and languages, which can make it challenging to develop a sentiment analysis algorithm that works well across diverse populations.

4.6. Bias in training data: Sentiment analysis algorithms can perpetuate bias if the training data is not diverse enough or if the data is biased towards certain demographics or perspectives.

Addressing these challenges is important for improving the accuracy and effectiveness of audio sentiment analysis algorithms.

5. Applications

Audio sentiment analysis has many applications

5.1. Speech Recognition Enhancement

Audio sentiment analysis enhances speech recognition systems by not only transcribing words but also capturing the emotional context, improving the overall accuracy and understanding of spoken communication.

5.2. Call Centre Analysis

In call centres', audio sentiment analysis is employed to assess customer emotions during interactions. This allows businesses to evaluate customer satisfaction, identify issues, and enhance the quality of customer service.

5.3. Customer Feedback Analysis

For customer feedback, audio sentiment analysis automates the evaluation of sentiments expressed in recorded reviews or feedback calls. This provides businesses with a nuanced understanding of customer opinions.

5.4. Media Monitoring

Audio sentiment analysis contributes to media monitoring by assessing public reactions to broadcasts. This capability helps media outlets gauge audience sentiments and tailor content accordingly.

5.5. Political Analysis

Understanding sentiment in political speeches or public discussions is crucial for political analysis. Audio sentiment analysis provides insights into public opinion and emotional responses to political discourse.

5.6. Social Media Analysis

In social media analysis, audio sentiment analysis complements text-based sentiment analysis by capturing sentiments expressed in audio-visual content. This comprehensive approach provides a deeper understanding of social media conversations.

6. Literature Review

Speech emotion is a critical paralinguistic aspect in communication, marked by significant subjectivity. Differences in feelings are observed in various languages and cultures, highlighting the significance of examining emotions, as previous studies have indicated. [1] A database called the "emotional speech ground truth database" has been created, which includes statements that have strong emotional or semantic content and are grouped into five different sentiment categories. Kim and Hansen [2] suggest a method for detecting angry speech effectively, which takes into account the structure of the spoken content. Their model incorporates a classifier using an "emotional language" framework, which merges a score from this framework with acoustic traits like the Teager Energy Operator (TEO) and Mel Frequency Cepstral Coefficients (MFCC). Combining TEO and MFCC features led to a 6.23% enhancement in the Equal Error Rate (EER). In [3], a model called the Bidirectional Emotional Recurrent Unit (BiERU) is introduced for analyzing sentiment in conversations, comparing it to single-sentence sentiment analysis. This model prioritizes text over speech or audio, using a neural tensor model with dual-channel classifiers to examine conversation context and conduct sentiment analysis. BiERU surpasses other models including Dialogue-RNN, Dialogue-CNN, and AGHMN in sentiment classification accuracy. The model obtained an average accuracy rate of 66% and a 64% F1 score when analyzing conversational text data. Garg and Sharma [4] study sentiment analysis using various Twitter data representations, while [5] investigates word embedding techniques for sentiment detection. Salehin and colleagues [6] utilized Support Vector Machines (SVM) and OpenCV for examining student emotions in a virtual learning setting. In addition, Moung and colleagues [7] employed a variety of techniques to identify emotions in pictures. However, novel approaches are still needed for sentiment analysis from audio data. Mahima et al. [8] review techniques for identifying multiple emotions from textual data.

Bertero and Fung [9] suggest an instantaneous method using CNN for identifying emotions in speech. Their study concentrates on developing a model by utilizing audio recordings from "TED Talks," which are categorized into three emotions: "angry," "happy," and "sad" through manual labeling. The model achieves a 66.1% average accuracy, which is 5% better than a baseline SVM model that uses feature-based methods. Kantipud and Kumar [10] propose an effective learning model for categorizing audio signal characteristics, utilizing filters that activate at different frequencies based on amplitude-related feature recognition. Chen and Luo [11] present a sentiment detection technique based on audio using a deep neural network driven by utterances. This method utilizes both CNN and an ASV to achieve its goals. Audio inputs are analyzed by creating a spectral graph and then passing it through an LSTM model to process the utterances. Conventional acoustic characteristics like spectral centroid and Mel-frequency cepstral coefficients (MFCC) are utilized. The accuracy of the model is reported to be 57.74%. Maghilnan and Kumar [12] introduce a model for analyzing sentiments in audio by using speech data distinguished by the speaker. The model has two starting points: one for differentiating speakers and another for recognizing speech. The focus of section 13 is on sentiment analysis within customer relationship management (CRM), highlighting the significance of cultivating lasting customer loyalty post the first transaction. The MALSA model in [14] suggests a multi-aspect sentiment analysis for CRM that detects sentiment and offers recommendations to assist customers in their buying process. The model is constructed based on four methods: detecting frequency, detecting syntax, utilizing supervised and unsupervised learning, and implementing hybrid models for sentiment analysis, all with a sole focus on textual data.

Rotovei and Negru [15] further develop the research conducted in [13] by incorporating a module for B2B engagements, and merging a recommendation system into the current framework. Rotovei and Negru's approach to aspect term extraction involves four techniques: recognizing common nouns and noun phrases, studying opinion-target connections, utilizing supervised learning, and employing topic modeling. Moreover, their framework also includes aspect aggregation. Aspect-based sentiment analysis produces a lineup of aspects in sentences written in natural language [16]. The suggested study presents an interactive multi-task learning (IMU) model that conducts aspect analysis for tokens and documents concurrently. A new method is created for IMU training, and it is evaluated against several different models. Reportedly, the new model has been found to surpass current models with an accuracy of 83.89% and an F1 score of 59.18.

Identifying sentiment polarity of specific targets within a context is crucial in aspect-level sentiment classification [17]. The suggested study presents a method that takes into account both objectives and their immediate environment, focusing on giving individual attention to each. This is accomplished by utilizing the interactive attention network (IAN) to understand the relationship between context and targets. The IAN model outperformed LSTM, TD-LSTM, and AE-LSTM models with a 4% to 5% increase in accuracy, achieving a final accuracy of 72.1%. The dataset used for training and evaluation in [18] was the SemEval 2014 dataset. Aspect-level sentiment analysis is a key task in natural language processing that focuses on determining the sentiment of particular aspect terms within a sentence [19]. This study expands on [17], overcoming drawbacks with enhanced and supplementary characteristics. An improvement made in sentiment analysis is the addition of grammatical rules, a feature that was missing in earlier research. The model proposed in the SemEval dataset outperforms the method in [17] by 2%, achieving an accuracy of 74.89%. It was also evaluated in comparison to other models like LSTM, AE-LSTM, and attention-based LSTM with aspect embedding (ATAE-LSTM).

Sentiment analysis goes beyond just recognizing sentiment in a sentence and can also be utilized to categorize intricate transactions as either "won" or "lost" [14]. The classification model utilizes various techniques, such as frequency, lexicon-based, and syntax-based detection, machine learning techniques, and hybrid methods. Extensive research has been conducted on the utilization of deep learning methods for sentiment prediction, regardless of whether the input is textual, speech, or audio. This research delves into different techniques like deep neural networks, LSTM, autoencoders, word embeddings, CNN, RNN, and attention mechanisms, with an emphasis on document-level sentiment analysis and alternative strategies. Zhang and colleagues [20] offer an extensive examination of deep learning methods for sentiment analysis, covering various features, traits, and methods of implementation. Analyzing sentiments is essential for measuring customer satisfaction and viewpoints. Capuano and colleagues suggest employing hierarchical attention networks to examine the sentiment hierarchy within customer feedback. Their approach utilizes a combined incremental learning system to enhance forecasts using reviews from CRM systems. In their study, a prototype of the model was created with a substantial dataset of labeled items, resulting in an average F1-score of 0.85.

"Interactive sentiment analysis," which was mentioned in [22], is a developing subcategory in the wider field of natural language processing (NLP). Progress in this area is currently limited by the lack of labeled datasets tailored for interactive sentiment analysis. In [22], the researchers created a new conversational database called Scenario SA, which was manually annotated to improve the efficiency of interactive sentiment detection. Due to the rise in smartphone usage and social media platforms, communication via chats, comments, and product reviews has become common. Lexicons commonly drive sentiment analysis in product reviews and are crucial. Creating these dictionaries is essential for precise identification of emotions. In [23], a technique was suggested for creating a sentiment lexicon specific to a domain automatically. This method focused on identifying words that express feelings from product reviews through a relational data algorithm. Social media sites such as Facebook and Quora offer high-resolution (HR) data that can provide valuable customer relationship management (CRM) insights through analyzing conversations about business events [24]. Conventional sentiment analysis in text has mainly depended on methods based on lexicons. Research in [25] examined customer behavior in product buying by analyzing feedback sentiments. In [26], a novel approach called "MMSTR" was presented for extracting and analyzing short text reviews and feedback to assess emotions using customer reviews. Swift sentiment analysis is essential for promptly addressing customer feedback. In section 27, a method called the "hierarchical approach" was introduced for sentiment analysis. This method involved analyzing word embeddings from reviews using Word2Vec and utilizing an extreme gradient boosting classifier (xgboost) to differentiate between positive and negative sentiments. Using the Doc2Vec approach, this model accurately categorized 12 different classes with an overall accuracy of 71.16%. Due to the growing popularity of online shopping, the significance of customer product reviews has seen a rise.Methods such as Online Analytical Processing (OLAP) and Data Cubes have been employed to examine sentences in these reviews. In the field of NLP, sentiment analysis and opinion mining continue to be important tasks, with the objective of identifying if a person's sentiment is positive, negative, or neutral. Different methods were analyzed, including logistic regression, the hybrid bag-boost algorithm, and support vector machines (SVM), to assess their performance. Sentiment analysis on social media platforms such as Twitter has also been studied in [30] and [31], in which tweets' sentiments are classified as positive, negative, or neutral. Concerning the processing of audio data, a pre-processing system, like the one detailed in [32], was put into action. This process included loading, padding, extracting features, and normalizing audio data, with the stored features such as MFCC, chroma, Mel spectrogram, contrast, and tonnetz in a pickle file format. Next, the deep neural network was utilized to train the model on the RAVDASS dataset.

7. Conclusion

Speech emotion and sentiment analysis has advanced through the integration of traditional acoustic features like MFCC and deep learning techniques such as CNN, RNN, and LSTM, significantly improving emotion detection in both speech and conversational contexts. Models like BiERU and innovations in processing data from TED Talks and CRM systems have surpassed previous benchmarks, demonstrating notable accuracy enhancements. Despite progress, challenges persist in effectively analyzing emotions from audio data, especially in multi-lingual and cross-cultural contexts. Future research must focus on developing robust multimodal models and comprehensive datasets to further enhance real-time emotion detection and its practical applications across various domains.

8. Research Gaps and Challenges

The field of audio sentiment analysis faces several challenges and research gaps that present opportunities for further innovation. Key areas include improving robustness to speech variability, as human speech exhibits diverse tones and accents that complicate accurate sentiment detection. Addressing noise resilience is essential since background noise can significantly impact model performance; hence, techniques for noise reduction tailored to sentiment analysis are needed. Data scarcity poses a challenge in acquiring large, labeled datasets for training, necessitating strategies like data augmentation and transfer learning. Additionally, enhancing contextual understanding is vital to accurately interpret sentiments within varied speaker characteristics and conversational contexts. Cross-cultural and multilingual sentiment analysis remains important as emotional expressions differ across cultures, demanding models that generalize well. Addressing bias and fairness concerns in sentiment analysis models is paramount to prevent perpetuating existing biases. Real-time analysis capabilities are crucial for capturing evolving sentiments in dynamic environments, supported by adaptive algorithms. Lastly, the development of interpretable models and standardized benchmarking for evaluation metrics will foster better comparisons across studies, facilitating advancements in the field. By addressing these gaps, the audio sentiment analysis domain can evolve into more accurate, robust, and ethically sound systems applicable across various sectors.

References:

[1] N. Vryzas, L. Vrysis, R. Kotsakis, and C. Dimoulas, “Speech emotion recognition adapted to multimodal semantic repositories,” in *2018 13th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, Sep. 2018, pp. 31–35, doi: 10.1109/SMAP.2018.8501881.

[2] W. Kim and J. H. L. Hansen, “Angry emotion detection from real-life conversational speech by leveraging content structure,” in *2010 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2010, pp. 5166–5169, doi: 10.1109/ICASSP.2010.5495021.

[3] W. Li, W. Shao, S. Ji, and E. Cambria, “BiERU: bidirectional emotional recurrent unit for conversational sentiment analysis,” *Neurocomputing*, vol. 467, pp. 73–82, Jan. 2022, doi: 10.1016/j.neucom.2021.09.057.

[4] N. Garg and D. K. Sharma, “Sentiment analysis of events on social web,” *International Journal of Innovative Technology and Exploring Engineering*, vol. 9, no. 6, pp. 1232–1238, Apr. 2020, doi: 10.35940/ijitee.F3946.049620.

[5] A. Samih, A. Ghadi, and A. Fennan, “Deep graph embeddings in recommender systems: a survey,” *Journal of Theoretical and Applied Information Technology*, vol. 99, no. 15, pp. 3812–3823, 2021.

[6] I. Salehin *et al.*, “Analysis of student sentiment during video class with multi-layer deep learning approach,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 4, pp. 3981–3993, Aug. 2022, doi: 10.11591/ijece.v12i4.pp3981- 3993.

[7] E. G. Moung, C. C. Wooi, M. M. Sufian, C. K. On, and J. A. Dargham, “Ensemble-based face expression recognition approach for image sentiment analysis,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 3, pp. 2588–2600, Jun. 2022, doi: 10.11591/ijece.v12i3.pp2588-2600.

[8] M. A. Mahima, N. C. Patel, S. Ravichandran, N. Aishwarya, and S. Maradithaya, “A text-based hybrid approach for multiple emotion detection using contextual and semantic analysis,” in *2021 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSES)*, Sep. 2021, pp. 1–6, doi: 10.1109/ICSES52305.2021.9633843.

[9] D. Bertero and P. Fung, “A first look into a convolutional neural network for speech emotion detection,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Mar. 2017, pp. 5115–5119, doi: 10.1109/ICASSP.2017.7953131.

[10] M. V. V. P. Kantipud and S. Kumar, “A computationally efficient learning model to classify audio signal attributes,” *International Journal of Electrical and Computer Engineering (IJECE)*, vol. 12, no. 5, pp. 4926–4934, Oct. 2022, doi: 10.11591/ijece.v12i5.pp4926-4934.

[11] F. Chen and Z. Luo, “Learning robust heterogeneous signal features from parallel neural

network for audio sentiment analysis," *arXiv preprint arXiv:1811.08065*, Nov. 2018.

[12] S. Maghilnan and M. R. Kumar, "Sentiment analysis on speaker specific speech data," in *2017 International Conference on Intelligent Computing and Control (I2C2)*, Jun. 2017, pp. 1–5, doi: 10.1109/I2C2.2017.8321795.

[13] D. Rotovei, "Multi-agent aspect level sentiment analysis in CRM systems," in *2016 18th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, Sep. 2016, pp. 400–407, doi: 10.1109/SYNASC.2016.068.

[14] D. Rotovei and V. Negru, "Improving lost/won classification in CRM systems using sentiment analysis," in *2017 19th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, Sep. 2017, pp. 180–187, doi: 10.1109/SYNASC.2017.00038.

[15] D. Rotovei and V. Negru, "Multi-agent recommendation and aspect level sentiment analysis in B2B CRM systems," in *2020 22nd International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC)*, Sep. 2020, pp. 238–245, doi: 10.1109/SYNASC51798.2020.00046.

[16] R. He, W. S. Lee, H. T. Ng, and D. Dahlmeier, "An interactive multi-task learning network for end-to-end aspect-based sentiment analysis," in *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019, pp. 504–515, doi: 10.18653/v1/P19-1048.

[17] D. Ma, S. Li, X. Zhang, and H. Wang, "Interactive attention networks for aspect-level sentiment classification," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, Aug. 2017, pp. 4068–4074, doi: 10.24963/ijcai.2017/568.

[18] N. Vryzas, L. Vrysis, R. Kotsakis, and C. Dimoulas, "Speech emotion recognition adapted to multimodal semantic repositoriess," in *2018 13th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, Sep. 2018, pp. 31– 35, doi: 10.1109/SMAP.2018.8501881.

[19] Q. Lu, Z. Zhu, G. Zhang, S. Kang, and P. Liu, "Aspect-gated graph convolutional networks for aspect-based sentiment analysis," *Applied Intelligence*, vol. 51, no. 7, pp. 4408–4419, Jul. 2021, doi: 10.1007/s10489-020-02095-3.

[20] L. Zhang, S. Wang, and B. Liu, "Deep learning for sentiment analysis: A survey," *WIREs Data Mining and Knowledge Discovery*, vol. 8, no. 4, Jul. 2018, doi: 10.1002/widm.1253.

[21] N. Capuano, L. Greco, P. Ritrovato, and M. Vento, "Sentiment analysis for customer relationship management: an incremental learning approach," *Applied Intelligence*, vol. 51, no. 6, pp. 3339–3352, Jun. 2021, doi: 10.1007/s10489-020-01984-x.

[22] Y. Zhang, Z. Zhao, P. Wang, X. Li, L. Rong, and D. Song, "ScenarioSA: a dyadic conversational database for interactive sentiment analysis," *IEEE Access*, vol. 8, pp. 90652–90664, 2020, doi: 10.1109/ACCESS.2020.2994147.

[23] J. Feng, C. Gong, X. Li, and R. Y. K. Lau, "Automatic approach of sentiment lexicon generation for mobile shopping reviews," *Wireless Communications and Mobile Computing*, pp. 1–13, Aug. 2018, doi: 10.1155/2018/9839432.

[24] S. E. Griesser and N. Gupta, "Triangulated sentiment analysis of Tweets for social CRM," in *2019 6th Swiss Conference on Data Science (SDS)*, Jun. 2019, pp. 75–79, doi: 10.1109/SDS.2019.000-4.

[25] G. Chaubey, P. R. Gavhane, D. Bisen, and S. K. Arjaria, "Customer purchasing behavior prediction using machine learning classification techniques," *Journal of Ambient Intelligence and Humanized Computing*, Apr. 2022, doi: 10.1007/s12652-022-03837- 6.

[26] A. Suriya, "Psychology factor based sentiment analysis for online product customer review using multi-mixed short text ridge analysis," *Journal of Critical Reviews*, vol. 6, no. 6, pp. 146–150, 2019, doi: 10.22159/jcr.06.06.20.

[27] M. Seyfioğlu and M. Demirezen, "A hierarchical approach for sentiment analysis and categorization of Turkish written customer relationship management data," in *Proceedings of the 2017 Federated Conference on Computer Science and Information Systems*, Sep. 2017, pp. 361–365, doi: 10.15439/2017F204.

[28] M. R. Yaakub, Y. Li, and J. Zhang, "Integration of sentiment analysis into customer relational model: the importance of feature ontology and synonym," *Procedia Technology*, vol. 11, pp. 495–501, 2013, doi: 10.1016/j.protcy.2013.12.220.

[29] S. K. Pathuri, N. Anbazhagan, and G. B. Prakash, "Feature based sentimental analysis for prediction of mobile reviews using hybrid bag-boost algorithm," in *2020 7th International Conference on Smart Structures and Systems (ICSSS)*, Jul. 2020, pp. 1–5, doi: 10.1109/ICSSS49621.2020.9201990.

[30] N. N. Alabid and Z. D. Katheeth, "Sentiment analysis of Twitter posts related to the COVID-19 vaccines," *Indonesian Journal of Electrical Engineering and Computer Science (IJEECS)*, vol. 24, no. 3, pp. 1727–1734, Dec. 2021, doi: 10.11591/ijeecs.v24.i3.pp1727-1734.

[31] S. Yousefinaghani, R. Dara, S. Mubareka, A. Papadopoulos, and S. Sharif, "An analysis of COVID-19 vaccine sentiments and opinions on Twitter," *International Journal of Infectious Diseases*, vol. 108, pp. 256–262, Jul. 2021, doi: 10.1016/j.ijid.2021.05.059.

[32] A. Katti and M. Sumana, "Pipeline for pre-processing of audio data," in *Smart Innovation, Systems and Technologies*, vol. 312, Springer Nature Singapore, 2023, pp. 191–198.