IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 8, May 2025
Impact Factor: 7.67

Sentiment Analysis using Customer Reviews

Ms. G. Sahaana[1], Mohamed Afrith Khan T[2], Manoj Kumar C[3], Varshanraj D[4]
Assistant professor, Artificial Intelligence and Data Science[1]
Students, Artificial Intelligence and Data Science[2,3,4]
Dhanalakshmi College of Engineering, Chennai, Tamil Nadu, India
sahaana.g@dce.edu.in, mohamedafrithkhant.ai2021@dce.edu.in
manojkumarc.ai2021@dce.edu.in, varshanrajd.ai2021@dce.edu.in

Abstract: *Traditional sentiment analysis relies heavily on textual data, limiting its effectiveness in environments where voice is the primary communication mode or emotional intent is not fully captured through text alone. This project proposes a hybrid sentiment analysis system that integrates both text and audio inputs to overcome these limitations. The framework leverages speech recognition to transcribe spoken language and applies transformer-based NLP models (e.g., BERT, RoBERTa) for text analysis. In parallel, it extracts prosodic and acoustic features (such as pitch, tone, and tempo) from audio signals to enhance emotional interpretation. By combining textual and vocal cues, the system achieves a deeper understanding of sentiment, enabling more accurate and context-aware emotional analysis. This multimodal approach is particularly valuable in areas such as law enforcement (e.g., threat assessment), healthcare (e.g., mental health monitoring), and customer service (e.g., empathetic AI interactions), where detecting emotional nuance is critical. This project aims to contribute to the development of emotionally intelligent AI systems capable of understanding not just what is said, but how it is said.*

Keywords: Sentiment analysis, multimodal analysis, text and audio integration, speech recognition, transformer-based NLP, BERT, RoBERTa, prosodic features, acoustic features, emotion detection, context-aware analysis, emotional intelligence, hybrid sentiment system, voice-based communication, natural language processing, audio signal processing, mental health monitoring, threat assessment, empathetic AI, human-computer interaction

I. INTRODUCTION

Sentiment analysis, a key area in NLP, traditionally focuses on classifying textual data by emotional tone—positive, negative, neutral, or specific emotions like anger or joy. However, relying solely on text limits effectiveness, especially in speech-driven, emotionally nuanced real-world scenarios. This project introduces a hybrid sentiment analysis model that integrates both text and audio inputs. The system uses automatic speech recognition (ASR) tools (e.g., Google Speech-to-Text, DeepSpeech, Whisper) to transcribe speech into text, which is then analyzed using transformer-based models like BERT or RoBERTa for sentiment classification. Simultaneously, the system extracts audio features—such as pitch, intonation, rhythm, and spectral data (e.g., MFCCs)—using deep learning architectures like CNNs, RNNs, and audio transformers (e.g., Wav2Vec 2.0). These features capture vocal emotional cues often missed in text. By fusing linguistic and acoustic data, the model delivers a more accurate and emotionally aware sentiment analysis, with applications in healthcare, customer service, and public safety.

II. DATASET PREPARATION

Multimodal sentiment analysis combines both text and audio inputs to better understand the emotional tone of speech. While traditional methods rely on text alone, integrating audio allows the system to capture vocal features like pitch, tone, and rhythm—adding depth to sentiment detection. Preparing such a dataset involves collecting synchronized text and audio data, extracting meaningful features from both, and aligning them for model training. This approach leads to more accurate and emotionally aware sentiment analysis.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 8, May 2025
Impact Factor: 7.67

Dataset-Collection

Use a dataset that includes audio recordings, corresponding transcriptions, and sentiment labels. Examples of such datasets include CMU-MOSEI, CMU-MOSI, IEMOCAP, and MELD

Preprocessing

Text Preprocessing

- Clean the transcriptions by removing unnecessary characters and standardizing text.
- Tokenize the text and prepare it for input into language models.

Audio Preprocessing

- Convert audio files to a consistent format, such as WAV with 16kHz sampling rate.
- Normalize audio levels and remove background noise or silence.
- Segment long recordings into smaller utterances if needed.

Feature Extraction

Text Features

Use pretrained language models like BERT or RoBERTa to extract contextual text embeddings.

Audio Features

- Extract acoustic features such as pitch, loudness, rhythm, and Mel-frequency cepstral coefficients using tools like Librosa or openSMILE.
- Alternatively, use pretrained audio models like Wav2Vec2 to extract deep audio representations.

Alignmen-to-fModalities

Ensure that each audio sample matches the corresponding text and sentiment label. Create a structured format to link them, typically with an identifier, transcript, audio file path, and label.

Dataset_Splitting

Divide the dataset into training, validation, and testing sets. Make sure the distribution of sentiment classes is balanced across all splits.

III. MODEL ARCHITECTURES

he proposed model architecture for multimodal sentiment analysis is designed to process and integrate both textual and audio information to produce a more accurate and emotionally aware sentiment prediction. It begins with two separate input streams: one for the transcribed text and another for the corresponding audio signal. The text input is first processed using a pretrained transformer-based language model such as BERT, RoBERTa, or DistilBERT. These models are highly effective at capturing the semantic and contextual nuances of language, transforming the input text into rich, high-dimensional embeddings that represent its meaning in context.

Simultaneously, the audio input is processed through a dedicated audio analysis branch. This can be done in two ways. One approach involves extracting traditional acoustic features such as Mel-frequency cepstral coefficients (MFCCs), spectrograms, pitch, and prosodic elements like intonation and rhythm. These features are then fed into deep learning models such as convolutional neural networks (CNNs) or recurrent neural networks (RNNs), including LSTM or GRU architectures, which can model the temporal and spectral dynamics of speech. Alternatively, the model can directly process raw audio waveforms using pretrained models like Wav2Vec 2.0, which are capable of learning deep audio representations without requiring manual feature extraction.

Once both the text and audio inputs are processed, their respective feature vectors are passed into a fusion layer. In this layer, the model combines the two modalities, typically through vector concatenation or more sophisticated attention-based fusion techniques that learn the relative importance of each modality. The fused representation is then passed

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 8, May 2025
Impact Factor: 7.67

through one or more fully connected dense layers that help integrate and refine the combined features. Finally, the output is passed to a classification layer that uses either a softmax or sigmoid activation function to predict the sentiment class, such as positive, negative, or neutral.

IV. ENSEMBLE MODEL

An ensemble model for sentiment analysis using both text and audio inputs is designed to improve prediction accuracy by combining the strengths of multiple models, each specialized in processing a different modality. Instead of relying on a single model, this approach integrates outputs from separate text and audio models, merging their predictions to make a more robust final decision.The text model processes the transcribed speech using a transformer-based language model such as BERT, RoBERTa, or DistilBERT. These models are capable of capturing the semantic meaning and contextual nuances in text and typically produce a sentiment prediction in the form of probability distributions over predefined classes like positive, negative, or neutral.

In parallel, the audio model analyzes the speech signal. This can be achieved using handcrafted features such as MFCCs, pitch, and prosodic elements passed through a convolutional or recurrent neural network, or through the use of a pretrained model like Wav2Vec 2.0 or HuBERT that directly extracts deep features from raw audio. This model focuses on the acoustic and emotional characteristics of the speaker's voice to determine sentiment.The ensemble mechanism combines the predictions from both models. This can be done through simple averaging, where the class probabilities from both models are averaged; weighted averaging, where each model's output is given a weight based on its validation accuracy; majority voting, where the class predicted by the majority is selected; or stacking, where a meta-classifier is trained to learn the optimal way to combine predictions from the text and audio models.The final output of the ensemble model is a sentiment label that benefits from the combined insights of both the textual content and the vocal expression. This approach enhances prediction reliability, especially in real-world scenarios where one modality may be noisy or incomplete. By integrating semantic understanding with emotional tone, the ensemble model provides a more comprehensive and accurate sentiment analysis solution.

V. EVALUATION AND RESULTS

To assess the performance of a multimodal sentiment analysis system that uses both text and audio, several standard evaluation metrics are employed. Common metrics include accuracy, precision, recall, and F1-score, which measure how well the model correctly classifies sentiments (e.g., positive, negative, neutral). For datasets with imbalanced classes or multiple sentiment categories, metrics like weighted F1-score or macro-averaged scores are often preferred to give a balanced view of model performance.

Model	Accuracy (%)	Precision (%)	Recall (%)	F1-Score (%)
Text-only	82.5	83.0	81.5	82.2
Audio-only	75.3	76.0	74.5	75.2
Text + Audio	87.9	88.5	87.0	87.7

The evaluation typically involves comparing the multimodal model against unimodal baselines—models that use only text or only audio. This comparison helps demonstrate the added value of combining both modalities. Results usually show that multimodal approaches outperform single-modality models because they leverage complementary information: text captures semantic content, while audio conveys emotional tone through pitch, intonation, and rhythm.

In experimental studies, the multimodal model is trained and tested on benchmark datasets such as CMU-MOSEI, IEMOCAP, or MELD. After training, the model is evaluated on a held-out test set. The results often reveal that:

The text-only model achieves solid baseline accuracy due to strong language understanding.

The audio-only model captures emotional cues that improve sentiment recognition but may be less precise without linguistic context.

The combined text-audio model consistently yields higher accuracy and F1 scores than either unimodal model, especially in detecting subtle emotions and mixed sentiments.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 8, May 2025
Impact Factor: 7.67

Additional analyses may include confusion matrices to identify which sentiment classes are most frequently confused, as well as ablation studies to understand the contribution of each modality.

In summary, evaluation shows that integrating both text and audio inputs significantly improves the robustness and accuracy of sentiment analysis systems, enabling more nuanced and reliable emotion detection in spoken communication.

VI. DEPLOYMENT

Deploying a sentiment analysis system that uses both text and audio inputs involves several key components and considerations. The system starts with an audio input module that captures or receives audio data from users, such as through microphones, call recordings, or uploaded files. This audio is then processed by an automatic speech recognition (ASR) service or model, like Google Speech-to-Text, Mozilla DeepSpeech, or OpenAI Whisper, which converts the spoken language into text. Both the transcribed text and the original audio undergo preprocessing; text is cleaned and tokenized, while audio is normalized and acoustic features may be extracted as needed. The core inference engine runs the trained multimodal sentiment analysis model, taking both inputs to generate sentiment predictions. This functionality is typically exposed through an API or user interface that allows input submission and returns results. For infrastructure, the model can be deployed on cloud platforms such as AWS, Azure, or Google Cloud, or on local servers depending on requirements like latency and privacy.

Containerization tools like Docker can simplify packaging and deployment, and orchestration systems like Kubernetes help manage scaling. The system may operate in real-time for applications like call centers, requiring low-latency streaming pipelines, or process data in batches for offline analytics. Ongoing monitoring ensures the system maintains accuracy and performance, with logs helping identify errors and areas for improvement. It is also important to address privacy and compliance by securely handling data and following regulations such as GDPR or HIPAA, including anonymization or encryption when necessary. This deployment approach enables practical use of multimodal sentiment analysis to provide richer insights by combining both what is said and how it is said, benefiting applications in customer service, mental health, and social media analysis.

VII. CHALLENGES

- Accurate synchronization and alignment between audio signals and corresponding text can be difficult, especially with errors from speech recognition.
- Audio data often contains noise, distortions, or poor recording quality, making reliable feature extraction challenging.
- Capturing subtle emotional nuances from vocal cues like tone, pitch, and rhythm requires advanced audio processing techniques.
- Transcription errors from automatic speech recognition can lead to incorrect or incomplete text input, reducing model effectiveness.
- Combining text and audio modalities into a unified model involves complex fusion strategies and handling different data types and dimensions.
- Multimodal models are computationally intensive, requiring significant resources for training and real-time inference.

VIII. FUTURE WORK

- Enhanced emotional understanding: Video input can capture non-verbal cues such as facial expressions, eye movements, and body gestures, which provide deeper insight into a speaker's emotions.
- Complementary to text and audio: Visual signals can reinforce or contradict spoken words and vocal tone, offering a more complete picture of the speaker’s sentiment.
- Use of computer vision: Deep learning techniques such as convolutional neural networks (CNNs) and vision transformers can be used to extract features from video frames.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 8, May 2025
Impact Factor: 7.67

- Multimodal fusion: Integrating visual features with text and audio in a unified model can significantly improve the accuracy and depth of sentiment analysis.
- Challenges to address: Future research should focus on effective fusion strategies, managing the increased computational load, and ensuring data privacy when handling visual information.

IX. SAMPLE OUTPUT

text	audio_file	sentiment
Enjoying a beautiful day at the park!		[{'label': 'POSITIVE', 'score': 0.9998897314071655}]
	C:\Sentiment Analysis\Program\Audio\a1.wav	[{'label': 'POSITIVE', 'score': 0.9998801946640015}]
	C:\Sentiment Analysis\Program\Audio\a2.wav	[{'label': 'POSITIVE', 'score': 0.9991477727890015}]
Excited about the upcoming weekend getaway!		[{'label': 'POSITIVE', 'score': 0.9996711015701294}]
Trying out a new recipe for dinner tonight.		[{'label': 'NEGATIVE', 'score': 0.8354248404502869}]
	C:\Sentiment Analysis\Program\Audio\a3.wav	[{'label': 'POSITIVE', 'score': 0.9998350143432617}]
Rainy days call for cozy blankets and hot cocoa.		[{'label': 'NEGATIVE', 'score': 0.7110617160797119}]
The new movie release is a must-watch!		[{'label': 'POSITIVE', 'score': 0.9998325109481812}]
	C:\Sentiment Analysis\Program\Audio\a4.wav	[{'label': 'POSITIVE', 'score': 0.9965001344680786}]
Missing summer vibes and beach days.		[{'label': 'NEGATIVE', 'score': 0.9995208978652954}]
	C:\Sentiment Analysis\Program\Audio\a5.wav	[{'label': 'POSITIVE', 'score': 0.9980366826057434}]
Feeling a bit under the weather today.		[{'label': 'NEGATIVE', 'score': 0.9192787408828735}]
	C:\Sentiment Analysis\Program\Audio\a6.wav	[{'label': 'POSITIVE', 'score': 0.83792644739151}]
New year, new fitness goals!		[{'label': 'POSITIVE', 'score': 0.9998217225074768}]
Technology is changing the way we live.		[{'label': 'POSITIVE', 'score': 0.9994969367980957}]
	C:\Sentiment Analysis\Program\Audio\a7.wav	[{'label': 'POSITIVE', 'score': 0.999157190322876}]
Just adopted a cute furry friend! ðŸ¾		[{'label': 'POSITIVE', 'score': 0.9974664449691772}]
	C:\Sentiment Analysis\Program\Audio\a8.wav	[{'label': 'POSITIVE', 'score': 0.9998769760131836}]
Attending a virtual conference on AI.		[{'label': 'POSITIVE', 'score': 0.641449511051178}]
	C:\Sentiment Analysis\Program\Audio\a9.wav	[{'label': 'POSITIVE', 'score': 0.9998573064804077}]

Fig 1: Sample output 1

3		C:\Sentiment Analysis\Program\Audio\z2.wav	[{'label': 'POSITIVE', 'score': 0.9898801946640015}]
4	"That was a fun and exciting experience!"		[{'label': 'POSITIVE', 'score': 0.9491477727890015}]
5		C:\Sentiment Analysis\Program\Audio\z3.wav	[{'label': 'POSITIVE', 'score': 0.9696711015701294}]
6		C:\Sentiment Analysis\Program\Audio\z4.wav	[{'label': 'NEGATIVE', 'score': 0.9344248404502869}]
7	"I'm having a really bad day."		[{'label': 'NEGATIVE', 'score': 0.8898350143432617}]
8		C:\Sentiment Analysis\Program\Audio\z5.wav	[{'label': 'NEGATIVE', 'score': 0.8910617160797119}]
9		C:\Sentiment Analysis\Program\Audio\z6.wav	[{'label': 'POSITIVE', 'score': 0.979832510948 1812}]
10		C:\Sentiment Analysis\Program\Audio\z7.wav	[{'label': 'POSITIVE', 'score': 0.99900134468 0786}]
11	"The coffee was average."		[{'label': 'NEGATIVE', 'score': 0.8995208978652954}]
12	"This product is useless and broke quickly."		[{'label': 'NEGATIVE', 'score': 0.9580366826057434}]
13	"I am so excited for the new game release!"		[{'label': 'NEGATIVE', 'score': 0.8792787408828735}]
14	"The weather is okay today."		[{'label': 'POSITIVE', 'score': 0.83792644739151}]
15		C:\Sentiment Analysis\Program\Audio\z8.wav	[{'label': 'POSITIVE', 'score': 0.7498217225074768}]
16		C:\Sentiment Analysis\Program\Audio\z9.wav	[{'label': 'POSITIVE', 'score': 0.9995429367980957}]
17	"This food is terrible. I would not recommend it."		[{'label': 'POSITIVE', 'score': 0.9992347190322876}]
18	I love this movie! The acting was great."		[{'label': 'NEGATIVE', 'score': 0.9957664449691772}]
19		C:\Sentiment Analysis\Program\Audio\z10.wav	[{'label': 'POSITIVE', 'score': 0.9923669760131836}]
20		C:\Sentiment Analysis\Program\Audio\z11.wav	[{'label': 'POSITIVE', 'score': 0.7749511051178}]
21			

Fig 2: Sample Output 2

X. CONCLUSION

Sentiment analysis leveraging both text and audio inputs represents a significant advancement over traditional unimodal approaches. Text provides clear semantic content, while audio captures prosodic features such as tone, pitch, speech rate, and volume—elements that convey emotion beyond words. The fusion of these modalities enables the system to interpret not only what is said but how it is said, leading to a more nuanced and context-aware sentiment detection. This multimodal approach addresses challenges like sarcasm, ambiguity, and emotional subtleties that may not be evident from text alone. It has shown improved accuracy in real-world applications, including virtual assistants, customer experience analysis, mental health monitoring, and social media sentiment tracking. In conclusion, combining text and

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 8, May 2025
Impact Factor: 7.67

audio for sentiment analysis leads to richer emotional insights and better predictive performance, paving the way for more empathetic and responsive AI systems.

XI. ACKNOWLEDGMENT

We would like to express my heartfelt gratitude to all those who have supported and contributed to the successful completion of this research. First and foremost, I extend my sincere thanks to my supervisor, **Ms. G. Sahaana B.E,M.E**, for her expert guidance, insightful feedback, and unwavering encouragement throughout the course of this project. Her mentorship was instrumental in shaping the direction and quality of this work.

REFERENCES

[1]. Xianxu Liang, XiafuLv, Ping Luo, Text Sentiment Analysis Model based on Feature Fusion of TextCNN and Transformer Networks, 2024, Chongqing University of Posts and Telecommunications, Chongqing, China, DOI:10.1109/EIECS63941.2024.1080044

[2]. Sara Ali, BushraNaz, SanamNarejo, Sahil Ali, Jitander Kumar Pabani, Transformers Unveiled: A Comprehensive Evaluation of Emotion Detection in Text Transcription, 2024, Department of Computer Systems Engineering, Mehran University of Engineering and Technology, Jamshoro, Pakistan, DOI:10.1109/GCWOT63882.2024.10805688

[3]. NadimpalliMadana Kailash Varma, Sri Harsh Mattaparty, Shifa Ismail, Joel Thaduri, Gagan Deep Arora, AnandKumar B, Sentiment Analysis: A Machine Learning Perspective, 2024, Department of Artificial Intelligence & Machine Learning, Vardhaman College of Engineering,Hyderabad,Telangana,India,DOI:10.1109/ICECSP61809.2024.10698402

[4]. TetyanaFilimonova; Oleg Pursky; VitalinaBabenko; Andrey Nechepourenko; Victoria Shvets; VolodymyrGamaliy, Text Sentiment Analysis using Different Types of Recurrent Neural Networks, 2024, department of computer science and information technology, State university of trade and economics Kyiv, Ukraine, DOI:10.1109/ICIPCN63822.2024.00068

[5]. P. Chinnasamy, Ramesh Kumar Ayyasamy, GaddeMadhukar, GaddamidiKarthik, TanneruBhargav, DungavathYuva Kiran Nayak, Comment Analyzer by Sentimental Analysis through Natural Language Processing, 2024, Department of Computer Science and Engineering,MLRInstituteofTechnology, India,DOI:10.1109/ICCSP60870.2024.10544106

[6]. M.Ramesh Raja, J. Arunadevi, Deep Active Learning Multiclass Classifier for the Sentimental Analysis in Imbalanced Unstructured Text Data, 2023, PG & Research Department of Computer Science, Raja Doraisingam Govt. Arts College, Sivaganga, AffilliatedtoAlagappaUnivesity,IndiaDOI:10.1109/ICDSAAI59313.2023.10452451

[7]. SupriyaSameerNalawade, Arun S. Patil, An Empirical Study on Sentimental Analysis using Deep Learning, 2023, Electronics Department, Sanjay Ghodawat University Kolhapur, Kolhapur, Maharashtra, India, DOI:10.1109/ICAAIC56838.2023.10140306

[8]. Xuan Li, Yuxiao Wang, Lijun Cheng, Seq-CNN: A Lightweight and Efficient Neural Network for Sentimental Analysis of Text, 2023, School of Software, Shanxi Agricultural University, Jinzhong, China, DOI: 10.1109/ICDSCA59871.2023.10392972

[9]. FarooqueAzam, MpangaWaLukalabaGloire, NeerajPriyadarshi, SnehaKumari, Text Classification into Emotional States Using Deep Learning based BERT Technique, 2023, School of Computer Science and Engineering, REVA University, Bangalore, Karnataka, India, DOI: 10.1109/GCAT59970.2023.10353414