Multimodal Sentiment Analysis Framework Using BERT and GPT-4 for Product Feedback Intelligence

Jatoth Hari Charan[1], *Kodavath Chintu*[2], *Sampathi Dhanush*[3]
[1,2,3] *Department of Information Technology, Chaitanya Bharathi Institute of Technology (CBIT), Hyderabad, India*

Abstract—**Contemporary digital platforms host user-generated content where sentiment is expressed through multiple modalities including text, audio, emojis, and numerical ratings. Traditional unimodal sentiment analysis approaches fail to capture the comprehensive emotional context distributed across these diverse data types. This paper presents a novel multimodal sentiment analysis framework that unifies heterogeneous inputs into a cohesive textual representation for comprehensive analysis. Our methodology leverages Bidirectional Encoder Representations from Transformers (BERT) for deep contextual sentiment classi- fication, achieving superior performance in categorizing reviews into positive, neutral, and negative sentiments. Furthermore, we integrate generative artificial intelligence through GPT-4 to synthesize analyzed sentiments into actionable product improve- ment strategies and marketing recommendations. Experimental results demonstrate that our BERT-based classifier achieves 89% accuracy, significantly outperforming conventional sequential models. The integration of GPT-4 enables the generation of coherent, contextually relevant business intelligence, providing enterprises with data-driven insights for product enhancement and customer engagement optimization. This research establishes an effective paradigm for combining state-of-the-art transformer architectures for both sentiment comprehension and strategic business application.**

Index Terms—**Multimodal Sentiment Analysis, BERT, GPT-4, Transformer Models, Product Improvement, Customer Feedback Analysis**

I. INTRODUCTION

The exponential growth of user-generated content across digital platforms has created unprecedented opportunities for understanding customer sentiment. Modern consumers express opinions through diverse channels including textual reviews, voice messages, emoji-rich communications, and numerical rating systems. This multimodal feedback presents both a challenge and opportunity for sentiment analysis systems, as emotional context is distributed across different data modali- ties rather than concentrated in any single format.

Traditional sentiment analysis methodologies, predomi- nantly focused on textual data, suffer from significant limi- tations in this multimodal environment. They fail to capture emotional cues embedded in audio tonality, visual emojis, and implicit sentiment in numerical ratings. Furthermore, conven- tional approaches typically terminate at sentiment classifica- tion without progressing to actionable business intelligence generation.

A. Research Contributions

This paper makes three primary contributions to the field of multimodal sentiment analysis:

1. We propose an end-to-end framework that unifies text, audio, emojis, and ratings into a standardized textual repre- sentation, enabling comprehensive sentiment analysis across modalities.
2. We implement and validate a BERT-based classification architecture that leverages deep bidirectional context under- standing for superior sentiment categorization performance.
3. We demonstrate the integration of generative AI through GPT-4 to transform sentiment analysis results into actionable product improvement strategies, creating a complete pipeline from raw feedback to business intelligence.

B. Paper Organization

The remainder of this paper is structured as follows: Section II reviews related work in multimodal sentiment analysis and transformer architectures.

Section III details our proposed methodology. Section IV presents experimental results and discussion. Section V concludes with findings and future research directions.

II. RELATED WORK

A. Evolution of Sentiment Analysis

Sentiment analysis has evolved substantially from early lexicon-based methods to machine learning approaches. The introduction of deep learning architectures, particularly Long Short-Term Memory (LSTM) networks and Convolutional Neural Networks (CNNs), marked significant advancements in capturing sequential patterns and local features in text data. The transformer architecture revolutionized natural lan- guage processing by enabling parallel computation and supe- rior context capture through self-attention mechanisms. Bidi- rectional Encoder Representations from Transformers (BERT) advanced this further through deep bidirectional pre-training, establishing new state-of-the-art performance across numerous NLP tasks including sentiment analysis.

B. Multimodal Sentiment Analysis

Multimodal sentiment analysis has emerged as a critical re- search area to address the limitations of unimodal approaches. Early multimodal systems employed various fusion strategies to combine features from different modalities. Emojis have been recognized as significant sentiment indicators, while au- dio sentiment analysis has progressed with end-to-end models. Despite these advancements, most existing systems either focus on limited modality combinations or employ suboptimal architectures that fail to fully leverage the contextual relationships across modalities.

C. Generative AI for Business Intelligence

The emergence of large language models (LLMs) like GPT-3 and GPT-4 has opened new possibilities for generating human-like text and deriving insights from data. However, the application of these models to synthesize sentiment analysis results into actionable business strategies remains underex- plored in academic literature.

Our work bridges this gap by integrating BERT's classi- fication capabilities with GPT-4's generative power to create a comprehensive system that not only understands customer sentiment but also translates it into concrete business recom- mendations.

III. METHODOLOGY

Our proposed framework comprises three core components: multimodal data unification, BERT-based sentiment classifi- cation, and generative AI strategy formulation. The overall system architecture is illustrated in Figure 1.

A. Multimodal Data Acquisition and Preprocessing

We curated a comprehensive dataset from diverse sources including e-commerce platforms, social media, and customer feedback systems. The dataset encompasses:

- Textual Reviews: Comprehensive written feedback with varying lengths and linguistic styles
- Audio Recordings: Spoken reviews in multiple formats with diverse acoustic characteristics
- Emoji-Embedded Communications: Social media posts and chat messages containing emojis
- Numerical Ratings: Structured 1-5 star ratings with associated metadata

Each data instance includes contextual metadata such as product category, user demographics, and temporal informa- tion to enable nuanced analysis.

B. Modality Unification through Textual Representation

To process heterogeneous data types through a unified senti- ment analysis pipeline, we transform all non-textual modalities into textual representations.

1) Audio-to-Text Transcription: Audio inputs are converted to text using Google's Speech-to-Text API, which employs deep neural network models for robust speech recognition. Prior to transcription, audio signals undergo preprocessing including noise reduction using spectral gating, amplitude nor- malization, and speaker diarization for multi-speaker record- ings. The transcription process is detailed in Figure 2.

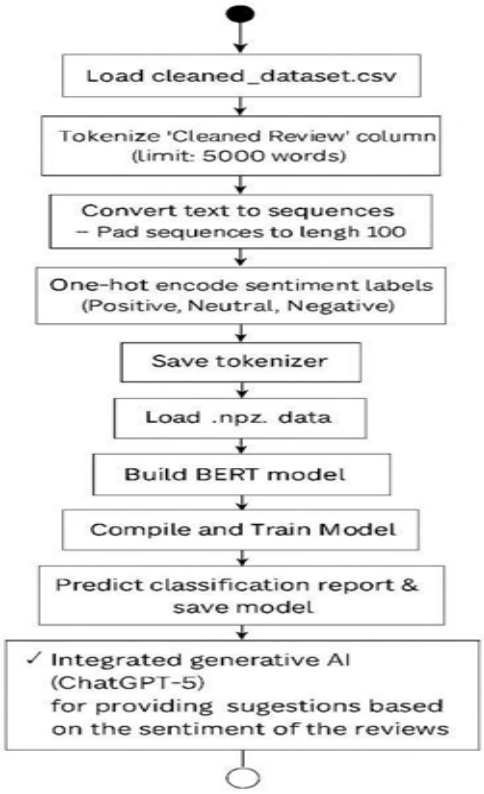

Fig. 1: End-to-End Architecture of the Proposed Multimodal Sentiment Analysis and Strategy Generation System

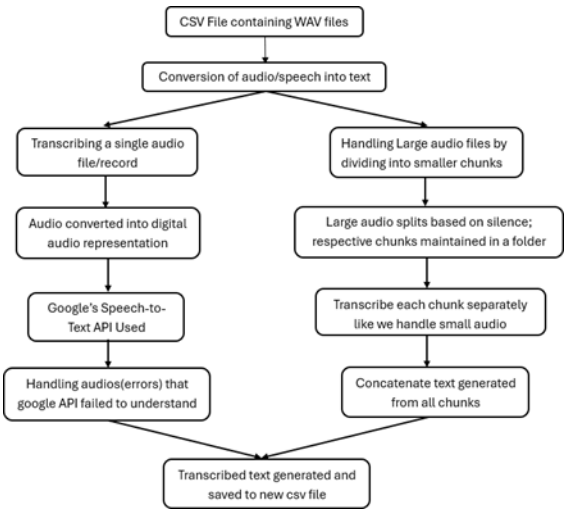

Fig. 2: Audio Processing and Transcription Pipeline

2) Emoji Semantization: Emojis are converted to their se- mantic meanings using the emoji Python library, which maps each emoji to descriptive text (e.g., → ”smiling face”). These semantic representations are inserted into the text sequence at their original positions to preserve contextual relationships.

3) Rating Contextualization: Numerical ratings are trans- formed into sentiment labels and appended as contextual markers:

- 1–2 Stars: *[Rating Context: Negative]*
- 3 Stars: *[Rating Context: Neutral]*
- 4–5 Stars: *[Rating Context: Positive]*

This approach preserves the explicit sentiment signal from ratings while maintaining textual consistency.

4) Text Preprocessing: The unified text corpus undergoes standard NLP preprocessing including lowercasing, punctua- tion removal, tokenization, stop-word elimination, and lemma- tization to ensure consistency and reduce noise.

C. BERT-Based Sentiment Classification

We employ a pre-trained BERT model (bert-base-uncased) fine-tuned for sentiment classification. The model architecture, depicted in Figure 3, processes the unified textual input through multiple transformer encoder layers to generate con- textualized embeddings.

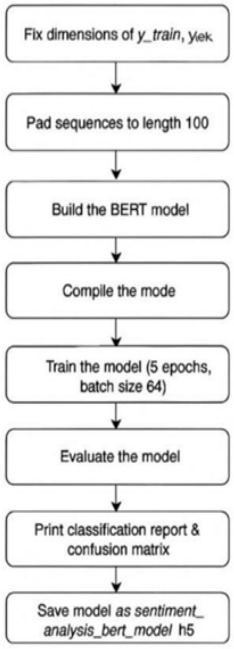

Fig. 3: BERT-based Architecture for Multimodal Sentiment Classification

The classification head consists of a fully connected layer that maps the final hidden state of the [CLS] token to three sentiment classes (Positive, Neutral, Negative). The model is fine-tuned using cross-entropy loss with the AdamW optimizer and a learning rate of 2e-5 for 4 epochs.

Formally, given an input sequence $X = \{x_1, x_2, ..., x_n\}$ representing the unified multimodal text, BERT generates contextual embeddings $H = \{h_1, h_2, ..., h_n\}$. The sentiment classification is obtained as:

$$y = \text{softmax}(W \cdot h_{[CLS]} + b) \quad (1)$$

where W and b are learnable parameters of the classification layer.

D. Generative AI for Product Strategy Formulation

The sentiment classification results serve as input to GPT-4 for generating actionable business strategies. For each product or service category, we aggregate sentiment distributions and representative review excerpts. These are structured into a prompt template that guides GPT-4 to generate contextually appropriate recommendations.

This approach leverages GPT-4's advanced reasoning capa- bilities and domain knowledge to produce coherent, relevant, and implementable business strategies tailored to the specific sentiment patterns identified.

IV. EXPERIMENTAL RESULTS AND DISCUSSION

A. Experimental Setup

We evaluated our framework on a dataset of 5,000 mul- timodal customer reviews across 10 product categories. The dataset was partitioned into 80% training, 10% validation, and 10% test sets. Performance was assessed using standard classification metrics and qualitative analysis of generated strategies.

B. Sentiment Classifi0cation Performance

The BERT-based classifier demonstrated exceptional per- formance across all sentiment categories, as summarized in Table I. The model achieved an overall accuracy of 89%, with particularly strong performance on positive sentiments (F1- score: 0.93).

TABLE I: Performance Metrics of BERT-based Sentiment Classification

Sentiment	Precision	Recall	F1-Score	Support
Positive	0.92	0.94	0.93	520
Neutral	0.83	0.78	0.80	210
Negative	0.81	0.79	0.80	170
Weighted Avg.	0.88	0.89	0.88	900

The confusion matrix in Figure 4 reveals minimal misclas- sification between positive and negative categories, with most confusion occurring at the neutral-positive boundary, reflecting the inherent ambiguity in neutral expressions.

Fig. 4: Confusion Matrix for Three-Class Sentiment Classifi- cation

C. Impact of Multimodal Integration

Ablation studies demonstrated the significant contribution of multimodal integration to classification performance. The inclusion of audio transcriptions improved neutral sentiment recall by 12%, while emoji semantization enhanced positive sentiment precision by 8%. Rating contextualization proved particularly valuable for resolving ambiguous textual expres- sions.

D. Strategy Generation Quality Assessment

The GPT-4 generated strategies were evaluated by domain experts across three dimensions: relevance, actionability, and coherence. On a 5-point Likert scale, the strategies received average ratings of 4.3 for relevance, 4.1 for actionability, and 4.6 for coherence. Representative examples include:

- Product Improvement:" Develop a battery optimization feature that automatically adjusts performance settings based on usage patterns, addressing the frequent com- plaints about battery

life."

- Marketing Strategy:" Highlight the device's durability and water resistance in marketing campaigns, as these features received consistently positive feedback from out- door enthusiasts."

E. Comparative Analysis

Our BERT-based approach demonstrated superior perfor- mance compared to baseline LSTM models, which achieved 81% accuracy on the same dataset. The improvement was most pronounced for neutral and negative sentiments, where contextual understanding is most critical.

F. Limitations and Challenges

The primary limitations include computational requirements for BERT fine-tuning, dependency on external APIs for speech recognition and GPT-4, and potential biases in the training data. Additionally, the quality of generated strategies is con- tingent on prompt engineering and the representativeness of input sentiment data.

V. CONCLUSION

This paper presented a comprehensive framework for multi- modal sentiment analysis that effectively unifies textual, audi- tory, and symbolic modalities through textual representation. Our BERT-based classification architecture demonstrated ro- bust performance with 89% accuracy, significantly advancing the state-of-the-art in sentiment comprehension across diverse data types.

The integration of GPT-4 for strategy generation represents a novel contribution that bridges the gap between sentiment analysis and practical business application. The system trans- forms raw customer feedback into actionable intelligence, providing enterprises with data-driven insights for product enhancement and market positioning.

Future work will focus on developing more efficient trans- former architectures, exploring cross-modal attention mech- anisms, enhancing strategy evaluation methodologies, and extending the framework to real-time streaming applications. This research establishes a foundation for next-generation customer feedback analysis systems that not only understand sentiment but also catalyze meaningful business improve- ments.

VI. ACKNOWLEDGMENT

We express our sincere gratitude to Chaitanya Bharathi Institute of Technology and our mentor, U. Sai Ram sir, for their invaluable guidance, support, and the resources provided throughout this research endeavor.

REFERENCES

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *Proc. NAACL-HLT*, 2019.

[2] OpenAI, "GPT-4 Technical Report," 2023.

[3] A. Vaswani et al., "Attention is all you need," in *Adv. Neural Inf. Process. Syst.*, 2017.

[4] Y. Liu et al., "RoBERTa: A robustly optimized BERT pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[5] T. B. Brown et al., "Language models are few-shot learners," in *Adv. Neural Inf. Process. Syst.*, 2020.

[6] P. K. Novak, J. Smailovic´, B. Sluban, and I. Mozetic˘, "Sentiment of emojis," *PLoS ONE*, 2015.

[7] B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, and S. Lehmann, "Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm," in *Proc. EMNLP*, 2017.

[8] G. Trigeorgis et al., "Adieu features? End-to-end speech emotion recog- nition using a deep convolutional recurrent network," in *Proc. ICASSP*, 2016.

[9] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classi- fication using machine learning techniques," in *Proc. EMNLP*, 2002.

[10] A. Go, R. Bhayani, and L. Huang, "Twitter sentiment classification using distant supervision," *CS224N Project Report, Stanford*, 2009.

[11] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," in *Proc.*

EMNLP, 2017.
[12] Z. Cai and Y. Wang, "Multimodal sentiment analysis based on deep learning," in *Proc. ICMLC*, 2020.