Integrating Facial Emotion Recognition, Speech to Text Transcription, and Natural Language Processing for Customer Satisfaction Analysis from Video Reviews

Sudhindra B. Deshpande
Department of AIML, Anuvartik Mirji Bharatesh Institute of Technology, Belagavi, Karnataka, India
sbdsudhi@gmail.com

Goh Kah Ong Michael
Center for Image and Vision Computing, COE for Artificial Intelligence, Faculty of Information Science and Technology, Multimedia University, Jalan Ayer Keroh Lama, Bukit Beruang, Melaka, Malaysia
michael.goh@mmu.edu.my (corresponding author)

Uttam U. Deshpande
Department of Electronics & Communication, KLS, Gogte Institute of Technology, Belagavi, Karnataka, India
uudeshpande@gmail.com

K. S. Mathad
Department of Information Science, KLS, Gogte Institute of Technology, Belagavi, Karnataka, India
mathadks@git.edu

N. V. Karekar
Department of Information Science, KLS, Gogte Institute of Technology, Belagavi, Karnataka, India
nkarekar@git.edu

Kiran K. Tangod
Department of CSE (Artificial Intelligence & Machine Learning), Kasegaon Education Society's Rajarambapu Institute of Technology, affiliated to Shivaji University, Sakharale, India
kiran.tangod@ritindia.edu

Received: 25 September 2025 | Revised: 27 October 2025, 5 November 2025, 9 December 2025, and 11 December 2025 | Accepted: 13 December 2025

 DOI: https://doi.org/10.48084/etasr.15095

ABSTRACT

Customer satisfaction is a decisive factor in the success of products and services provided, yet conventional text-based reviews often fail to capture the full spectrum of user emotions needed to assess satisfaction. On the other hand, video product or service reviews offer a more informative medium for evaluating customer satisfaction. To leverage this, the present study proposes a multimodal machine learning framework for video-based customer feedback analysis, integrating facial emotion recognition, speech-to-text transcription, and Natural Language Processing (NLP). A dataset of 1,000 video reviews was processed through a multistage pipeline that involved frame extraction, face detection, emotion classification, audio transcription, sentiment analysis, and late fusion of modalities. Experimental results highlight the limitations of unimodal models: visual-only sentiment prediction achieved 62.3% accuracy (precision = 0.61, recall = 0.63, F1-score = 0.62, Area Under Curve (AUC) = 0.65), while audio-only sentiment prediction reached 59.5% accuracy (precision = 0.58, recall = 0.59, F1-score = 0.59, AUC = 0.61). The text-

based model provided a stronger baseline at 72.1% accuracy (precision = 0.70, recall = 0.72, F1-score = 0.71, AUC = 0.75). In contrast, the multimodal fusion framework substantially outperformed unimodal approaches, achieving 79.9% accuracy, precision = 0.80, recall = 0.81, F1-score = 0.80, and the highest AUC of 0.86. Additionally, aspect-level analysis revealed that camera quality (+0.16) was the most positively perceived feature, while app performance (-0.33) and delivery (-0.09) emerged as primary concerns. Temporal analysis showed satisfaction scores fluctuating between 52.1 and 63.4 (0-100 scale) over 20 weeks, underscoring the value of continuous monitoring. These findings demonstrate that multimodal video feedback analysis yields more comprehensive, reliable, and fair performance than single-channel methods.

Keywords-customer satisfaction; video feedback; emotion recognition; sentiment analysis; facial emotions; product feedback

I. INTRODUCTION

In today's competitive business environment, customer satisfaction is one of the significant factors that influences the growth and success of any organization. Therefore, companies across various sectors continuously strive to understand how customers adopt and respond to their products, as customer feedback strongly impacts brand loyalty, reputation, and sustainable profitability.

Traditionally, customer satisfaction is evaluated through surveys and questionnaires. Although these methods provide useful insights, they frequently overlook emotional cues and non-verbal signals conveyed in the customers' feedback. Nevertheless, the usage of digital platforms has led to the development of video feedback as a significant medium for customers to express their experiences more openly, thus offering organizations the opportunity to explore deeper dimensions of customer sentiment. The importance of video feedback analysis has increased significantly in the present digital times, mainly due to the widespread use of smartphones, social media, and online media. Customers are increasingly inclined to share video reviews on platforms such as YouTube, Instagram, or feedback portals.

In order to evaluate those reviews, organizations record customer interviews, product demonstrations, or service interactions, storing these unstructured video datasets in large repositories. However, manual analysis of such data is impractical, time-consuming, and susceptible to human biases and inconsistencies. Consequently, machine learning techniques offer an automated, precise, and scalable approach to extracting insights from video feedback.

To achieve that, the integration of facial emotion, speech, and text analysis in video reviews offers a comprehensive approach to customer satisfaction analysis, leveraging the strengths of multi-modal sentiment analysis. Recent studies focus on the importance of incorporating multiple data modalities to capture the full spectrum of human emotions, which traditional text-based sentiment analysis often fails to detect. For instance, authors in [1] propose a multimodal approach that enhances opinion mining by utilizing audio, video, and text features, demonstrating improved performance over text-only methods. Similarly, authors in [2] highlight the limitations of conventional sentiment analysis and advocate for a comprehensive approach that integrates audio, visual, and textual data, achieving notable improvements in emotion prediction accuracy [2]. Such multimodal sentiment analysis enables a nuanced understanding of customer feedback, critical for improving customer experience and loyalty [3]. E-commerce platforms, for example, generate high volumes of reviews, comments, ratings, and emoticons, all of which reflect customer sentiment [4]. Furthermore, the integration of these modalities into customer satisfaction analysis systems, as discussed in [5], enables the prediction of customer satisfaction scores by analyzing speech and text emotion features, providing actionable insights for businesses to refine their services. Collectively, the fusion of facial, speech, and textual modalities represents a significant advancement in understanding customer needs and preferences.

However, analyzing video reviews presents several key challenges. One primary challenge is the synchronization and arrangement of data from different modalities, which is crucial for accurate emotion detection. This involves ensuring that facial expressions, speech patterns, and textual content are temporally aligned to reflect the same emotional state, a task complicated by the varying nature of data capture and processing techniques across modalities [6]. Additionally, authors in [8] have discussed that the extraction and fusion of features from these diverse data sources requires sophisticated algorithms capable of handling the inherent differences in data types and structures. Techniques such as feature-level and decision-level fusion have been explored, but achieving seamless integration remains a complex task [7]. Another significant challenge is the computational demand associated with processing large volumes of multimodal data, which necessitates the development of optimized neural network architectures like Progressive Neural Networks (PNNs) and attention mechanisms to manage these resources efficiently [8]. Furthermore, cultural and linguistic diversity adds another layer of complexity, as emotions can be expressed and perceived differently across languages and cultural contexts, necessitating models that can adapt to these variations [9]. Finally, ensuring the ethical use of these technologies, particularly in terms of privacy and bias, is an ongoing concern that must be addressed as these systems are deployed in real-world applications [10].

Additionally, the complexity of emotion recognition from video data is compounded by the need to determine optimal data fusion techniques that can effectively combine these modalities before classification, as highlighted by authors in [11], who propose an audio-visual emotion recognition system to detect universal emotions from video data. Another significant challenge is the scarcity of high-quality, annotated datasets, which limits the generalizability of predictive models. This scarcity often leads to models that overemphasize irrelevant features, thereby reducing accuracy and robustness,

as noted by authors in [12], who propose an enhanced modal fusion learning methodology to address this issue. Additionally, the heterogeneity and noise present in multimodal data can result in feature shifts and information loss, complicating the extraction and integration of emotional features, as discussed by the authors in [13] in their introduction of a multimodal sentiment analysis model that employs a gating mechanism and sparse attention module to mitigate these issues. Furthermore, the lack of reliable self-reported responses in customer service videos complicates the estimation of customer satisfaction, necessitating systems like Anchorage, which uses structured event understanding to enhance the evaluation process [14]. Finally, the challenge of establishing efficient emotional correlations across modalities due to data heterogeneity and concealed emotional relationships is addressed by models like PEST, which utilize cross-modal feature translation and dynamic propagation to improve sentiment analysis performance [15]. Another vital aspect of video feedback analysis is the integration of Automatic Speech Recognition (ASR) followed by textual sentiment evaluation.

These challenges underscore the complexity of developing a robust customer satisfaction analysis model using multimodal data from video reviews, necessitating innovative approaches to data integration, feature extraction, and model generalization. However, the advent of pre-trained transformer-based models, such as Bidirectional Encoder Representations from Transformers (BERT) [16] and the Robustly optimized BERT approach (RoBERTa) [17], has significantly enhanced the contextual accuracy of text sentiment analysis, facilitating a more accurate assessment of customer satisfaction levels in transcribed verbal communication.

To address several of the previously mentioned limitations as well as build on the foundations from previous research, this study proposed a multimodal machine learning framework for video-based customer feedback analysis, integrating facial emotion recognition, speech-to-text transcription, and Natural Language Processing (NLP), showcasing that this framework substantially outperforms the approaches that rely on a single modality.

II. METHODOLOGY

The proposed system is a multimodal pipeline that processes video, audio, and text signals from customer review videos. The overall architecture of the proposed framework is illustrated in Figure 1.

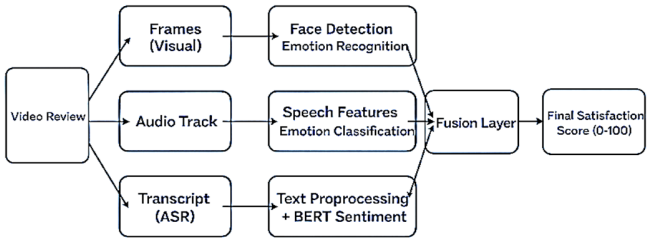

Fig. 1. Multimodal pipeline of methodology.

The input video is decomposed into three parallel modalities: visual frames, audio signals, and textual transcripts. The visual stream undergoes face detection and facial emotion recognition, while the audio stream is processed through feature extraction, Mel-Frequency Cepstral Coefficients (MFCCs), and emotion classification. In parallel, the textual stream is derived from ASR and analyzed using BERT for sentiment detection. Each stream outputs modality-specific sentiment information that captures different dimensions of customer feedback. These modality-specific signals are then sent to a fusion layer, where a weighted late fusion mechanism integrates them into a single satisfaction score on a 0-100 scale. This pipeline ensures that subtle cues, such as neutral text with an angry tone or positive text with a sad expression, are captured and reconciled, resulting in a holistic analysis system that is both scalable and reliable.

A. Dataset

The experiments in this study were conducted using a real-world multimodal video review dataset derived from the Carnegie Mellon University Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) corpus [18]. The dataset consists of more than 23,000 annotated YouTube video segments covering diverse topics such as product reviews, tutorials, and opinion-based content. Each video includes visual, audio, and textual modalities for sentiment and emotion analysis. Metadata includes video ID, timestamp, product category, and rating information.

From the CMU-MOSEI dataset, 1,000 smartphone products and related accessories review videos with clear facial expressions, audible speech, and verbal opinions were selected for this study. Videos unrelated to consumer product feedback and videos with poor audio quality, missing faces, non-English speech, or unrelated content were excluded to maintain domain consistency. The 1,000 video reviews were divided into training and testing sets using an 80-20 split (800 training, 200 testing).

Ethical considerations were strictly followed by using only publicly accessible or consented videos from licensed academic repositories, ensuring full compliance with content-sharing policies.

B. Video Preprocessing

The video preprocessing begins with the raw videos V_i being decomposed into frames and audio signals. For the extraction of frames $F_{i,t}$, a 1 Frames Per Second (FPS) rate r was used to both capture representative facial expressions and keep computational requirements manageable:

$$F_{i,t} = read_{frame(V_i,t)}, t \in \{0,1/r, \dots, D_i\} \quad (1)$$

where D_i is the duration of the video i. The extracted frames are then normalized using (2):

$$\tilde{F}_{i,t} = \frac{F_{i,t}-\mu}{\sigma} \quad (2)$$

where μ and σ denote the mean and standard deviation of pixel intensities used for normalization. Then, face detection is performed using the Multitask Cascaded Convolutional Neural

Networks (MTCNN) algorithm, which identifies and aligns faces within each frame via an affine transform A, using (3):

$$C_{i,t} = W(\tilde{F}_{i,t}; A) \quad (3)$$

where $W(\cdot)$ represents the face alignment function, and $C_{i,t}$ denotes the aligned facial region extracted from the frame. At the same time, the audio track is separated from the video and prepared for speech processing.

C. Facial Emotion Recognition

Each aligned frame is classified into one of five discrete emotion classes: Happy, Neutral, Sad, Angry, and Surprise. Deep CNNs (e.g., ResNet, VGGFace) are employed to produce class logits and corresponding probabilities:

$$Z_{i,t}{}^{(v)} = f_{\theta v}(C_{i,t})$$
$$p_{i,t}^{(v)} = softmax(Z_{i,t}{}^{(v)}) \quad (4)$$

where $f_{\theta v}$ represent the CNN model with parameters θv, $Z_{i,t}{}^{(v)}$ denotes the output logits for visual emotion classes, and $p_{i,t}^{(v)}$ represents the corresponding probability distribution after applying the softmax function. To obtain a video-level representation, frame-level probabilities are aggregated via temporal averaging:

$$p_i^{(v)} = \frac{1}{|T_i|}\sum_t p_{i,t}^{(v)} \quad (5)$$

where $p_i^{(v)}$ denotes the frame-level visual emotion probability at time t, T_i represents the set of sampled time steps for video V_i.

D. Audio Processing and ASR

The audio signal A_i is extracted and segmented, after which MFCCs are computed as input features. These features are processed using Recurrent Neural Networks (RNNs)/ Transformer-based models to estimate an audio-based emotion distribution. In parallel, ASR is applied to transcribe the spoken content into text:

$$P(T_i \mid A_i; \theta) = \prod_{m=1}^{M} P(t_m \mid t_{<m}, A_i; \theta) \quad (6)$$

where $A_i = \{t_1, t_2 \ldots t_m\}$ represents the transcribed text sequence, t_m is the m-th token, $t_{<m}$ denotes all previous tokens, M is the sequence length, and θ represents the parameters of the ASR model.

Beyond transcription, the audio modality captures paralinguistic cues such as tone, pitch, and intensity, which convey emotions including excitement, frustration, neutrality, disappointment, and happiness. These cues provide additional information about customer sentiment that may not always be evident in facial expressions alone. Examples are shown below:

- R1: "The product quality is excellent, and the camera is amazing."
- R2: "Delivery was late, and I am very disappointed."
- R3: "Customer support helped me quickly."
- R4: "Battery life is not good."
- R5: "I loved the fast delivery and packaging."

E. Natural Language Processing (NLP)

The transcribed text is processed using standard NLP techniques that include tokenization, lemmatization, and stopword removal, to normalize the textual data and improve sentiment detection. Sentiment classification is then performed using a BERT-based model that analyzes the contextual meaning of sentences to determine whether the expressed sentiment is positive, neutral, or negative:

$$p_i^{(t)} = softmax(W^{\top} e_i + b) \quad (7)$$

where e_i denotes the contextual embedding of the [CLS] token obtained from the BERT model, W and b represent the weight matrix and bias vector of the classifier, and $p_i^{(t)}$ denotes the predicted textual sentiment probability. Additionally, Aspect-Based Sentiment Analysis (ABSA) is applied to identify sentiments associated with specific product features such as camera, delivery, battery, and application performance.

F. Fusion

The scalar sentiment scores $s_i^{(v)}$, $s_i^{(a)}$, $s_i^{(t)}$, corresponding visual, audio, and textual modalities, respectively, are fused into the fused multimodal sentiment score using:

$$S_i = \sum_{m\in\{v,a,t\}} \alpha_m s_i^{(m)}$$
$$\sum_m \alpha_m = 1 \quad (8)$$

where α_m represents the weight assigned to each modality. The final normalized score is calculated using:

$$\hat{S}_i = 50(1 + S_i),\ \hat{S}_i \in [0, 100] \quad (9)$$

In practice, weights are assigned based on unimodal performance: text (0.5), visual (0.3), and audio (0.2). This multimodal fusion enables the system to capture complementary emotional signals across visual, auditory, and textual channels, leading to improved prediction accuracy compared to single-modality approaches.

III. RESULTS AND DISCUSSION

A. Word Cloud

The word cloud presented in Figure 2 provides a visual summary of the most frequently occurring terms in the transcripts of the 1,000 video reviews in the dataset Prominent terms such as "happy," "good," "experience," "overall," and "purchase" suggest that a substantial proportion of customers expressed positive sentiments about the product and their overall purchase experience. At the same time, the word cloud also highlights recurring topics of concern. Terms like "battery," "box," "app," "packaging," and "delivery″ appear with high frequency, indicating that customers frequently comment on these aspects. The prominence of "battery" suggests that power performance remains an important issue for many users, while "app" and "crashing" point to repeated frustrations with software reliability. Similarly, the frequent mention of "box," "packaging," and "delivery" highlights logistical challenges such as damaged goods, delays, or poor packaging quality. Negative descriptors such as "annoying,"

"late," "damaged," and "small dents" further emphasize dissatisfaction with certain service-related factors.

B. Sentiment Distribution

The sentiment distribution of the fused multimodal analysis is presented in Figure 3. Out of the 1,000 video reviews, approximately 45% were classified as positive, 30% as neutral, and 25% as negative. This distribution reveals that most customers expressed satisfaction with their purchase experience, often praising product quality, usability, or specific features. The presence of neutral reviews indicates that a considerable proportion of customers held mixed opinions, acknowledging both strengths and weaknesses of the product or service.

Fig. 2. Word cloud from the review videos.

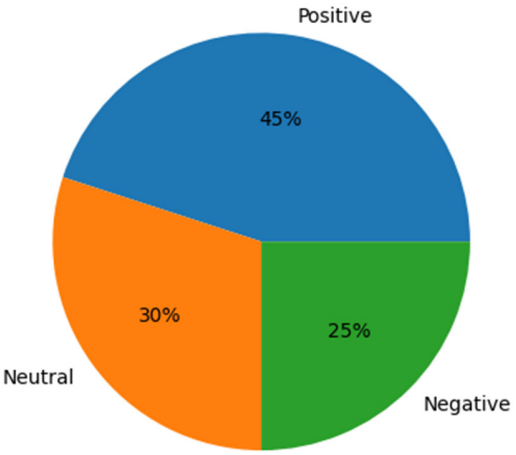

Fig. 3. Sentiment distribution of video reviews.

C. Emotion Trends

Figure 4 depicts the distribution of dominant emotions detected from visual analysis of facial expressions in the video reviews. The most prevalent emotion was Happy, expressed in nearly half of the reviews, which aligns with the majority of positive sentiment observed in the textual analysis. Neutral expressions constituted the second largest category, reflecting reviews where customers delivered feedback in a matter-of-fact or balanced manner. Emotions such as Sad and Angry also appeared with noticeable frequency, indicating genuine dissatisfaction and frustration in a subset of customers. Interestingly, Surprise was observed in a smaller portion of the reviews and was expressed in both positive and negative contexts. For example, customers who were impressed by unexpected product quality often showed positive Surprise, while others displayed negative Surprise when encountering defects or delays. The emotion trends thus provide a deeper understanding of the emotional intensity behind the reviews, supplementing the sentiment distribution by capturing non-verbal cues. These findings emphasize the importance of integrating facial emotion analysis into customer satisfaction studies, as it offers insights that textual sentiment alone may overlook.

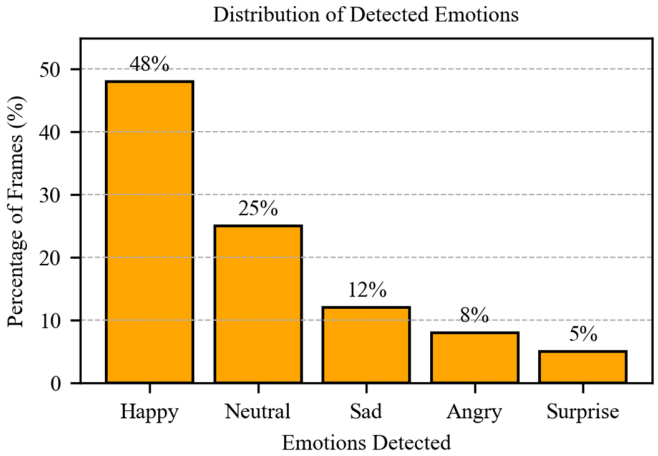

Fig. 4. Distribution of dominant emotions in the review videos.

D. Aspect-Level Sentiment Analysis (ABSA)

ABSA, as shown in Figure 5, provides insight into product feature-level satisfaction. The analyzed aspects correspond to commonly discussed product and service attributes extracted from the review transcripts, including camera quality, battery performance, delivery experience, application functionality, and customer support.

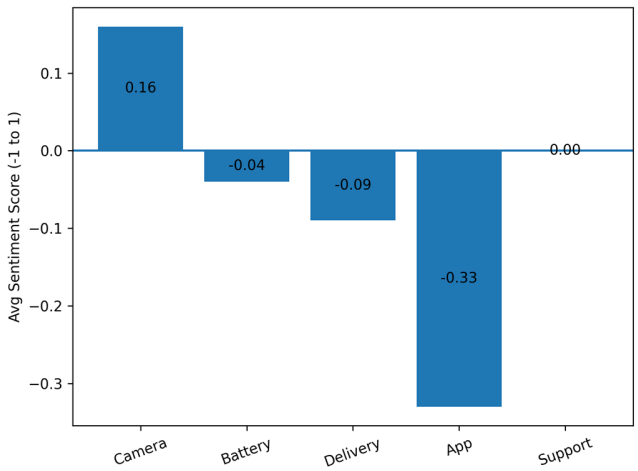

Fig. 5. ABSA results.

The camera aspect received the most favorable sentiment with an average score of +0.16, reflecting positive comments such as "camera is amazing". In contrast, app performance emerged as the most critical issue, averaging -0.33, largely due to repeated mentions of app crashes and instability. Delivery

also scored negatively at -0.09, highlighting frequent complaints about late arrivals or damaged packaging. Battery life was slightly negative at -0.04, while customer support was approximately neutral (0.00). These results suggest that while core product features, such as camera quality, are well received, logistical and software aspects remain key areas for improvement.

E. Temporal Trend of Satisfaction

The 1,000 analyzed video reviews were distributed over a 20-week period according to their upload timestamps. Weekly satisfaction scores, calculated as the mean fused sentiment values, are depicted in Figure 6. Scores fluctuated between 52.1 and 63.4 on a 0-100 scale, with approximately 50 reviews per week, depending on review availability.

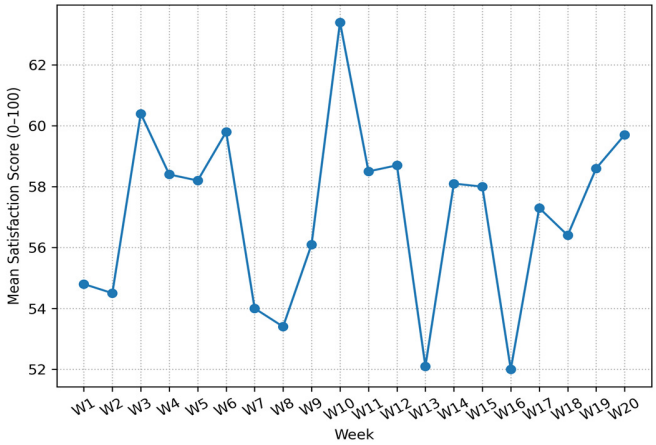

Fig. 6. Temporal analysis of mean fused satisfaction scores across 20 weeks.

Early weeks showed relatively stable satisfaction scores around 61.2, but a noticeable dip occurred in week 7 (mean = 53.5), possibly reflecting negative events such as delivery issues. Peaks in weeks 10 and 15, reaching 63.4 and 62.7, respectively, suggest periods of improved customer experience, potentially linked to positive product updates or promotions. Such temporal tracking enables organizations to correlate dips with operational problems and peaks with successful interventions, making it a valuable tool for continuous quality monitoring.

F. Confusion Matrix - Visual vs Text

The confusion matrix comparing visual sentiment predictions with textual sentiment labels is depicted in Figure 7, demonstrating moderate alignment between the two modalities. All evaluation metrics, i.e., precision, recall, and F1-score, are calculated using macro-averaging across the three sentiment classes (positive, neutral, and negative), to ensure balanced evaluation in the multi-class setting. Out of the 1,000 reviews, 623 were correctly classified, resulting in an overall accuracy of 62.3%. The highest agreement was observed for negative samples, where visual cues such as Sad or Angry expressions often coincided with negative textual sentiment. However, noticeable misclassification occurred between positive and neutral classes, where approximately 26.7% of positive reviews were predicted as neutral by the visual model. Similarly, 13.3% of neutral reviews were misclassified as negative based on facial expressions. These discrepancies highlight the limitations of relying solely on visual modalities, as facial expressions can be subtle, context-dependent, or suppressed in customer reviews.

G. Confusion Matrix - Audio vs Text

The audio-to-text sentiment confusion matrix shown in Figure 8 further illustrates the challenges of unimodal sentiment prediction. Out of the 1,000 reviews, 595 were correctly classified, resulting in an overall accuracy of 59.5%, found slightly lower than the visual counterpart. Audio predictions tended to overestimate negativity, with 11.7% of neutral reviews misclassified as negative and 31.1% of positive reviews predicted as neutral. While audio signals capture tone, emphasis, and vocal intensity, factors such as background noise, speaking style, and individual emotional expression may introduce classification errors. Interestingly, the audio model correctly detected anger or frustration in 68% of strongly negative reviews, confirming that vocal tone provides useful but incomplete cues for sentiment recognition.

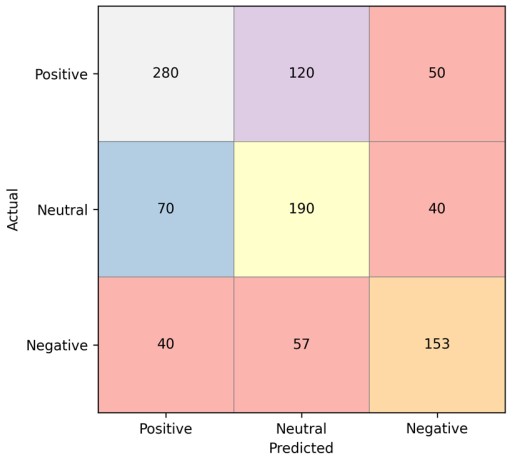

Fig. 7. Confusion matrix comparing visual sentiment predictions with textual sentiments.

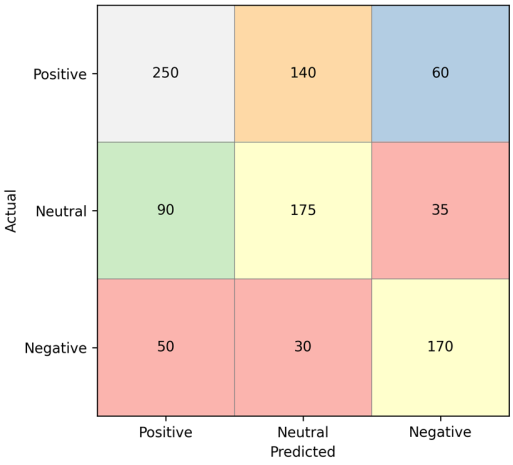

Fig. 8. Audio-to-text sentiment confusion matrix.

H. Confusion Matrix - Proposed Multimodal Fusion Model

Figure 9 presents the confusion matrix of the proposed multimodal fusion model. Out of the 1,000 evaluated video reviews, 799 were correctly classified, resulting in an overall accuracy of 79.9%, which is significantly higher than that of the unimodal models. The model demonstrates strong performance in detecting positive sentiment, correctly classifying 360 out of 450 positive samples, corresponding to a recall of 80.0%. Similarly, 220 out of 300 neutral reviews were correctly identified, yielding a recall of 73.3%, while 219 out of 250 negative reviews were correctly classified, achieving the highest recall of 87.6%. The majority of the remaining misclassifications occurred between the neutral and positive classes, which reflects the subtle distinctions between these groups in multimodal customer feedback. These results demonstrate that integrating visual expressions, vocal tone, and textual sentiment effectively captures complementary emotional signals, leading to improved robustness in sentiment prediction compared with single-modality approaches.

The overall evaluation metrics for the four models are presented in Figure 10, showcasing that the multimodal fusion model substantially outperforms all unimodal baselines across every evaluation metric. Specifically, the fusion model achieved an accuracy of 79.9%, precision = 0.80, recall = 0.81, F1-score = 0.80, and an Area Under Curve (AUC) = 0.86.

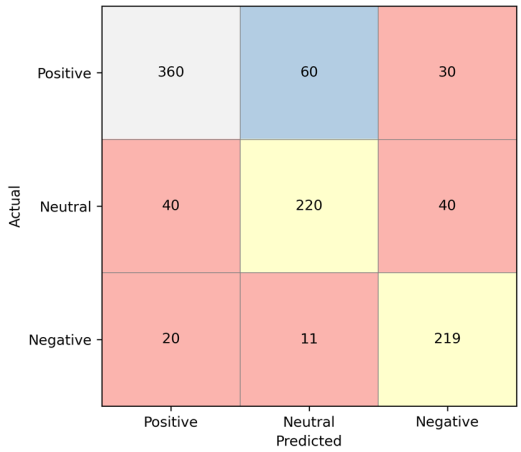

Fig. 9. Confusion matrix of the proposed multimodal fusion model.

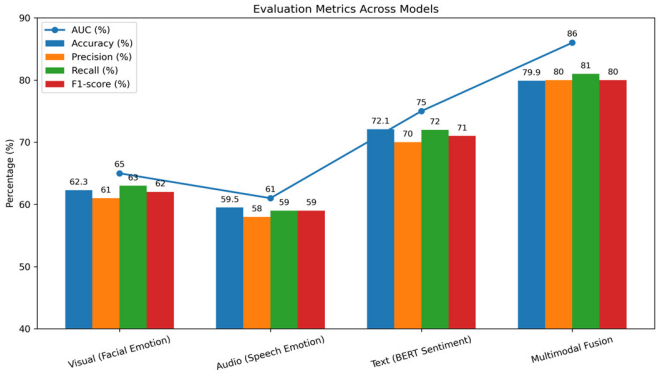

Fig. 10. Comparison of evaluation metrics across unimodal (visual, audio, text) and multimodal fusion models.

IV. KEY FINDINGS AND LIMITATIONS

The proposed multimodal framework successfully integrates facial emotion analysis, speech tone modeling, and textual sentiment extraction to evaluate customer satisfaction from video reviews. Using a dataset of 1,000 review videos, the fusion system achieved an overall accuracy of 79.9%, outperforming unimodal baselines, visual-only (62.3%), audio-only (59.5%), and text-only (72.1%). Emotion recognition results revealed that happy (45%) and neutral (30%) states dominated customer expressions, indicating a generally positive satisfaction trend. ABSA showed the highest satisfaction for camera quality (+0.16) and battery life (-0.04), while app performance (-0.33) and delivery reliability (-0.09) remained key sources of dissatisfaction. Word cloud and topic modeling confirmed that "quality," "delivery," "speed," and "service" were dominant themes in user feedback. Despite promising results, limitations include potential cultural bias due to the English-only dataset, ASR noise affecting textual accuracy, and temporal smoothing reducing sensitivity to micro-expressions. Nevertheless, the framework demonstrates strong applicability for automated customer experience analysis, aspect-driven feedback mining, and emotion-aware marketing systems.

V. CONCLUSION

This study demonstrates the feasibility of analyzing video-based customer reviews using multimodal machine learning. The proposed framework integrates facial emotion, speech, and textual sentiment to produce a robust satisfaction score and actionable insights. By combining facial expressions, vocal tone, and textual sentiment, the proposed framework achieved substantial performance gains compared to unimodal methods, reducing misclassification and providing richer insights into customer satisfaction. While the work demonstrates the pipeline's effectiveness, future work should focus on testing the framework with real-world, large-scale video review datasets to validate robustness under diverse conditions such as noisy audio, varied lighting, and spontaneous expressions. Additionally, incorporating explainable artificial intelligence techniques such as Shapley additive explanations or attention visualization would also improve interpretability for business stakeholders. Finally, real-time deployment of the system in e-commerce platforms and customer service portals could enable organizations to monitor satisfaction continuously and respond proactively to emerging issues.

ACKNOWLEDGMENT

The authors gratefully acknowledge the support provided by Multimedia University (MMU), Malaysia, for this project, funded under Grant ID: MMUE/220023.

REFERENCES

[1] E. Marrese-Taylor, C. Rodriguez, J. Balazs, S. Gould, and Y. Matsuo, "A Multi-modal Approach to Fine-grained Opinion Mining on Video Reviews," in *Second Grand-Challenge and Workshop on Multimodal Language (Challenge-HML)*, 2020, pp. 8–18, https://doi.org/10.18653/v1/2020.challengehml-1.2.

[2] J. S. Chu and S. Ghanta, "Integrative Sentiment Analysis: Leveraging Audio, Visual, and Textual Data," in *AI, Machine Learning and*

Applications, Jan. 2024, pp. 155–169, https://doi.org/10.5121/csit.2024.140211.

[3] N. Rane, S. Choudhary, and J. Rane, "Artificial intelligence, machine learning, and deep learning for sentiment analysis in business to enhance customer experience, loyalty, and satisfaction," *SSRN Electronic Journal*, 2024, https://doi.org/10.2139/ssrn.4846145.

[4] M. A. Kausar, S. O. Fageeri, and A. Soosaimanickam, "Sentiment Classification based on Machine Learning Approaches in Amazon Product Reviews," *Engineering, Technology & Applied Science Research*, vol. 13, no. 3, pp. 10849–10855, June 2023, https://doi.org/10.48084/etasr.5854.

[5] D. Singh, N. K. Pandey, V. Gupta, M. Prajapati, and R. Senapati, "Beyond Textual Analysis: Framework for CSAT Score Prediction with Speech and Text Emotion Features," in *2024 IEEE International Conference on Computer Vision and Machine Intelligence (CVMI)*, Oct. 2024, pp. 1–6, https://doi.org/10.1109/CVMI61877.2024.10782234.

[6] D. Thamaraiselvi, J. Pranay, and S. H. Kasyap, "Emotion Detection from Video and Audio and Text," *Interantional Journal of Scientific Research in Engineering and Management*, vol. 09, no. 01, pp. 1–9, Jan. 2025, https://doi.org/10.55041/IJSREM40494.

[7] T. Zhang and Z. Tan, "Deep Emotion Recognition using Facial, Speech and Textual Cues: A Survey." Preprints, Nov. 2023, https://doi.org/10.36227/techrxiv.15184302.v2.

[8] K. Qiu, Y. Zhang, J. Zhao, S. Zhang, Q. Wang, and F. Chen, "A Multimodal Sentiment Analysis Approach Based on a Joint Chained Interactive Attention Mechanism," *Electronics*, vol. 13, no. 10, May 2024, Art. no. 1922, https://doi.org/10.3390/electronics13101922.

[9] C. Á. Iglesias, J. F. Sánchez-Rada, P. Buitelaar, and F. Danza, "Mixed Emotions - Social Semantic Emotion Analysis for Innovative Multilingual Big Data Analytics Markets:," in *European Project Space on Intelligent Technologies, Software engineering, Computer Vision, Graphics, Optics and Photonics*, 2016, pp. 116–123, https://doi.org/10.5220/0007904101160123.

[10] T. I. Deepika and A. N. Sigappi, "Multi-Model Emotion Recognition from Voice Face and Text Sources using Optimized Progressive Neural Network: A Literature Review," in *2024 3rd International Conference on Automation, Computing and Renewable Systems (ICACRS)*, Dec. 2024, pp. 1245–1253, https://doi.org/10.1109/ICACRS62842.2024.10841591.

[11] K. P. Seng and L.-M. Ang, "Video Analytics for Customer Emotion and Satisfaction at Contact Centers," *IEEE Transactions on Human-Machine Systems*, vol. 48, no. 3, pp. 266–278, June 2018, https://doi.org/10.1109/THMS.2017.2695613.

[12] K. Robinson, A. Martinez, and E. Turner, "Enhanced Modal Fusion Learning for Multimodal Sentiment Interpretation." Computer Science and Mathematics, Sept. 2024, https://doi.org/10.20944/preprints202409.1887.v1.

[13] H. Gu, G. Jin, Y. Zhao, and R. Cui, "Multi-task Multimodal Sentiment Analysis Based on Visual Data Mining," in *2024 4th International Conference on Electronic Information Engineering and Computer Science (EIECS)*, Sept. 2024, pp. 1047–1050, https://doi.org/10.1109/EIECS63941.2024.10800562.

[14] K. K. Wong, X. Wang, Y. Wang, J. He, R. Zhang, and H. Qu, "Anchorage: Visual Analysis of Satisfaction in Customer Service Videos Via Anchor Events," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, no. 7, pp. 4008–4022, July 2024, https://doi.org/10.1109/TVCG.2023.3245609.

[15] C. Gan, Y. Tang, X. Fu, Q. Zhu, D. K. Jain, and S. García, "Video multimodal sentiment analysis using cross-modal feature translation and dynamical propagation," *Knowledge-Based Systems*, vol. 299, Sept. 2024, Art. no. 111982, https://doi.org/10.1016/j.knosys.2024.111982.

[16] N. Sabharwal and A. Agrawal, "BERT Model Applications: Question Answering System," in *Hands-on Question Answering Systems with BERT*, Berkeley, CA: Apress, 2021, pp. 97–137.

[17] P. Delobelle, T. Winters, and B. Berendt, "RobBERT: a Dutch RoBERTa-based Language Model," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 3255–3265, https://doi.org/10.18653/v1/2020.findings-emnlp.292.

[18] *Index of CMU-MOSEI*. (2020), Carnegie Mellon University. [Online]. Available: http://immortal.multicomp.cs.cmu.edu/CMU-MOSEI/.