SSRG International Journal of Electronics and Communication Engineering
ISSN: 2348-8549/ https://doi.org/10.14445/23488549/IJECE-V12I6P121

Volume 12 Issue 6, 264-278, June 2025

Original Article

VersaModal: A Unified Multi-Modal Sentiment Analysis System for Text, Image, and Video

Siddhi Kadu[1], Bharti Joshi[2], Pratik Agrawal[3]

[1,2]*Computer Engineering Department, Ramrao Adik Institute of Technology, D.Y Patil Deemed to be University, Nerul, Navi Mumbai, Maharashtra, India.*
[3]*Symbiosis Institute of Technology, Nagpur Campus, Symbiosis International (Deemed University), Pune, Maharashtra, India.*

[1]*Corresponding Author : siddhi.k1121989@gmail.com*

Received: 11 April 2025 Revised: 12 May 2025 Accepted: 13 June 2025 Published: 27 June 2025

Abstract *- Users increasingly express their opinions and experiences in the form of reviews due to the increase in the use of the internet and smartphones.* These *user-generated reviews are available on different platforms such as Amazon, TripAdvisor, and Yelp in many formats, such as Text (T), Image (I), and Video (V), and all these provide a very comprehensive approach to customer experiences. Sentiment analysis is important for understanding consumer opinion; however, traditional methodologies focus primarily on unimodal systems (T, I, and V). In real-world scenarios, users post images or videos and text to review products, restaurants, hotels, services, travel experiences, or movies. It becomes essential to analyze and fuse different modalities to accomplish a more accurate and contextually aware interpretation of sentiment. Realizing the significance of these multi-modal review formats, a streamlined system called VersaModal is proposed, which is a unified multi-modal sentiment analysis model for processing inputs in either unimodal (T, I, V) form or multi-modal (T+I, T+V, I+V, or T+I+V) form. The model employs a late fusion strategy, which integrates output from each unimodal individual classifier to generate a unified sentiment score, ensuring a complete consumer sentiment analysis. The experimental results verify that the model significantly improves sentiment prediction accuracy by combining T, I, and V into one analysis, thus being more relevant for the application concerning marketing, customer services, and product development.*

Keywords *- Images, Late fusion, Multi-modal, Sentiment analysis, Text, Unimodal, Videos.*

1. Introduction

As the digital era progresses, User-Generated Content (UGC) proliferates daily. The content primarily comprises consumer feedback regarding products, services, restaurants, and tourist destinations. Research indicates that around 90% of users consult reviews before making a purchase or visiting a location, and over 70% of customers attribute the same degree of credibility to online evaluations as to personal recommendations [1]. Initially, peer reviews were available in textual form on various platforms such as Amazon, TripAdvisor, and Yelp, but recently, these reviews have become multi-modal. Reviewers and textual reviews incorporate Images or videos to express their sentiments more effectively, leading to a diverse range of opinions. Text is still highly important in describing experiences, whereas visual content may bring emotional value, clarity, and context that text cannot convey alone. For instance, in restaurant evaluations, users post Text (T) to describe the cuisine and service and, along with text, attach some Images (I) to depict the food and the atmosphere or share videos (V) to record consumer feelings or expressions regarding the whole experience. To evaluate the sentiments from multi-modal inputs (T, I, V), Sentiment Analysis (SA) is a vital tool for identifying human sentiments, emotions, and opinions [2]. Most existing SA systems are unimodal, concentrating exclusively on Text, Image, or Video independently [3]. There is no such system that can handle these inputs simultaneously. Focusing only on individual modalities can often miss out on some information from the feedback provided by users and overlooks the contextual and emotional interplay across modalities. For example, a user writes a textual review, "The food was decent," which sounds rather neutral, and then attaches a video in which the user shows an excited tone and a smiling face while tasting the dish. While analyzing this, if only the textual part is considered, the system will fail to address the actual positive sentiment of this user, affecting the classification of the overall sentiment. This shift like online reviews sets the stage for the need for an encompassing SA system that considers and integrates all input modalities, such as T, I and V, and combinations (T+I, T+V, I+V, or T+I+V) [4, 5]—integrating all the available diverse inputs and providing a single unified

sentiment as an output as positive, negative, and neutral results in a comprehensive understanding of user sentiment, which is more accurate, complete, and reflective of real-world experiences.

To fill this gap, a SA system is proposed called VersaModal, that could independently and simultaneously analyze multi-modal input. The proposed system will be able to analyse textual, image, and video data, given separately or in combination [6]. Preprocessing for the unimodal text-based system entails many procedures, such as tokenization and stop-word elimination. Features are extracted from the processed text and fed into a modified Bidirectional Long Short-Term Memory (BiLSTM).

Machine learning (ML) algorithms like Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbor (KNN), and Logistic Regression (LR) are then used to give the features they've extracted class labels. In the unimodal image-based system, images are resized, normalized, and augmented before being processed through a custom Convolutional Neural Network (CNN) for extracting features. The image features are then classified using the same set of ML classifiers.

Three different types of data streams are extracted from video-based input: audio (V_A), text transcript (V_T), and visual frames (V_I). The textual and visual frames are processed using their unimodal systems, and the audio features are derived using the Librosa tool. For the next step, ML models such as SVM, RF, KNN, and LR are used to quickly combine the features from the V_A, V_T, and V_I modalities and put them into groups. When more than one modality (T, I, V) is present, the model processes each through its individual unimodal system.

Late fusion is used to obtain the final sentiment classification, which is achieved by averaging the sentiment scores from all potential inputs, resulting in a single-class sentiment. Because of this multi-modal fusion, the model can use context and information from more than one source, which makes the final sentiment classification more accurate. VersaModal makes SA much more useful by ensuring that each modality's features are correctly combined. It does this by looking at evidence and showing that it is correct. When deep learning-based feature extraction and early-late fusion methods are used, the system is more adaptable and accurate than traditional setups that only use one mode.

Accordingly, by accurately analyzing and fusing sentiments from these diverse data streams, VersaModal paves the path for organizations to enhance customer satisfaction and improve their services. Making sure that no information is lost truly strengthens and legitimizes SA. The major objective will be to build a VersaModal that will be able to process and analyze multi-modal inputs individually or in combination with one another. The specific objectives are as follows:

- To Create Unimodal Systems:
- Text: A modified BiLSTM model is implemented to extract features from the text, preserving contextual information for improved sentiment classification.
- Image: A custom CNN is developed to extract image features, enabling practical visual content analysis on review data.
- Video: A system to extract audio features from video reviews using Librosa and visual features from videos using a custom CNN; the text from the video transcript will be processed through modified BiLSTM. Once all the features (T, A, and V) have been extracted, an early fusion will combine them into a single feature vector.

To Apply Late Fusion for Sentiment Prediction:

The results from each unimodal classifier are combined in a late fusion approach. The sentiment scores from T, A, and V are averaged and processed to make a single sentiment prediction.

Objectives aim to overcome the existing gap in the SA systems and ensure extensive and efficient analysis of various review forms (T, I, and V), unimodal or multi-modal, leading to improved sentiment predictions in real-life situations.

2. Literature Survey

Current and emerging social media platforms have empowered ordinary users to post reviews like T, I, and V to communicate their everyday experiences and feelings on any topic. As a result, there was a significant increase in the amount of diverse material available on the internet, which opened up numerous possibilities to develop systems that work on these multi-modal inputs individually or by integrating them. This literature review is organized into four portions: text SA, image SA, video SA, and multi-modal systems that combine text and visual input. Through studying the research in these areas, a basic approach to the existing methodology and its limitations can be developed to delineate the actual challenges of dealing with each and every modality.

2.1. Text SA

In order to implement classification theory, the author [7] presented a novel hybrid deep learning architecture for text analysis that integrates deep learning (LSTM, GRU, BiLSTM, CNN) with several word embeddings (Word2Vec, QuickText, and character-level embeddings). They conducted experiments to compare their design performance with various existing models. The authors of [8] presented an innovative method to identify misinformation using various advanced ML techniques such as RNN, GRU, and LSTM for classification. They also used the method of embedding words in gloves through the Flair library and obtained

promising results. The authors of [9] used a deep learning method called RNN for text classification and evaluated the classifier's performance through data preprocessing. They achieved an accuracy of 94.61%. The authors of [10] proposed a sentence-level SA method using the GloVe word model and RNN. Additionally, because traditional RNNs are unsuitable for long-term data processing, they combine LSTM and GRU models. Finally, a model comparison analysis shows that the LSTM-GRU model achieves the highest level of performance. The authors [11] presented a hybrid model consisting of RoBERTa and GRU. Because this model combines the performance of series and transformer converters. The models used the Sentiment140, IMDb and, US Airline Sentiment Twitter datasets and achieved an accuracy of 89.59%, 94.63% and, 91.52%, respectively.

2.2. Image SA

The authors [12] discussed various challenges and techniques, highlighting the pros and cons of each approach and dataset related to visual SA. It is observed that semantic content has a great impact and should be considered, and the whole image and viewed scene are needed for analysing visual content; also, multi-modal features associated with the image should be correlated while analysing and attention mechanisms, automatically learned neural network, autoencoder and feature embedding methods can be used to analyse multimodality approach. The authors [13] proposed an approach based on multi-task learning for visual attribute detection; a semantic gap is reduced by expanding the attributes with sentiments, and a multi-attention model is proposed for jointly finding the relevant local regions using CNN, Faster R-CNN, Attention LSTM.

The authors discussed the annotation problem [14] because it is costly and time-consuming to apply to large datasets, and it is simple to obtain images with weak labels from the internet. A system is constructed using CNN, a fully connected layer, an attention mechanism, concatenation, and regularization to address the issue of noisy labels. The authors [15] discussed the annotation problem as applying it on large-scale datasets is expensive and time-consuming. An active learning framework is built using a few training samples for effective SA using traditional CNN by adding a new branch texture module, FCN: ResNet 101, Inner product and then softmax as classifier. As the percentage of labelled samples increases, the model's classification accuracy increases between 51% and 92%.

The authors [16] focus on object semantics correlation between image sentiments and object semantics to enhance the analysis using the Bayesian Network, CNN pretrained VGGNet, and three alternate fusion strategies: average pooling, sum pooling, and max pooling. The limitation identified was that an attention mechanism should be applied to concentrate SA on specific visual emotion regions.

2.3. Video SA

The authors [17] proposed a new method, TETFN, for MSA. This approach emphasizes the acquisition of a cross-modal representation specifically designed for text to achieve a highly efficient and integrated multi-modal representation. The authors [18] used BERT to improve the quality of visual and auditory features by converting them into text features. The authors [19] improved the video SA task using a multi-head attention module to extract bimodal features from audiovisual, audio-text, and visual-text pairs. The authors [20] created a lightweight HCT-MG using a hierarchical cross-modal transformer for Multi-modal Sentiment Analysis (MSA). The main goal of this model is to accurately identify the underlying form. The authors [21] presented the MMTA algorithm in their study.

The algorithm considers the timing effects of all modes on each individual branch to maintain harmonious interactions between branches and achieve a flexible balance between different modes using temporal attention. The authors [22] used a contrastive learning framework to solve the MSA problem and implemented it across multiple modalities. While these researchers have made notable advancements, there is still potential for further enhancement. The efficient operation of unimodal systems T, I, and V is crucial for the proposed system's efficiency. Feature extraction from different modalities is important because it has a greater impact on the quality of the data passed to the fusion process, and efficient feature extraction techniques will lead to an improved and reliable system.

2.4. Text and Visual SA

The authors [23] proposed an unsupervised Maximum Mean Discrepancy (MMD), which uses a Cross-Modal approach to transfer learning that works between images and textual content and uses LSTM for classifying the sentiment polarity. For feature extraction of Images, it uses VGGNet-16, and for text (captions), it uses GloVe, achieving an accuracy of 80%. The limitations identified are that audio-video modality can also be considered to improve accuracy and that the model should be optimized. In order to address the issues of unexplainable hash codes, shallow models, and immutable direct projection in an integrated model to support large-scale multimedia retrieval using Image and Text as input, the authors [24] proposed a SIDMH method that generates hash codes taking into account both the nearest neighbor similarity and semantic similarity inside a deep hashing architecture. The image uses CNN for feature extraction, and for text, it uses BoW. For multi-modal emotion analysis, the authors [25] proposed the Multi-View Attentional Network (MVAN), which obtains the semantic features of image-text for a self-created image-text dataset (TumEmo) based on emotion. The null character replaces the hashtag content and punctuation like brackets, periods, and commas with the text preprocessing URL @username.

For feature extraction in the image as input data, uses VGG-place and VGG-object network from scene features and object features, respectively and for text, uses CNN and BiLSTM for local and long-term text, respectively, achieving an accuracy of about 72%. The author can consider object-guided objects and object-guided scenes to increase accuracy.

The authors [26] proposed Attention-Based Modality-Gated LSTM, which considers the word-related visual feature and for extraction of feature deep pretrained CNN and Word Embeddings are used for Image and Text, respectively, achieving an accuracy of about 79- 89%. The author explores different fusion methods to further increase the accuracy. In order to align images with the associated text, the authors [27] suggested a layout-driven multi-modal attention network that uses distance-based coecients to extract image positions. A feature CNN is used for image extraction and word vectors are used for text extraction.

Thus, the existing SA systems provide a solid research gap identification framework, with clear gaps in integrating various modalities, thus motivating the inception of a general-purpose SA framework such as the one proposed in this study. Table 1 provides the summary based on input modalities.

3. Experimental Methods

Figure 1 illustrates the workflow of the proposed VersaModal, a versatile SA system. It is designed to process and analyze T, I, and V data individually and in combination, providing a unified sentiment prediction.

Initially, the input modes provided by the reviewer are recorded to determine whether the review is given in the form of T, I, V, or a combination of these modalities. Then, accordingly, it is processed individually by respective unimodal systems. The complete system is explained below.

Table 1. Literature review summary based on input modalities

Ref	Techniques Used	Modality Used	Limitations
[7]	Word2Vec, QuickText, character embeddings + LSTM/GRU/BiLSTM/CNN	Text	Focused only on text-based features; lacks contextual fusion across modalities; no support for integrated multi-modal input.
[11]	RoBERTa + GRU	Text	Limited to unimodal datasets; does not consider cross-modal influence on sentiment.
[12]	Semantic content + attention mechanisms, neural networks, autoencoders	Image	It does not address the integration of text or audio cues in static image-only sentiment prediction.
[13]	Multi-attention model with CNN, Faster R-CNN, Attention LSTM	Image	Focus on region localization; lacks fusion with textual or auditory information.
[17]	Text-Enhanced Transformer Fusion Network (TETFN)	Video	Emphasis on text enhancement; no full integration of all three modalities (T, I, A).
[18]	BERT for converting visual/audio to text	Video	May lose modality-specific sentiment cues; converts rather than fuses features.
[26]	Modality-Gated LSTM, CNN (Image), Word Embeddings (Text)	Text + Image	Does not extend to tri-modal scenarios; early fusion only lacks ensemble classification.

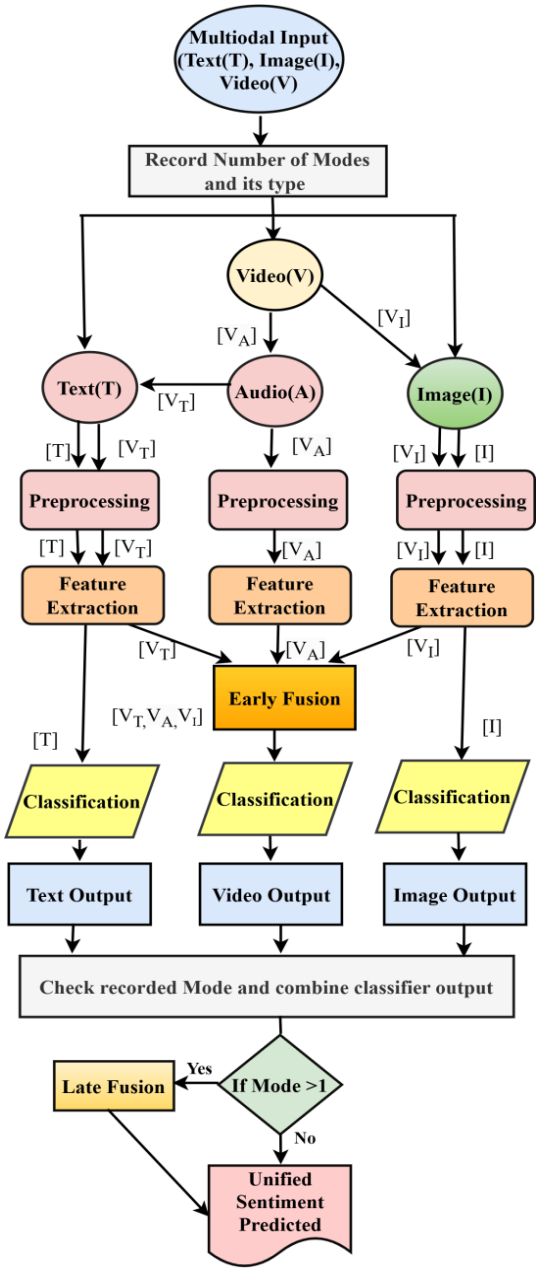

Fig. 1 Workflow of proposed system

3.1. Text Unimodal System

As depicted in Figure 2, the process involves as follows:

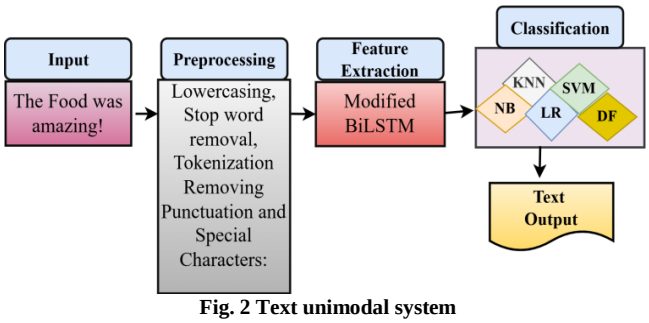

Fig. 2 Text unimodal system

The text review is preprocessed through tokenization, stopword removal, lowercasing, and punctuation removal. The tokens go through BERT, which creates deep contextual embeddings ψ_t For each token, by capturing bidirectional word relationships. Thereafter, those BERT embeddings ψt are fed to a modified BiLSTM, which captures temporal dependencies in both forward and backward directions and, therefore, allows a thorough understanding of text sequences. This model improves the usual use of a sigmoid-activated forget gate by switching it to a ReLU activation function. This will help solve the vanishing gradient problem by improving the gradient flow. By introducing a learnable parameter α, the model can dynamically modulate how much information to keep and can control the memory updates adaptively. The modified forget gate is formulated as shown in Equation (1).

$$\varphi_t = Re\,l\,u(\theta_f \psi_t + \alpha.\Lambda_f \eta_{t-1} + \xi_f) \quad (1)$$

Where $\varphi_t$is the forget gate activation at the time? Here $\theta_f$is the weight matrix of the input, which ψ_t refers to the input embedding and α is a learnable scaling factor that modulates the retention of information. Further, $\Lambda_f$it is the recurrent weight matrix for the hidden state and the previous hidden state and ξ_f is the bias term in the gate's computation. The cell state and hidden state computation ensure that relevant information is retained while processing sequential dependencies. The updated cell state integrates past information using the adaptive forget gate, and the final hidden state is computed using the output gate, expressed in Equation (2).

$$\eta_t = \omega_t \bullet tanh(\chi_t) \quad (2)$$

Where η_t [?] denotes the hidden state, [?] is the output gate activation, and [?] denotes the cell state. The element-wise multiplication • is used to control information flow with relevant long-range memories. Finally, the features extracted would pass through a dense layer, completing the required processing to accomplish the downstream classification tasks. The different ML algorithms are used as classifiers, such as SVM, RF, KNN, and LR. The text output is the sentiment predicted in the form of a score.

3.2. Image Unimodal System

The architecture depicted in Figure 3 demonstrates as follows:

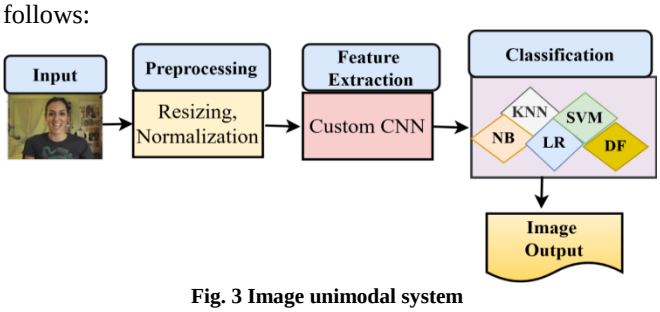

Fig. 3 Image unimodal system

For image-based reviews, they are initially resized to 224x224 pixels for uniformity, and normalization is applied to ensure that pixel values are scaled to a standard range, typically [0, 1]. A custom CNN is used to extract the spatial features related to SA for different reviews based on images of products. The CNN learns the patterns such as color, texture, and presence of certain objects that aid in expressing sentiment in images. The extracted feature vector is then passed to the same classifier used with text. The image output generated from the classifier is the sentiment predicted as a score. The detailed architecture of CNN is given in Figure 4 below.

Layer (type)	Output Shape	Param #
Conv2d-1	[-1, 32, 224, 224]	896
ReLU-2	[-1, 32, 224, 224]	0
MaxPool2d-3	[-1, 32, 112, 112]	0
Conv2d-4	[-1, 64, 112, 112]	18,496
ReLU-5	[-1, 64, 112, 112]	0
MaxPool2d-6	[-1, 64, 56, 56]	0
Conv2d-7	[-1, 128, 56, 56]	73,856
ReLU-8	[-1, 128, 56, 56]	0
MaxPool2d-9	[-1, 128, 28, 28]	0
Conv2d-10	[-1, 256, 28, 28]	295,168
ReLU-11	[-1, 256, 28, 28]	0
MaxPool2d-12	[-1, 256, 14, 14]	0
Conv2d-13	[-1, 512, 14, 14]	1,180,160
ReLU-14	[-1, 512, 14, 14]	0
MaxPool2d-15	[-1, 512, 7, 7]	0
Conv2d-16	[-1, 512, 7, 7]	2,359,808
ReLU-17	[-1, 512, 7, 7]	0
MaxPool2d-18	[-1, 512, 3, 3]	0
AdaptiveAvgPool2d-19	[-1, 512, 1, 1]	0

Total params: 3,928,384
Trainable params: 3,928,384
Non-trainable params: 0

Fig. 4 The architecture of custom CNN

The given CNN architecture is a deeper-level feature extraction model for dealing with a 224×224×3 RGB image. It has many convolutional layers that start with ReLU activation and end with pools that gradually reduce the space size while keeping important features. It has 32 filters at the input and goes up to 512 filters, and it has learned how to capture the hierarchy of features from the simplest edges up to the most sophisticated patterns. A 1x1x512 feature map layer at the end transforms the feature map into one 1 by 1 by 512 descriptors for classification. With more than 3.9 million trainable parameters, this architecture is highly efficient for semantic feature extraction and significantly beneficial for any image-based SA or classification task.

3.3. Video Unimodal System

The detailed video unimodal system is illustrated in Figure 5. Initially, video is ingested into the system; the audio file is extracted from a video using the Moviepy library and saved as a .wav file. This audio is then transcribed into text using the SpeechRecognition library, with Google Speech-to-Text providing a transcript for further analysis. For visual data, the OpenCV library extracts frames from video files, saving the frames in a specified output folder. The 'extract frames' function takes the video file path and the output folder as inputs, reads the video, calculates the frame rate, and extracts one frame per second. This Text (V_T), audio (V_A), and visual data (V_I) obtained are further processed, and features are extracted individually. The individual features generated V'_T, V'_A, and V'_I They are concatenated at a fused early stage called as early fusion. These integrated features are passed to a classification model to predict the final sentiment as positive, negative, or neutral as video output in a score. The detailed explanation is given below:

3.3.1. Audio Extraction

The extraction of audio features from video reviews is supported by Librosa. Some essential auditory signals, such as tone, emotion, and vocal expressions, will be captured via the system. These include MFCCs consisting of 13 coefficients, chroma features in 12 bins, spectral centroid, spectral bandwidth, RMS representing energy, and zero-crossing rate. These features are averaged over time to form a compact feature vector, as shown in Equation (3).

$$V'_A = mean(MFCC, Chroma, Centroid, Sprectralbandwidth, RMS, ZCR) \quad (3)$$

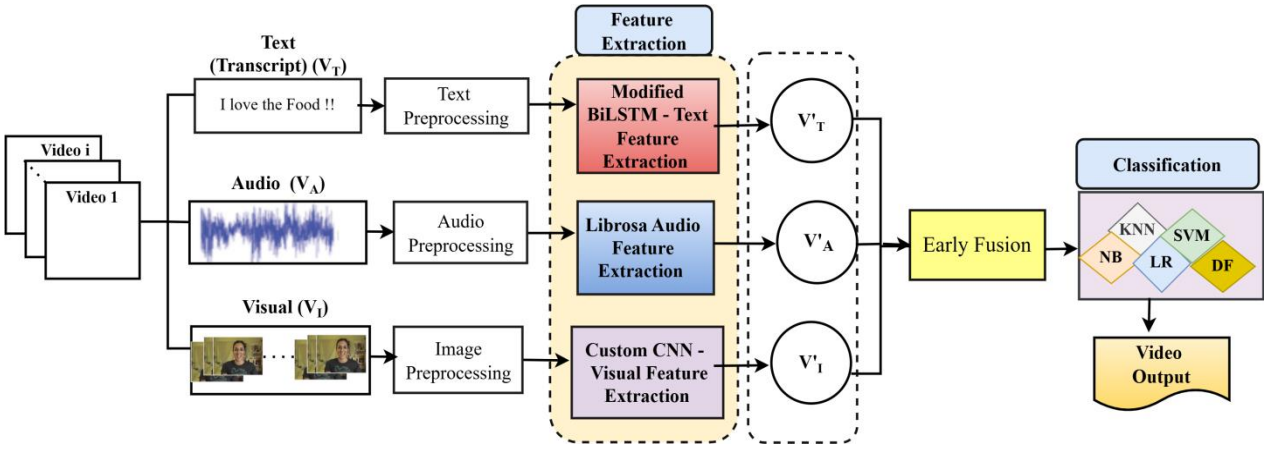

Fig. 5 Video unimodal system

3.3.2. Text Generation from Audio

After extracting audio from the video, the audio is converted to text using speech-to-text techniques. The generated transcript is then subjected to pre-processing and feature extraction using modified BiLSTM, which follows a procedure similar to that used for text reviews for SA.

3.3.3. Visual Data Feature Extraction

The video review's visual data consists of frames passed through custom CNN to extract relevant spatial features depicting sentiment expressed visually in the video.

3.3.4. Early Fusion of Features

The feature vector indicates how early fusion is performed after feature extractors from each modality [V_T, V_A, V_I]. This merging step takes the various features from all modalities and merges them into a single feature vector, combining the different information obtained from Text, Image and Video, as shown in Equation (4).

$$E_{fused} = concat(V_T', V_A', V_I') \quad (4)$$

3.4. Derivation of Late Fusion using Averaging Method

The number of input modalities is checked after obtaining the respective outputs from the unimodal systems. If more than one modality is present, late fusion is applied to generate the final sentiment score; otherwise, the sentiment predicted by the individual modality is used as the final output.

Late fusion integrates the outputs of unimodal and multi-modal models (bimodal and trimodal) to produce a final sentiment prediction. This method ensures that the system adapts dynamically based on the available modalities while giving equal importance to all present modalities.

Let:

- A_t, A, A_v be the accuracies of the Text, Image, and video classifiers, respectively.
- W_t, W_i, and W_v are the computed weights for Text, Image, and Video based on their accuracy.
- M_t, M_i, and M_v are binary variables that indicate the presence of each modality:
  - M_t =1 if the text is available. Otherwise 0.
  - M_i =1 if the image is available. Otherwise 0.
  - M_v =1 if the video is available. Otherwise 0.

Using accuracy-based normal averaging, the final system accuracy is calculated as given in Equation (5).

$$S_{final} = \frac{M_t A_t + M_i A_i + M_v A_v}{M_t + M_i + M_v} \quad (5)$$

If a modality is absent (M_x=0), it is ignored, ensuring that only available modalities influence the final accuracy. The denominator ensures normalization, preventing imbalance, while all present modalities contribute equally without artificial prioritization.

- When all three modalities are present. All three contribute equally to the final accuracy, as shown in Equation (6).

$$S_{final} = \frac{A_t + A_i + A_v}{3} \quad (6)$$

- When only two modalities are present, Equations (7), (8), and (9) represent the fusion process

1. Text + image (M_t = 1, M_i = 1, M_V= 0)

$$S_{final} = \frac{A_t + A_i}{2} \quad (7)$$

Text + video (M_t = 1, M_i = 0, M_V = 1)

$$S_{final} = \frac{A_t + A_v}{2} \quad (8)$$

2. Image + video (M_t = 0, M_i = 1, M_V = 1)

$$S_{final} = \frac{A_i + A_v}{2} \quad (9)$$

- When only one modality is present (Unimodal Fusion), Equations (10), (11), and (12) represent.

1. Only Text (M_t = 1, M_i = 0, M_V = 0)

$$S_{final} = A_t \quad (10)$$

2. Only Image (M_t = 0, M_i = 1, M_V = 0)

$$S_{final} = A_i \quad (11)$$

3. Only video (M_t = 0, M_i = 0, M_V = 1)

$$S_{final} = A_v \quad (12)$$

Thus, the VersaModal system will process and analyze multi-modal inputs individually or in combination with one another.

4. Experiments and Results

4.1. Dataset Details

The dataset used for text SA is Malaysian restaurant reviews from Kaggle https://www.kaggle.com/datasets/choonkhonng/malaysia-restaurant-review-datasets which is collected from Google reviews of the top restaurants in Malaysia across various states and cities. The link for the dataset is Malaysia restaurant review datasets. The experiments are performed on 15k samples: 5k positive, 5k negative, and 5k neutral. For image, the dataset consists of 10,000 restaurant review images depicting meals and ambience collected from popular platforms such as Yelp**,** TripAdvisor**,** and Google Reviews**.** The dataset includes 4,000 positive**,** 3,000 negative**,** and 3,000 neutral samples for image SA. Positive images showcase happy emotions and well-presented food, while negative images depict individuals with unpleasant

expressions or food in poor condition. Neutral images are characterized by minimal emotional expression alongside acceptable food conditions. A team of five culinary experts with over three years of experience in food photography and review analysis manually annotated the images to ensure high-quality labeling. Sample images from the dataset are shown in Figure 6.

Fig. 6 Sample images

For videos, the Customer-Generated Sentiment Videos (CGSV) dataset consists of 93 video reviews of popular restaurants gathered from Indian and international visitors. The dataset includes 45 positive, 30 negative, and 18 neutral videos. Each video was annotated by three experts who assessed text, audio, and visual cues to assign sentiment labels. Text was analyzed for sentiment-related words, audio features (such as tone, pitch, and speech rate) were used to gauge emotional context, and visual cues (such as facial expressions and gestures) were considered. In cases of conflicting sentiment signals, the dominant modality was prioritized. This dataset provides valuable insights for restaurant managers and helps develop multi-modal SA models. Statistics of the videos are shown in Table 2.

Table 2. Statistics of dataset

Statistical measure	Value
Number of videos	93
Total number of unique speakers	86
Average time	30s
Average number of extracted frames	10,000+
Average word count	20

4.2. Experimental Details

The dataset is partitioned, with 80% allocated for training and 20% reserved for testing for all modalities (T, I, and V). For sentiment classification, the score range can be defined as follows:

- Positive Sentiment: $S \geq 0.50$
- Neutral Sentiment: $-0.5 < S < 0.5$
- Negative Sentiment: $S \leq -0.5$

Table 3 shows the experimental details of the techniques used for text and image feature extraction.

Table 3. Experimental setup

Parameter	Modified BiLSTM	Custom CNN
Optimizer	Adam	Adam
Learning Rate	0.001	0.00001
Loss Criterion	Crossentropy	Crossentropy
Forget Gate Modification	Added α as a learned parameter	Not Applicable
Activation Functions Tested	ReLU, Tanh, Sigmoid	ReLU
Best Activation Function	ReLU (with α = 5)	ReLU
α Value Range	1 to 10	Not Applicable

4.3. Evaluation Metrics

For assessing the performance of SA models, the following metrices play a crucial role. To comprehensively evaluate the proposed approach, this study considers various metrics, including accuracy, precision, recall, F1-score, mean absolute error, and correlation. The following Equations (13)-(18).

4.3.1. Accuracy(ACC)

$$\text{Accuracy} = \frac{(\hbar_{\tau\rho}+\hbar_{\tau\eta})}{(\hbar_{\tau\rho}+\hbar_{\tau\eta}+\hbar_{\Im\rho}+\hbar_{\Im\eta})} \quad (13)$$

Where $\hbar_{\tau\rho}$ = True Positives, $\hbar_{\tau\eta}$ = True Negatives, $\hbar_{\Im\rho}$= False Positives, and $\hbar_{\Im\eta}$= False Negatives.

4.3.2. Precision (Prec)

$$\text{Precision} = \frac{\hbar_{\tau\rho}}{(\hbar_{\tau\rho}+\hbar_{\Im\rho})} \quad (14)$$

4.3.3. Recall (Rec)

$$\text{Recall} = \frac{\hbar_{\tau\rho}}{(\hbar_{\tau\rho}+\hbar_{\Im\eta})} \quad (15)$$

4.3.4. F1-Score (F1)

$$\text{F1} = 2 \times \frac{(Precision \times Recall)}{(Precision + Recall)} \quad (16)$$

4.3.5. Mean Absolute Error (MAE)

$$\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|\Lambda_i - \hat{\Lambda}_i\right| \quad (17)$$

Where $\mathcal{N}$ = total number of samples, Λ_i = actual sentiment label, and $\widehat{\Lambda_i}$ = predicted sentiment label.

4.3.6. Correlation (Corr)

$$\text{Corr} = \frac{\sum_{k=1}^{m}(\gamma_k - \hat{\gamma})(\hat{\gamma}_k - \bar{\hat{\gamma}})}{\sqrt{\sum_{k=1}^{m}(\gamma_k - \hat{\gamma})^2} \cdot \sqrt{\sum_{k=1}^{m}(\hat{\gamma}_k - \bar{\hat{\gamma}})^2}} \quad (18)$$

Here, γ_k the actual observation, the predicted observation, and the mean of predicted values are denoted, and m represents the total number of instances.

4.4. Results and Discussion

To evaluate the effectiveness of the proposed model, an extensive experiment is conducted with VersaModal, a versatile system for SA, capable of processing and analyzing multi-modal inputs, including T, I, and V, either unimodal or in combination. The unimodal systems were evaluated individually for the restaurant review dataset, with results compared against the respective baseline models, thereby improving accuracy. The modalities were integrated, and several bimodal and multi-modal combinations were analyzed. Ultimately, a comprehensive evaluation of VersaModal, wherein all three modalities were considered, was performed compared to other multi-modal models, pointing to its higher accuracy and effectiveness in sentiment classification.

4.4.1. Text Unimodal System

A dataset of restaurant reviews was used to evaluate the unimodal text system. Textual data were preprocessed with BERT embeddings before feature extraction to conduct the experiments. $\mathcal{Acc}$, $\boldsymbol{\mathcal{Prec}}$, $\boldsymbol{\mathcal{Rec}}$, and F1 were utilized to measure classification performance.

The comparative results against baseline text sentiment classification models show a significant gain in sentiment classification $\mathcal{Acc}$ due to the modified BiLSTM architecture. It shows that the proposed model has effectively captured sentiment variations in textual reviews.

The accuracy of various feature extraction methodologies (CNN, RNN, GRU, LSTM, BiLSTM, and Modified BiLSTM) is presented in Figure 7 for multiple classifiers: LR, DF, KNN, NB, and SVM. The trend of increasing $\mathcal{Acc}$ with advanced feature extraction is evident, with modified BiLSTM giving the top accuracy. BiLSTM and LSTM improve upon CNN, RNN, and GRU, thus signifying the importance of bidirectional processing in encoding contextual information. The best classifiers by far are DF and SVM, while the worst one is NB, which shows poor $\mathcal{Acc}$ due to its inability to tackle complex feature representation. The results amplify the skills of advanced deep learning approaches in producing better sentiment classification accuracy. When the Text unimodal SA model is compared to other methods, it does better on all of the evaluation metrics shown in Table 4. The traditional ML models, such as NB, LR, DT, and KNN, also have lower ACC, Prec, Rec and F1. These are all strong factors that affect how well they can represent complex linguistic patterns. The GRU, LSTM, and BiLSTM deep-learning models all greatly improve performance. However, the BiLSTM model does better with two-way contextual learning than the simple recurrent models. This wealth of techniques reaches heights of accuracy with a CNN-LSTM hybrid architecture that embeds spatial and sequential features. The proposed text-unimodal model has achieved the best $\mathcal{Acc}$ (91.50%) and F1-score (91.25%), indicating its efficacy in rich text representation and superior sentiment classification compared to existing models.

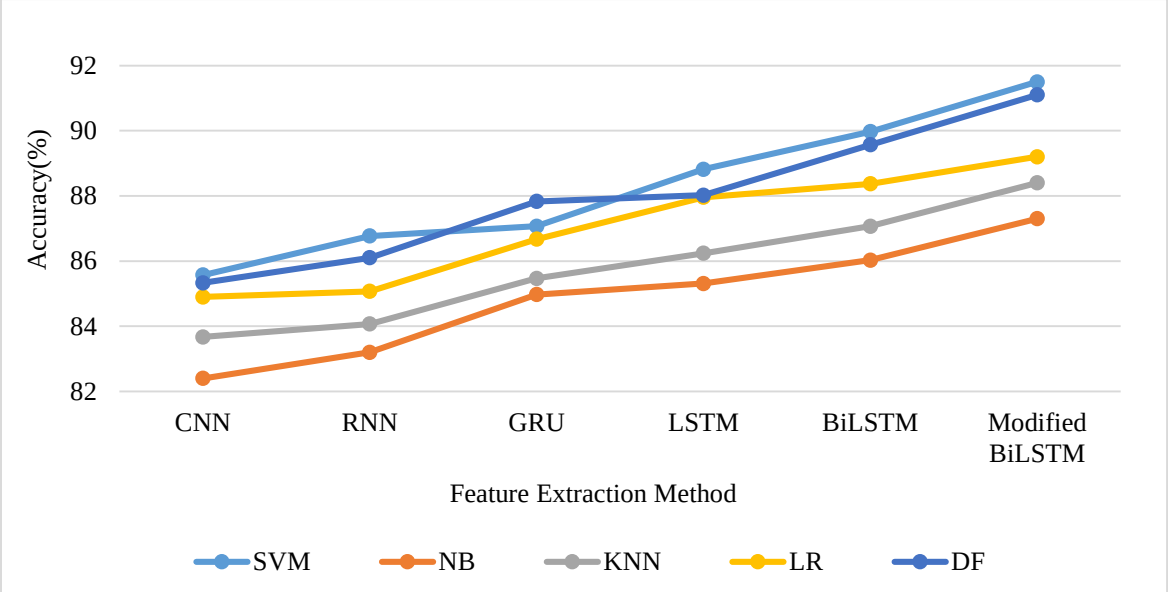

Fig. 7 Classifier accuracy for different methods for feature extraction

Table 4. Comparison with text (unimodal) existing models

Methods	ACC (%)	Prec (%)	Rec (%)	F1 (%)
Naive Bayes [28]	78.00	77.00	76.50	76.75
Logistic Regression[32]	81.00	80.50	80.00	80.25
Decision Tree [29]	75.00	74.50	74.00	74.25
KNN [32]	79.00	78.50	78.00	78.25
AdaBoost[30]	82.00	81.50	81.00	81.25
GRU [31]	88.00	87.50	87.00	87.25
LSTM [31]	89.00	88.50	88.00	88.25
BiLSTM [33]	90.00	89.50	89.00	89.25
CNN-LSTM [34]	91.00	90.50	90.00	90.25
Proposed Text Unimodal Model	**91.50**	**90.65**	**90.22**	**91.25**

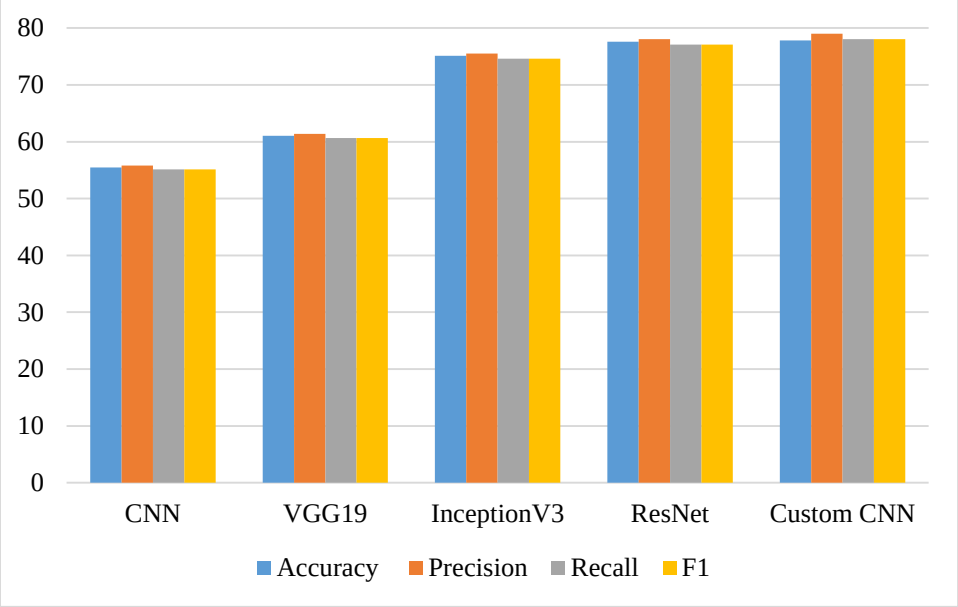

Fig. 8 Comparison of different models for feature extraction across performance metrics

4.4.2. Image Unimodal System

The image unimodal system is tested on a dataset of restaurant review images, dividing them into three categories: positive, neutral, and negative. The images were preprocessed by resizing and normalizing before being fed into a custom CNN for feature extraction. The extracted features were further classified using SVM, RF, KNN, and LR. The evaluation of model performance was conducted using metrics including Prec, Rec, and F1. The experiments demonstrated that the custom CNN outperformed the baseline models, emphasizing its prowess in extracting sentiment-laden visual features from food and people expressing while eating food. A comparative evaluation of different CNN-based models used for image sentiment classification is displayed in Figure 8, employing algorithms from ACC, Prec, Rec and F1. The basic CNN performed the worst on all metrics, showing it could not extract higher-level features. The intermediate layers of VGG19 help improve its performance by extracting features effectively. InceptionV3 performs far better thanks to its architecture, which captures fine details. ResNet further improves these results by solving the vanishing gradient issue through residual connections. The custom CNN performs best on all metrics, proving its ability to extract relevant sentiment features from restaurant review images.

The comparison of image-based SA models in Table 5 illustrates that the proposed image unimodal model outperforms all the existing methods. While item-oriented CNN works moderately, it does not harness its potential in higher feature extraction. CNN-based Inception-V3 achieved better ACC and Rec due to its deeper architecture. However, the proposed model achieved the highest ACC (77.84%) along with improved Prec, Rec and F1 over the other proposed models designed to extract sentiment-related features from restaurant review images. Hence, this improvement establishes why the custom CNN architecture in the proposed model works so well.

Table 5. Comparison with Image-based(unimodal) existing models

Models	ACC (%)	Prec (%)	Rec (%)	F1 (%)
Item-oriented CNN [35]	67.82	67.00	67.32	65.80
CNN-based Inception-v3 [36]	73.00	67.25	68.00	68.22
Proposed Image Unimodal Model	**77.84**	**79.01**	**78.07**	**78.07**

Table 6. Comparison with different classifiers

Classifier	ACC (%)	Prec (%)	Rec (%)	F1 (%)	MAE	Corr
SVM	**80.12**	**79.40**	**80.80**	**79.20**	**0.18**	**0.85**
NB	72.34	71.80	70.30	72.50	0.24	0.71
KNN	74.89	74.50	75.00	73.80	0.22	0.75
LR	76.50	75.90	76.20	75.30	0.21	0.78
DF	73.78	72.90	71.62	74.00	0.23	0.72

Table 7. Comparison with Video-based existing models

Model	ACC (%)	Prec (%)	Rec (%)	F1 (%)
MFN [37]	77.40	75.80	74.60	75.20
MARN[38]	77.10	75.80	74.60	77.80
TFN [41]	76.40	75.30	74.20	74.80
GME-LSTM(A) [40]	75.70	72.20	73.30	73.90
BC-LSTM [39]	74.60	75.33	74.43	74.50
Proposed Video Unimodal Model	**80.12**	**79.40**	**80.80**	**79.20**

4.4.3. Video Unimodal System

The video unimodal system was tested on a dataset of restaurant review videos, where audio, transcript (Text), and visual frames were considered for sentiment classification. The experimental results for the unimodal system proved that including multi-modal features considerably enhanced sentiment classification accuracy rather than unimodal analysis. The results confirm that multiple modalities are important for better sentiment comprehension in video reviews.

Table 6 shows the performance comparison of various classifiers for SA. Out of all the models tested, the SVM is found to be the most useful classifier due to its excellent score of 80.12% for $\mathcal{Acc}$ and 0.85 for Corr. The KNN and LR classifiers performed next best, with ACC scores of 74.89% and 76.50%, respectively. The NB and DF classifiers performed less effectively, with accuracy scores of 72.34% and 73.78%. The lower MAE value for SVM (0.18) further proves its strength, thereby establishing it as a strong candidate in sentiment classification as contrasted with other classifiers.

For comparative analysis of video-based SA models, experiments were performed on the standard dataset CMU-MOSI [42] that the proposed unimodal video model has surpassed the pre-existing models with the highest $\mathcal{Acc}$ (80.12%) and $\mathcal{F}_1$ (79.20%), as shown in Table 7. Classic ones like TFN, MFN, and MARN have lower performance, scoring less in Rec and Prec. The BC-LSTM and GME-LSTM(A) models scale less in ACC and overall performance. The proposed model got great results, which shows that it can read deeper sentiment-related cues from video data and is, therefore, a more reliable way to classify sentiment.

4.4.4. Integrated System Performance

This subsection assesses the functionality of the proposed VersaModal system for MSA evaluated using bimodal (T+I) and trimodal (T+I+V) combinations. The first step is to assess how well the model performs in the bimodal case where the (T+I) inputs are combined. This assessment is now compared with earlier models using both modalities. The final sentiment classification is done using late fusion, where sentiment scores of the individual modalities are averaged. The resulting accuracy for the bimodal system (T+I) is computed by substitution in Equation (5).

Table 8. Comparison Between VersaModal and Existing Models

Modality as Input	Compared Model	Model ACC (%)	VersaModal ACC (%)
Text Only (T)	CNN-LSTM [34]	91.00	**91.50**
Image Only (I)	CNN-based Inception-v3 [36]	73.00	**77.84**
Video Only (V)	BC-LSTM [39]	74.60	**80.12**
T+I	MVAN: [25]	72.00	**84.67**

$$S_{\text{final}}(T+I) = \frac{91.50 + 77.84}{2} = 84.67\%$$

Next is the evaluation of the integrated VersaModal system's performance, which comprises all three modalities (T+I+V). The final outcome of integration leverages the strengths of all unimodal systems through their combined output, which is more inclusive towards the overall sentiment classification. The final accuracy of the system is evaluated by taking an average of each modality's individual accuracy. In this case, substituting the accuracy values obtained for Text (A_t=91.50%), Image (A_i=77.84), and Video (A_v=80.12) helps us get the overall system accuracy, as shown below, by substituting the values in Equation (4).

$$S_{\text{final}}(T+I+V) = \frac{91.50 + 77.84 + 80.12}{3} = 83.15\%$$

While existing studies have explored MSA combining text and images, limited research has incorporated video as an additional modality. To meet this gap, a system that integrates textual, image, and video modalities is proposed to allow more thorough sentiment classification. The system processes each modality individually and fuses the outputs late to achieve the final sentiment score. The evaluation shows that the fusion of these three modalities gives a better understanding of sentiments considering the textual context, visual expressions, and audiovisual cues, which is a more holistic view than any unimodal or multi-modal classification.

The goal of this study is not to establish that an integrated multi-modal system is superior to unimodal systems but to demonstrate its feasibility in real-world SA scenarios. Since user-generated reviews can be expressed through different formats-T, I, or V-the VersaModal system is designed to analyze and integrate the available inputs dynamically. Each unimodal system has been evaluated and validated individually against existing benchmarks, ensuring its effectiveness. Additionally, the system can handle bimodal (T+I, T+V, I+V) and trimodal (T+I+V) combinations, allowing SA based on the available data. While no prior studies have considered T, I, and V as separate inputs for integration, a hypothesis is made that combining these modalities provides a richer sentiment representation as it can gather varied contextual cues in the separate modalities. The VersaModal system can adapt to any input combination, making it flexible and holistic.

4.5. Comparison Between VersaModal and Existing Models

The performance study confirms the effectiveness of the proposed VersaModal system in different input settings, as shown in Table 8. VersaModal has slightly better performance under the condition of text-only SA (91.50% Vs 91.00%), and compared with the CNN-LSTM, it is owing to a BiLSTM, which can capture much deeper context dependencies. When using images only, VersaModal achieves superior results (77.84%) than the Inception-V3-based CNN model (73.00%) by focusing on enhanced spatial and frequency domain feature extraction. For video input, VersaModal achieves an accuracy of 80.12% higher than BC-LSTM (74.60%) as it can process and efficiently fuse features - from audio, transcript, and visual frames separately. VersaModal obtains an accuracy of 84.67% for T+I (Text + Image), which is significantly higher than the MVAN model (72.00%), benefiting from that its high-level modality-specific processing and fusion schemes are learned. The advantages of VersaModal in integrating modalities and obtaining sentiment information are shown by these outcomes.

Thus, the main reason for the improved results of the proposed VersaModal system is its modality-specific feature extraction architecture that allows for specialised processing for each input type-Text, image, and video. For textual inputs, a variant BiLSTM architecture is utilized that also includes a deep embedding layer to enhance semantic embedding by capturing local and sequential dependencies. The raw review images are initially processed using a customized CNN designed to extract sentiment-specific visual aspects, including ambience, presentation, and facial expression, with higher performance than general pretrained models. For video input, the model breaks up the content into three streams: Text (V_T), visual frames (V_I), and audio (V_A). These are passed through their respective pipelines: BiLSTM for Text, CNN for visual frames, and Librosa-based analysis to extract tonal and pitch-based emotional features from audio. This decoupled strategy guarantees that each modality's fine-grained, attention-based weighted features are preserved. Multi-modal Input Integration: A score-level late fusion combines multi-modal inputs (T+I, T+V, I+V, T+I+V). This

fusion is enabled such that the degree of the modality is considered in dynamically learning the influence of other modalities, allowing sentiment scores to be normalized for balanced fusion. Besides yielding higher sentiment classification accuracy, this design strengthens the system with robustness and the capability of coping with real-world multi-modal user-generated reviews.

4.6. Case Study

If a reviewer has provided (T+V+I) as a review, the final sentiment score is determined by integrating the available modalities. The system follows these steps:

- Text sample - "A perfect dining experience! The flavors were rich, and the desserts were exceptional."
- Image sample -

- Video Sample
  a. Transcript:" The ambience was breathtaking, and the staff was extremely polite. This place is a must-visit!"
  b. Visual Expression: Smiling and animated facial expressions.
  c. Audio Tone: Enthusiastic with a rising intonation.

- Score Generated

1. Text score: 0.85 – Positive
2. Image score- 0.78 – Positive
3. Video score: 0.97– Positive
4. Total Integrated (T+V+I) score: 0.86 – Positive

Thus, the final sentiment score is 0.86, which suggests an overall positive sentiment when considering the combined influence of T, I, and V reviews.

5. Conclusion

A unified SA system called VersaModal has been developed to operate in a way that allows T, I, and V inputs to be fed to the system one by one or in combination with each other. Unlike existing applications focusing only on unimodal and bimodal SA, this system combines various modalities to boost sentiment classification within real-life situations. Results from experiments demonstrated that each unimodal system outperformed other benchmarks, leading to gains in accuracy. Furthermore, modalities are sourced dynamically to capture a more complete sentiment.

The novelty of the research is that T, I, and V are handled as separate input sources instead of lifting all modalities from a single video review, as done in earlier research. None of the existing systems directly fuse T, I, and V in this manner. So, the unimodal implementation evaluation proves that the idea is good by showing that adding them to useful unimodal models will make it easier to understand how people feel. Further work will focus on refining fusion strategies, incorporating more datasets, and investigating advanced deep-learning architectures to optimize MSA. With the VersaModal system, MSA is well-positioned for applications in experience-driven analytics, customer feedback evaluations, social media monitoring, and more.

References

[1] Linan Zhu et al., "Multimodal Sentiment Analysis Based on Fusion Methods: A Survey," *Information Fusion*, vol. 95, pp. 306-325, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[2] Ramandeep Kaur, and Sandeep Kautish, "Multimodal Sentiment Analysis: A Survey and Comparison," *International Journal of Service Science, Management, Engineering, and Technology*, vol. 10, no. 2, pp. 1846-1870, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[3] Jitendra V. Tembhurne, and Tausif Diwan, "Sentiment Analysis in Textual, Visual and Multimodal Inputs Using Recurrent Neural Networks," *Multimedia Tools and Applications*, vol. 80, pp. 6871-6910, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[4] Ananya Pandey, and Dinesh Kumar Vishwakarma, "Progress, Achievements, and Challenges in Multimodal Sentiment Analysis Using Deep Learning: A survey," *Applied Soft Computing*, vol. 152, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[5] Keyuan Qiu et al., "A Multimodal Sentiment Analysis Approach Based on a Joint Chained Interactive Attention Mechanism," *Electronics*, vol. 13, no. 10, pp. 1-23, 2024. [CrossRef] [Google Scholar] [Publisher Link]

[6] Ankita Gandhi et al., "Multimodal Sentiment Analysis: A Systematic Review of History, Datasets, Multimodal Fusion Methods, Applications, Challenges and Future Directions," *Information Fusion*, vol. 91, pp. 424-444, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[7] Mehmet Umut Salur, and Ilhan Aydin, "A Novel Hybrid Deep Learning Model for Sentiment Classification," *IEEE Access*, vol. 8, pp. 58080-58093, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[8] Sebastian Kula et al., "Sentiment Analysis for Fake News Detection by Means of Neural Networks," *Proceedings, Part IV 20th International Conference Computational Science*, Amsterdam, Netherlands, pp. 653-666, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[9] Priya Patel, Devkishan Patel, and Chandani Naik, "Sentiment Analysis on Movie Review Using Deep Learning RNN Method," *Conference Proceedings Frontiers in Intelligent Computing: Theory and Applications*, Surathkal, India, vol. 2, pp. 155-163, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[10] Ru Ni, and Huan Cao, "Sentiment Analysis Based on GloVe and LSTM-GRU," *2020 39th Chinese Control Conference (CCC) 2020*, Shenyang, China, pp. 7492-7497, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[11] Kian Long Tan, Chin Poo Lee, and Kian Ming Lim, "RoBERTa-GRU: A Hybrid Deep Learning Model for Enhanced Sentiment Analysis," *Applied Sciences*, vol. 13, no. 6, pp. 1-16, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[12] Alessandro Ortis, Giovanni Maria Farinella, and Sebastiano Battiato, "Survey on Visual Sentiment Analysis," *IET Image Processing*, vol. 14, no. 8, pp. 1440-1456, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[13] Zhuanghui Wu, Min Meng, and Jigang Wu, "Visual Sentiment Prediction with Attribute Augmentation and Multi-Attention Mechanism," *Neural Processing Letters*, vol. 51, pp. 2403-2416, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[14] Luo-yang Xue et al., "NLWSNet: A Weakly Supervised Network for Visual Sentiment Analysis in Mislabeled Web Images," *Frontiers of Information Technology & Electronic Engineering*, vol. 21, pp. 1321-1333, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[15] Jie Chen, Qirong Mao, and Luoyang Xue, "Visual Sentiment Analysis with Active Learning," *IEEE Access*, vol. 8, pp. 185899-185908, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[16] Jing Zhang et al., "Object Semantics Sentiment Correlation Analysis Enhanced Image Sentiment Classification," *Knowledge-Based Systems*, vol. 191, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[17] Di Wang et al., "TETFN: A Text Enhanced Transformer Fusion Network for Multimodal Sentiment Analysis," *Pattern Recognition*, vol. 136, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[18] Bo Yang et al., "Multimodal Sentiment Analysis with Two-Phase Multi-Task Learning," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 2015-2024, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[19] Ting Wu et al., "Video Sentiment Analysis with Bimodal Information-Augmented Multi-Head Attention," *Knowledge-Based Systems*, vol. 235, pp. 1-12, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[20] Zheng Wang, Peng Gao, and Xuening Chu, "Sentiment Analysis from Customer-Generated Online Videos on Product Review Using Topic Modeling and Multi-Attention BLSTM," *Advanced Engineering Informatics*, vol. 52, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[21] Zhipeng He et al., "Advances in Multimodal Emotion Recognition Based on Brain–Computer Interfaces," *Brain Sciences*, vol. 38, no. 10, pp. 1-29, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[22] Sijie Mai et al., "Hybrid Contrastive Learning of Tri-Modal Representation for Multimodal Sentiment Analysis," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 2276-2289, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[23] Ke Zhang et al., "Transfer Correlation between Textual Content to Images for Sentiment Analysis," *IEEE Access*, vol. 8, pp. 35276-35289, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[24] Xu Lu et al., "Semantic Driven Interpretable Deep Multi-Modal Hashing for Large-Scale Multimedia Retrieval," *IEEE Transactions on Multimedia*, vol. 23, pp. 4541-4554, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[25] Xiaocui Yang et al., "Image-Text Multimodal Emotion Classification via Multi-View Attentional Network," *IEEE Transactions on Multimedia*, vol. 23, pp. 4014-4026, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[26] Feiran Huang et al., "Attention-Based Modality-Gated Networks for Image-Text Sentiment Analysis," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 16, no. 3, pp. 1-19, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[27] Wenya Guo et al., "LD-MAN: Layout-Driven Multimodal Attention Network for Online News Sentiment Recognition," *IEEE Transactions on Multimedia*, vol. 23, pp. 1785-1798, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[28] Young Gyo Jung et al., "Enhanced Naive Bayes Classifier for Real-Time Sentiment Analysis with SparkrR," *2016 International Conference on Information and Communication Technology Convergence (ICTC)*, Jeju, Korea (South), pp. 141-146, 2016. [CrossRef] [GoogleScholar] [Publisher Link]

[29] Arman S. Zharmagambetov, and Alexandr A. Pak, "Sentiment Analysis of a Document Using Deep Learning Approach and Decision Trees," *2015 Twelve International Conference on Electronics Computer and Computation (ICECCO)*, Almaty, Kazakhstan, pp. 1-4, 2015. [CrossRef] [Google Scholar] [Publisher Link]

[30] Kian Long Tan et al., "RoBERTa-LSTM: A Hybrid Model for Sentiment Analysis with Transformer and Recurrent Neural Network," *IEEE Access*, vol. 10, pp. 21517-21525, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[31] Sagar Hossen et al., "Hotel Review Analysis for the Prediction of Business Using Deep Learning Approach," *2021 International Conference on Artificial Intelligence and Smart Systems (ICAIS)*, Coimbatore, India, pp. 1489-1494, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[32] Tanushree Dholpuria, Y.K. Rana, and Chetan Agrawal, "A Sentiment Analysis Approach through Deep Learning for a Movie Review," *2018 8th International Conference on Communication Systems and Network Technologies (CSNT)*, Bhopal, India, pp. 173-181, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[33] Apar Garg, and Rohit Kumar Kaliyar, "PSent20: An Effective Political Sentiment Analysis with Deep Learning Using Real-Time Social Media Tweets," *2020 5th IEEE International Conference on Recent Advances and Innovations in Engineering (ICRAIE)*, Jaipur, India, pp. 1-5, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[34] Praphula Kumar Jain, Vijayalakshmi Saravanan, and Rajendra Pamula, "A Hybrid CNN-LSTM: A Deep Learning Approach for Consumer Sentiment Analysis Using Qualitative User-Generated Contents," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 20, no. 5, pp. 1-15, 2021. [CrossRef] [Google Scholar] [Publisher Link]

[35] Quoc-Tuan Truong, and Hady W. Lauw, "Visual Sentiment Analysis for Review Images with Item-Oriented and User-Oriented CNN," *Proceedings of the 25th ACM International Conference on Multimedia*, California USA, pp. 1274-1282, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[36] Gaurav Meena, Krishna Kumar Mohbey, and Sunil Kumar, "Sentiment Analysis on Images Using Convolutional Neural Networks Based Inception-V3 Transfer Learning Approach," *International Journal of Information Management Data Insights*, vol. 3, no. 1, pp. 1-13, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[37] Amir Zadeh et al., "Memory Fusion Network for Multi-view Sequential Learning," *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, LA, USA, pp. 5634-5641, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[38] Amir Zadeh et al., "Multi-Attention Recurrent Network for Human Communication Comprehension," *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, New Orleans, LA, USA, pp. 5642-5649, 2018. [CrossRef] [Google Scholar] [Publisher Link]

[39] Soujanya Poria et al., "Context Dependent Sentiment Analysis in User-Generated Videos," *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada, vol. 1, pp. 873-833, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[40] Minghai Chen et al, "Multimodal Sentiment Analysis with Word-Level Fusion and Reinforcement Learning," *Proceedings of the 19th ACM International Conference on Multimodal Interaction*," Glasgow UK, pp. 163-171, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[41] Amir Zadeh et al., "Tensor Fusion Network for Multimodal Sentiment Analysis," *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, pp. 1103-1114, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[42] Amir Zadeh et al., "MOSI: Multimodal Corpus of Sentiment Intensity and Subjectivity Analysis in Online Opinion Videos," *Arxiv*, pp. 1-10, 2016. [CrossRef] [Google Scholar] [Publisher Link]