Ensemble Pretrained Models for Multimodal Sentiment Analysis using Textual and Video Data Fusion

Zhicheng Liu
The University of Sydney
School of Computer Science
Camperdown, NSW, Australia
zliu5126@uni.sydney.edu.au

Ali Braytee*
University of Technology Sydney
School of Computer Science
Ultimo, NSW, Australia
ali.braytee@uts.edu.au

Ali Anaissi
The University of Sydney
School of Computer Science
Camperdown, NSW, Australia
ali.anaissi@sydney.edu.au

Guifu Zhang
The University of Sydney
School of Computer Science
Camperdown, NSW, Australia
gzha0010@uni.sydney.edu.au

Lingyun Qin
The University of Sydney
School of Computer Science
Camperdown, NSW, Australia
lqin2678@uni.sydney.edu.au

Junaid Akram
The University of Sydney
School of Computer Science
Camperdown, NSW, Australia
jakr7229@uni.sydney.edu.au

ABSTRACT

We introduce an ensemble model approach for multimodal sentiment analysis, focusing on the fusion of textual and video data to enhance the accuracy and depth of emotion interpretation. By integrating three foundational models—IFFSA, BFSA, and TBJE—using advanced ensemble techniques, we achieve a significant improvement in sentiment analysis performance across diverse datasets, including MOSI and MOSEI. Specifically, we propose two novel models—IFFSA and BFSA, which utilise the large language models BERT and GPT-2 to extract the features from text modality and ResNet and VGG for video modality. Our work uniquely contributes to the field by demonstrating the synergistic potential of combining different modal analytical strengths, thereby addressing the intricate challenge of nuanced emotion detection in multimodal contexts. Through comprehensive experiments and an extensive ablation study, we not only validate the superior performance of our ensemble model against current state-of-the-art benchmarks but also reveal critical insights into the model's capability to discern complex emotional states. Our findings underscore the strategic advantage of ensemble methods in multimodal sentiment analysis and set a new precedent for future research in effectively integrating multimodal data sources.

CCS CONCEPTS

• **Computing methodologies** → **Ensemble methods**; • **Information systems** → **Multimedia information systems**.

KEYWORDS

Multimodal sentiment analysis; multimodality learning; data fusion; stacking ensemble; emotion interpretation

*Corresponding author

WWW '24 Companion, May 13–17, 2024, Singapore, Singapore

ACM ISBN 979-8-4007-0172-6/24/05.
https://doi.org/10.1145/3589335.3651971

ACM Reference Format:
Zhicheng Liu, Ali Braytee, Ali Anaissi, Guifu Zhang, Lingyun Qin, and Junaid Akram. 2024. Ensemble Pretrained Models for Multimodal Sentiment Analysis using Textual and Video Data Fusion. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion), May 13–17, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3589335.3651971

1 INTRODUCTION

In the digital age, the rapid proliferation of social media and online platforms has resulted in an unprecedented wealth of multimodal content, blending text, audio, and visual data[20]. This fusion of modalities presents a unique opportunity and challenge for interpreting the nuanced spectrum of human emotions embedded within[24]. The emergence of Multimodal Sentiment Analysis (MSA) marks a significant evolution in artificial intelligence and natural language processing, aiming to decode these complex emotional expressions by transcending the limitations of text-only analysis [2]. MSA leverages the synergistic potential of combining diverse data types to offer a more holistic and accurate portrayal of sentiments, acknowledging the critical role of non-verbal cues—such as tone, inflection, and facial expressions—in emotion conveyance [1].

However, despite technological advancements, MSA faces substantial hurdles, especially in scenarios marked by data scarcity. Integrating emotional signals across modalities poses a formidable challenge, critically impacting the efficacy of sentiment analysis in practical applications where precision and reliability are indispensable [13]. This study introduces an ensemble model approach designed to surmount these obstacles. Utilising a stacking ensemble strategy [28], this method synergistically harnesses the strengths of multimodal models, enhancing the nuanced recognition and accuracy of emotional state analysis [21]. This work addresses the pivotal challenge of fusing multimodal data effectively, striking a balance between model complexity and computational efficiency.

The primary objective of this work is to augment the accuracy and practical applicability of MSA models, offering significant theoretical and practical contributions. By innovating with ensemble model methodologies that merge various data modalities, this study

seeks to resolve pertinent research questions, including the effective implementation of ensemble methods to optimize MSA models, the most efficacious strategies for data modality fusion, and the performance implications of these optimized models in data-limited environments [6]. Adopting a multimodal sentiment analysis framework that utilises a stacking ensemble strategy, this work aims to develop a robust and accurate tool for sentiment analysis. By employing advanced machine learning algorithms to extract and integrate features from textual and visual data, we leverage each data type's unique strengths to forge a comprehensive sentiment understanding. In this work, we proposed two novel models *Intermediate Feature Fusion Sentiment Analysis (IFFSA)* and *Bilinear Fusion Sentiment Analysis (BFSA)*, which considered the backbone models of the ensembling strategy, in addition to a model brought from the literature named as A Transformer-based joint-encoding (TBJE) [4]. The novel IFFSA and BFSA methods utilise the large language models BERT and GPT-2 to extract features from the text modality and ResNet and VGG for the video modality, respectively. This work makes significant contributions to multimodal sentiment analysis (MSA), particularly in enhancing content moderation mechanisms for the betterment of online communities. Our contributions can be detailed as follows:

- Our study enhances Multimodal Sentiment Analysis (MSA) precision, reliability, and accuracy, advancing content moderation for safer online environments.
- We present two novel models, Intermediate Feature Fusion Sentiment Analysis (IFFSA) and Bilinear Fusion Sentiment Analysis (BFSA), to fuse data from multiple modalities. Additionally, we employ a stacking ensemble, combining diverse models for enhanced sentiment analysis performance in real-world multimodal data, advancing its applicability.
- We compare our proposed methods to the state-of-the-art and present an ablation study to evaluate the various ensembling strategies.

2 RELATED WORK

The inception of multimodal sentiment analysis (MSA) marks a transformative development in artificial intelligence and natural language processing, with the ambitious goal of decoding complex emotional expressions by amalgamating textual, audio, and visual data. The interdisciplinary nature of MSA is celebrated for its wide-ranging applications, from revolutionizing customer service to facilitating mental health assessments, by offering profound insights into consumer behavior and preferences via the analysis of varied communicative forms [2]. Central to the evolution of MSA is the strategic focus on ensemble models. These models harness the collective power of diverse algorithms to significantly enhance predictive accuracy while overcoming the limitations of individual models [21]. Incorporating techniques such as bagging, boosting, and stacking, ensemble models provide a comprehensive framework for the intricate exploration of multimodal data, ensuring a detailed and reliable sentiment analysis [12].

The success of MSA research critically depends on the utilisation of extensive datasets that encompass a broad spectrum of text, audio, and visual information, facilitating the intricate learning of emotional expressions across various modalities [8]. Noteworthy datasets such as IEMOCAP, DEAP, CMU-MOSI, CMU-MOSEI, and MELD have become instrumental in offering unparalleled insights into human emotional dynamics through different modalities [3, 10, 18, 30, 33]. Technological advancements, particularly the introduction of BERT and its subsequent iterations like DistilBERT, RoBERTa, and ALBERT, underscore the significant strides made in processing complex linguistic patterns [5, 14, 15, 22]. Concurrently, the rise of Convolutional Neural Networks (CNNs) has propelled MSA forward, demonstrating their efficacy in tasks spanning computer vision and natural language processing [7, 11].

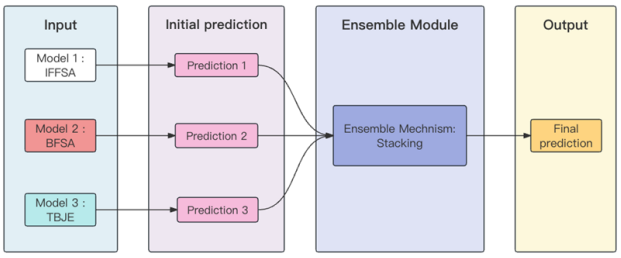


Figure 1: Our proposed stacking ensemble strategy for multimodal sentiment analysis.

Fusion methodologies play a pivotal role in integrating multimodal data, with strategies ranging from early fusion to late fusion, alongside attention mechanisms and cross-modal retrieval. These approaches are critical for achieving a holistic representation of data, effectively amalgamating disparate sources of multimodal information to capture the entirety of emotional expressions [1, 13, 19, 26, 27, 33]. The deployment of ensemble models within MSA has shown tremendous potential, particularly in refining the analysis of sentiments conveyed through multimodal content such as videos [31]. The efficacy of these models in merging features from different modalities highlights their capacity for delivering more precise and robust sentiment predictions, paving the way for advancements in sentiment analysis [17].

3 METHODOLOGY

This study introduces a novel approach to multimodal sentiment analysis by employing a stacking ensemble strategy aimed at significantly enhancing the accuracy and reliability of sentiment predictions. This method integrates the strengths of various models, capitalising on their unique abilities to process and analyse multimodal data. At the core of our methodology is the stacking ensemble strategy, depicted in Figure 1, which orchestrates the integration of multiple models to achieve superior sentiment analysis performance. This strategy facilitates a comprehensive and nuanced understanding of emotional expressions by leveraging diverse feature extraction and fusion techniques.

We have developed two foundational models, IFFSA and BFSA, alongside incorporating a high-accuracy SOTA model, TBJE. Each model targets specific aspects of multimodal data, ensuring a thorough extraction of emotional features from both textual and visual inputs.

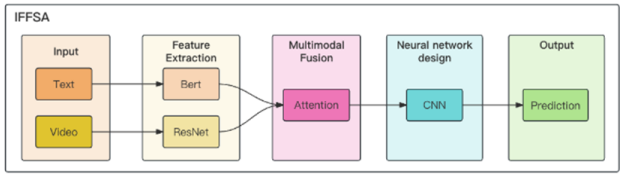


Figure 2: Architecture of the IFFSA model, illustrating the fusion of BERT and ResNet for enhanced multimodal sentiment analysis.

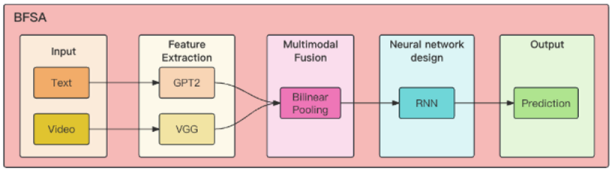


Figure 3: The BFSA model's structure showcases the integration of GPT-2 and VGG for multimodal sentiment analysis.

3.1 IFFSA Model (BERT and ResNet)

We present the IFFSA model that combines the linguistic analytical power of BERT with the visual pattern recognition capabilities of ResNet, as shown in Figure 2.

- **BERT:** Deployed for processing textual data, BERT's transformer-based architecture excels in capturing contextual nuances and semantic relationships within text.
- **ResNet:** Utilized for visual feature extraction, ResNet leverages deep convolutional layers and residual learning to identify complex image patterns effectively.
- **Feature Fusion:** An attention mechanism is applied to refine the integration of textual and visual features, focusing on critical emotional cues.

3.2 BFSA Model (GPT-2 and VGG)

We also propose the BFSA model that leverages the generative capabilities of GPT-2 and the image processing proficiency of VGG, as illustrated in Figure 3.

- **GPT-2:** Tasked with extracting textual context features, GPT-2's transformer-based model is adept at handling generative tasks and processing complex language structures.
- **VGG:** A renowned convolutional neural network for image classification, VGG excels in extracting detailed visual features.
- **Feature Fusion:** Bilinear pooling is employed to amalgamate textual and visual features, enhancing the model's ability to discern the intricate interplay between text and image elements.

3.3 TBJE Model (GloVe and CNN)

The TBJE model synergizes the textual processing power of GloVe with the visual analysis capabilities of CNNs, designed to provide a comprehensive analysis of multimodal sentiment data, as depicted in Figure 4.

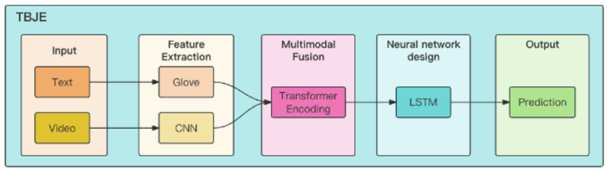


Figure 4: The structural design of the TBJE model illustrates the integration of GloVe and CNN for advanced multimodal sentiment analysis.

- **GloVe:** Utilized for text data processing, GloVe excels in capturing intricate semantic relationships between words, laying the groundwork for in-depth textual analysis.
- **CNN:** Specialized in visual data processing, CNN employs convolutional and pooling layers to extract significant features and patterns from images.
- **Feature Fusion:** The model employs a combination of Transformer encoding and LSTM networks to fuse features effectively, enabling the capture of temporal dynamics in multimodal data.

3.4 Stacking Ensemble Frameowrk

The stacking ensemble strategy is an advanced method designed to enhance the predictive performance of our multimodal sentiment analysis framework. This sophisticated approach leverages the strengths of individual models by combining their predictions in a structured manner to achieve a more accurate and robust final prediction. The process involves the following detailed steps:

(1) **Data Preparation:** We begin by systematically loading the validation and test datasets. These datasets align with the output structure of the three foundational models (IFFSA, BFSA, and TBJE), ensuring a seamless integration process. This preparation phase is critical for maintaining data consistency and compatibility across models.

(2) **Feature Matrix Preparation:** In this pivotal step, we intricately process the predictive outputs from the foundational models. The outputs from IFFSA and BFSA are meticulously averaged to leverage their complementary predictive insights. This averaged output is then thoughtfully merged with the predictions from TBJE, creating a comprehensive feature matrix X. This matrix serves as the enriched input for the subsequent meta-learner training phase, encapsulating the distilled essence of the foundational models' predictive capabilities.

(3) **Target Vector Preparation:** The integrity of our training process is underpinned by the precise construction of the target vector y. Binary true labels, derived from the dataset corresponding to the initial model (IFFSA), form the basis of this vector. This ensures the meta-learner's training is anchored to a reliable and accurate representation of the desired output.

(4) **Meta-Learner Training:** A logistic regression model, renowned for its efficiency and effectiveness in classification tasks, is meticulously selected as the meta-learner. This choice is predicated on its ability to synergize with the feature matrix X, undergoing a rigorous training process with the validation dataset. This phase is instrumental in refining the meta-learner's ability to accurately interpret and integrate the foundational models' predictions, culminating in an enhanced overall prediction accuracy.
(5) **Model Integration:** The culmination of our ensemble strategy is marked by the integration phase, where the trained meta-learner assimilates the predictive insights from IFFSA, BFSA, and TBJE. This process yields a unified composite prediction, embodying the combined analytical strengths of the foundational models. The meta-learner, through its sophisticated understanding of the feature matrix X, effectively harmonizes these insights, delivering a prediction that surpasses the capabilities of the individual models.

This work utilises several advanced computational techniques to optimise multimodal data analysis. The Attention Mechanism component aids the model in focusing on crucial information by assigning varying attention weights. By doing so, it boosts the efficiency of processing multimodal data. Another critical aspect is Bilinear Pooling, which effectively combines features from different modalities, such as text and image, creating a unified feature space. This integration strengthens the model's analytical capabilities. Furthermore, the model utilises Transformer Encoding, leveraging multi-head attention mechanisms to navigate long-range dependencies within complex textual structures. This aspect plays a pivotal role in managing the intricacies of textual data. Finally, the model incorporates an LSTM Network, specialised in analysing time series data. LSTM networks excel in processing and predicting sequential events, enhancing the model's ability to handle the temporal aspects inherent in multimodal data analysis.

4 EXPERIMENTAL EVALUATION

4.1 Datasets

To rigorously assess our multimodal sentiment analysis framework, we utilised two preeminent datasets, each renowned for its comprehensive coverage of multimodal content: CMU-MOSI and CMU-MOSEI. These datasets are paramount in facilitating the experimental validation of sentiment analysis methodologies within artificial intelligence and machine learning spheres.

4.1.1 CMU-MOSI [34]. The CMU-MOSI dataset amalgamates video, audio, and textual data from YouTube movie reviews, offering a rich tapestry for analysing emotional expressions and opinions. Its detailed emotional labels and time-aligned multimodal data make CMU-MOSI a critical resource for advancing research across AI, machine learning, and human-computer interaction domains.

4.1.2 CMU-MOSEI [35]. Building upon the MOSI dataset, CMU-MOSEI extends the research frontier by incorporating a wider array of emotions and sentiments from diverse YouTube content. Its comprehensive annotations and open access significantly contribute to emotion recognition and sentiment analysis.

4.2 Experimental Setup

Our experimental design is meticulously crafted to rigorously validate the effectiveness of our novel multimodal sentiment analysis models, including IFFSA, BFSA, TBJE, and the EnsembleModel. This setup is methodically divided into several critical steps, each pivotal to ensuring the robustness and reliability of our findings:

(1) **Data Loading and Preprocessing:** We initiate our experiment by loading the datasets, which include visual paths, textual paths, and labels, using the h5py library for efficient handling of large-scale, complex data structures. This step encompasses video frame extraction and textual content tokenization, alongside normalization processes to ensure uniform data readiness for analysis. Special attention is given to handling visual and textual modalities to preserve the integrity and richness of multimodal data. Video frame extraction involves selecting key frames from the video content to represent the visual data adequately. This process is crucial for maintaining the semantic richness of the videos while ensuring compatibility with CNN-based feature extraction techniques.
(2) **Feature Extraction:** Advanced machine learning techniques are employed for feature extraction. BERT, renowned for its deep contextual understanding, is utilized for processing textual data, capturing the nuances embedded within the language. Simultaneously, Convolutional Neural Networks (CNNs) are applied to visual data, adept at extracting salient features from images, thereby ensuring a comprehensive analysis of the emotional content. This dual approach ensures that both textual and visual modalities are thoroughly analyzed, capturing the depth and breadth of the emotional cues present.
(3) **Handling Varied Video Lengths:** In our preprocessing steps, we address the challenge of varying video lengths through strategic padding and truncation. This ensures that all video data fed into the model maintains a uniform structure, facilitating efficient learning and analysis. By standardizing video input sizes, we mitigate potential biases and improve the model's ability to generalize across different video lengths and content types.
(4) **Model Training and Validation:** The models undergo meticulous training focusing on parameter optimization, leveraging a train-test split approach to ensure a thorough evaluation. A 70-30 partition is used for creating training and testing sets, enabling robust cross-validation and overfitting prevention strategies such as early stopping to enhance model generalization.
(5) **Emotion Classification Tasks:** Our models are rigorously tested against binary classification tasks, distinguishing between positive and negative emotions and seven-category classification tasks to delve deeper into the spectrum of human emotions. This comprehensive assessment underscores our models' versatility and precision in emotion recognition.
(6) **Model Performance Evaluation:** A comparative analysis with state-of-the-art models is conducted to benchmark our models' performance, utilizing accuracy and F1 scores as primary metrics. This evaluation not only highlights our

models' competitive edge but also offers critical insights into their scalability and adaptability across various datasets.

(7) **In-depth Result Interpretation:** We undertake a detailed examination of the experimental outcomes, providing a nuanced understanding of the models' strengths and pinpointing areas for further refinement. This analysis facilitates a deeper comprehension of the models' capabilities in navigating the complexities of multimodal sentiment analysis.

4.3 Evaluation Metrics and compared methods

A comprehensive benchmarking was conducted to evaluate our models against MSA the recent state-of-arts methods, including LMF [16], TFN [32], MTAG [29], ICCN [23], MulT [25], and MISA [9]. Three evaluation metrics were incorporated, namely Binary Classification Accuracy (ACC-2), Seven-category Classification Accuracy (ACC-7), and F1 Score. ACC-2 measures the models' ability to differentiate between positive and negative emotions. ACC-7 assesses the models' capability to classify emotions across a more nuanced spectrum i.e. 7 classes, and F1 Score balances precision and recall.

5 RESULTS AND DISCUSSION

This section delves into the empirical results from several experiments conducted using our foundational models (IFFSA, BFSA, and TBJE) and the integrated EnsembleModel across the MOSI and MOSEI datasets.

As presented in Table 1, the EnsembleModel showcases promising performance across multiple metrics compared to various state-of-the-art methods such as LMF, TFN, MTAG, ICCN, MulT, and MISA on two benchmark datasets, MOSI and MOSEI. On the MOSI dataset, it achieved an unparalleled ACC-7 of 38.64%, an ACC-2 of 84.1%, and an F1 score of 86.7%, underscoring its superior capability in analysing complex multimodal data streams. Continuing its impressive performance, the EnsembleModel on the MOSEI dataset achieved an ACC-2 of 84.0% and an F1 score of 89.1%, outstripping the existing models and highlighting its adaptability and precision across diverse datasets. These results affirm the efficacy of our EnsembleModel in the realm of multimodal sentiment analysis, providing a robust framework for the accurate interpretation of emotions across varied data types.

Integrating IFFSA, BFSA, and TBJE into the EnsembleModel signifies a leap forward in multimodal sentiment analysis, showcasing superior performance on critical metrics across the MOSI and MOSEI datasets. This advancement underscores the efficacy of amalgamating varied analytical methodologies to forge a sentiment analysis tool that is not only comprehensive but also remarkably precise.

5.1 Performance evaluation of our individual models

As presented in Table 1, IFFSA achieved an ACC-7 of 23.18%, an ACC-2 of 61.4%, and an F1 score of 60.8%, demonstrating its strength in binary classification tasks and the nuanced dissection of sentiments. BFSA presented an improvement with an ACC-7 of 26.59%, an ACC-2 of 62.7%, and an F1 score of 63.1% on MOSI, indicating superior feature extraction capabilities and a refined grasp of complex emotional dynamics. TBJE exceeded the performance of its counterparts on MOSI, TBJE recorded an ACC-7 of 32.17%, an ACC-2 of 75.9%, and an F1 score of 77.6%, showcasing its exceptional analytical proficiency in capturing intricate emotional nuances.

5.2 Binary vs. Seven-Class sentiment classification

A distinct variance between binary (ACC-2) and multi-class (ACC-7) classification accuracies, especially evident in the TBJE and the EnsembleModel, highlights the complex nature of multi-class sentiment classification. This disparity underscores the imperative for advanced algorithms adept at deciphering the complex spectrum of human emotions, furthering the pursuit of models capable of nuanced emotion differentiation.

The evolution from the foundational models to the EnsembleModel encapsulates a deliberate and strategic enhancement in model development. This progression is marked by notable improvements in accuracy, robustness, and the capacity to unravel the complexities of sentiment analysis, embodying a methodical advancement in the field. In summation, the EnsembleModel emerges as a paragon of innovation within multimodal sentiment analysis. By harmoniously merging diverse analytical techniques, it presents an unparalleled approach to accurately interpreting and understanding the intricate layers of emotional data.

5.3 Error Analysis using confusion matrices

Confusion matrices serve as a pivotal tool in classification tasks, offering a detailed visualisation of model performance across various predicted categories. We present an error analysis that provides insights into the accuracy and misclassifications made by our models on MOSI dataset, which is crucial for understanding their strengths and areas for improvement.

In the following sections, we present the error analysis based on confusion matrices of the individual proposed models on the MOSI dataset.

5.3.1 IFFSA analysis. As illustrated in figure 5a, the IFFSA model demonstrates a relatively balanced classification capability across the sentiment spectrum, particularly excelling in identifying neutral sentiments (category 0). However, the model encounters challenges in distinguishing between closely related sentiment categories, notably between slight negativity (-1) and neutrality (0), as well as neutrality (0) and slight positivity (1). The difficulty in accurately classifying extreme sentiments (-3 and 3) suggests a need for enhanced sensitivity to strong emotional expressions.

5.3.2 BFSA analysis. As illustrated in figure 5b, BFSA, like IFFSA, shows commendable accuracy in pinpointing neutral sentiments. Yet, it exhibits a marginal increase in misclassifications between adjacent positive categories (1 and 2), indicating a nuanced challenge in differentiating levels of positivity. Like IFFSA, BFSA struggles with recognising extreme sentiments, underscoring an area ripe for methodological enhancements.

5.3.3 TBJE analysis. As illustrated in figure 5c, TBJE shifts the paradigm by achieving higher accuracy in classifying slightly positive sentiments (category 1). This model demonstrates an improved capacity in parsing positive sentiments, albeit with persistent confusion amongst negative categories. Notably, TBJE marks a slight

Table 1: Multi-modality sentiment classification results on MOSI and MOSEI datasets.

Method	MOSI			MOSEI		
	ACC-7	ACC-2	F1	ACC-7	ACC-2	F1
LMF [16]	33.20	-/82.5	-/82.4	48.00	-/82.0	-/82.1
TFN [32]	34.90	-/80.8	-/80.7	50.20	-/82.5	-/82.1
MTAG [29]	38.90	-/82.3	-/81.6	-	-	-
ICCN [23]	39.00	-/83.0	-/83.0	51.60	-/84.2	-/84.2
MulT [25]	-	81.50/84.10	80.60/83.90	-	-/82.5	82.67/83.97
MISA [9]	-	80.79/82.10	80.77/82.03	-	82.59/84.23	82.53/85.30
Model 1: IFFSA	23.18	61.4	60.8	41.00	78.9	84.3
Model 2: BFSA	26.59	62.7	63.1	40.33	80.7	86.2
Model 3: TBJE	32.17	75.9	77.6	44.39	80.4	85.8
EnsembleModel	**38.64**	**84.1**	**86.7**	**45.29**	**84.0**	**89.1**

advancement in recognizing extreme sentiments compared to its predecessors, albeit still below optimal performance levels.

5.3.4 EnsembleModel analysis. As illustrated in figure 5, the Ensemble Model introduces a distinct shift in performance dynamics, as evidenced by its confusion matrix. This section delves into the nuanced behavior of the Ensemble Model, drawing comparisons with the foundational models to highlight its unique strengths and areas requiring further investigation. Unlike the individual models, the Ensemble Model showcases a markedly different pattern in its confusion matrix. A notable absence of predictions within negative sentiment categories suggests a potential issue in the model's ability or in the processing of results, warranting further examination. In contrast, the model exhibits a significant accuracy in identifying strong positive sentiments (category 4), surpassing the performance of the individual models and indicating a marked improvement in capturing positive emotional expressions. However, the Ensemble Model appears to experience increased confusion among neutral and positive categories (0, 1, 2), a divergence from the clearer distinctions made by the individual models.

5.3.5 Comprehensive analysis. The base models, particularly IFFSA and BFSA, demonstrate proficiency in recognising neutral sentiments, with a well-balanced classification performance across a spectrum of categories. TBJE, in particular, shows enhanced capability in identifying positive sentiments, presenting a more evenly distributed predictive accuracy. A common challenge across all models, including the EnsembleModel, is the difficulty in accurately classifying extreme sentiments, compounded by a tendency to confuse adjacent sentiment categories. The EnsembleModel's specific shortfall in predicting negative categories underscores a critical area for improvement. While the EnsembleModel advances in detecting strong positive sentiments, its inability to classify negative sentiments necessitates a thorough investigation to uncover and rectify the underlying causes. The diminished ability of the EnsembleModel to engage with negative sentiment categories highlights a reduction in the performance spectrum, emphasising the need for methodological enhancements to broaden its predictive capacity. This detailed analysis underscores the EnsembleModel's strategic advancements and delineates specific areas where methodological and algorithmic refinements are imperative.

Table 2: Results on several ensemble strategies on MOSEI

Ensemble Method	F1 Score	ACC-2
Stacking	**0.892**	**0.840**
Bagging	0.891	0.839
Boosting	0.886	0.835
Voting	0.891	0.839

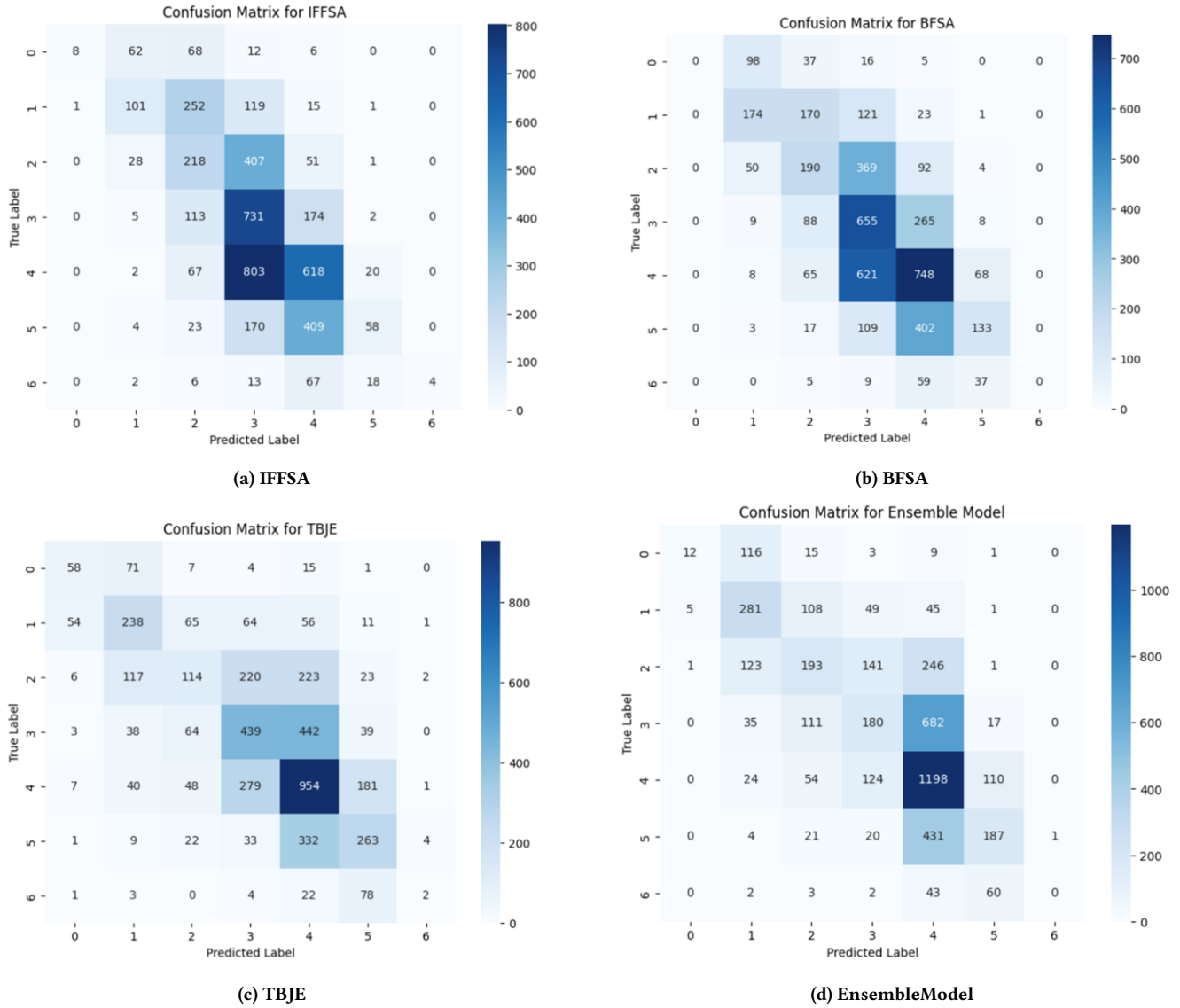
Table 3: 2-class classification (Stacking) on MOSEI

Model	F1 Score	ACC-2
IFFSA	0.843	0.789
BFSA	0.862	0.807
TBJE	0.858	0.804
IFFSA and BFSA	0.886	0.821
IFFSA and TBJE	0.891	0.832
BFSA and TBJE	0.887	0.824
EnsembleModel	**0.891**	**0.840**

5.4 Ablation Study

Ablation studies play a pivotal role in multimodal sentiment analysis by dissecting the contributions of various ensemble learning strategies and model combinations towards task performance on MOSEI dataset(see tables 2, 3 and 4). This analysis delves into the comparative effectiveness of ensemble techniques such as stacking, bagging, boosting, and voting, alongside evaluating the synergy within the trio of foundational models: IFFSA, BFSA, and TBJE.

5.4.1 Ensemble Methodology Performance Insights. As presented in Table 2, the stacking approach emerged as the most effective, achieving F1 and ACC-2 scores of 0.892 and 0.840, respectively. This underscores stacking's capability to adeptly amalgamate different models' predictions, thereby refining classification accuracy significantly. While bagging showcased a performance closely trailing



(a) IFFSA **(b) BFSA**

(c) TBJE **(d) EnsembleModel**

Figure 5: Confusion matrices on MOSI dataset

Table 4: 7-class classification (Stacking) on MOSEI

Model	F1 Score	ACC-7
IFFSA	0.347	0.410
BFSA	0.338	0.403
TBJE	0.425	0.444
IFFSA and BFSA	0.412	0.434
IFFSA and TBJE	0.425	0.434
BFSA and TBJE	0.424	0.438
EnsembleModel	**0.439**	**0.453**

stacking, particularly in F1 scores, its slightly lower ACC-2 hints at a potential inefficiency in multimodal sentiment analysis contexts. Conversely, the voting method, despite paralleling bagging in efficacy, fell short of stacking's benchmark, suggesting that a straightforward majority vote may not always suffice for optimal results. Notably, boosting registered an ACC-2 of 0.835, hinting at a susceptibility to noise and a potentially reduced robustness across diverse data distributions.

5.4.2 Impact of Model Integration. Based on the stacking approach, we observed from the results in Table 3, establishing IFFSA as the baseline, with an ACC-2 of 0.789, the integration narratives unfold intriguingly. BFSA's standalone improvement over IFFSA, with heightened F1 and ACC-2 scores, sets a precedent for the efficacy of feature extraction and sentiment discernment optimizations. The amalgamation of IFFSA and BFSA catalyzes a notable leap in

performance, culminating in F1 and ACC-2 enhancements. This collaboration underscores the complementary strengths of the models, enriching the classification outcomes. Conversely, the IFFSA and TBJE combination did not manifest a similar level of performance elevation, spotlighting the nuanced dynamics of model synergies.

The EnsembleModel's prowess in binary classification tasks (ACC-2 of 0.840) juxtaposed with its diminished performance in seven-class tasks (ACC-7 of 0.453) illustrates the escalating challenge of nuanced emotion differentiation as classification complexity escalates.

6 CONCLUSION

In conclusion, our investigation into multimodal sentiment analysis (MSA) through an ensemble model approach represents a significant stride in artificial intelligence and natural language processing. By adeptly fusing textual and visual data, our method not only transcends traditional sentiment analysis limitations but also showcases a remarkable improvement in emotion interpretation accuracy across varied datasets. Integrating foundational models IFFSA, BFSA, and TBJE, via advanced ensemble techniques underlines the potent synergy achievable through combining diverse modal analytical strengths. Our extensive experimentation and ablation study validate the ensemble model's superior performance against state-of-the-art benchmarks, highlighting its strategic advantage in MSA. These findings not only contribute to the academic discourse but also set a new precedent for future research aimed at integrating multimodal data sources more effectively. Ultimately, this work illuminates the path toward developing more accurate, robust, and practical tools for understanding and interpreting human emotions in digital communication, fostering advancements in content moderation, customer feedback analysis, and beyond.

ACKNOWLEDGEMENT

We acknowledge Zeao Zhang, Yunzhe Wang, and Hao Chen for their invaluable contributions to the project.

REFERENCES

[1] P. K. Atrey, M. A. Hossain, A. El Saddik, and M. S. Kankanhalli. 2010. Multimodal Fusion for Multimedia Analysis: A Survey. *Multimedia Systems* 16, 6 (2010), 345–379.

[2] T. Baltrušaitis, C. Ahuja, and L. P. Morency. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019).

[3] C. Busso, M. Bulut, C. C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan. 2008. IEMOCAP: Interactive Emotional Dyadic Motion Capture Database. *Language Resources and Evaluation* 42, 4 (2008), 335–359.

[4] Jean-Benoit Delbrouck, Noé Tits, Mathilde Brousmiche, and Stéphane Dupont. 2020. A transformer-based joint-encoding for emotion recognition and sentiment analysis. *arXiv preprint arXiv:2006.15955* (2020).

[5] J. Devlin, M. W. Chang, K. Lee, and K. Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *arXiv preprint arXiv:1810.04805* (2018).

[6] T. G. Dietterich. 2000. Ensemble Methods in Machine Learning. In *International Workshop on Multiple Classifier Systems*. Springer, Berlin, Heidelberg, 1–15.

[7] I. Goodfellow, Y. Bengio, and A. Courville. 2016. *Deep Learning*. MIT Press.

[8] Abdelhamid Haouhat, Slimane Bellaouar, Attia Nehar, and Hadda Cherroun. 2023. Towards Arabic Multimodal Dataset for Sentiment Analysis. *arXiv preprint arXiv:2306.06322* (2023).

[9] Devamanyu Hazarika, Roger Zimmermann, and Soujanya Poria. 2020. Misa: Modality-invariant and-specific representations for multimodal sentiment analysis. In *Proceedings of the 28th ACM international conference on multimedia*. 1122–1131.

[10] S. Koelstra, C. Muhl, M. Soleymani, Jong-Seok Lee, A. Yazdani, T. Ebrahimi, T. Pun, A. Nijholt, and I. Patras. 2011. DEAP: A Database for Emotion Analysis; Using Physiological Signals. *IEEE Transactions on Affective Computing* 3, 1 (2011), 18–31.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton. 2012. Imagenet Classification with Deep Convolutional Neural Networks. *Advances in Neural Information Processing Systems* 25 (2012), 1097–1105.

[12] L. I. Kuncheva and C. J. Whitaker. 2003. Measures of Diversity in Classifier Ensembles and Their Relationship with the Ensemble Accuracy. *Machine Learning* 51, 2 (2003), 181–207.

[13] D. Lahat, T. Adali, and C. Jutten. 2015. Multimodal Data Fusion: An Overview of Methods, Challenges, and Prospects. *Proc. IEEE* 103, 9 (2015), 1449–1477.

[14] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. 2019. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. arXiv preprint arXiv:1909.11942.

[15] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, and V. Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv preprint arXiv:1907.11692.

[16] Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2018. Efficient low-rank multimodal fusion with modality-specific factors. *arXiv preprint arXiv:1806.00064* (2018).

[17] S. Poria, E. Cambria, and A. Gelbukh. 2017. Deep Convolutional Neural Networks Text-based Emotion Recognition. *IEEE Signal Processing Letters* 24, 4 (2017), 523–527.

[18] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea. 2019. MELD: A Multimodal Multi-party Dataset for Emotion Recognition in Conversations. In *ACL 2019*.

[19] A. Radford, L. Metz, and S. Chintala. 2015. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. arXiv preprint arXiv:1511.06434.

[20] Anwar Ur Rehman, Zobia Rehman, Waqar Ali, Munam Ali Shah, and Muhammad Salman. 2018. Statistical topic modeling for Urdu text articles. In *2018 24th International Conference on Automation and Computing (ICAC)*. IEEE, 1–6.

[21] O. Sagi and L. Rokach. 2018. Ensemble Learning: A Survey. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8, 4 (2018), e1249.

[22] V. Sanh, L. Debut, J. Chaumond, and T. Wolf. 2019. DistilBERT, a Distilled Version of BERT: Smaller, Faster, Cheaper and Lighter. arXiv preprint arXiv:1910.01108.

[23] Zhongkai Sun, Prathusha Sarma, William Sethares, and Yingyu Liang. 2020. Learning relationships between text, audio, and video via deep canonical correlation for multimodal language analysis. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 8992–8999.

[24] Arsalan Tahir. 2018. Lexicon and heuristics based approach for identification of emotion in text. In *2018 International conference on frontiers of information technology (FIT)*. IEEE, 293–297.

[25] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, Vol. 2019. NIH Public Access, 6558.

[26] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, et al. 2017. Attention is All You Need. In *Advances in Neural Information Processing Systems*, Vol. 30.

[27] W. Wang, R. Arora, K. Livescu, and J. Bilmes. 2016. On Deep Multi-view Representation Learning. In *International Conference on Machine Learning*. 1083–1092.

[28] D. H. Wolpert. 1992. Stacked Generalization. *Neural Networks* 5, 2 (1992), 241–259.

[29] Jianing Yang, Yongxin Wang, Ruitao Yi, Yuying Zhu, Azaan Rehman, Amir Zadeh, Soujanya Poria, and Louis-Philippe Morency. 2020. Mtag: Modal-temporal attention graph for unaligned human multimodal language sequences. *arXiv preprint arXiv:2010.11985* (2020).

[30] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L. P. Morency. 2016. Tensor Fusion Network for Multimodal Sentiment Analysis. arXiv preprint arXiv:1707.07250.

[31] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L. P. Morency. 2017. Tensor Fusion Network for Multimodal Sentiment Analysis. In *Emnlp*. 1103–1114.

[32] Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. *arXiv preprint arXiv:1707.07250* (2017).

[33] A. Zadeh, P. P. Liang, S. Poria, E. Cambria, and L. P. Morency. 2018. Multimodal Language Analysis in the Wild: CMU-MOSEI Dataset and Interpretable Dynamic Fusion Graph. In *ACL 2018*.

[34] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems* 31, 6 (2016), 82–88.

[35] AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2236–2246.