A Novel Dual-Pipeline based Attention Mechanism for Multimodal Social Sentiment Analysis

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

Andy Shueh-Chih Yang*
The University of Sydney
School of Computer Science
Camperdown, NSW, Australia
ayan4778@uni.sydney.edu.au

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

Kunal Chaturvedi
University of Technology Sydney
School of Computer Science
Ultimo, NSW, Australia
kunal.chaturvedi@uts.edu.au

Mukesh Prasad
University of Technology Sydney
School of Computer Science
Ultimo, NSW, Australia
mukesh.prasad@uts.edu.au

ABSTRACT

Traditionally, sentiment analysis methods rely solely on text or image data. However, most user-generated social media content includes both textual and image content. In this study, we propose a novel Dual-Pipeline based Attentional method that uses different modalities of data, including text and images, to analyse and interpret emotions and sentiments expressed in tweets. Our proposed method simultaneously extracts meaningful local and global contextual features from multiple modalities. Local fusion layers within each pipeline combine modality-specific features using an attention mechanism to enrich the joint multimodal representation. A global fusion layer consolidates the collective sentiment representation by seamlessly intermixing the outputs of both pipelines. We evaluate our proposed method using performance metrics such as accuracy and F1-score. Through extensive experimentation on the MVSA dataset, our method demonstrates superior performance compared to state-of-the-art techniques in identifying the sentiment conveyed in social media data.

CCS CONCEPTS

• **Computing methodologies → Neural networks**; • **Information systems → Multimedia information systems**.

KEYWORDS

Multimodal sentiment analysis; attention mechanism; deep learning; feature fusion

ACM Reference Format:
Ali Braytee, Andy Shueh-Chih Yang*, Ali Anaissi, Kunal Chaturvedi, and Mukesh Prasad. 2024. A Novel Dual-Pipeline based Attention Mechanism for Multimodal Social Sentiment Analysis. In *Companion Proceedings of the ACM Web Conference 2024 (WWW '24 Companion), May 13–17, 2024, Singapore, Singapore.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3589335.3651967

1 INTRODUCTION

With the rapid advancement of artificial intelligence, sentiment analysis has received a lot of attention among researchers, driven by the increasing recognition of its potential applications. Sentiment analysis finds applications in various fields, including emotion recognition, customer feedback analysis, social media sentiment analysis, mental health monitoring, customer service enhancement, and human-computer interaction [6]. Most of the current techniques in sentiment analysis rely solely on textual information. However, the rise of social media has opened up new possibilities for multimodal sentiment analysis (MSA). Multimedia data from social media can provide supplementary information streams, enhancing and extending sentiment analysis beyond text-based methods.

Multimodal sentiment analysis uses text, visual, audio, and other modalities to extract meaningful insights for improved performance. It enhances the precision of sentiment analysis, making it applicable across a broad spectrum of use cases [4]. This approach proves invaluable in comprehending human behavior, refining products and services, and addressing real-world challenges. Machine learning plays a key role in multimodal sentiment analysis by facilitating data fusion from different sources to understand and categorise emotions and sentiments [21]. The earlier stage of sentiment analysis focused on single modality text [22]. Still, in recent years, computer vision and natural language processing have been integrated into neural networks, opening the era for converging multimodal sentiment analysis [22]. Several methods have been proposed to learn features from multiple modalities [5]. However, several challenges persist. One notable challenge is the difficulty in recognising diverse patterns within each modality, including text, visuals, and audio. Existing approaches may fail to capture these intricate patterns effectively, hindering their ability to provide comprehensive sentiment analysis. For instance, SentiBank + SentiStrength and CNN-Multi, encounter difficulties extracting meaningful features from multimodal data, leading to suboptimal performance in sentiment analysis tasks [19]. DNN-LR, lacks the

[*] Both authors contributed equally to this research.

WWW '24 Companion, May 13–17, 2024, Singapore, Singapore
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0172-6/24/05
https://doi.org/10.1145/3589335.3651967

versatility and adaptability needed to accommodate diverse models for each modality [8]. Also, the challenge of generalisation arises from struggles in identifying local and global contextual information across different modalities [11, 16].

To this end, we propose a Dual-Pipeline based Attentional method incorporating attention mechanism into each modality pipeline to provide enhanced representations as input to fusion. Each modality is structured with dual pipelines by creating two parallel neural network models. Subsequently, we implement a content-based attention mechanism on the text pipeline, which combines the soft attention mechanism with an extended location-based function. This attention mechanism enables each neural network model to focus on the most relevant information from text modality. For the image pipeline, we implement a spatial attention mechanism. Our approach is designed with a high degree of flexibility, allowing us to tailor our model selection to the unique characteristics of each modality. While originating from diverse sources, multimodal signals possess shared intentions and objectives expressed by the user. The representations from both textual and visual modality networks comprehensively interpret the multimodal data, and their fusion is used for improved sentiment analysis. The fusion mechanism combines modality-specific features to enrich the joint multimodal representation.

The novel contributions of this study can be summarised as:

- Develop a novel Dual-Pipeline based Attentional method that captures local and global features from distinct modalities through a multimodal deep-learning pipeline integrated with an attention mechanism.
- Conduct comprehensive experiments on single and multimodal data to demonstrate the method's proficiency in extracting features from multimodal sources, enhancing sentiment analysis performance.
- Create a detailed ablation study to choose the appropriate models.

2 RELATED WORK

Sentiment analysis, a domain dedicated to extracting sentiments and opinions about a subject [15], has traditionally concentrated on textual content. Nonetheless, due to technological progress, sentiment analysis has broadened its scope to encompass audio, images, and videos, giving rise to multimodal sentiment analysis.

2.1 Machine learning for MSA

In the field of multimodal machine learning, which aims to construct models capable of processing and relating information from multiple modalities such as natural language, visual signals, and vocal signals, (Amir et al., 2017) propose a taxonomy of five core technical challenges: representation, translation, alignment, fusion, and co-learning. Addressing these challenges is essential for advancing the field. The growing importance of Multimodal Sentiment Analysis in natural language processing was discussed in (Ankita et al., 2022). Multimodal Sentiment Analysis utilises machine learning and deep learning to analyse user sentiment toward products or services using multiple modalities, including videos. The article examines various Multimodal fusion architectures and their relative strengths and limitations. It also proposes several interdisciplinary applications and future research directions. Several representative works also demonstrate the concept of multimodal models. Beginning with a basic multimodal approach [7] that separates text from images using an optical character recogniser and then aggregates the independently processed image and text sentiment scores. The MultiSentiNet model employs a deep semantic network to extract deep semantic features from images using salient detectors and a visual feature-guided attention LSTM model to extract important words for sentiment analysis [10]. This model was tested on two public sentiment datasets and demonstrated high correlations with human sentiments, highlighting the importance of considering both visual and textual content in sentiment analysis. Another valuable contribution to this field is the Contrastive Learning and Multi-Layer Fusion (CLMLF) model. This method aligns and fuses token-level features of text and images. It also includes two contrastive learning tasks to help the model learn common features related to sentiment in multimodal data.

2.2 MSA methods based attention mechanism

Multimodal sentiment analysis leverages attention mechanisms to effectively capture emotional cues from diverse data sources such as text, images, videos, audio, acoustics, etc. The attention mechanism allows the model to focus on the most relevant information within each modality while considering their interactions, enhancing its ability to decipher complex emotions. A paper proposed a method for extracting sentiment characteristics using a multi-head attention mechanism from visual, audio, and text data [17]. Another paper introduces an innovative approach and model tailored for handling emotions through video content, focusing on addressing the temporal delay and hysteresis features inherent in multimodal data over time [14]. A paper presents deep generalised canonical correlation analysis with an attention mechanism as an advanced technique for recognising emotions in multimodal data. The attention mechanism empowers a neural network to learn fusion weights for diverse modalities adaptively, leading to enhanced multimodal fusion and superior performance in emotion recognition [9].

3 OUR METHOD

We propose a Dual-Pipeline based Attentional architecture within each modality separately, which incorporates an attention mechanism to empower the model to learn and extract more refined and nuanced local features, ultimately leading to improved data comprehension. By treating each modality separately, our method can effectively capture and leverage the specific characteristics and inherent patterns within the data, resulting in enhanced feature extraction and modeling capabilities. Our proposed method comprises the following components presented in Fig. 1.

3.1 Dual-Pipeline architecture

Each modality is structured with dual pipelines within the multimodal feature extraction process. This entails that each pipeline within a specific modality receives the same input and processes it in parallel. Treating each modality as an independent entity, achieved either through utilising different neural network models or by fine-tuning the parameters of the same neural network model

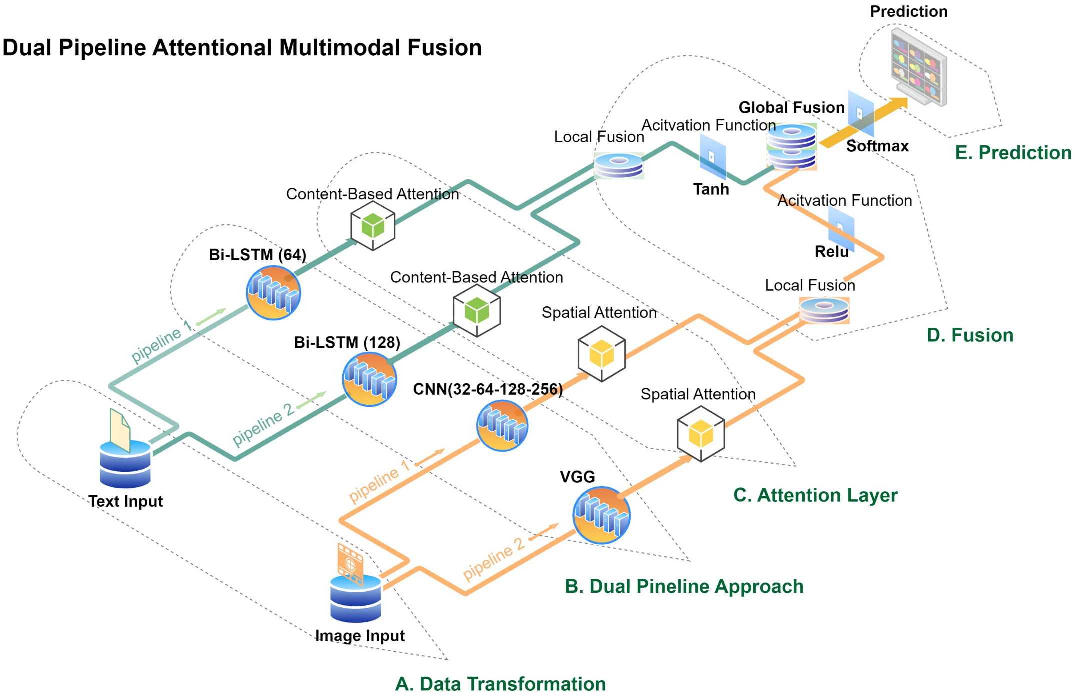

Figure 1: Our proposed Dual-Pipeline based Attentional method

to extract distinct features, allows us to capture modality-specific features and patterns effectively.

3.1.1 Text modality. The proposed method for text classification uses a two-stream structure. Given that a single bidirectional LSTM (BiLSTM) stream cannot capture the hierarchy of features, another BiLSTM stream is added. BiLSTM, an improved variant of LSTM, is commonly used for sequence modelling tasks such as natural language processing. The key difference between BiLSTM and LSTM lies in their ability to capture information from both past and future contexts. BiLSTM addresses the limitation of LSTM by enabling information processing in both forward and backward directions.

The proposed two-stream BiLSTM module takes a text input with a shape of 100 time steps. Specifically, the two BiLSTM layers use different configurations: the first BiLSTM layer has 64 units, and the second BiLSTM layer has 128 units. Both layers are preceded by an embedding layer with 10,000 input and 100 output dimensions. Dropout layers with rates of 0.25 and 0.5 are applied after the embedding layers to mitigate overfitting, respectively. Additionally, a content attention mechanism is employed after each BiLSTM layer. After processing, the outputs are concatenated, followed by a dense layer.

This innovative approach offers the dual advantage of simultaneous data processing through independent pipelines. By embracing separate BiLSTM models, the architecture optimally harnesses the potential of each pipeline, culminating in a cohesive framework designed to unveil rich insights from complex data streams.

3.1.2 Image modality. The proposed image modality method employs a multi-layer Convolutional Neural Network (CNN) in conjunction with the VGG16 network. The multi-layer CNN architecture consists of four convolutional layers followed by max-pooling layers, whereas VGG16 consists of 16 layers, including 13 convolutional layers and three fully connected layers. VGGNet, a CNN architecture, is a prominent example in this category. Despite its relative simplicity, VGGNet demonstrates remarkable effectiveness in image recognition tasks. Notably, VGGNet is also a popular architecture for feature extraction. After thorough testing and comparison, we discovered that combining different state-of-the-art image classifiers in a Dual-Pipeline configuration could lead to overfitting issues. We found that employing a simpler CNN model and a state-of-the-art model yielded the best results. By integrating a simpler CNN model, we can effectively capture essential visual features while maintaining a balanced level of complexity.

3.2 Attention mechanism

3.2.1 Content-based attention. In the proposed method, we integrate two parallel BiLSTMs within their respective pipelines and incorporate content-based attention. The content-based attention mechanism combines the soft attention mechanism [1] with an extended location-based function [12]. This attention mechanism allows each neural network model to focus on the most relevant information from each modality while considering their interactions. The model accepts a text input of length 100. It then proceeds through two distinct BiLSTM layers, each configured differently. After each BiLSTM layer, the output is passed through a dense layer with a specified activation function (in this instance, Tanh).

The attention mechanism is employed for both BiLSTM outputs, allowing the model to focus on pertinent information. The score function of the soft attention mechanism approach is defined in Eq. 1.

$$\text{score}(s_t, h_i) = v_\alpha^T \tanh(W_\alpha [s_t; h_i]) \tag{1}$$

Where s_t is the hidden state of the decoder at time-step t, and h_i is a collection of these hidden states; v_α and W_α are the weight matrices to be learned by the alignment model.
The global attention mechanism is extended by introducing three additional score functions: general, location-based, and dot-product. The location-based function is described in Eq. 2.

$$\alpha_{t,i} = \text{softmax}(W_a S_t) \tag{2}$$

To mitigate overfitting, dropout layers are thoughtfully integrated. Subsequently, the outputs from the two BiLSTM layers are concatenated, followed by applying a final dense layer with a Tanh activation function. This dual content-based attention-BiLSTM configuration facilitates the extraction of diverse features, enhancing the model's capacity to decipher intricate nuances within the data.

3.2.2 Spatial attention. We implement Spatial Attention, as shown in Fig. 2, using a convolutional layer with a sigmoid activation function to enable the model to focus on relevant spatial locations and enhance the representation of important features. The sigmoid activation function ensures that the resulting attention weights are within the range of [0, 1], representing the importance or relevance of each spatial location. The Sigmoid function is expressed in Eq. 3.

$$\text{Sigmoid } S(x) = \frac{1}{1 + e^{-x}} \tag{3}$$

3.3 Fusion

3.3.1 Local fusion. At the local level, the extracted features from the BiLSTM (text modality) are locally fused using a concatenate layer followed by a dense layer with a hyperbolic tangent (tanh) activation function. Similarly, we merge representations from both the branches, i.e., Multi-Layer CNN and VGG16 (image modality) using concatenation followed by dense and dropout layers, along with regularization techniques. This ensures the model's capacity to learn complex patterns while avoiding overfitting. The local fusion step ensures that the model can effectively combine the modality-specific information captured by each component.

3.3.2 Global fusion. At the global level, the combined features from both modalities are further fused to create an overall joint representation. This fusion allows the model to leverage the complementary information from the extracted features in text and image data, enabling a more comprehensive understanding of the sentiment expressed in the tweets. Next, we use a softmax prediction layer to classify polarity (positive, neutral, and negative) based on the input data. The input data can be text, image or multi-view (image-text pair).

3.4 Prediction

The joint representation obtained from the global fusion followed by a Softmax activation function is used as input to a prediction model. The Softmax function allows us to interpret the output of the neural network as probabilities, indicating the likelihood of each class given the input data. The Softmax formula is described as follows Let Z be the input vector to the Softmax function, consisting of elements $Z_0, Z_1, \ldots, Z_K$.

The Softmax function is defined as:

$$\text{Softmax}(Z_i) = \frac{e^{Z_i}}{\sum_{j=1}^{K} e^{Z_j}} \tag{4}$$

where e^{Z_i} is the exponential function applied to each element of the input vector, and K is the number of classes in the multi-class classifier.

4 EXPERIMENTS

4.1 Datasets

The Multi-View Sentiment Analysis dataset (MVSA) [13] consists of sets of image-text pairs, each manually annotated and collected from the Twitter platform. MVSA is a popular dataset frequently used in multimodal fusion research. It focuses on analyzing data from multiple modalities to gain a comprehensive understanding of complex systems. In this paper, we use MVSA-Multiple and MVSA-Single datasets. The MVSA-Single contains 4,869 text-image pairs, whereas the MVSA-Multiple dataset contains 19,600 text-image pairs from Twitter. In the former dataset, each data instance is assigned a single label for each modality, labelled by one annotator. In the latter dataset, each data instance is labelled by three annotators.

4.2 Data preprocesssing

Notably, for the MVSA-Multiple dataset, we consider labels valid only if at least two of the three annotators agree on the exact label[19]. Furthermore, for both datasets, we consider the data instance invalid if the text and image labels have opposing sentiment labels. The tweets containing such inconsistent labels are removed to ensure high-quality data.

In the data preprocessing phase, several steps are taken to standardise and enhance the quality of the text. First, all characters are converted to lowercase to ensure consistency and minimise discrepancies. URLs are systematically identified and removed to eliminate unnecessary noise in the data. User handles, denoted by '@users', are eliminated to reduce irrelevant information. Tags are deleted to streamline the content, emphasising key texts. Contractions are expanded to their full form for clarity and comprehensive understanding. Punctuation marks are removed to prevent interference with the analysis. The data is then tokenised, breaking it down into individual words or 'tokens' for easy manipulation and analysis. Lastly, stemming is applied using a stemmer to reduce words to their root forms, thereby improving the processing efficiency.

For image data, we employ specific image preprocessing techniques, including resizing the images to a standardised dimension, normalising the pixel values to a common range, and deleting the corrupted images. These preprocessing steps help to standardise the images and enhance the model's ability to extract relevant features. This enables us to extract meaningful insights and patterns from the data during the modelling and fusion stages.

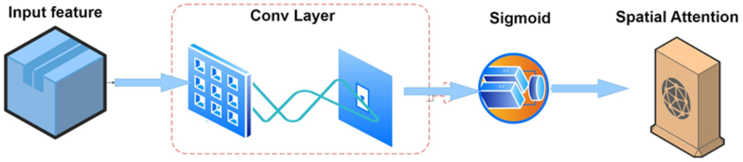

Figure 2: The spatial attention mechanism for the image modality

4.3 Experiment settings

The MVSA datasets were divided into three sets: the training set (80%), the validation set (10%), and the test set (10%). The hyper-parameters were carefully selected to govern the training process, with a significant influence on the model's convergence and ultimate performance. The Adam optimiser, configured with a learning rate 1e-3, was utilised to facilitate the model's weight updates during training. We employed the categorical cross-entropy loss function to guide the optimisation process. Various evaluation metrics, including the F1-score, assess model performance, especially in situations with class imbalance or varying class sizes.

5 RESULTS

Method	**MVSA-Single**		**MVSA-Multiple**	
	Acc	**F1**	**Acc**	**F1**
SentiBank (image only) [2]	45.22	43.80	55.02	51.15
SentiStrength (text only) [2]	49.86	48.45	50.57	55.36
SentiBank + Strength [2]	52.05	50.08	65.62	55.36
CNN-Multi [3]	61.20	58.37	66.39	64.19
DNN-LR [20]	**61.42**	61.03	67.86	66.33
HSAN [18]	-	**66.90**	-	67.76
Our method	57.29	56.76	**73.18**	**69.76**

Table 1: Classification Results (%) of our proposed method compared with the state-of-the-art.

In this section, we present a comparison of our proposed method with several state-of-the-art approaches on the MVSA-Single and MVSA-Multiple datasets. As shown in Table 1, the proposed approach is compared against six different methods, including SentiBank&Strength [2], CNN-Multi [3], DNN-LR [20], and HSAN [18]. Although our results closely align with the top-performing DNN-LR in the MVSA single dataset, our proposed method demonstrates superior performance on the MVSA-Multiple dataset, securing the highest accuracy of 73.18 and F1 score of 69.76. Our proposed Dual-Pipeline indicates its robustness in handling multiple sentiment analysis, which is often considered more challenging due to the complexity of the data. The results indicate that our proposed method, with its unique combination of features and classifiers, excels particularly in scenarios where multiple sentiments are expressed, making it a promising approach for complex sentiment analysis tasks.

5.1 Evaluating extracted feature robustness in the presence of class imbalance

We further investigate the robustness of our extracted features at the class imbalance. The results presented in Table 2 indicate a noteworthy impact of class imbalance on the performance metrics. While the overall accuracy stands at 0.7153, suggesting a reasonable level of correctness in predictions, the macro-average values for precision, recall, and F1 score are considerably lower at 0.4822, 0.4277, and 0.4408, respectively. This discrepancy in the macro-average values highlights the challenge posed by the class imbalance, as it affects the model's ability to consider and evaluate all classes equally. On the other hand, the weighted average values, accounting for the class distribution, show improvements with precision at 0.6788, recall at 0.7318, and F1 score at 0.6976. This suggests that the model's performance is more favorable when considering the class imbalance, but the disparities between macro and weighted averages underscore the importance of addressing and mitigating the effects of class imbalance in the feature extraction process for a more robust and balanced model evaluation.

Class	Precision	Recall	F1-Score
(Neutral)	0.2263	0.1220	0.1586
(Positive)	0.7981	0.9144	0.8523
(Negative)	0.4222	0.2468	0.3115
Accuracy			0.7153
Macro Avg	0.4822	0.4277	0.4408
Weighted Avg	0.6788	0.7318	0.6976

Table 2: The detailed classification results of our proposed method on MVSA-Multiple dataset

6 ABLATION STUDY

We tested many supervised models from text and image modalities to find the appropriate ones for our proposed Dual-Pipeline architecture, including BiLSTM, VGG16, CNN, RCNN, DenseNet, and many more. In the following sections, we will provide an ablation study to justify our selection of BiLSTM+BiLSTM and CNN+VGG16 for our proposed method.

6.1 Model selection based on text modality

This section evaluates different text classifiers on the text data only. As shown in Table 3, it can be inferred that LSTM, BiLSTM, RCNN, GRU, CNN and SimpleRNN received the F1-score of 57.91%, 57.49%, 54.00%, 54.83%, 54.83%, and 57.91% on MVSA single text modality whereas F1-score of 63.35%, 63.21%, 62.57%, 63.43%, 63.07%, and 62.26% is achieved on MVSA-Multiple text modality. Based on the F1-score, BiLSTM is the best choice for both MVSA-Single and MVSA-Multiple datasets.

Text Classifier	**Single**		**Multiple**	
	Acc	**F1**	**Acc**	**F1**
LSTM	57.91	58.00	65.31	63.35
BiLSTM	57.49	58.35	64.13	63.21
RCNN	54.00	55.08	66.29	62.57
GRU	54.83	52.41	67.47	63.43
CNN	54.83	55.60	68.76	63.07
SimpleRNN	57.91	58.43	62.29	62.26

Table 3: Classification results (%) of various methods using text modality only.

Image Classifier	**Single**		**Multiple**	
	Acc	**F1**	**Acc**	**F1**
VGG16	58.93	53.00	68.92	56.00
Resnet50	55.24	41.00	68.31	58.00
DenseNet	56.47	41.00	68.93	56.00
SimpleDNN	56.47	42.00	68.93	56.00
CNN	56.47	41.00	68.92	56.00
Vision Transformer	57.29	43.00	68.93	56.00
Inception	58.11	48.00	68.86	57.00

Table 4: Classification results (%) of various methods using image modality only.

6.2 Model selection based on image modality

In the image modality, we explored popular deep neural network architectures such as ResNet50, VGG16, DenseNet, SimpleDNN, CNN, Inception, and Vision Transformer (ViT), known for their excellence in extracting visual features and patterns from image data. Table 4 compares the different image classifiers used in this study. By evaluating these architectures' performance using F1 score, we found that VGG16 is the best choice for both MVSA-Single and Multiple datasets.

6.3 Compare Dual-Pipeline to baseline w/wo. attention

This experiment is a foundation for evaluating the effectiveness and performance of the Dual-Pipeline architecture compared to the baseline BiLSTM when utilizing only the text modality. As shown in Table 5, our proposed Dual-Pipeline method outperforms the baseline BiLSTM in both the MVSA-Single and MVSA-Multiple datasets.

6.4 Model selection of Dual-Pipeline against single pipeline methods for text and image

We focused on various popular models used in text and image classifiers, including LSTM, BiLSTM, GRU, CNN, ResNet50, VGG16, and ViT. These models served as baselines for benchmarking. We conducted experiments with various combinations under the same architecture to select the best combination for multi-modality, including our Dual-Pipeline method. As detailed in Table 6, our proposed Dual-Pipeline comprises dual BiLSTM for the text modality and CNN and VGG16 for the image modality. Compared to single-model approaches for each modality, our dual model achieves the highest accuracy and nearly a 10% improvement in F1-score on the MVSA-Single dataset, along with a nearly 3% increase in F1-score on the MVSA-Multiple dataset. These results unequivocally demonstrate significant enhancements in multimodal fusion tasks, including substantial improvements in accuracy and F1-score metrics over single-modality models. This validates the effectiveness of Dual-Pipeline with attention as a powerful strategy for multi-modal fusion, with the potential to advance applications relying on information fusion from multiple modalities.

Model	**Single**		**Multiple**	
	Acc	**F1**	**Acc**	**F1**
(BiLSTM)	56.26	57.08	64.44	62.71
Dual-Pipeline (BiLSTM)	59.34	59.49	65.42	63.62
Dual-Pipeline (BiLSTM + ATT)	59.14	59.36	65.36	63.14

Table 5: Classification results (%) of baseline (BiLSTM) against the Dual-Pipeline method with and without attention (ATT) on text modality only

Classifier		**Single**		**Multiple**	
Text	**Image**	**Acc**	**F1**	**Acc**	**F1**
RCNN	VGG16	54.83	49.20	**75.13**	65.51
BiLSTM	ResNet50	52.77	46.51	74.51	65.20
DenseNet	DenseNet	54.41	42.82	72.92	66.96
GPT2+LSTM	ResNet50	51.45	47.24	74.51	65.45
BERT	ResNet50	53.80	48.59	74.87	66.06
BiLSTM+BiLSTM	**CNN+VGG16**	**57.29**	**56.76**	73.18	**69.76**

Table 6: Classification results (%) of our proposed method with the single pipeline models on MVSA datasets

6.5 Dual-Pipeline model selection using text and image modalities

In Table 7, we compare various dual models' performance. Among the models considered for MVSA datasets, our dual-pipeline method, incorporating BiLSTM, CNN, and VGG16, emerged as the top performer. It achieved the highest F1 score, reaching 56.76% for the MVSA-Single dataset and an impressive 69.76% for the MVSA-Multiple dataset. In addition to accuracy analysis, a BiLSTM architecture is chosen over a standard LSTM due to its ability to capture bidirectional context information, which is particularly valuable for large datasets. The BiLSTM layers are configured with specific numbers of units (e.g., 64 and 128) to balance model complexity and performance. A CNN was chosen as the baseline model for image processing due to its simplicity and effectiveness, especially when combined with a more advanced model like VGG. VGG is known for its deep architecture and superior performance in image recognition tasks.

Classifier		**Single**		**Multiple**	
Text	**Image**	**Acc**	**F1**	**Acc**	**F1**
BiLSTM+GRU	ResNet50+VGG16	54.62	54.81	74.51	68.15
LSTM+LSTM	CNN+VGG	54.41	53.54	74.05	67.36
BiLSTM+BiLSTM	CNN+EfficientNet	53.59	54.89	73.64	67.64
BILSTM+GRU	CNN+ViT	**59.14**	55.24	**74.92**	67.95
BiLSTM+BiLSTM	CNN+ResNet50	58.52	54.36	72.56	68.07
BiLSTM+BiLSTM	**CNN+VGG16**	57.29	**56.76**	73.18	**69.76**

Table 7: Classification results (%) of our proposed method compared with various dual models.

7 CONCLUSION

In conclusion, our study introduces a novel Dual-Pipeline based Attentional method that captures local and global features from distinct modalities for multimodal sentiment analysis. Our method's performance is rigorously evaluated using accuracy and F1-score. Through comprehensive experimentation and analysis, our approach consistently outperforms state-of-the-art techniques in effectively discerning sentiment conveyed in tweets. However, our proposed method is constrained by limited modalities, focusing solely on image and text data. Additionally, the interpretability of decisions made by our method is not available. Moreover, the computational demands of the dual pipeline approach surpass those of a single pipeline, highlighting the need for efficient resource management. To address these shortcomings, potential areas for improvement include integrating advanced models such as BERT, GPT, Vision Transformer (ViT), and EfficientNet to augment performance across modalities. Lastly, more modalities, including video and audio data, should be considered in MSA.

ACKNOWLEDGEMENT

We acknowledge Jiahe Song, Xinran Zhou, Shi Su, Yuqi Sun and Jianfan Chen for their invaluable contributions to the project.

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural Machine Translation by Jointly Learning to Align and Translate. arXiv:1409.0473 [cs.CL]

[2] Damian Borth, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang. 2013. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In *Proceedings of the 21st ACM International Conference on Multimedia* (Barcelona, Spain) *(MM '13)*. Association for Computing Machinery, New York, NY, USA, 223–232. https://doi.org/10.1145/2502081.2502282

[3] Guoyong Cai and Binbin Xia. 2015. Convolutional Neural Networks for Multimedia Sentiment Analysis. In *Natural Language Processing and Chinese Computing*, Juanzi Li, Heng Ji, Dongyan Zhao, and Yansong Feng (Eds.). Springer International Publishing, Cham, 159–167.

[4] Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. 2022. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion* (2022).

[5] Wei Han, Hui Chen, Alexander Gelbukh, Amir Zadeh, Louis-philippe Morency, and Soujanya Poria. 2021. Bi-bimodal modality fusion for correlation-controlled multimodal sentiment analysis. In *Proceedings of the 2021 International Conference on Multimodal Interaction*. 6–15.

[6] Ramandeep Kaur and Sandeep Kautish. 2022. Multimodal sentiment analysis: A survey and comparison. *Research Anthology on Implementing Sentiment Analysis Across Multiple Disciplines* (2022), 1846–1870.

[7] Akshi Kumar and Geetanjali Garg. 2019. Sentiment analysis of multimodal twitter data. *Multimedia Tools and Applications* 78 (2019), 24103–24119.

[8] Songning Lai, Xifeng Hu, Haoxuan Xu, Zhaoxia Ren, and Zhi Liu. 2023. Multimodal sentiment analysis: A survey. *Displays* (2023), 102563.

[9] Yu-Ting Lan, Wei Liu, and Bao-Liang Lu. 2020. Multimodal emotion recognition using deep generalized canonical correlation analysis with an attention mechanism. In *2020 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 1–6.

[10] Zhen Li, Bing Xu, Conghui Zhu, and Tiejun Zhao. 2022. CLMLF: a contrastive learning and multi-layer fusion method for multimodal sentiment detection. *arXiv preprint arXiv:2204.05515* (2022).

[11] Kuan Liu, Yanen Li, Ning Xu, and Prem Natarajan. 2018. Learn to Combine Modalities in Multimodal Deep Learning. arXiv:1805.11730 [stat.ML]

[12] Minh-Thang Luong, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. arXiv:1508.04025 [cs.CL]

[13] Teng Niu, Shiai Zhu, Lei Pang, and Abdulmotaleb El Saddik. 2016. Sentiment analysis on multi-view social data. In *MultiMedia Modeling: 22nd International Conference, MMM 2016, Miami, FL, USA, January 4-6, 2016, Proceedings, Part II 22*. Springer, 15–27.

[14] Qingfu Qi, Liyuan Lin, and Rui Zhang. 2021. Feature extraction network with attention mechanism for data enhancement and recombination fusion for multimodal sentiment analysis. *Information* 12, 9 (2021), 342.

[15] Zaher Salah, Abdel-Rahman F Al-Ghuwairi, Aladdin Baarah, Ahmad Aloqaily, Bar'a Qadoumi, Momen Alhayek, and Bushra Alhijawi. 2019. A systematic review on opinion mining and sentiment analysis in social media. *International Journal of Business Information Systems* 31, 4 (2019), 530–554.

[16] Nan Wu, Stanisław Jastrzębski, Kyunghyun Cho, and Krzysztof J. Geras. 2022. Characterizing and overcoming the greedy nature of learning in multi-modal deep neural networks. arXiv:2202.05306 [cs.LG]

[17] Chen Xi, Guanming Lu, and Jingjie Yan. 2020. Multimodal sentiment analysis based on multi-head attention mechanism. In *Proceedings of the 4th international conference on machine learning and soft computing*. 34–39.

[18] Nan Xu. 2017. Analyzing multimodal public sentiment based on hierarchical semantic attentional network. In *2017 IEEE International Conference on Intelligence and Security Informatics (ISI)*. 152–154. https://doi.org/10.1109/ISI.2017.8004895

[19] Nan Xu and Wenji Mao. 2017. Multisentinet: A deep semantic network for multimodal sentiment analysis. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 2399–2402.

[20] Yuhai Yu, Hongfei Lin, Jiana Meng, and Zhehuan Zhao. 2016. Visual and Textual Sentiment Analysis of a Microblog Using Deep Convolutional Neural Networks. *Algorithms* 9, 2 (2016). https://doi.org/10.3390/a9020041

[21] 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).

[22] Linan Zhu, Zhechao Zhu, Chenwei Zhang, Yifei Xu, and Xiangjie Kong. 2023. Multimodal sentiment analysis based on fusion methods: A survey. *Information Fusion* 95 (2023), 306–325.