Multimodal Sentiment Analysis using Deep Learning Fusion Techniques and Transformers

Muhaimin Bin Habib[1], Md. Ferdous Bin Hafiz[2], Niaz Ashraf Khan[3], Sohrab Hossain[4]*
Department of Computer Science and Engineering, East Delta University, Chattogram, Bangladesh[1, 4]
Department of Computer Science and Engineering, University of Liberal Arts Bangladesh, Dhaka, Bangladesh[2, 3]

***Abstract*—Multimodal sentiment analysis extracts sentiments from multiple modalities like text, images, audio, and videos. Most of the current sentiment classifications are based on single modality which is less effective due to simple architecture. This paper studies multimodal sentiment analysis by combining several deep learning text and image processing models. These fusion techniques are RoBERTa with EfficientNet b3, RoBERTa with ResNet50, and BERT with MobileNetV2. This paper focuses on improving sentiment analysis through the combination of text and image data. The performance of each fusion model is carefully analyzed using accuracy, confusion matrices, and ROC curves. The fusion techniques implemented in this study outperformed the previous benchmark models. Notably, the EfficientNet-b3 and RoBERTa combination achieves the highest accuracy (75%) and F1 score (74.9%). This research contributes to the field of sentiment analysis by showing the potential of combining textual and visual data for more accurate sentiment analysis. This will lay the groundwork for researchers in the future to work on multimodal sentiment analysis.**

***Keywords*—Multimodal sentiment analysis; deep learning; transfer learning; natural language processing; image processing; BERT**

I. Introduction

Sentiment analysis is an important part of natural language processing which determines emotions in text [1]. A decade ago, sentiment analysis was performed using text data only, but now advanced technology and programming languages have allowed researchers to combine text with images, audio and video to generate sentiments [2]. Textual data, which contains words, phrases and sentences, provides necessary information about emotions. In visual data (i.e. images and videos) there are facial expressions, body language, and scenes, which can be indicators about sentiments. In Audio Data, a person's voice tone, pitch and intonation carry emotional information. This variation of information modality states the need for multimodal sentiment analysis where information from multiple modalities like textual and visual components are combined. This has inspired researchers to utilize multiple modalities to understand complex emotions.

Multimodal sentiment analysis is inspired from human communication where people use both words and pictures to grasp feelings. Analyzing only one modality i.e text provides a limited view. Focusing solely on text can miss many underlying emotions. By mixing textual and image data, deeper layers of sentiment can be analyzed using facial expressions, scene context, and color tones. Modern AI models are very powerful, and they can combine features of many modalities as well as can handle large datasets. At present almost everyone posts on various social media platforms and these posts contain text, images and emojis. Determining the sentiment of a post requires all these modalities. In healthcare, analyzing the patient's voice, facial expression can assist to diagnose the patient efficiently. Multimodal sentiment analyis can also be applicable in marketing. Researchers believe that combining data of various modalities can reveal deeper layers of sentiment.

The main goal of this paper is to inspect various combinations of different text and image processing models to determine the best fusion technique for multimodal senitment analysis through rigorous comparative analysis. Three model pairings i.e. RoBERTa with EfficientNet, RoBERTa with ResNet50, and BERT with MobileNetV2 are explore in this study. At first, various data preprocessing techniques are explored. Then, different models are combined and trained. Finally, the results are assessed using various evaluation techniques. This includes examining the effectiveness of the chosen model combinations and the preprocessing techniques employed. The main contribution of the study can be outlined as follows:

- Demonstrating how multimodal fusion techniques enhance sentiment classification accuracy.
- Outlining a comprehensive methodology for data preprocessing, model integration, training, and evaluation.
- Developing a framework that future researchers can use and build upon to advance multimodal sentiment analysis.

The rest of the research is arranged in the following manner: Section II discusses related works to this work. Section III concisely presents the datasets used in this investigation. The proposed methodology and the detailed approach, including the use of deep learning models and preprocessing techniques, are also described in this section. The results are presented in Section IV and the discussions are described in Section V. Finally, the paper concludes in Section VI, summarizing the findings and suggesting directions for future research.

II. Literature Review

A. Evolution and Methodological Innovations

Early sentiment analysis used traditional machine learning techniques to detect sentiments from text. Sentiment analysis was first introduced by Pang et al. [3]. In this paper, sentiments of movie reviews were identified using three machine learning

techniques: Naive Bayes, maximum entropy classification, and support vector machine. Turney [4] extended this research by utilizing an unsupervised learning algorithm for classification based on semantic orientation. Kumar et al. mined customer reviews from amazon and used Naive Bayes, Logistic Regression and SentiWordNet to classify the reviews [5]. The introduction of deep learning has significantly changed the sentiment analysis process, producing better sentiment classification and analysis than traditional machine learning models.

B. The Relationship of NLP and Computer Vision

In recent years, significant development in NLP and computer vision has been driven by deep learning. The invention of transformer models has notably improved text analysis, with prominent examples being BERT [6]. There is also an optimized version of BERT and its optimized version, RoBERTa [7]. A decade ago, Word2Vec was widely used for word embedding using a simple neural network [8]. In the realm of computer vision, several powerful architectures, such as ResNet [9] and EfficientNet [10] have been introduced, significantly enhancing image classification and analysis. These advances in deep learning models have laid the groundwork for more sophisticated approaches to understanding and integrating image and text information.

C. Multimodal Sentiment Analysis

Multimodal sentiment analysis was first introduced in 2011. Morency et al. addressed the growing need to harvest relevant information from the vast amount of multimodal data available online, particularly from social websites [11]. The resarch demonstrated that a joint model integrating visual, audio, and textual features could effectively identify sentiment in web videos. A comprehensive survey of multimodal machine learning is provided in [12], presenting a new taxonomy that goes beyond the typical early and late fusion approaches. The authors introduced a novel deep learning architecture for multimodal sentiment analysis, the Gated Multimodal Embedding Long Short-Term Memory (LSTM) with Temporal Attention (GME-LSTM(A)) model, which performs modality fusion at the word level [13]. This model addresses the challenges of noisy modalities by employing gated multimodal embedding and temporal attention mechanisms, achieving good results on the CMU-MOSI dataset. Furthermore, The Attention-based Multimodal Sentiment Analysis and Emotion Recognition (AMSAER) model was developed in study [14], which proposed the hybrid LXGB Model. This model combines the strengths of LSTM and XGBoost classifiers to capture nuanced emotions from diverse data sources like text, images, and audio. It achieved an exceptional accuracy of 97.18% on its dataset. Huang et al. demonstrated a text-centered fusion network with cross-modal attention (TeFNA), which models unaligned multimodal timing information [15]. TeFNA uses text as the primary modality and maximizes mutual information between modality pairs to preserve task-related emotional information. The fusion of ResNet 50 and RoBERTa was utilized in study [16] for multimodal fake news detection, on the FACTIFY dataset. This study combined OCR information and text using models like Bi-directional LSTM and LightGBM for classification, achieving a weighted average F1 score of 0.7428. Peng et al. [17] introduced the Fine-grained Modal Label-based Multi-Stage Network (FmlMSN), which addresses the challenge of handling various sentiments within a video by using seven sentiment labels. This study proposed a discriminative joint multi-task framework (DJMF) to simultaneously perform sentiment prediction and emotion recognition.

Despite significant advancements in natural language processing (NLP) and image processing, there remains a notable gap in the literature concerning the integration of these two modalities through fusion techniques. While some studies have explored multimodal approaches, they predominantly focus on combining text and image data for tasks such as captioning, visual question answering, or sentiment analysis. Our paper aims to address these gaps by developing and implementing advanced fusion techniques for integrating natural language processing (NLP) and image data.

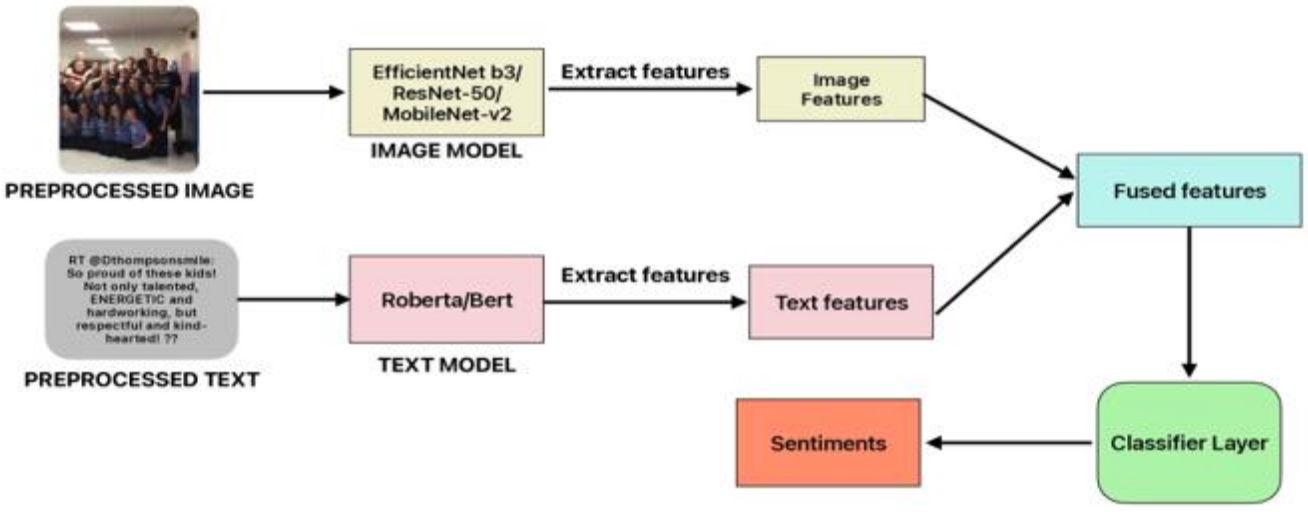

Fig. 1. Block diagram of model.

III. METHODOLOGY

In this study, a comprehensive approach to multimodal sentiment analysis is analyzed by integrating both image and text data using deep learning fusion techniques. The proposed methodology leverages state-of-the-art pre-trained models to extract rich features from both modalities and subsequently combines these features to predict sentiments with high accuracy. Specifically, EfficientNet-B3, ResNet-50, and MobileNet-V2 for image feature extraction, and RoBERTa, and BERT for text feature extraction were utilized. As illustrated in Fig. 1, the preprocessed image and text inputs are fed into their respective models to extract meaningful features. These features are then fused and passed through a classifier layer to generate sentiment predictions. This approach aims to make use of the complementary strengths of visual and textual data, providing a robust framework for sentiment analysis on multimodal datasets. The following sections outline the dataset used, preprocessing steps, and the proposed methodology for feature extraction, fusion, and sentiment classification, along with the techniques used.

A. Dataset

The dataset used in this research is 'MVSA-Single' which was introduced in [18]. A publicly accessible dataset in the field of multimodal sentiment analysis, the MVSA-Single was gathered using Twitter. On the social networking platform Twitter, users can post tweets that include text, photos, hashtags, and other content. Every text-image pair has a unique sentiment label associated with it. The sentiment labels are positive, neutral, and negative. MVSA-Single has 4869 image-text pairs [19]. In case of images there are 2708 positive images, 1223 negative images, and 938 neutral images. In case of texts, there are 1731 positive texts, 1217 negative texts, and 1921 neutral texts.

B. Dataset Preprocessing

All the Images and texts are preprocessed before passing them into the text and image modals. Firstly, the sentiment labels which are initially in text format are converted to numerical labels using dictionary mapping. The positive, neutral, and negative labels were converted to 0,1, and 2 respectively.

Secondly, tokenization is performed on the text using RobertaTokenizer/BertTokenizer depending on the text modal used in the fusion technique. The process of breaking a sentence into smaller pieces (tokens) is called tokenization. These tokens are then converted to numerical identifiers. This process is performed so that each text is converted to a format understood by the machine learning models. The tokenizers add special tokens that are used by the model to understand the structure of the text, such as beginning of sentence and end of sentence markers. To make each text of the same length, tokenizers add padding to tokens. Tokenization is demonstrated in Fig. 2.

The function 'encode_plus' is used to tokenize the text, which also adds special tokens like start or end of sequence markers. This function also pads the sequence to a fixed length with padding tokens if the text is shorter. Attention masks are generated to identify important parts of the sequence. For model compatibility, everything is converted to PyTorch tensors.

In case of images, they are converted to RGB if necessary and then they are resized to a fixed size. Normalization techniques are also applied to images. Images are converted to PyTorch tensors and pixel values are normalized to a specific range for better model performance during training.

These preprocessing techniques ensure the data is in a format suitable for the chosen pre-trained models and the deep learning framework used for training. Train-Test Split has been used in the dataset and 90% is kept for training and 10% is kept for testing.

C. Proposed Methodology

The main goal of this thesis is to use the multimodal dataset and predict the sentiments and get a good accuracy and F1 score using different deep learning fusion techniques. Fig. 1 shows that firstly preprocessed images and text are passed to an image modal and a text modal. This is done to extract features from text and images. The features are fused in the model and then it is classified. Finally, the model generates sentiments as outputs. The entire process can be shown in a few equations.

$$F_T = R(T) \tag{1}$$

$$F_I = E(I) \tag{2}$$

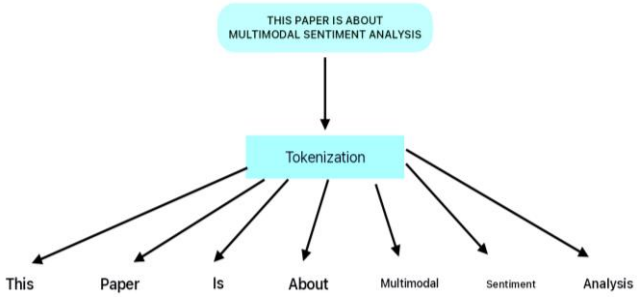

Fig. 2. Tokenization architecture.

In Eq. (1), a text input T is passed into the text model R, which extracts text features F_T. In Eq. (2), an image input I, is passed into the Image model E, which extracts image features F_I. Feature fusion and classification can be represented by a series of equations.

$$F = C(F_T, F_I) \tag{3}$$

$$F' = \sigma(L_I\,(\text{Dropout}(F))) \tag{4}$$

$$S = L_2\,(F') \tag{5}$$

In Eq. (3), the extracted image and text features F_T and F_I are fused (C) into a feature F. In Eq. (4) and Eq. (5), the fused feature F is then passed through linear transformations (L) and activation functions (σ) to produce the final sentiment prediction S. L1 and L2 are layers that convert the feature F to the desired dimensions. The activation function σ is a ReLU that introduces non-linearity.

$$\mathcal{L}(Y, \acute{Y}) = -\sum_i Y_i \log(\acute{Y}) \tag{6}$$

The model is trained on a dataset D using a cross-entropy loss function, $\mathcal{L}$, with the aim to minimize the difference between predicted sentiment labels, $\acute{Y}$, and true labels, Y as shown in Eq. (6). During training, optimisation is performed

using the Adam optimiser. A learning rate scheduler has been used to update the model parameters.

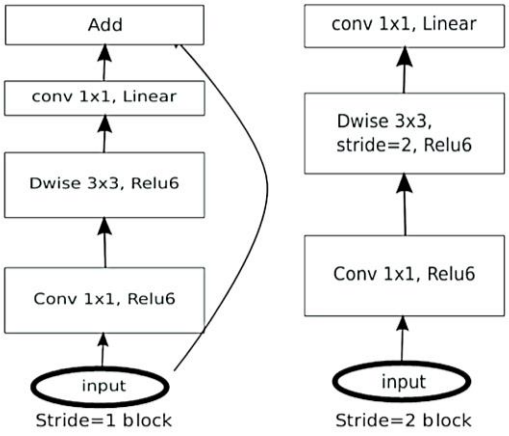

Fig. 3. MobileNet V2 architecture [20].

D. EfficientNet-B3

EfficientNet is a type of convolutional neural network architecture. This architecture was first introduced by a team of researchers at Google AI as a new approach that involves scaling the CNN architectures to achieve high accuracy and efficiency [10]. The EfficientNet-b3 belongs to a family of EfficientNet models. To obtain optimal performance, EfficientNet-B3 uses a new scaling method that scales the dimension of components like depth, width, and resolution. To extract features, it combines pooling layers, activation functions, and convolutional layers. To improve feature representation, it uses squeeze and excitation block technique.

E. MobileNet V2

MobileNetV2 is an improved version of MobileNet. MobileNetV2 was also developed by Google Researchers in 2018. It was specifically designed for mobile and embedded devices. Researchers made this model for image classification & feature extraction tasks [20]. The model is based on an inverted residual structure with shortcut connections in between the thin bottle-neck layers. The intermediate expansion layer filters act as a source of non-linearity using lightweight depth wise convolutions. This model is mainly used for devices which have less resources. Fig. 3 portrays the model architecture.

F. ResNet-50

ResNet-50 and other models of the ResNet family were introduced in 2016 by a group of researchers. ResNet-50 is a deep convolutional neural network architecture. It is known for its depth and effectiveness in image classification tasks. ResNet-50 has 50 layers which help to get high accuracy on image classification tasks [9]. ResNet-50 implements a type of learning called residual learning. This type of learning enables layers to add information to the output of previous layers. Skip connections are used for this task. This helps to solve the vanishing gradient problem. The ResNet-50 architecture is shown in Fig. 4.

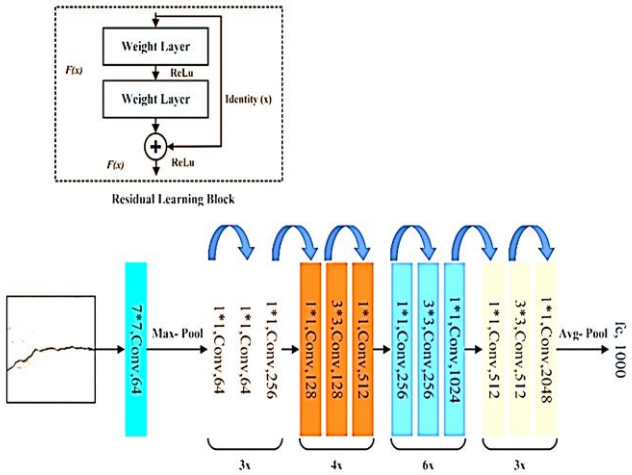

Fig. 4. ResNet-50 architecture [9].

G. BERT

The full form of BERT is 'Bidirectional Encoder Representations from Transformers'. BERT is a powerful pre-trained language model. It was developed by a group of researchers at Google AI in 2018. This model is based on transformers. An encoder only architecture is used by this model [6]. It can work with massive amounts of data. BERT can be fine tuned by adding additional layers. This model also connects all output elements with all input elements.

H. RoBERTa

The full form of Roberta is 'Robustly Optimized BERT Pretraining Approach'. Roberta is a pre-trained language model. It is an improvement to BERT. Roberta was first introduced by a group of researchers at Facebook AI [7]. RoBERTa uses a new training approach where only selective tokens are masked in each training step. The primary difference between BERT and RoBERTa is that there are changes in the main hyperparameters of RoBERTa. Moreover, the next sentence prediction objective that is used in BERT pre-training, is not used in RoBERTa. Larger batch sizes can be used for RoBERTA and this helps to improve training efficiency.

IV. Experimental Results

A. Performance Evaluation

For the three different fusion techniques RoBERTa+EfficientNet b3, MobileNetv2+BERT & ResNet50+RoBERTa, many evaluation measures were used. These measures are Accuracy, F1 score, Confusion matrix, ROC curve and classification report. Evaluation measures were performed on the testing set. There are 487 image-text pairs in the testing set and 4382 image-text pairs in the training set.

B. RoBERTa+EfficientNet b3 Results

RoBERTa+EfficientNet b3 is the first fusion technique that have been applied. In this fusion technique, the best result in accuracy, roc curve, confusion matrix and F1 score were achieved.

The confusion matrix of the RoBERTa+EfficientNet b3 model is shown in Fig. 5. One of the approaches for assessing a classification model's performance is the confusion matrix. The diagonal part of the confusion matrix shows the correctly predicted sentiments by the model. There are 1278 true positives, 1462 true neutrals and 907 true negatives predictions. The other values of the confusion matrix are the incorrect predictions made by the model. From the confusion matrix, it can be inferred that the model was more accurate in predicting neutral class and less accurate in predicting negative class. The classification report in Table I shows precision, recall, f1-score, support & accuracy. The precision indicates the proportion of positive identifications that were correct. For instance, positive class has a precision of 0.77, meaning 77% of predictions were correct. Precision, recall and F1 scores are relatively consistent across all three classes (positive, neutral and negative). This suggests that the model is performing uniformly across different types of data. The recall for the negative class is slightly lower (0.70) compared to the positive and neutral class. The support values indicate the number of instances for each class in the dataset. The 'accuracy' value of 0.75 is the overall accuracy of the model. Plotting the true positive rate (TPR) against the false positive rate (FPR) at each threshold setting creates the ROC curve.

From Fig. 6, it can be inferred that for the positive ROC curve, the model has a slightly better performance in classifying true positives and false positives, as indicated by its higher area under the curve (AUC) of 0.83. Neutral and Negative ROC curves have an AUC of 0.80, which is slightly lower than the positive curve.

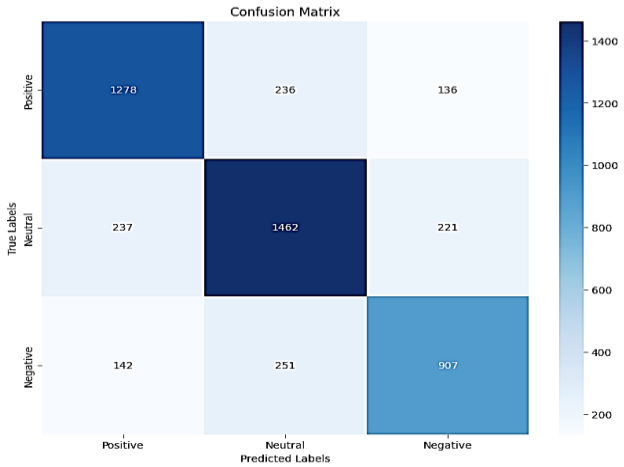

Fig. 5. Confusion matrix of RoBERTa+EfficientNet b3 model.

TABLE I. EFFICIENTNET B3+ROBERTA CLASSIFICATION REPORT

	Precision	Recall	F1-score	Support
Positive	0.77	0.77	0.77	1650
Neutral	0.75	0.76	0.76	1920
Negative	0.72	0.70	0.71	1300
Accuracy			0.75	4870
Macro avg	0.75	0.74	0.75	4870
Weighted Avg	0.75	0.75	0.75	4870

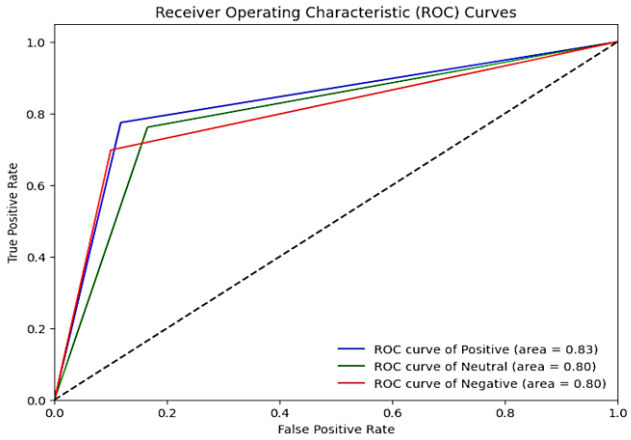

Fig. 6. ROC curve of RoBERTa+EfficientNet b3 model.

C. *MobileNetV2+BERT RESULT*

MobileNetV2+Bert is the second fusion technique implemented. While it achieved good results in accuracy, ROC curve, confusion matrix, and F1 score, the EffcientNet b3 + RoBERTA combination yielded superior performance.

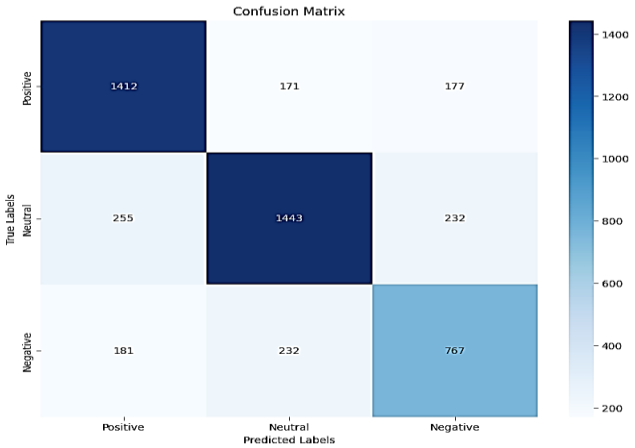

Fig. 7. Confusion matrix of MobileNetV2+BERT model.

From Fig. 7, it can be inferred that the model correctly identified 1412 instances as positive which is true positive. The model correctly identified 1443 instances as neutral which is true neutral. The model correctly identified 767 instances as negative which is true negative. The other values of the confusion matrix are the incorrect predictions made by the model. From this analysis, it appears that the model has a strong performance in identifying neutral sentiments. From Table II, it can be inferred that the model has the best performance with the positive class and the worst performance with the negative class. The model's precision is highest with neutral class, suggesting it is most reliable when predicting this class. The macro and weighted averages being close to the overall accuracy suggests a balanced dataset.

From Fig. 8, it can be inferred that, for the positive ROC curve the model has a good performance in classifying true positives and false positives, as indicated by its higher area under the curve (AUC) of 0.83. This means that the model has 83% chance of correctly distinguishing between positive and

non-positive instances. For the neutral ROC curve, the model has an AUC of 0.81, which is slightly lower than the positive model. This suggests that the model has an 81% chance of correctly distinguishing between neutral and non-neutral instances. For the negative ROC curve, the model has the lowest AUC of 0.77. In summary, the model performs best when predicting positive classes and worst when predicting negative classes.

TABLE II. MOBILENETV2+BERT CLASSIFICATION REPORT

	Precision	Recall	F1-score	Support
Positive	0.76	0.80	0.78	1760
Neutral	0.78	0.75	0.76	1930
Negative	0.65	0.65	0.65	1180
Accuracy			0.74	4870
Macro avg	0.73	0.73	0.73	4870
Weighted Avg	0.74	0.74	0.74	4870

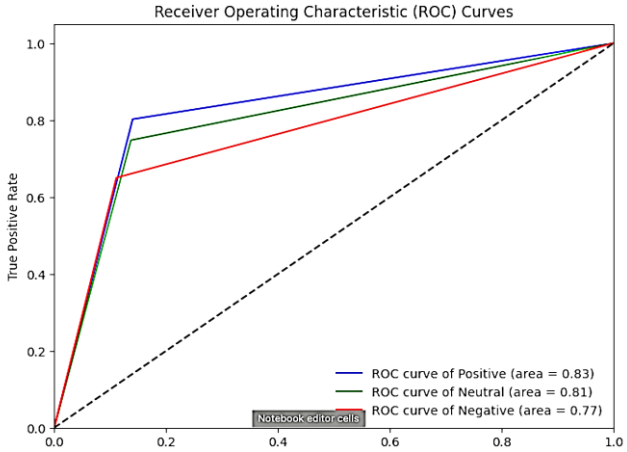

Fig. 8. ROC curve of BERT+MobileNetV2.

D. ResNet-50+RoBERTa Result

ResNet-50+RoBERTa is the final fusion technique that has been applied. The result of this fusion was not as satisfactory as the previous two fusion techniques. From Fig. 9, it can be inferred that the model correctly identified 1244 positive instances, 1272 neutral instances, and 997 negative instances. The other values of the confusion matrix are the incorrect predictions made by the model. From this analysis, it appears that the model has a strong performance in identifying positive sentiments. There are still a significant number of misclassifications, especially for the positive and neutral classes being incorrectly predicted as negative. This suggests that the model struggles with distinguishing between these classes. From Table III, it can be inferred that overall accuracy, macro average, and weighted average are all at 0.72 which means consistent performance across all classes. The model has the best performance with positive class and the worst with negative class. The model's precision is highest with neutral which means that it is most reliable when predicting this class.

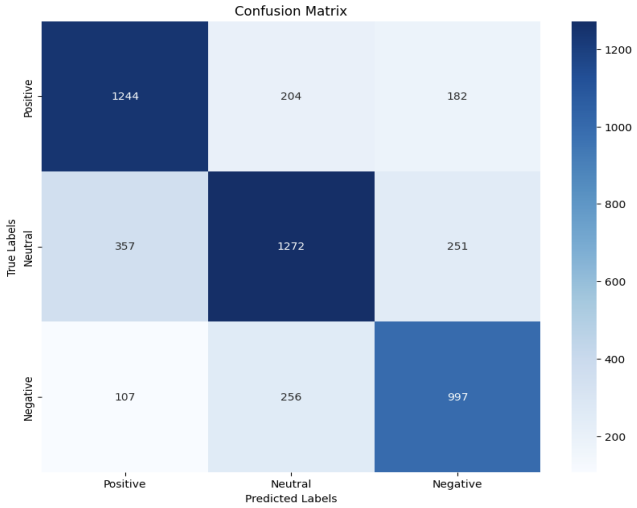

Fig. 9. Confusion matrix of ResNet-50+RoBERTa.

TABLE III. CLASSIFICATION REPORT OF ROBERTA+RESNET-50

	Precision	Recall	F1-score	Support
Positive	0.73	0.76	0.75	1630
Neutral	0.73	0.68	0.70	1880
Negative	0.70	0.73	0.71	1360
Accuracy			0.72	4870
Macro avg	0.72	0.72	0.72	4870
Weighted Avg	0.72	0.72	0.72	4870

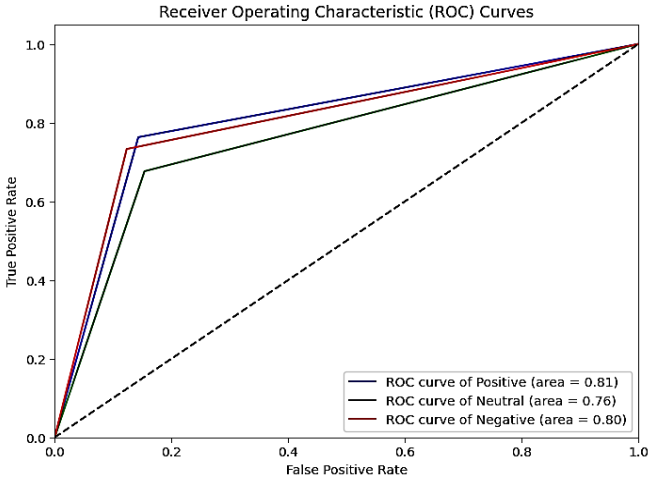

Fig. 10. ROC curve of RoBERTa+ResNet-50.

From Fig. 10, it can be inferred that for the positive ROC curve, the model performs well in classifying true positives and false positives, as indicated by its higher area under the curve (AUC) of 0.81. For the Neutral ROC curve, the model has the lowest AUC of 0.76. In summary, the model performs best when predicting positive classes and worst when predicting neutral classes.

V. DISCUSSION

The study compared the performance of three deep learning fusion techniques for multimodal sentiment analysis: RoBERTa with EfficientNet-b3, MobileNetV2 with BERT, and ResNet-50 with RoBERTa. Among the three techniques, RoBERTa with EfficientNet-b3 achieved the best overall performance, with an accuracy of 75% and an F1 score of 74.9%. This suggests that the combination of EfficientNet-b3 for image feature extraction and RoBERTa for text feature extraction is particularly effective for multimodal sentiment analysis.

Looking deeper into the results for RoBERTa + EfficientNet-b3, it is evident that precision, recall, and F1-scores are relatively consistent across all three classes (positive, neutral, and negative). This indicates that the model performs well on all sentiment categories. The ROC curve analysis further supports this, with AUC values around 0.8 for all classes, suggesting good performance in distinguishing between true positives and false positives. MobileNetV2 with BERT achieved an accuracy of 74%, with the best performance for the positive class and the worst for the negative class. The model's precision is highest for the neutral class, indicating good reliability in predicting neutral sentiment. ResNet-50 with RoBERTa had the lowest accuracy (72%) among the three techniques (see Table III). While it performed well on the positive class, it struggled with distinguishing between positive and neutral classes, as indicated by a significant number of misclassifications in the confusion matrix.

A comparison study with different benchmark models from other papers were conducted. The comparison is presented in Table IV where accuracy and F1 score of all the models are compared. It can be seen that EfficientNet b3+RoBERTa is better than the other fusion techniques as the accuracy of EfficientNet b3+RoBERTa is highest. The F1 scores of all the fusion techniques were almost similar to the accuracy but they are slightly lower than the accuracy. They outperformed the models from other papers.

TABLE IV. ACCURACIES AND F1 SCORES FOR VARIOUS MODELS

Models	Accuracy	F1
CNN-Multi [21]	61.2	58.4
BDMLA [22]	61.7	62.8
DNN-LR [23]	61.4	61.0
LATE-RMNN [24]	67	66.5
CoMN [25]	70.5	70
MultiSentiNet [26]	69.8	69.6
DMAF [27]	70.1	71.7
ResNet-50+RoBERTa	72	72.1
MobileNetV2+BERT	74	73.4
EfficientNet b3+RoBERTa	**75**	**74.9**

VI. CONCLUSION

This paper explores multimodal sentiment analysis through the application of three distinct fusion techniques: EfficientNet b3 + RoBERTa, MobileNetV2 + BERT, and RoBERTa + ResNet-50. This research contributes to the field of sentiment analysis by demonstrating the potential of combining text and image data using deep learning fusion techniques to achieve superior sentiment analysis accuracy compared to traditional methods that rely on a single modality. The approach utilized various convolutional neural networks (CNNs) for image feature extraction, while leveraging two distinct transformers for textual feature extraction. Following feature extraction from text and images through fusion, the model underwent training and testing. The evaluation results showed that EfficientNet-b3 + RoBERTa achieved the best accuracy (75%) and F1 score (74.9%) among the three fusion techniques. This suggests that the combination of EfficientNet-b3 for image analysis and RoBERTa for text analysis is particularly effective for multimodal sentiment analysis.

The primary limitation is the use of a relatively small dataset (MVSA-Single), which might restrict the model's ability to generalize to unseen data. Future work will involve utilizing a new, larger dataset. The creation of a new, expansive multimodal dataset sourced from Facebook is planned for future work. Additionally, the implementation of the latest EfficientNet version is envisioned. The utilization of more powerful computing resources to accommodate larger models is anticipated for future work. Furthermore, advancements in Recurrent Neural Networks (RNNs) are expected to further contribute to the development of multimodal sentiment analysis.

REFERENCES

[1] A. Das, M. M. Hoque, O. Sharif, M. A. A. Dewan, and N. Siddique, "TEmoX: Classification of Textual Emotion Using Ensemble of Transformers," IEEE Access, vol. 11, pp. 109803–109818, 2023, doi: 10.1109/ACCESS.2023.3319455.

[2] G. A.V., M. T., P. D., and U. E., "Multimodal Emotion Recognition with Deep Learning: Advancements, challenges, and future directions," Information Fusion, vol. 105, p. 102218, 2024, doi: https://doi.org/10.1016/j.inffus.2023.102218.

[3] B. Pang, L. Lee, and S. Vaithyanathan, "Thumbs up? Sentiment classification using machine learning techniques," arXiv preprint cs/0205070, 2002.

[4] P. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, P. Isabelle, E. Charniak, and D. Lin, Eds., Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, Jul. 2002, pp. 417–424. doi: 10.3115/1073083.1073153.

[5] K. L. S. Kumar, J. Desai, and J. Majumdar, "Opinion mining and sentiment analysis on online customer review," in 2016 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), 2016, pp. 1–4. doi: 10.1109/ICCIC.2016.7919584.

[6] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), J. Burstein, C. Doran, and T. Solorio, Eds., Minneapolis, Minnesota: Association for Computational Linguistics, Jun. 2019, pp. 4171–4186. doi: 10.18653/v1/N19-1423.

[7] Y. Liu et al., "Roberta: A robustly optimized bert pretraining approach," arXiv preprint arXiv:1907.11692, 2019.

[8] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv:1301.3781, 2013.

[9] K. He, X. Zhang, S. Ren, and J. Sun, Deep Residual Learning for Image Recognition. 2016. doi: 10.1109/CVPR.2016.90.

[10] M. Tan and Q. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," in Proceedings of the 36th International Conference on Machine Learning, K. Chaudhuri and R. Salakhutdinov, Eds., in Proceedings of Machine Learning Research, vol. 97. PMLR, Jun. 2019, pp. 6105–6114. [Online]. Available: https://proceedings.mlr.press/v97/tan19a.html

[11] L.-P. Morency, R. Mihalcea, and P. Doshi, "Towards multimodal sentiment analysis: harvesting opinions from the web," in Proceedings of the 13th International Conference on Multimodal Interfaces, in ICMI '11. New York, NY, USA: Association for Computing Machinery, 2011, pp. 169–176. doi: 10.1145/2070481.2070509.

[12] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal Machine Learning: A Survey and Taxonomy," IEEE Trans Pattern Anal Mach Intell, vol. 41, no. 2, pp. 423–443, 2019, doi: 10.1109/TPAMI.2018.2798607.

[13] M. Chen, S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh, and L.-P. Morency, "Multimodal sentiment analysis with word-level fusion and reinforcement learning," in Proceedings of the 19th ACM International Conference on Multimodal Interaction, in ICMI '17. New York, NY, USA: Association for Computing Machinery, 2017, pp. 163–171. doi: 10.1145/3136755.3136801.

[14] A. Aslam, A. B. Sargano, and Z. Habib, "Attention-based multimodal sentiment analysis and emotion recognition using deep neural networks," Appl Soft Comput, vol. 144, p. 110494, 2023, doi: https://doi.org/10.1016/j.asoc.2023.110494.

[15] C. Huang, J. Zhang, X. Wu, Y. Wang, M. Li, and X. Huang, "TeFNA: Text-centered fusion network with crossmodal attention for multimodal sentiment analysis," Knowl Based Syst, vol. 269, p. 110502, 2023, doi: https://doi.org/10.1016/j.knosys.2023.110502.

[16] W. Bai, "Greeny at Factify 2022: Ensemble model with optimized roberta for multi-modal fact verification," in Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEUR, 2022.

[17] J. Peng et al., "A fine-grained modal label-based multi-stage network for multimodal sentiment analysis," Expert Syst Appl, vol. 221, p. 119721, 2023, doi: https://doi.org/10.1016/j.eswa.2023.119721.

[18] T. Niu, S. Zhu, L. Pang, and A. El Saddik, "Sentiment Analysis on Multi-View Social Data," in MultiMedia Modeling, Q. Tian, N. Sebe, G.-J. Qi, B. Huet, R. Hong, and X. Liu, Eds., Cham: Springer International Publishing, 2016, pp. 15–27.

[19] H. Wang, X. Li, Z. Ren, M. Wang, and C. Ma, "Multimodal Sentiment Analysis Representations Learning via Contrastive Learning with Condense Attention Fusion," Sensors, vol. 23, no. 5, 2023, doi: 10.3390/s23052679.

[20] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. Chen, "MobileNetV2: Inverted Residuals and Linear Bottlenecks," in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2018, pp. 4510–4520. doi: 10.1109/CVPR.2018.00474.

[21] G. Cai and B. Xia, "Convolutional Neural Networks for Multimedia Sentiment Analysis," in Natural Language Processing and Chinese Computing, J. Li, H. Ji, D. Zhao, and Y. Feng, Eds., Cham: Springer International Publishing, 2015, pp. 159–167.

[22] J. Xu et al., "Visual-textual sentiment classification with bi-directional multi-level attention networks," Knowl Based Syst, vol. 178, May 2019, doi: 10.1016/j.knosys.2019.04.018.

[23] Y. Yu, H. Lin, J. Meng, and Z. Zhao, "Visual and Textual Sentiment Analysis of a Microblog Using Deep Convolutional Neural Networks," Algorithms, vol. 9, p. 41, Jun. 2016, doi: 10.3390/a9020041.

[24] N. Xu and W. Mao, "A residual merged neutral network for multimodal sentiment analysis," in 2017 IEEE 2nd International Conference on Big Data Analysis (ICBDA), 2017, pp. 6–10. doi: 10.1109/ICBDA.2017.8078794.

[25] N. Xu, W. Mao, and G. Chen, "A Co-Memory Network for Multimodal Sentiment Analysis," in The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, in SIGIR '18. New York, NY, USA: Association for Computing Machinery, 2018, pp. 929–932. doi: 10.1145/3209978.3210093.

[26] N. Xu and W. Mao, "MultiSentiNet: A Deep Semantic Network for Multimodal Sentiment Analysis," in Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, in CIKM '17. New York, NY, USA: Association for Computing Machinery, 2017, pp. 2399–2402. doi: 10.1145/3132847.3133142.

[27] F. Huang, X. Zhang, Z. Zhao, J. Xu, and Z. Li, "Image–text sentiment analysis via deep multimodal attentive fusion," Knowl Based Syst, vol. 167, pp. 26–37, 2019, doi: https://doi.org/10.1016/j.knosys.2019.01.019.