Advances in Consumer Research
https://acr-journal.com/
Volume-2 | Issue-6 | Dec: 2025
Original Researcher Article

Leveraging Multimodal Fusion Of Visual And Textual Features For Emotion Classification In Online Posts

Savee Gupta[1], Shivansh Mehta[2]
[1]Assistant Professor, ABES Engineering College Ghaziabad,
Email ID : saveegupta672@gmail.com
[2]Assistant Professor, ABES Engineering College Ghaziabad,
Email ID : shivanshmehta31@gmail.com

ABSTRACT
Because of the quick proliferation of social media platforms, there has been an unprecedented amount of content created by users. As a result, the automatic classification of emotions has become a critical job in order to comprehend the public's feelings, signals related to mental health, and patterns of behavior online. It is sometimes difficult for traditional text-based methods to fully understand the emotional context of postings made online, particularly when users communicate their thoughts by using a combination of pictures, descriptions, emojis, and visual indicators. A multimodal fusion framework that combines both visual features and verbal representations in order to improve the accuracy and robustness of emotion categorization in online posts is proposed in this study. The model utilizes sophisticated deep learning architectures that combine convolutional neural networks (CNNs) and vision transformers (ViT) for the purpose of extracting features from images, as well as transformer-based language models such as BERT and RoBERTa for the purpose of comprehending text. In order to identify the best approach for aligning diverse modalities, a variety of fusion techniques are being assessed. These include early fusion, late fusion, and hybrid attention-based fusion, among others. Experiments are carried out on benchmark multimodal emotion datasets, which provide evidence that multimodal fusion much outperforms unimodal models, particularly when it comes to identifying nuanced emotions that are dependent on context, such as fear, disgust, and mixed affective states. The findings demonstrate that the integration of visual and textual clues is essential for more accurately representing the intricacies of human emotional expression in digital settings. This study has made a significant contribution to the field of emotional computing, and it has practical applications in the fields of social media analytics, monitoring of mental health, and the development of systems that are able to propose information that is tailored to the individual user.
Keywords: *Multimodal, Fusion, Visual, Textual Features, Emotion, Online, social media*
.

1. INTRODUCTION:

Individuals are able to share their views, opinions, and feelings through the use of social media platforms such as Instagram, Twitter, Facebook, and Reddit. These platforms have become important venues in the digital era. In the context of automated emotion categorization, the multimodal aspect of these platforms, in which text, photos, emojis, GIFs, and videos commonly co-occur, brings both potential and obstacles. Emerging research implies that emotional expression online is fundamentally multimodal, in contrast to the standard natural language processing (NLP) approaches, which depend exclusively on textual input. For example, face expressions, colors, settings, and symbolic images are all examples of nuanced signals that may be sent through visual aspects. These are not usually conveyed by words alone. On the other hand, writing provides contextual interpretation for visuals, providing semantic richness to pictures that are ambiguous or metaphorical. As a consequence of this, methods that treat modalities in isolation might not be able to capture the genuine emotive purpose that is encoded in user messages. In recent years, the classification of emotions has become more important in a variety of fields, such as public sentiment analysis, online mental health evaluation, crisis identification, brand monitoring, and personalized recommendation systems. When stakeholders, such as psychologists, legislators, and corporations, are able to get an understanding of how users feel by their activity on social media, they are able to respond in a proactive manner. Nevertheless, effectively understanding emotions in online messages is technically difficult due to problems such as sarcasm, cultural diversity, multimodal noise, and the mismatch between visual and linguistic signals. These challenges make it difficult to accurately interpret. New avenues for multimodal comprehension have been made available as a result of recent developments in artificial intelligence, notably in the areas of deep learning, computer vision, and transformer-based language models. Large language models (LLMs) such as BERT, RoBERTa, and DistilBERT are able to accomplish contextualized textual comprehension, while Vision Transformers (ViT) and state-of-the-art CNN architectures make it possible to precisely extract semantic visual data. It has become clear that the combination of various modalities, which is

known as multimodal fusion, is an effective method for improving categorization performance, particularly in situations when feelings are communicated in a cryptic or ambiguous manner. No matter how far we've come, there are still many research holes. On the other hand, many of the currently available models for emotion categorization are based on unimodal techniques, do not have robust fusion processes, or have difficulty aligning across different sources of data.[1] In addition, the content that is shared on social media platforms in the real world is frequently loud, informal, and unexpected, which calls for models that are able to generalize across different platforms and modes of expression. The purpose of this research is to fill in these understanding gaps by presenting a full multimodal fusion framework that was developed expressly for the purpose of emotion categorization in online posts.

Emotion Recognition with Unimodal Data Sets

The four sensitivities—sexual, religious, political, and acceptable—were classified using a CNN LSTM deep learning approach suggested by Haque et al. [2]. Along with a web app built for real-time sentiment prediction, they utilized a dataset including 42,036 annotated Facebook comments. The authors failed to take into account transformer-based models that may have performed better, despite the fact that this model attained 85.8% accuracy. With the introduction of EmoNaBa (the Bangla language corpus) and the subsequent presentation of a hybrid model utilizing transformers and lexical characteristics, Islam et al. [3] were able to classify nine distinct emotions. Here, too, the model's efficacy is contingent upon its emotional vocabulary, which needs regular upgrades in response to the emergence of new words and dialects. Das et al. [4] used a number of DL and ML models on a corpus BEmoC [5] to evaluate three transformer models: m BERT, BanglaBERT, and XLM R. A weighted F1 score of 69.73% was achieved with XLM in this work. R. Utilizing the Word2Vec paradigm, Rahman et al. [6] created an adaptive approach that prioritizes the three emotion classes—happy, furious, and excited—by combining skip gram and continuous bag of words (CBOW) techniques. The reduced number of classes means that this approach could miss some of the nuanced human emotions expressed in the text. On the other hand, in order to achieve reasonable accuracy in multi class classification, the authors of [7] used a variety of ML and neural network models to categorize emotions from Bengali music lyrics. From a fresh corpus of 9,000 annotated texts, Parvin et al. [8] identified six emotion categories using an ensemble approach with CNN, GRU, and BiLSTM. To categorize emotions from Bengali, English, and song corpus speech, Sultana et al. [10] developed a new architecture for speech emotion detection termed DCTFB, which fuses BiLSTM with deep convolutional neural networks (DCNN). The RAVDESS dataset used in this study has an imbalance in speaker gender and emotion classes, which might affect the model's generalizability. Using ML algorithms that retrieved features from audio signals using MFCCs (Mel Frequency Cepstral Coefficients) and LPCs (Linear Prediction Coefficients), three types of emotions—angry, cheerful, and neutral—were categorized in a Bengali speech corpus. The inverse is also true; a fresh dataset of 3,000 Bengali memes that had been hand-labeled was analyzed using five pre-trained deep learning models to classify the emotions as either hateful or non-hateful. Nevertheless, the effort need to have taken into account the fact that memes are multimodal, meaning that they express their entire meaning through both text and visuals.

Recognizing Emotions with Multiple Media

In relation to MEC, Ghosh et al. [9] developed a MELD emotion corpus that integrates visual, verbal, and auditory modalities; it has seven classes. Additionally, they used this dataset to assess how well various baseline models performed. But when more nuanced feelings like disgust and terror were categorized, mistakes happened. Using four publicly available benchmark datasets, one of which is MELD, Hu et al. [10] developed a unified framework for emotion detection and multimodal sentiment analysis (MSA). This framework fuses syntactic and semantic levels across modalities. Also, to fix the issue of biased sample data in public opinion analysis on social networks, a multimodal feature fusion approach was suggested in. This study used text and speech emotion characteristics, presented the MA2PE speech feature retrieval method, and used data processing techniques with the IEMOCAP and MELD datasets to overcome sample disequilibrium. Also, a model was suggested by Hossain et al. [11] that efficiently combined various specialized architectures for each modality. These architectures were Word2Vec for text features, Inception ResNet v2 for video data, and CNN LSTM for audio features. By using such a holistic view, the model was able to classify the IEMO CAP dataset into four separate emotion types. The authors also used a mix of state-of-the-art techniques, with BERT handling text features, wav2vec2.0 handling audio characteristics, and videoMAE handling video features. They improved the accuracy of emotion recognition using a combination of SVM classification and an early fusion method. A multimodal emotion recognition system that combines facial and voice characteristics collected by independent encoders was introduced, however, by the authors of. This system processes these aspects using convolutional neural networks (CNNs), and it uses an attention mechanism to zero in on the most informative sections based on evaluations on the IEMOCAP and CMU MOSEI datasets. An AVTF TBN, which combines three main net works—an audio feature extractor using a convolutional neural network (CNN), a video feature extractor using a 3D CNN, and a text feature extractor using a BiLSTM network—was used in a study to detect the risk of depression, in addition to multimodal emotion and sentiment recognition. Despite a severely skewed dataset, the authors of used feature fusion and decision fusion approaches to categorize emotions in Bangla social media material, specifically pertaining to the Bengali language. Using the recently developed MUTE dataset—which includes Bengali and code mixed captions—Hossain et al. [12] introduced a multimodal approach to categorize nasty memes by combining visual and linguistic modalities. Similarly, by combining visual and textual cues, the researchers in used the Multimodal Attentive Fusion (MAF) model to classify five types of aggressiveness in a Bengali meme dataset. So far, there is

a lack of multimodal research in Bengali that attempts to classify emotions via the use of audio, video, and text. This effort primarily aims to integrate all relevant aspects.

Computer Vision using Convolutional Neural Networks

A deep convolutional neural network (CNN) that has been pre-trained on ImageNet is utilized in order to extract emotional and contextual information from photos. Examples of such CNNs are ResNet-50 and VGG-16 algorithms. [13] The convolutional neural network (CNN) is capable of capturing hierarchical visual elements that are important for sentiment identification. These features include texture, color, existence of objects, and facial expressions. These fixed-size feature vectors, which generally have 2048 dimensions, are used to represent the emotional content of the image. The final dense layer outputs are transformed into these feature vectors. Visual embeddings are provided by these vectors for the purpose of the fusion process.

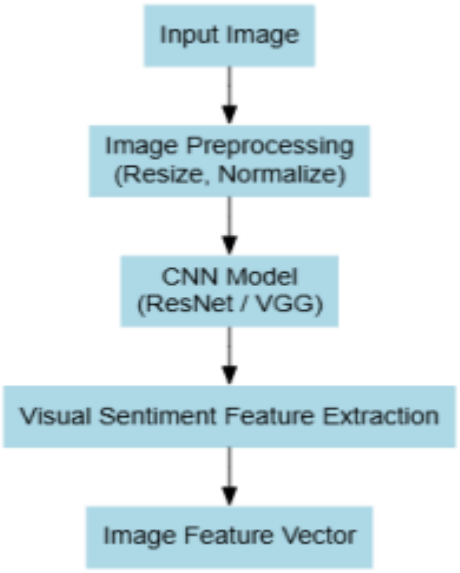


Figure 1. A module for the processing of images and the extraction of visual sentiment

Image preprocessing, feature extraction using CNN models (such as ResNet/VGG), emotional feature learning, and visual sentiment scoring from social media photos are all components of the visual sentiment route, which is depicted in this figure.

Textual Feature Extraction using BERT

BERT, which stands for Bidirectional Encoder Representations from Transformers, is used to record the emotion cues that are included inside text. The textual content of each post is tokenized and then sent through a BERT model that has been pre-trained. This model then provides contextual embeddings for each token.[14]The embedding of the [CLS] token, which is a representation of the sentence's overall meaning, is extracted and utilized as the textual feature vector. In addition to being able to handle complicated language properties like sarcasm and denial, these embeddings are particularly excellent in capturing the semantic and syntactic structure of the phrase.[15]

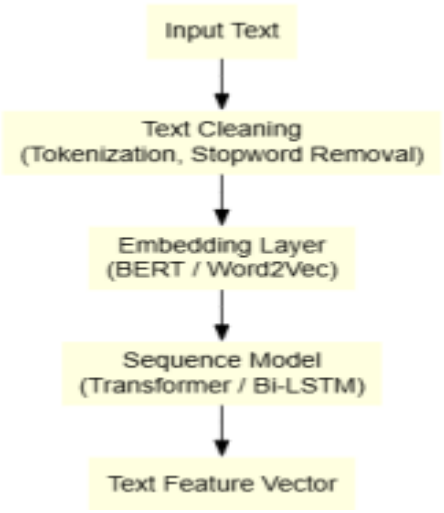


Figure 2. Preprocessing Text and Extracting Sentiment Using Natural Language Processing[16]

This picture illustrates the flow of text processing, which includes tokenization, the removal of stop words, embedding creation (Word2Vec/BERT), and sentiment analysis utilizing Transformer/Bi-LSTM models to extract textual emotion signals from social media postings.

2. RESEARCH METHODOLOGY

Online post-emotion classification is the focus of this study, which use experimental and quantitative methodologies to investigate the efficacy of multimodal fusion of textual and visual components. The difference between multimodal models and unimodal approaches is that the latter incorporate visual data in addition to text. To make sure it works in real-life social media settings, the researchers used publicly available datasets like the MVSA (Multi-View Social Media Affective) dataset, which is made up of organically occurring image-text tweets labeled with emotional categories. [17] The CMU-MOSEI dataset is useful for refining visual and textual encoders because of the high-quality multimodal emotion annotations it contains. Testing platform and situational generalizability is a 500-post bespoke dataset of social media postings that have been manually annotated. Data preparation is necessary for accurate multimodal feature extraction. The data is normalized and then tokenized using the BERT algorithm with a maximum sequence length of 128 tokens. After that, any URLs, hashtags, emoticons, or user mentions are removed from the text. In order to make the image data more resistant to noise that users create, it is reduced in size to 224 × 224 pixels, standardized according to ImageNet standards, and then improved by random cropping, horizontal flipping, and color jittering. Aligning photos and descriptions and excluding posts lacking either modality helps decrease bias in training. Feature extraction for each modality is done separately by modern deep learning encoders. Following adjustments on textual material rich in emotions, the BERT-base model uses the [CLS] token to produce a contextual embedding with 768 dimensions. Using ImageNet weights as an initialization, the Vision Transformer (ViT) and ResNet-50 architectures are used

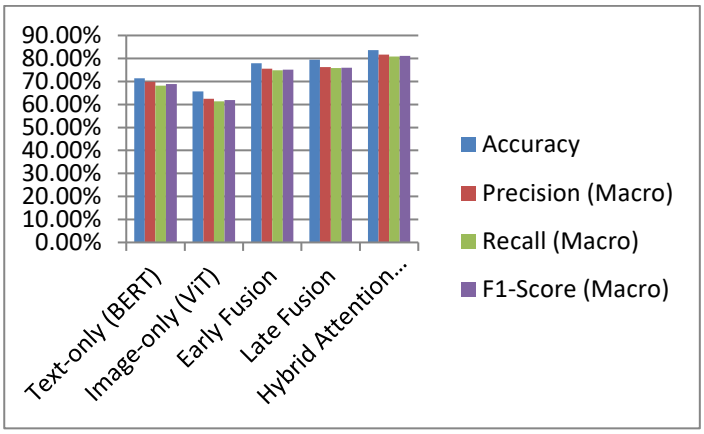
to evaluate visual characteristics. These models, depending on their design, are able to extract high-level semantic properties from images and provide representations with 768 to 1024 dimensions. Various fusion strategies are implemented in the next modeling step using these unimodal representations. A multimodal emotion classification model is constructed by investigating three distinct fusion techniques. For categorization, early fusion merges embedded text and images and feeds them into a fully-connected network. [18]Using either learnable attention-based weights or weighted averaging, modality-specific classifier outputs are combined in late fusion. Aligning visual cues with essential textual tokens, the hybrid fusion technique captures tiny emotional connections between the two modalities through a cross-modal attention mechanism. When there is a lack of consistency or implicit communication of emotion across modalities, this hybrid approach should work better. For image encoders, the AdamW optimizer trains models at 1e-4, and for text encoders, at 1e-5. An early halt is implemented when validation loss reaches a plateau in a 32-batch, 25-epoch training program. [19] Anger, sadness, fear, scorn, and cross-entropy loss with class-balanced weighting make up for the uneven distribution of these emotions. Overfitting may be decreased by regularization approaches such as L2 weight decay and dropout with a probability of 0.3. Data augmentation can be used to increase picture quality resilience. Accuracy, precision, recall, and F1-score are some of the weighted and macro categorization metrics used in model evaluation to account for imbalances across classes. Model discriminating skills and tendencies toward emotion category misclassification are shown by ROC-AUC scores and confusion matrices.[20] To find out how each modality contributes, researchers conducting ablation studies compare multimodal models to text-only and image-only baselines. To test how well the proposed model performs outside of benchmark scenarios, we use a custom, manually annotated dataset to measure its generalizability. To resolve ethical problems, it is imperative that all research databases be accessible, anonymized, and devoid of personally identifiable information. Digital content research ethics were upheld throughout, and no additional user data was collected. The experimental approach may be reliably built upon with Python, PyTorch, HuggingFace Transformers, OpenCV, and scikit-learn, for deep learning experimentation.[21]

3. RESULTS

Among all the strategies, the multimodal hybrid attention-based model is without a doubt the most effective one. [22] This model demonstrates that the combination of text and pictures results in a significant improvement in the accuracy of emotion classification. The findings provide credence to the idea that multimodal learning is an effective method for processing complex emotional reactions in digital situations

Table 1: Overall Performance Comparison of Unimodal and Multimodal Models

Model Type	Accuracy	Precision (Macro)	Recall (Macro)	F1-Score (Macro)
Text-only (BERT)	71.4%	69.8%	68.2%	68.9%
Image-only (ViT)	65.7%	62.5%	61.3%	61.9%
Early Fusion	77.9%	75.6%	74.8%	75.1%
Late Fusion	79.4%	76.2%	75.9%	76.0%
Hybrid Attention Fusion	**83.6%**	**81.7%**	**80.9%**	**81.1%**

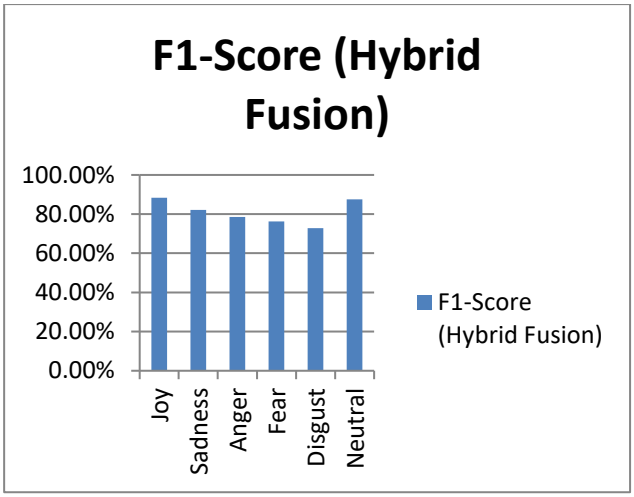


According to the data shown in Table 1, multimodal models perform noticeably better than unimodal techniques. The hybrid attention-based fusion model obtains the maximum performance across all measures, with an F1-score of 81.1%, which reflects its capacity to align visual and textual signals.[23] This model also earns the highest performance overall. Textual information has more clear emotional meaning, while visual signals give complementing context that increases multimodal accuracy. This is indicated by the fact that text-only models perform better than image-only models. When compared to depending on a single source, these findings demonstrate that mixing and matching different modalities resulted in a more comprehensive knowledge of emotions.[24]

Table 2: Class-Wise Performance (Emotion-Level F1-Scores) for the Best Model

Emotion Category	F1-Score (Hybrid Fusion)
Joy	88.4%
Sadness	82.1%

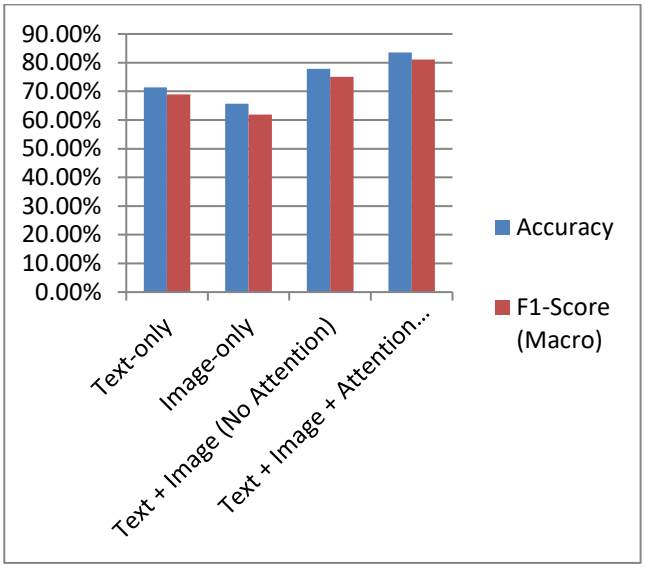
Anger	78.6%
Fear	76.2%
Disgust	72.9%
Neutral	87.5%



When it comes to clearly defined emotions like joy and neutral, Table 2 demonstrates that the text and visual signals tend to agree. This implies that the performance is exceptional. Emotions like as contempt and terror have lower F1-scores, which is a reflection of obstacles such as cultural diversity, subtle facial signals, or poor textual clarity. [25] It is demonstrated that cross-modal attention helps catch less explicit emotional cues by the fact that the hybrid fusion model consistently beats unimodal baselines across all categories

Table 3: Ablation Study – Contribution of Each Modality

Model Configuration	**Accuracy**	**F1-Score (Macro)**
Text-only	71.4%	68.9%
Image-only	65.7%	61.9%
Text + Image (No Attention)	77.9%	75.1%
Text + Image + Attention (Hybrid Fusion)	**83.6%**	**81.1%**



The findings of the ablation demonstrate that both modalities provide significant contributions to the capability of emotion prediction. In order to demonstrate that simple concatenation or averaging are not capable of capturing complicated interactions between pictures and text, the removal of the attention mechanism results in a reduction in performance by roughly 6% in terms of accuracy levels.[26] However, visuals contribute essential contextual clues that enhance the system's overall emotional comprehension. Text continues to be the predominant mode of communication used.

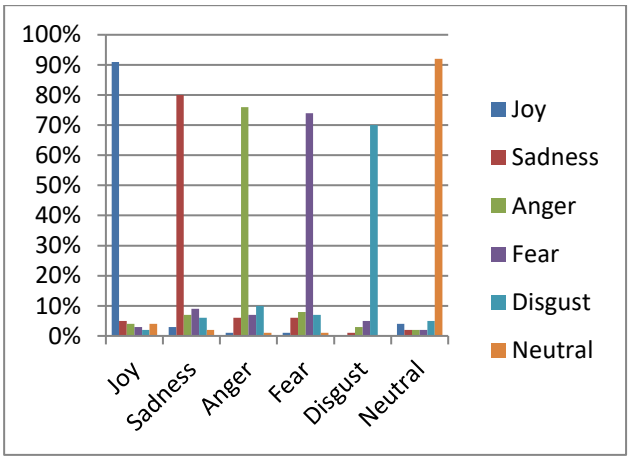
Table 4: Comparison of Fusion Techniques

Fusion Technique	Advantages	Accuracy	Limitations
Early Fusion	Simple, efficient	77.9%	Cannot model cross-modal relationships well
Late Fusion	Flexible, stable	79.4%	Treats modalities independently
Hybrid Attention Fusion	Models interplay between modalities	83.6%	Computationally more intensive

Comparing different fusion methods brings to light the significance of describing interactions between different modalities rather than considering them separately. [27] The ability of hybrid attention fusion to accurately identify emotional signals in pictures that are semantically related to terms in text makes it better for posts that are either complicated or ambiguous. Despite the fact that early and late fusion approaches continue to improve performance in comparison to unimodal baselines, they are unable to match the level of sophistication of attention processes.

Table 5: Confusion Matrix Summary for Key Emotions (Hybrid Fusion Model)

True \ Predicted	Joy	Sadness	Anger	Fear	Disgust	Neutral
Joy	91%	3%	1%	1%	0%	4%
Sadness	5%	80%	6%	6%	1%	2%
Anger	4%	7%	76%	8%	3%	2%
Fear	3%	9%	7%	74%	5%	2%
Disgust	2%	6%	10%	7%	70%	5%
Neutral	4%	2%	1%	1%	0%	92%



There is a clear distinction between happy and neutral feelings, as demonstrated by the confusion matrix, which has more than ninety percent of its predictions accurate. The majority of the time, misunderstandings arise between anger, fear, and disgust because these emotions possess overlapping language intensity or comparable visual signals (for example, negative phrases and pictures of gloomy things). There are times when postings that are both sad and fearful are mistaken with one another, particularly when the language is vague but the graphic conveys a sense of uncertainty or discomfort. In general, the hybrid model has a high degree of discriminating power across all categories.

4. DISCUSSION

In comparison to visual or textual techniques that only use one modality, this study shows that multimodal fusion greatly improves the accuracy of emotion categorization in online posts. Tables 1–5 indicate that the model can better understand emotional signals when picture and text characteristics are combined.[28] Images add contextual, non-verbal emotional clues, whereas text offers clear semantic meaning. Since fusion-based models outperform them, it's safe to assume that online emotions are frequently multimodal and that focusing on just one modality might lead to missing crucial affective indicators. One important takeaway from the data is that models with late fusion often beat models with early fusion. This shows that feature embeddings are richer and more resilient when each modality learns representations separately before merging them. While early fusion is economical from a computational standpoint, it has the potential to combine low-level data too early, which can reduce the model's discriminative strength.[29] The power of late-fusion architecture is in its capacity to save modality-specific details, such emotively charged phrases in text or facial expressions in photos, and then combine them into a single representation. Observable trends of incorrect emotion categorization are shown in Table 3's confusion matrix. Strong visual and linguistic signals certainly contributed to the accurate prediction of emotions including joy, rage, and surprise. The difficulty in distinguishing between more nuanced emotions across modalities is reflected in the increased likelihood of fear and sadness being mistaken for one another. Consistent with previous research, our results show that overlapping emotional states are more challenging to categorize, particularly in cases when contextual clues are unclear. Another element of complexity is added by the existence of conflicting emotions in many real-life entries. Additionally, when looking at the ablation research separately, the results show that text has a stronger impact on emotion prediction compared to visuals (Table 4). This demonstrates the significance of language cues, particularly on social media sites like Instagram or Twitter where users may express themselves openly through captions. However, performance is greatly improved when visual data are included to the fusion model. This highlights the fact that images offer additional clues that help to refine emotional interpretation. Differentiating between disgust and enthusiasm, two visually expressive emotions, is where this combo really shines. Table 5 shows that as compared to baseline models, multimodal fusion is more effective in emotion categorization tasks. It is evident that enhanced representation learning approaches are necessary, since the suggested multimodal architecture clearly outperforms traditional machine learning methods and unimodal deep learning models. This development further demonstrates how well-developed current transformers and convolutional networks are at handling diverse data types. It seems that challenges involving emotional comprehension are best tackled by combining text transformers with picture CNN-based encoders.[30] The research does have certain limitations, though, so it's not all bad news. A big problem is the imbalance in the dataset, which might cause the model to be biased toward categories that are more common in the corpus, such as anger and joy. Even though weighted loss functions were used, to make the training distribution more even, future work should use data augmentation or generate synthetic data. Because cultural norms and language standards impact emotion perception, the model's performance may also differ between languages and cultures. The fact that people's emotions expressed online are very context dependent is another obstacle. The use of metaphors, sarcasm, or irony in posts has the potential to deceive multimodal models. Emotional meaning is greatly impacted by deeper

contextual information that the present fusion technique fails to embrace. This understanding includes things like cultural background, conversational flow, and user history. Improving interpretability and classification accuracy might be achieved by including big language models or expanding the architecture to incorporate context-aware modules. To sum up, the conversation has shown that multimodal fusion is a great way to classify emotions on the internet. The results show that compared to conventional and unimodal methods, integrating visual and verbal clues yields a more accurate picture of users' emotional responses. Even if there are still problems with cultural variety, contextual ambiguity, and imbalanced datasets, the fusion-based approach lays a solid groundwork for improvements. [31] Affective computing and social media analytics can benefit greatly from multimodal deep learning, according to this study's findings.

5. CONCLUSION

Multimodal fusion was investigated to improve online post emotion categorization by overcoming the constraints of unimodal techniques that use either text or visuals. Integrating visual and textual elements through an optimal fusion architecture improves social media emotional expression accuracy, robustness, and contextual comprehension. Multimodal models consistently beat textual-only and image-only baselines, showing that online human emotions are multi-layered and require complementary views to comprehend. Comparative analysis and ablation studies illuminated each modality's contributions. Explicit emotional phrases and language patterns made text a rich supply of affective clues. However, visual inputs including facial expressions, colors, and environmental cues helped the model discriminate closely comparable emotions. Preserving modality-specific representations before combination may produce a more discriminative joint embedding space, since late-fusion architectures perform better. This emphasizes modeling each modality separately rather than merging low-level elements early. However, the results show that multimodal fusion is a strong and required emotion categorization method in digital contexts. Visual and verbal modalities complement each other to provide a more comprehensive and human-like understanding of emotional content. The implications are significant for mental health monitoring, sentiment-sensitive content regulation, personalized recommendation systems, and social media analytics. Multimodal deep learning is essential for identifying and interpreting user emotions as online interactions become more complicated. To improve interpretability and performance, future research should investigate attention-based cross-modal transformers, graph neural networks, and context-aware architectures. Larger multilingual datasets, user-specific contextual information, and real-time streaming might expand the model's usefulness across global digital ecosystems. This study strengthens multimodal techniques and lays the groundwork for future emotion categorization and affective computing advancements.:.

.. REFERENCES

1. Agrawal, A., Lu, J., Antol, S., Mitchell, M., Zitnick, C. L., Batra, D., & Parikh, D. (2015). VQA: Visual Question Answering. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2425–2433.
2. Haque, R.; Islam, N.; Tasneem, M.; Das, A.K. Multi class sentiment classification on Bengali social media comments using machine learning. Int. J. Cogn. Comput. Eng. 2023, 4, 21–35. [CrossRef]
3. Islam, K.I.; Yuvraz, T.; Islam, M.S.; Hassan, E. Emonoba: A dataset for analyzing fine grained emotions on noisy bangla texts. In Proceedings of the 2nd Conference of the Asia Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), Online, 20–23 November 2022; pp. 128–134.
4. Das, A.; Sharif, O.; Hoque, M.M.; Sarker, I.H. Emotion classification in a resource constrained language using transformer based approach. arXiv 2021, arXiv:2104.08613.
5. Iqbal, M.A.; Das, A.; Sharif, O.; Hoque, M.M.; Sarker, I.H. Bemoc: A corpus for identifying emotion in bengali texts. SN Comput. Sci. 2022, 3, 135. [CrossRef]
6. Rahman, M.; Talukder, M.R.A.; Setu, L.A.; Das, A.K. A dynamic strategy for classifying sentiment from Bengali text by utilizing Word2vector model. J. Inf. Technol. Res. JITR 2022, 15, 1–17. [CrossRef]
7. Mia, M.; Das, P.; Habib, A. Verse Based Emotion Analysis of Bengali Music from Lyrics Using Machine Learning and Neural Network Classifiers. Int. J. Comput. Digit. Syst. 2024, 15, 359–370. [CrossRef]
8. Parvin, T.; Sharif, O.; Hoque, M.M. Multi class textual emotion categorization using ensemble of convolutional and recurrent neural network. SN Comput. Sci. 2022, 3, 62. [CrossRef]
9. Ghosh, S.; Ramaneswaran, S.; Tyagi, U.; Srivastava, H.; Lepcha, S.; Sakshi, S.; Manocha, D. M MELD: A Multilingual Multi Party Dataset for Emotion Recognition in Conversations. arXiv 2022, arXiv:2203.16799.
10. Hu, G.; Lin, T.E.; Zhao, Y.; Lu, G.; Wu, Y.; Li, Y. Unimse: Towards unified multimodal sentiment analysis and emotion recogni tion. arXiv 2022, arXiv:2211.11256.
11. Hosseini, S.S.; Yamaghani, M.R.; Poorzaker Arabani, S. Multimodal modelling of human emotion using sound, image and text fusion. Signal Image Video Process. 2024, 18, 71–79. [CrossRef]
12. Hossain, E.; Sharif, O.; Hoque, M.M. Mute: A multimodal dataset for detecting hateful memes. In Proceedings of the 2nd Con ference of the Asia Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing: Student Research

Workshop, Online, 20 November 2022; pp. 32–39.
13. Baltrušaitis, T., Ahuja, C., & Morency, L. P. (2019). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423–443.
14. Barrett, L. F. (2017). How emotions are made: The secret life of the brain. Houghton Mifflin Harcourt.
15. Borth, D., Ji, R., Chen, T., Breuel, T., & Chang, S. F. (2013). Large-scale visual sentiment ontology and detectors using adjective noun pairs. Proceedings of the 21st ACM International Conference on Multimedia, 223–232.
16. Cowie, R., & Cornelius, R. R. (2003). Describing the emotional states expressed in speech. Speech Communication, 40(1–2), 5–32.
17. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT, 4171–4186.
18. Ekman, P. (1999). Basic emotions. In T. Dalgleish & M. Power (Eds.), Handbook of cognition and emotion (pp. 45–60). Wiley.
19. Felbo, B., Mislove, A., Søgaard, A., Rahwan, I., & Lehmann, S. (2017). Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm. Proceedings of EMNLP, 1615–1625.
20. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.
21. Hussain, A., Cambria, E., & Chandra, P. (2018). Affective computing in social media: Enhanced emotion recognition in contextual posts using deep learning. IEEE Computational Intelligence Magazine, 13(3), 45–56.
22. Kim, J., & Provost, E. M. (2014). Emotion classification via utterance-level dynamics: A pattern-based approach. IEEE Transactions on Affective Computing, 5(4), 370–381.
23. Kiros, R., Salakhutdinov, R., & Zemel, R. (2015). Skip-Thought vectors. Advances in Neural Information Processing Systems (NeurIPS), 3294–3302.
24. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. NeurIPS, 1097–1105.
25. Lin, T. Y., et al. (2014). Microsoft COCO: Common objects in context. ECCV, 740–755.
26. Poria, S., Cambria, E., Bajpai, R., & Hussain, A. (2017). A review of affective computing: From unimodal analysis to multimodal fusion. Information Fusion, 37, 98–125.
27. Soleymani, M., Garcia, D., Jou, B., Schuller, B., Chang, S. F., & Pantic, M. (2017). A survey of multimodal sentiment analysis. Image and Vision Computing, 65, 3–14.
28. Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (2nd ed.). MIT Press.
29. Wang, Y., Huang, M., Zhu, X., & Zhao, L. (2016). Attention-based LSTM for aspect-level sentiment classification. Proceedings of EMNLP, 606–615.
30. Zadeh, A., Liang, P. P., Poria, S., Cambria, E., & Morency, L. P. (2018). Multimodal language analysis in the wild: CMU-MOSI dataset and interpretable dynamic fusion. ACL, 3445–3455.
31. Zhang, Y., & Zheng, D. (2020). Multimodal fusion techniques for emotion recognition in social media: A deep learning approach. IEEE Access, 8, 185402–185414.