Journal of Chemical Health Risks

www.jchr.org

JCHR (2023) 13(3), 1280-1291 | ISSN:2251-6727

Social Image Sentiment Analysis by Exploiting Multimodal Content and Heterogeneous Relations

Mohd Amer[1], Dr. Mohammed Jameel Hashmi[2], Dr. C. Kishor Kumar Reddy[3]

PG Scholar, Department of CSE, ISL Engineering College, Hyderabad, India.

Associate Professor, Department of CSE, ISL Engineering College, Hyderabad, India.

Associate Professor; Dept of CSE; Stanley College of Engineering & Technology for Women, Hyderabad, India.

(Received: 04 August 2023 Revised: 12 September Accepted: 06 October)

KEYWORDS

Brassica campestris,
GC-FID,
Erucic acid,
MP-AES,
Heavy metals

ABSTRACT:

Sentiment analysis, because to its potential in comprehending people's attitudes and emotions in the context of social big data, is gaining a lot of interest. As massive data emerges on social platforms with numerous manifestations, traditional sentiment analysis approaches that concentrate on a single modality become inadequate. It is argued in this article that multimodal learning techniques, which focus only on the channels rather than the regions, are necessary to capture the links between images and texts. Furthermore, social photos on the social platforms are interconnected in a variety of linkages that are likewise conducive to sentiment classification but are mostly ignored by current research. To enhance the effectiveness of multimodal sentiment analysis, we suggest a heterogeneous relational model based on the user's attention. To learn the combined image-text representation from the viewpoint of content information, we propose a progressive dual attention module to first capture the connections between picture and text. In this work, we suggest using a channel attention schema to draw attention to semantically rich picture channels, and we develop a region attention schema to draw focus to emotional regions based on the attended channels. We next build a heterogeneous relation network and augment a graph convolutional network to integrate content information from social settings as supplements to develop high-quality representations of social pictures. Experiment findings show that our suggested model is superior to the state of the art, and our approach is fully examined on two benchmark datasets.

1. Introduction

Multimodal social media platforms are gaining popularity as a result of the expansion of the Internet. Users are increasingly relying on data from many modalities, most often a mix of text and graphics, to convey their feelings. Sentiment analysis in social media offers great potential in areas including audience engagement modeling, product development, and market research. The combination of text and visuals has led to a new area of study in the field of sentiment analysis [1]. Early studies mostly dealt with sentiment analysis on one kind of data (text, photos, or videos). Analysis of the emotional tone of a text is called "text sentiment analysis" [2]. Integrating various low-level machine learning models, using neural networks to encode the text contextually, and transforming low-level vector representations into high-level vector representations are all common practices in deep learning-based text sentiment analysis [3, 4]. Sentiment analysis at the sentence and the chapter levels have both benefited greatly from this approach. The IAN, which was suggested in [4] and allows for interactive learning of target and contextual attention, was developed. A multi-attention network (MAN) was presented for aspect-based sentiment analysis in [5], which employs intralayer and interlayer attention methods to address the issue of information loss. Emotions may be read from people's faces and body language with the help of visual sentiment analysis, which employs image processing methods to do so. Sentiment analysis of

Journal of Chemical Health Risks

www.jchr.org

JCHR (2023) 13(3), 1280-1291 | ISSN:2251-6727

massive amounts of visual input has shown promising results in the literature [6,7]. Single-modal data sentiment analysis has been used for things like product recommendations and consumption forecasting, but it has limitations like a low recognition rate and poor generalization because to things like limited data and a lack of variety in expression. The data from many modalities often have a certain complementarity that may make up for the limitations of one another. There is a dearth of complementing modal information in the data for the sentiment analysis job of large-scale visual material in the literature, which results in a less comprehensive model representation. To enhance outcomes in sentiment analysis tasks, researchers in recent years have taken a new approach, concentrating on deep learning-related approaches and several modalities [8-10]. After inputting multimodal data, it is necessary to extract image and textual features, acquire modality-specific representation information, and finally choose the most suitable multimodal data fusion technique to fuse this information together. Sentiment analysis is carried out based on this understanding of the representations. The main flow is shown in Fig. 1.

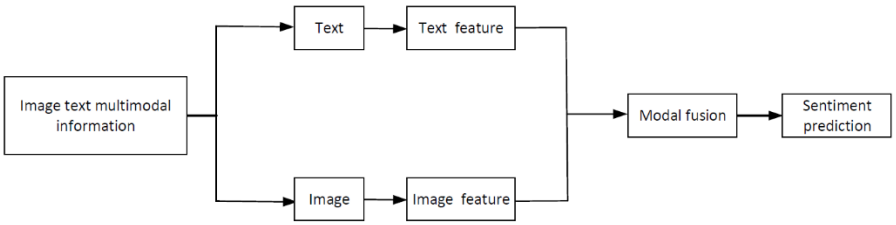


Figure 1. *Illustration of Fusion method.*

Cross-modality consistent regression was suggested by you and your colleagues. For the first time, our CCR model used a sentence-resolution-based semantic tree structure to properly align text and picture regions for analysis [11]. To efficiently fuse the bimodal input of text and pictures and predict overall sentiment, Chen et al. (2017) investigated an end-to-end deep fusion convolutional neural network for co-learning text and visual emotional representations [12]. Recursive neural networks with a tree topology were suggested in [13]. Semantic tree building using phrase parsing and attentional fusion. Using a memory network to facilitate information interaction across modalities and to collect deep semantic properties of pictures and text, the authors of [14] suggested a new multimodal emotion analysis model with encouraging findings.

There are three basic difficulties in multimodal sentiment analysis. First, it has issues with model performance, including slow training speeds, slow inference times, and inaccurate parameter estimates. Due to the increased complexity of modeling many modalities, multimodal models have worse performance. Second, there must be additional sources of labeling data. The Flickr30k dataset [15], the VQA dataset [16], and the CMU-MOSEI dataset [17] are the most popular multimodal datasets currently in use. In addition, a model can only be used effectively in its intended situation of training. This research focuses on the analysis of visual and textual representations of emotion. By using the association information between each modality, multimodal models are able to extract better and more comprehensive feature representations than single-modality models, leading to the desired outcome after fusion. Recently, contrastive learning has risen in favor because to its ability to learn more common features for sentiment analysis tasks, its solution to the issue of little or no labelling, and the construction of comparable examples as positive samples via data augmentation. The CLMLF model makes use of contrastive learning to acquire multimodal emotional representations. For this reason, we devised a supervised comparison learning task with additional loss to put samples of the same type as near as feasible and examples of different kinds as far apart as possible. Our goal is to increase the precision with which classification problems are solved, making the most of contrastive learning by avoiding the pitfalls associated with misclassifying negative samples in supervised sample settings. This research presents a multimodal sentiment analysis strategy to fusing the original Transformer structure using convolutional neural networks (CNNs), based on the problems encountered

Journal of Chemical Health Risks

www.jchr.org

JCHR (2023) 13(3), 1280-1291 | ISSN:2251-6727

in the past. We test the model on the MVSA-Single [16, 17], MVSA-Multiple [18, 19], and HFM datasets [20, 21]. The experimental findings validate the superior detection impact of our suggested model. The advantages of our proposed approach for problem-solving on multimodal sentiment analysis are shown by a thorough series of ablation experiments.

The following statements outline our principal contributions.

- Here we offer a convolutional neural network (CNN) and Transformer-based method for multimodal sentiment analysis to extract richer information from both text and images. Our approach combines local feature extraction from CNN with global feature capture from Transformer to get a more accurate picture of the data.
- For better results, our technique makes use of data augmentation and a supervised contrastive Learning strategy. We use the supervised contrastive learning loss to push embedding vectors within the same class closer together and those across classes farther away. This strategy enhances our system's resilience by more accurately characterizing intra-class similarity.

2. Related Work

2.1. Multimodal Learning

Foreshadowing the current multimodal learning research explosion, they offered a comprehensive overview and forecast for the field in [20]. Category-based Deep Canonical Correlation Analysis (C-DCCA) is a revolutionary deep learning model published in [21] for completing the matching and retrieval of heterogeneous information by projecting the modal information of picture and text into the same semantic space. Multimodal stochastic RNNs networks (MS-RNN) were suggested by Song et al. [22], which take semantic information from two modalities and transform it into text subtitles. Our work, such as multimodal sentiment analysis, is grounded on multimodal fusion, a popular subfield of multimodal learning [23]. Single-modal feature extraction, feature fusion, model classification, and output all make up the multimodal data fusion stage. The process may be broken down into three distinct phases, labeled "early fusion," "late fusion," and "intermediate fusion," respectively. While dot-product operations and series concatenation are effective for early fusion, improved temporal synchronization across multimodal features is desirable. Getting cross-correlation across modes is challenging. Late fusion increases the complexity of fusion by training many modalities with their matching models and then fusing their output findings. The benefits of both of these types of fusion are combined in intermediate fusion. It is preferable to use intermediate fusion techniques when working with deep learning models because of their adaptability and variety. Recent advances in deep learning have allowed for the successful use of structures like convolutional neural networks and attention to intermediate fusion [24].

The current gold standard for fusing graphics and text operates on the assumption that there is a one-to-one relationship between the two. However, in most online social situations, users attach many photographs in order to make the text more vibrant and completely express the emotional content. The addition of these visuals to the text enhances the ability to convey feelings. So, in this study, we look at sentiment analysis tools that don't have access to enough cross-modal fusion feature data [26]. Using a multi-headed attention mechanism and convolutional neural networks (CNN), the picture is matched with the text content, and the cross-modal information of visuals and text is thoroughly merged. In addition, present approaches for graphic emotion classification tend to ignore the special qualities of models in favor of data mining only the association interaction information across modalities.

2.2. Multimodal Sentiment Analysis

To complete user sentiment analysis, multimodal sentiment analysis combines verbal and non-verbal variables. By studying people's body language, facial expressions, conversation, music, and more, researchers in this subject hope to uncover hidden meanings and emotions [27]. There are several publicly available datasets in this area, but some of the more popular ones include CMUMOSEI [17], Memotion Analysis [28], CH-SIM [29], CMU-MOSEAS [30], B-T4SA [31], and MEMOTION 2 [32]. In order to extract deep semantic characteristics from photos, Xu et al. [9] developed the Multisentinet deep semantic network, which they

Journal of Chemical Health Risks

www.jchr.org

JCHR (2023) 13(3), 1280-1291 | ISSN:2251-6727

utilized to extract objects and scenes from photographs. When compared to conventional feature extraction approaches, Multisentinet's proposal excels thanks to the strong relationship between the derived semantic characteristics and human emotions. In the progress of this sector, many outstanding models are actively generated, leading to greater recognition rates and F1 values. In order to enhance performance by more than 5%, Poria et al. [33] introduced an LSTM-based multimodal sentiment analysis model that takes into account the video's context when classifying sentiment. Feature fusion is hierarchical, with no sequential link between distinct modes, and there is still opportunity for development in the underlying feature extraction process in text, audio, and video. Using hybrid fusion to take advantage of natural correlations between visual and textual characteristics for combined sentiment analysis, the model presented in [34] shows promising results even on poorly labeled and regular datasets. The visual aspect attention network (Vistanet) was studied by Truong et al. in 2019 [35]. This network uses attention to steer the model's focus on crucial sentences within the text rather than visual information. The attention-based heterogeneous relational model (AHRM) was suggested in [36] for use in the year 2020. The model improves its performance by integrating rich social information with a progressive dual attention module that captures the association between pictures and text. This allows the model to learn the image-text joint representation from the standpoint of content information. In 2021, Wu et al. [37] set out to create a Shared Private (TCSP) architecture with a primary emphasis on text. The system places special emphasis on text modalities, presents a novel approach for training multimodal sentiment analysis models with unlabeled data, and extracts shared and private semantic information from the other two modalities to enhance modal text information. For their study, Tan et al. [38] looked into the following aspects of multimodal emotion recognition: face and EEG; facial feature extraction with CNN; final classification with soft-max; multiple voting in the late fusion layer; final classification results of the two modalities combined with the threshold method; and statistical simulation methods to obtain the final multimodal emotion classification results. In [39], the authors proposed a social network-based, real-time traffic accident monitoring framework that makes use of sentiment analysis technology to determine the emotional tone of user comments in the aftermath of traffic accidents. The proposed word embedding approach takes both formally and informally written words and transforms them into low-dimensional vector representations to aid in classification task performance. Multi-layer fusion, contrastive learning, and multi-modal sentiment analysis are all elements that Li et al. Token-level feature fusion, it is claimed, is more suitable than the preceding feature fusion layer [40] for merging local information in pictures and text.

Currently, most image-text sentiment analysis algorithms successfully integrate multiple modalities by realizing the interaction between modal features by sharing the feature vectors of the neural network representation layer. While previous work has made progress in terms of accuracy, challenges such as sparse training data and insufficient support for multi-level feature information in each modality persist. In this study, we focus on addressing the issues of irrelevant information interference and incomplete fusion features, allowing the model to pick up on more shared emotional characteristics during the fusion phase. We suggested the MLFC module, and its efficacy was confirmed, based on the models found in the aforementioned literature and our study of the current difficulties.

Multimodal sentiment analysis is a synthesis of multimodal polarity analysis and multimodal emotion analysis. The problem of multimodal sentiment analysis is approached with the use of standard machine learning techniques (P'erez-Rosas et al., 2013; You et al., 2016). There has been recent progress in this area using deep learning models. Hazarika et al. (2020) built a novel framework, MISA, which projects each modality to two distinct subspaces: modalityinvariant and modality-specific subspaces; and Wang et al. (2020b) proposed a novel method, TransModality, to fuse multimodal features with end-to-end translation models for the video dataset. Due to the vast amounts of image-text data available on social media platforms, academics have been more interested in doing image-text multimodal sentiment analysis. For multimodal sentiment analysis, Xu et al. (2017) built three distinct networks: HSAN, MDSN, and Co-Mem (2018). Additionally, Yang et al. (2020) introduced MVAN for

Journal of Chemical Health Risks

www.jchr.org

JCHR (2023) 13(3), 1280-1291 | ISSN:2251-6727

multimodal emotion analysis and created TumEmo, an image-text emotion dataset.

2.2 Graph Neural Network

For text classification, multi-label identification, and multi-modal tasks specifically, the Graph Neural Network has shown encouraging results. Graph Neural Networks (GNNs) and its derivatives, such as Text GCN (Yao et al., 2019), TensorGCN (Liu et al., 2020), and TextLevelGNN (Huang et al., 2019), have been quickly developed and show improved performance over conventional approaches for text categorization. To represent the label dependencies, the GCN is also implemented in the multi-label image recognition job (Chen et al., 2019). These include Visual Question Answering (VQA) (Hudson and Manning, 2019; Khademi, 2020), Multimodal Fake News Detection (Wang et al., 2020a), and Visual Dialog (Guo et al., 2020; Khademi, 2020). Jiang et al. (2020) used a cutting-edge Knowledge- Bridge Graph Network (KBGN) to simulate the intricate connections between different types of data in a visual conversation. To model semantic representations for false news identification, Wang et al. (2020a) suggested a new Knowledge-driven Multimodal Graph Convolutional Network (KMGCN).

However, the KMGCN only used the visual words that were retrieved as visual information, rather than the whole picture. The Multimodal Neural Graph Memory Network (MNGMN) proposed by Khademi (2020) is a novel neural network architecture designed specifically for visual quality assessment (VQA). This model builds a visual graph network from the bounding-boxes, which might result in duplicate information due to overlap between nodes.

For the picture-text dataset, we discovered that certain co-occurrences of words in a text post and certain co-occurrences of objects or sceneries inside an image are indicative of particular emotions. Using a multichannel GNN, we explicitly model these dataset-wide features.

3. Materials and Methods

Our goal is to train a feature embedding network using tagged multimodal data. Embedding vectors of the same class ought to be spatially near together, whereas those of different classes ought to be widely spaced away [41]. Our technique makes use of supervised comparison learning to fine-tune the supervised categorization. Classification uses a set of input data to do data augmentation. The embedding vector for a given instance does not vary under varied text and picture data improvement, whereas the embedding vector for other examples may change. To acquire the feature sequence after fusing visuals and text, as shown in Figure 2, the original graphic example and the improved example must be input into the MLFC module. As a last step, we include the SCSupCon module into the feature space, positioning related features near together and dissimilar classes far apart. The embedding vectors are mapped to the output layer via the multilayer fusion module. The network's output is used to compute a cross-entropy loss and a supervised comparative learning loss. Here, we'll lay out the basic structure, explain how we keep an eye on the comparison loss and multilayer fusion modules, and describe how we optimize our multilayer convergence strategy.

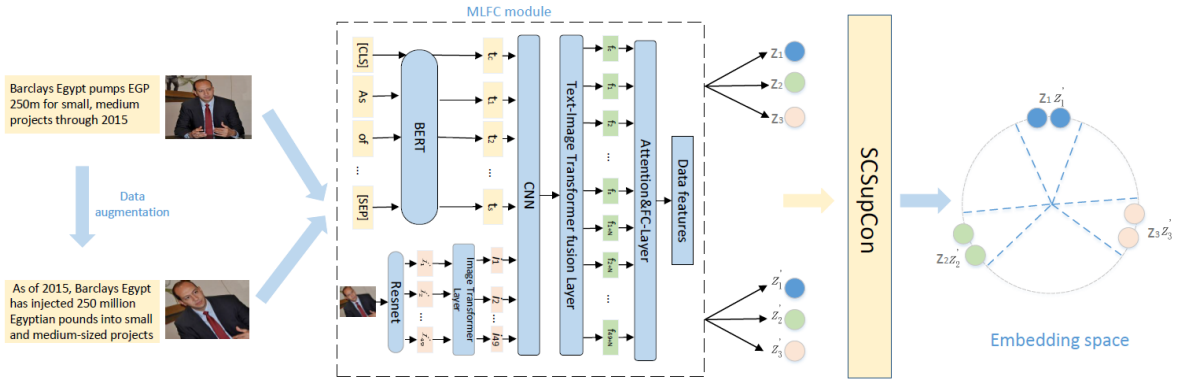


Figure 2: Frame work Diagram

Journal of Chemical Health Risks

www.jchr.org

JCHR (2023) 13(3), 1280-1291 | ISSN:2251-6727

3.1. Representation Learning Framework

Inspired by modern contrastive learning approaches, self-supervised learning utilizes a contrastive approach. To compensate for the increased focus on invariant characteristics after fusing data from many modalities, SCSupConLoss uses potential contrast loss to its fullest. These unaltered characteristics provide emotional underpinnings to guarantee that the user's intended meaning will remain consistent despite the textual shift. As can be seen in Figure 2, the overall model structure consists of four distinct parts.

3.1.1. Data Augmentation

In data augmentation, also known as data enrichment, existing data is used to create the impression of greater depth and breadth without actually expanding the dataset. Data augmentation helps your model be more stable and generalizable. Data augmentation is necessary due to the mixed-modal nature of the data (text and pictures), as well as the increased emphasis placed on fixed attributes after data fusion.

Since back-translation may provide alternative interpretations while keeping the original sentence's semantics intact, it can be used to improve text data [42]. The term "back translation" refers to the act of translating a statement into another language and then re-translating it back into the original language to compare the similarities and differences between the two versions. Because it is more succinct and straightforward to sample consistently from the same set of enhancement transformations [43], R and Augmentation was selected as the picture data enhancement approach. With R and Augmentation, data augmentation strategies are learned through a simple raster search, and the incremental simple grid search is greatly reduced by data augmentation, as it is integrated into the model training process rather than executed as a separate preprocessing task.

3.1.2. Encoder Network

While there are many different ways to interpret the input text and generate the picture, the encoder network is largely responsible for this step. Data sets of picture-text pairs are sent into the encoder network, and features are drawn from them using the image selection Resnet-50 model. Resnet's last convolutional layer yields the image feature M′c.

3.1.3. Multi-Layer Fusion Convolution Neural Network

The encoder output is transformed into the final multimodal representation R using the Multi-Layer Fusion Convolution Neural Network (MLFC), a multilayer fusion module. Its primary use is in the alignment and integration of data from different sources. MLFC consists of a Convolutional Neural Network, a Transformer Encoder, and an Attention Layer. The convolution operation is introduced before the Transformer encoder because the Transformer is constrained in that it can only collect long-range feature dependencies and ignore the specifics of local features. While the convolution technique is useful for extracting local characteristics, it falls short when trying to capture a representation of global features.

4. Experiment

The suggested model was developed in the Pytorch framework, and its major metrics are ACC and F1. One NVIDIA 3090 GPU and one A40 GPU were used in our study on a high-performance computing (HPC) node. While the parameters of the pre-trained BERT model and the Resnet-50 model remain constant during training, the parameters of the word embedding and the attribute embedding are continuously updated [19]. Glove [47] was used to train word and attribute embeddings on the Twitter dataset. We employ Weighted-F1 [48] on the MVSA dataset and Macro-F1 [50] on the HFM dataset.

4.1. Data Processing

Both the MVSA-Single and MVSA-Multiple datasets are freely accessible resources for researchers in the area of multimodal sentiment analysis; they were amassed from tweets that include text, photos, hashtags, and more. There is a separate emotion label for each set of words and pictures. The MVSA dataset is annotated with positive, neutral, and negative sentiment labels by hand. Separate from the 19,598 image-text pairings classified by three annotators with three sentiment labels is the MVSA-Multiple (MVSA-M) dataset, which is comprised of 4869 image-text pairs identified by a single annotator with a single sentiment label. The second part is MVSAMultiple, which uses 19,598 image-text pairings annotated by three annotators and three emotion labels.

Journal of Chemical Health Risks

www.jchr.org

JCHR (2023) 13(3), 1280-1291 | ISSN:2251-6727

We use the identical procedures for processing the two original MVSA datasets [9] describes. Using a split ratio of 8:1:3, we randomly segment the MVSA dataset into training, validation, and test sets. We employed the identical data preparation approach published in the literature [19] for HFM datasets, with the exception that we randomly partitioned the datasets. To confirm the correctness of the labels, both the training and validation sets underwent human inspection. The NLTK library is the go-to for parsing text into individual words, emojis, and labels. The labels are separated by the character #, and all capital letters are lowered to lowercase.

MLF is an attention alignment and fusion graphical features-based multi-layer fusion technique developed from Transformer-Encoder. MLFC is a fusion technique that builds on top of MLF by using convolutional techniques. Below is a table detailing the remaining hyperparameters. Our first solution was trained and tested on the MVSA dataset. In particular, we present the MVSA and HFM F1 and ACC values from our analysis. Each and every experiment is conducted in a controlled setting. The experimental setup is shown in full in Table 1. There are a few different ways that text and picture data may be improved, but two of the most common are back-translation and RandAugment. Sentiment analysis loss and supervised vs unsupervised learning loss are both included in the "SCSupCon" loss naming scheme for this design; the table additionally displays the distinct polarity of sentiment in each of the three datasets.

4.2. Comparative Experiments

We began with MVSA-Single, and then expanded to MVSA-Multiple and HFM to demonstrate the model's efficacy. Our model was able to get currents similar to the SOTA model, as shown by the findings. We evaluate our model against both the unimodal and multimodal normative data sets. The data are presented in tabular form below.

Table 1. Heterogeneous Relations Setting Display.

Heterogeneous Relations	**MVSA-Single**	**MVSA-Multiple**	**HFM**
Text data augmentation	back-translation	back-translation	back-translation
Image data augmentation	R and Augment	R and Augment	R and Augment
Emotional polarity	3	3	2
Loss	Conloss/SCSupConloss	Conloss/SCSupConloss	Conloss/SCSupConloss
Contrasting learning styles	Self-Supervised/Supervised	Self-Supervised/Supervised	Self-Supervised/Supervised
Text Encoder	BERT	BERT	BERT
Image Encoder	Resnet-50	Resnet-50	Resnet-50
Fuse model epoch	20	20	20
Epoch	30	30	30
Integration method	MLF/MLFC	MLF/MLFC	MLF/MLFC
Optimizer	Adams	Adams	Adams
Learning rate	2 × 10−5	2 × 10−5	2 × 10−5
Batch	32/32	64/64	128/48

The importance of the text model is shown by comparing the performance of the image-based model with the text adoption CNN. Instead of building a single graph for the whole corpus, the new GNN-based TGNN model builds graphs with shared global parameters for each input text. This technique removes the cost of intertextual dependencies while maintaining global information for online testing. The Resnet and the OSDA models are selected as the default ones for use with photos. Focusing on both targets and scenes, the OSDA model analyzes user emotions by using visual attributes from a variety of angles [14].

Journal of Chemical Health Risks

www.jchr.org

JCHR (2023) 13(3), 1280-1291 | ISSN:2251-6727

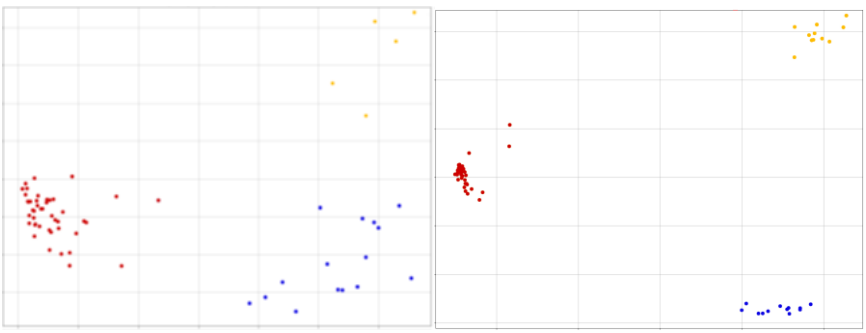
The MVSA dataset was used, and multimodal models MGNNS and CLMLF were selected. Sentiment analysis in images and texts may be performed using the MGNNS model, which is a multi-channel graph neural network. The fusion strategy of the model is accomplished by the multi-head attention mechanism, and the multi-channel graph neural network learns global properties of multimodal representation [48]. In order to align and fuse token-level properties of text pictures and two contrast, the CLMLF model employs multilayer fusion. The D&R Net paradigm provides both cross-modal contrast and semantic connection via the use of a decomposition and relational network. The deconstruction network depicts the similarities and contrasts between visuals and text, while the relationship network provides the semantic linkage in a cross-modal environment.

To begin gauging our model's performance on the MVSA dataset, we've included a table displaying the outcomes of our experiments. Here we have a hierarchy of models that are representational in terms of text, visuals, and multimodality. The low information density of images makes it difficult to capture feature information, and sentiment analysis has a particularly negative impact on picture mode. Our approach improves upon the state-of-the-art SOTA model on the MVSA-Single dataset by 1.10% on ACC and 2.15% on F1. The F1 improvement is 0.52%, which is on par with the MVSA-Multiple dataset, and the ACC impact is similar. Our concept is quite similar to SOTA in many respects.

Suggested model for multimodal sentiment analysis enhances performance by 1.21% and 1.35%, respectively. To begin with, the suggested Multi-Layer Fusion Convolution Neural Network (MLFC) is better able to capture the interplay between visuals and text. Then, a more precise representation of the same type may be obtained by supervising the contrast-loss sentiment classification and de-motivation encoder. To improve the accuracy of sentiment prediction in multimodal sentiment analysis, SCSupCon analyses samples at the decision boundary.

4.4. Data Visualization

We conduct feature space visualization experiments on the dataset to see whether our supervised comparison learning task can help the machine learn common features linked to sentiment in multimodal data. Data feature vectors may be seen by lowering their dimensionality in the last layer of the model. The model's cross-entropy loss is seen in Figure 3a, and its SCSupConLoss findings are visualized in Figure 3b. Figure indicates that by including supervised comparison learning, the gap between happy and sad feelings in the vector space widens and more information is collected. This shows that the model is able to differentiate between these data in vector spaces using common characteristics in the sentiment data. This shows that the model is able to differentiate between these data in vector spaces using common characteristics in the sentiment data. Our model's visualization collects together neutral sentiment data rather of distributing it over the vector space as if it were merely cross-entropy losses due to the tiny quantity of data available for neutral sentiment. All of this points to supervised contrastive learning as a means to boost the model's performance by teaching it to recognize common characteristics linked with emotions.



(**a**) Cross-entropy (**b**) SCSupConLoss

Figure 3. Loss function clustering visualization

Journal of Chemical Health Risks

www.jchr.org

JCHR (2023) 13(3), 1280-1291 | ISSN:2251-6727

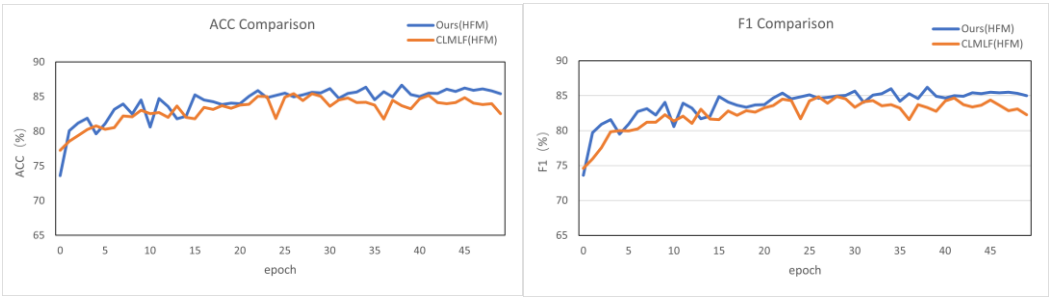


Figure 4. Accuracy and F1 value plot.

After 50 training epochs on the HMM test set, the accuracy and F1 value curves of our model and CLMLF are shown in Figure 4. In terms of accuracy as a percentage of the whole sample of accurate predictions, as shown in Figure 4a, our model has a shorter growth curve and greater accuracy than CLMLF. After 40 epochs, our suggested model's performance has less volatility than CLMLF. The F1 Score takes a harmonic mean of the classification model's accuracy and recall to provide a quantitative evaluation of the model's efficacy. As can be seen in Figure 4b, our model stabilizes at an F1 value higher than the CLMLF of 1.35% after 40 epochs, producing more accurate classifications than the baseline.

5. Conclusions

To improve the efficiency of multimodal sentiment analysis, we present a technique based on supervised contrastive learning. By using supervised contrastive learning and a multilayer fusion module, our model is able to solve the issues of inadequate multimodal fusion feature information and fusion feature interference information. To perform the sentiment analysis job without being impacted by extraneous data, our model blends convolutional neural networks (CNNs) with a multi-layer fusion module to extract local features. This allows us to more accurately gather psychological characteristics. In addition to using the Transformer to improve the fusion effect and generate more digestible graphical characteristics, we employ it in our model to gain global features. We use supervised contrastive learning and data augmentation to minimize noise and emphasize key characteristics. On both the MVSA and HFM datasets, our proposed model improves upon the detection performance as measured by Acc and F1.

Future work on the current encoder network will take into account the feature extraction strategy that is superior for multimodal data in an effort to address the feature redundancy issue for sentiment analysis applications. To address the issue of insufficient cross-modal feature fusion, we will also explore novel multimodal fusion techniques and think about including an audio modality.

References

[1] Kaur, R.; Kautish, S. Multimodal sentiment analysis: A survey and comparison. In Research Anthology on Implementing Sentiment Analysis Across Multiple Disciplines; IGI Global: Hershey, PA, USA, 2022; pp. 1846–1870.

[2] Balazs, J.A.; Velasquez, J.D. Opinion mining and information fusion: A survey. Inf. Fusion **2016**, 27, 95–110.

[3] Ke, Z.; Sheng, J.; Li, Z.; Silamu, W.; Guo, Q. Knowledge-guided sentiment analysis via learning from natural language

[4] explanations. IEEE Access **2021**, 9, 3570–3578.

[5] Ma, D.; Li, S.; Zhang, X.; Wang, H. Interactive attention networks for aspect-level sentiment classification. arXiv **2017**, arXiv:1709.00893.

[6] Xu, Q.; Zhu, L.; Dai, T.; Yan, C. Aspect-based sentiment classification with multi-attention network. Neurocomputing **2020**, 388, 135–143.

[7] Jindal, S.; Singh, S. Image sentiment analysis using deep convolutional neural networks with domain specific fine tuning. In Proceedings of the 2015 International Conference on

Journal of Chemical Health Risks

www.jchr.org

JCHR (2023) 13(3), 1280-1291 | ISSN:2251-6727

Information Processing (ICIP), Pune, India, 16–19 December 2015; pp. 447–451.

[8] Yang, J.; She, D.; Sun, M.; Cheng, M.M.; Rosin, P.L.; Wang, L. Visual sentiment prediction based on automatic discovery of affective regions. IEEE Trans. Multimed. **2018**, 20, 2513–2525.

[9] Xu, N. Analyzing multimodal public sentiment based on hierarchical semantic attentional network. In Proceedings of the 2017 IEEE International Conference on Intelligence and Security Informatics (ISI), Beijing, China, 22–24 July 2017; pp. 152–154.

[10] Xu, N.; Mao, W. Multisentinet: A deep semantic network for multimodal sentiment analysis. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, Singapore, 6–10 November 2017; pp. 2399–2402.

[11] Hafsa Fatima, Shayesta Nazneen, Maryam Banu, Dr. Mohammed Abdul Bar," Tensorflow-Based Automatic Personality Recognition Used in Asynchronous Video Interviews", Journal of Engineering Science (JES), ISSN NO:0377-9254, Vol 13, Issue 05, MAY/2022

[12] Mohammed Shoeb, Mohammed Akram Ali, Mohammed Shadeel, Dr. Mohammed Abdul Bari, "Self-Driving Car: Using Opencv2 and Machine Learning", The International journal of analytical and experimental modal analysis (IJAEMA), ISSN NO: 0886-9367, Volume XIV, Issue V, May/2022

[13] Mr. Pathan Ahmed Khan, Dr. M.A Bari,: Impact Of Emergence With Robotics At Educational Institution And Emerging Challenges", International Journal of Multidisciplinary Engineering in Current Research(IJMEC), ISSN: 2456-4265, Volume 6, Issue 12, December 2021,Page 43-46

[14] Mohammed Abdul Bari, Shahanawaj Ahamad, Mohammed Rahmat Ali," Smartphone Security and Protection Practices", International Journal of Engineering and Applied Computer Science (IJEACS) ; ISBN: 9798799755577 Volume: 03, Issue: 01, December 2021

[15] Yu, Y.; Lin, H.; Meng, J.; Zhao, Z. Visual and textual sentiment analysis of a microblog using deep convolutional neural networks. Algorithms **2016**, 9, 41.

[16] Dr. K. Shashidhar, B Benhur, Khaja Moinuddin, Rizwan Ahmad, Design And Implementation Of Iot Based Smart Health Care Monitoring System, International Journal of Multidisciplinary Engineering in Current Research - IJMEC Volume 8, Issue 1, January-2023, http://ijmec.com/, ISSN: 2456-4265.

[17] You, Q.; Luo, J.; Jin, H.; Yang, J. Cross-modality consistent regression for joint visual-textual sentiment analysis of social multimedia. In Proceedings of the Ninth ACM International Conference onWeb Search and Data Mining, San Francisco, CA, USA, 22–25 February 2016; pp. 13–22.

[18] Chen, X.;Wang, Y.; Liu, Q. Visual and textual sentiment analysis using deep fusion convolutional neural networks. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 1557–1561.

[19] You, Q.; Cao, L.; Jin, H.; Luo, J. Robust visual-textual sentiment analysis: When attention meets tree-structured recursive neural networks. In Proceedings of the 24th ACM International Conference on Multimedia, Amsterdam, The Netherlands, 15–19 October 2016; pp. 1008–1017.

[20] Mr. Saiful Islam, Shaik Aman, Ibrahim Naushad, Rahmath Baig, TRAFFIC CONTROLLER BASED ON RF, International Journal of Multidisciplinary Engineering in Current Research - IJMEC Volume 8, Issue 1, January-2023, http://ijmec.com/, ISSN: 2456-4265.

[21] Yang, X.; Feng, S.;Wang, D.; Zhang, Y. Image-text multimodal emotion classification via multi-view attentional network. IEEE Trans. Multimed. **2020**, 23, 4014–4026.

[22] Plummer, B.A.;Wang, L.; Cervantes, C.M.; Caicedo, J.C.; Hockenmaier, J.; Lazebnik, S. Flickr30k entities: Collecting region-tophrase correspondences for richer image-to-sentence models. In Proceedings of the IEEE

Journal of Chemical Health Risks

www.jchr.org

JCHR (2023) 13(3), 1280-1291 | ISSN:2251-6727

International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2641–2649.

[23] Narsaiah Domala, Radwan Jameel, Mohammed Yamin Ahmed, Mohammed Mujtaba Ahmed, Real Time Pantry Information System Using Iot, International Journal of Multidisciplinary Engineering in Current Research - IJMEC Volume 8, Issue 2, February-2023, http://ijmec.com/, ISSN: 2456-4265.

[24] Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; Parikh, D. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6904–6913.

[25] Zadeh, A.B.; Liang, P.P.; Poria, S.; Cambria, E.; Morency, L.P. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Melbourne, Australia, 15–20 July 2018; pp. 2236–2246.

[26] Dr Altaf C, Syed Muzammil Ahmed,Mir Mohammed Asad Ali , Rayan Razvi, Automatic Railway Gate Controlling Using IR Sensors And Microcontroller**,** International Journal of Multidisciplinary Engineering in Current Research - IJMEC Volume 8, Issue 3, March-2023, http://ijmec.com/, ISSN: 2456-4265.

[27] Dr. K . Shashidar, Mohammed Saakhib, Shaik Mansoor Basha, PV Soalr Panel Grid And Battery Voltage Monitoring Over IOT**,** International Journal of Multidisciplinary Engineering in Current Research - IJMEC Volume 8, Issue 3, March-2023, http://ijmec.com/, ISSN: 2456-4265.

[28] Niu, T.; Zhu, S.; Pang, L.; Saddik, A.E. Sentiment analysis on multi-view social data. In Proceedings of the International Conference on Multimedia Modeling; Springer: Cham, Switzerland, 2016; pp. 15–27.

[29] Cai, Y.; Cai, H.; Wan, X. Multi-modal sarcasm detection in twitter with hierarchical fusion model. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Florence, Italy, 28 July–2 August 2019; pp. 2506–2515.

[30] Ngiam, J.; Khosla, A.; Kim, M.; Nam, J.; Lee, H.; Ng, A.Y. Multimodal deep learning. In Proceedings of the ICML, Bellevue, WA, USA, 28 June–2 July 2011.

[31] Yu, Y.; Tang, S.; Aizawa, K.; Aizawa, A. Category-based deep CCA for fine-grained venue discovery from multimodal data. IEEE Trans. Neural Netw. Learn. Syst. **2018**, 30, 1250–1258.

[32] Song, J.; Guo, Y.; Gao, L.; Li, X.; Hanjalic, A.; Shen, H.T. From deterministic to generative: Multimodal stochastic RNNs for video captioning. IEEE Trans. Neural Netw. Learn. Syst. **2018**, 30, 3047–3058.

[33] Morency, L.P.; Mihalcea, R.; Doshi, P. Towards multimodal sentiment analysis: Harvesting opinions from the web. In Proceedings of the 13th International Conference on Multimodal Interfaces, Alicante, Spain, 14–18 November 2011; pp. 169–176.

[34] Valada, A.; Oliveira, G.L.; Brox, T.; Burgard, W. Deep multispectral semantic scene understanding of forested environments using multimodal fusion. In Proceedings of the International Symposium on Experimental Robotics, Nagasaki, Japan, 3–8 October 2016; Springer: Cham, Switzerland, 2016; pp. 465–477.

[35] Majumder, N.; Hazarika, D.; Gelbukh, A.; Cambria, E.; Poria, S. Multimodal sentiment analysis using hierarchical fusion with context modeling. Knowl.-Based Syst. **2018**, 161, 124–133.

[36] Xi, C.; Lu, G.; Yan, J. Multimodal sentiment analysis based on multi-head attention mechanism. In Proceedings of the 4th International Conference on Machine Learning and Soft Computing, Haiphong City, Vietnam, 17–19 January 2020; pp. 34–39.

Journal of Chemical Health Risks

www.jchr.org

JCHR (2023) 13(3), 1280-1291 | ISSN:2251-6727

[37] Li, Z.; Li, X.; Sheng, J.; Slamu, W. AgglutiFiT: Efficient low-resource agglutinative language model fine-tuning. IEEE Access **2020**, 8, 148489–148499.

[38] Sharma, C.; Bhageria, D.; Scott,W.; Pykl, S.; Das, A.; Chakraborty, T.; Pulabaigari, V.; Gamback, B. SemEval-2020 Task 8: Memotion Analysis—The Visuo-Lingual Metaphor! arXiv **2020**, arXiv:2008.03781.

[39] Yu, W.; Xu, H.; Meng, F.; Zhu, Y.; Ma, Y.; Wu, J.; Zou, J.; Yang, K. Ch-sims: A chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 5–10 July 2020; pp. 3718–3727.

[40] Zadeh, A.; Cao, Y.S.; Hessner, S.; Liang, P.P.; Poria, S.; Morency, L.P. CMU-MOSEAS: A multimodal language dataset for Spanish, Portuguese, German and French. In Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing, Online, 16-20 November 2020; NIH Public Access: Stroudsburg, PA, USA, 2020; Volume 2020, p. 1801.

[41] Lopes, V.; Gaspar, A.; Alexandre, L.A.; Cordeiro, J. An AutoML-based approach to multimodal image sentiment analysis. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 18–22 July 2021; pp. 1–9.

[42] Ramamoorthy, S.; Gunti, N.; Mishra, S.; Suryavardan, S.; Reganti, A.; Patwa, P.; Das, A.; Chakraborty, T.; Sheth, A.; Ekbal, A.; et al. Memotion 2: Dataset on Sentiment and Emotion Analysis of Memes. In Proceedings of De-Factify: Workshop on Multimodal Fact Checking and Hate Speech Detection, CEURc, Vancouver, BC, Canada, 22 February–1 March 2022.

[43] Poria, S.; Cambria, E.; Hazarika, D.; Majumder, N.; Zadeh, A.; Morency, L.P. Context-dependent sentiment analysis in usergenerated videos. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Vancouver, BC, Canada, 30 July–4 August 2017; pp. 873–883.

[44] Huang, F.; Zhang, X.; Zhao, Z.; Xu, J.; Li, Z. Image–text sentiment analysis via deep multimodal attentive fusion. Knowl.-Based Syst. **2019**, 167, 26–37.

[45] Truong, Q.T.; Lauw, H.W. Vistanet: Visual aspect attention network for multimodal sentiment analysis. In Proceedings of the AAAI Conference on Artificial Intelligence, Honolulu, HI, USA, 27 January–1 February 2019; Volume 33; pp. 305–312.