Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 4, Issue 2, 2026

SENTIMENT ANALYSIS IN SOCIAL MEDIA USING MULTI-MODAL DATA FUSION TECHNIQUES

[1]Muhammad Zarar, [2]Talha Nadeem, [3]Huzaifah Bin Khawar, [4]Hood Khizer
[1]*AI Researcher, Research and Development Department, Trouve Labs, AHOY DMCC and The Superior University Lahore, Pakistan*
[2]*Systems Architect AI/ML, Research and Development Department, Trouve Labs, AHOY DMCC and The Superior University Lahore, Pakistan*
[3]*AI Researcher, Research and Development Department, Trouve Labs, AHOY DMCC and The Superior University Lahore, Pakistan*
[4]*CEO Trouve Labs, CTO AHOY DMCC, Research and Development Department, Trouve Labs, AHOY DMCC, Dubai*
[1]muhammadzarar.ieee@gmail.com [2]talhanadeem.ieee@gmail.com
[3]huzaifahkhawar.ieee@gmail.com[4]hoodkhizer@gmail.com

DOI: https://doi.org/10.5281/zenodo.21458995

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

Article History
Received on: 10 Feb, 2026
Accepted on: 18 Feb, 2026
Published on 28 Feb, 2026

Copyright @Author
Corresponding Author

Abstract
In recent years, with the popularity of social media, users are increasingly keen to express their feelings and opinions in the form of pictures and text, which makes multimodal data with text and pictures the con tent type with the most growth. The bulk of information written on the social media by users is clearly sentimental in nature and multimodal sentiment analysis has emerged as a vital area of study. Receiving the text and image feature separately then integrating them to classify the sentiment has been the main focus of previous studies on multimodal sentiment analysis. These researches tend to overlook the interrelationship between text and images. Thus, the new model of multimodal sentiment analysis is suggested in this paper. The model initially removes noise interference in text related information and cleanses more significant image features. After that, under the feature-fusion step which is founded on the attention mechanism, the text and images teach the inner features of each other with the help of symmetry. Then the fusion features are implemented on sentiment classification problems. The proposed model proves to be effective as evident in the experimental findings of the two popular multimodal sentiment datasets.

1. INTRODUCTION

As social media continues to grow in popularity, people are becoming keen to share their opinions on different mediums. There is a great amount of multimodal data produced every day by hundreds of millions of data records, a large share of which are text and image mixed. Sentimental information can be extracted in this data, which can aid us in knowing the attitude and opinions of people towards different events through sentiment analysis. Multiple applications Multimodal sentiment analysis Multimedia Multimodal sentiment analysis has important applications in box office predictions, political campaigns, and stock market forecasting. The sentimental information presented in both the text and image of the graphic data that are presented in socio media is unique but complementary. The multimodal data are also more valuable in understanding the real feelings of users as compared to single-mode data. But there are a number of challenges in doing multimodal sentiment analysis. There is variation in sentimental information in various modal data and analysis of sentimental features cannot be done effectively without appropriate representation of sentimental features.

The human view of everything in an image as a source of sentiment expression is false, as contextual data is not all relevant to the sentiment. Thus, in the process of feature extraction, it is important to highlight the main sentimental regions of the images at the minimal cost of noise in written materials. Different modal data provide underlying features in different dimensions and attributes also. Statistically-based sentiment analysis of single-mode texts has traditionally been heavily dependent on statistical techniques, which are typically constrained by the quality of the extracted features. As an example, scholars have used data envelopment analysis as an instrument of word sense disambiguation and have tried to find out what words mean in context, and thus determine voter intentions towards political candidates (Rodger et al.). Likewise, sentiment analysis of single-mode images relies on how well manually developed extraction rules can work, and can cause redundant sentential information.

As machine learning and deep learning tools have developed, a variety of both novel and improved techniques to sentiment analysis with multimodal data have surfaced, which have demonstrated encouraging outcomes. The field of multimodal sentiment analysis is rapidly gaining relevance and complexity and several studies have been conducted to uncover its complexity. One of the major issues related to multimodal sentiment analysis is the effective exploitation of internal and interactive data of various modalities. The internal information is the data which can be mined in the individual mode and the interaction information is the product of relationships and links among modalities.

The choice of the way to utilize the interactive information using different modalities is what separates multimodal tasks and single-modal tasks. As a result, successful multimodal affective analysis should determine ways to incorporate the information in each mode effectively. Through their structural similarities with different modal data, researchers can capture necessary information that the traditional tasks tend to ignore and as a result of this, modality information can be fused with inter-modal interaction to its full potential.

But lots of past research have not sufficiently made use of both the internal and the interaction information of both the text and the images modes. Deep semantic networks have been suggested by some researchers in multimodal sentiment analysis, using long short-term memory (LSTM) models with attention mechanisms to get keywords. Although

this method only considers the directional impact of images on the text, it does not address the cause-and-effect relationship between text and images (Xu et al., 2019). Other research has also used so-called co-memory networks to learn the interaction information between text and image content, and then used the fused features to classify entitled sentiments. However, the rugged attention mechanism involved in those methods can be a detriment to certain features and cause possible loss of information (Xu et al., 2020).

In order to solve these issues, this paper presents a multimodal sentiment classification model that is based on a fine-grained mechanism of attention. First, denoising autoencoder is used to de-noise textual data by extracting features that depict the original text in a more accurate way. Second, an enhanced variational autoencoder, together with the attention mechanism, is used in deriving features of images. The fusion model using the attention mechanism is then proposed, which learns the interactive modal fusion representation vector, which is a successful combination of the only pertinent elements of the text and image data.

The findings of the paper can be summed up as the following:

- It uses denoising autoencoder and a better variational auto encoder with an attention to extract finer features of texts and images.
- A brand new multimodal cross-feature fusion model using the attention mechanism that supports successful fusion of various modal features providing more accurate sentiment classification information.
- The suggested approach was tested using two standard Twitter datasets. It was experimented comparatively and proved to be feasible and better than the proposed model.

The rest of the article follows a following set of arrangement: The discussion of current research in the field of multimodal sentiment analysis, the effective description of the suggested classification model, the commentary of the results of the experimental work with the two Twitter datasets and a conclusion of the findings and the recommendations made regarding future research in the field.

2. Related Work

The use of traditional sentiment analysis has been limited to text data. Nevertheless, the fast development of social media has enabled people to shop their views and opinions, using an assortment of formats, such as text, images, audio and video. This trend has resulted in heightened interest in multimodal sentiment analysis especially in the combination of text and image data.

Past studies primarily used conventional feature-based approaches to classification of data in sentiment classification. In one example, Borth et al. used machine learning to extract 1,200 adjectives noun pairs, and then added these mid-level features to existing image features to obtain sentimental scores. Likewise, Cao et al. used a sentiment dictionary to extract textual features, with adjective-noun pairs they used to create a visual sentiment ontology of images, then fused with text features to classify sensations. Poria et al. delved into sentiment analysis in which they derived text, voice, and video features and combined them into a support vector machine (SVM) classifier. Another method by Wang et al. seeks to build a coherent cross-media bag-of-words model to obtain features of both text and image and then use the logistic regression, SVM and Naive Bayes classifiers to do comparative analysis. The model by Baecchi et al. was a semi-supervised sentiment classifier named CBOW-DA-LR that built upon the Continuous Bag of Words (CBOW) model, employed a denoising autoencoder to obtain image features, and classified the latter

image features and textual features jointly, whether in an unsupervised or semi-supervised fashion.

With the development of machine learning and deep neural networks, multimodal sentiment analysis has made significant headway utilizing deep learning methods. As an example, you et al. introduced a cross-modal consistent regression (CCR) approach which states that there is a requirement of consistent sentimental tendencies among image and text and their features. This method brought about a constraint in performance of the fusion process to create more coherent feature vector used in sentiment classification. You et al. also described different research, a TreeLSTM-based model, using a tree structure and a visual attention system to evaluate the polarity of the sentiment of text and the images based on their internal relationships, and it was quite successful.

Chen et al. used the attention mechanism with SVM and convolutional neural networks (CNNs) to develop independent sentiment classification models of both text and images, exploring the effect of various fusion strategies on the results of sentiment classification. Yu et al. used CNNs to derive text features at different steps such as word, phrase and sentence features and then combines the features data with image information at the feature layer to classify sentiments. Cai et al. employed two independent CNNs to learn features of texts and images, and then this input was added in another CNN to learn the connection between the two features.

In a more elaborate method Xu et al. created a deep semantic network in multimodal sentiment analysis, where an LSTM network exhibited attention to acquire keywords using image features. Co-memory network to learn the interaction information of text and image content was also suggested by Xu et al. using the resulting fusion features to recognize sentiments.

Interestingly, video data can also be considered multimodal data; Poria et al. developed a video sentiment classification framework that used multi-core learning (MKL) as a method to combine video, audio and text features. Zadeh et al. proposed the use of tensor fusion to extract joint feature representations on video data, which essentially captures single-mode, dual-mode and triple-mode features, and they learn the interaction information between modalities.

3. Proposed Model

We present our multimodal sentiment analysis model: CFF-ATT in this part. The text and picture information in the social media are usually concomitant, however, in some cases one text can be associated with several pictures. Our assumed symmetry of multimodal data, i.e., assuming a text to have a corresponding image, is to simplify processing multimodal data. The processing of text and an image are equivalent in the proposed model, and lastly, merge their features. This fully takes advantage of their symmetry. We set an image–text pair as (X, I), where X denotes a single text, X = {X1, X2, . . . , XM}, and I denotes a single image. Our model aims to avail the sentimental polarity of image-text pairs u ∈ {positive, negative, neutral} correctly. The overall framework of the model presented in this paper is shown in Figure 1. The bottom layer of the model is the text-feature extraction module and image-feature extraction module, transforming the text from X = {x1, x2, . . . , xM} to Y = {y1, y2, . . . , yM} ∈ Rdw ×M, where dw is the dimension of the word vector. The photo was converted to a set-size vector. Next in the model is a feature fusion module, which has an attention mechanism, it learns shared implicit text and image representations that are then interactively learned to combine the features in a cross-modal manner. The full connectivity layer is placed on the top of the model which links with

two different main and secondary features of the fusion module as a portion of the input in order to complete the sentiment classification.

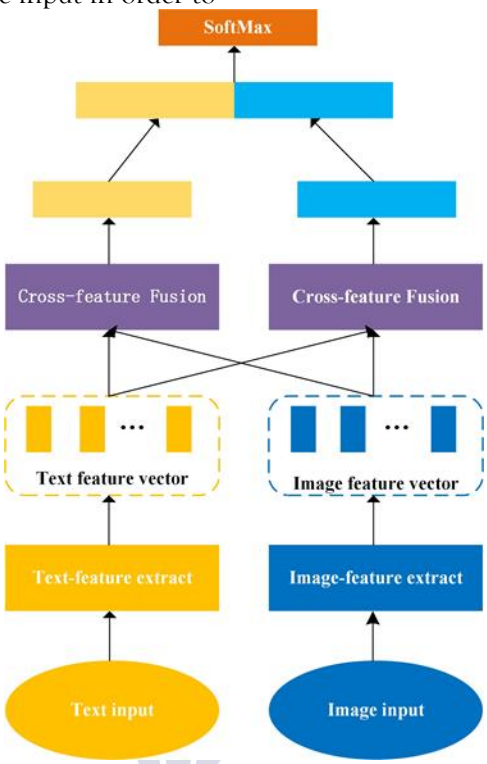

Figure 1. Overall structure of the multimodal sentiment analysis model.

3.1. Text-Feature Extraction Module

The role of the text-feature extraction layer is to find a low-dimensional mapping of every word to a low-dimensional vector, or word embedding. There is a lot of noise in the textual data of social media, which would influence the quality of feature extraction. A denoising autoencoder (DAE) was used to extract text features to remove noise interference and get more robust features [23]. The denoising autoencoder erases the original input matrix with a certain probability distribution (usually using the binomial distribution), that is, each value is randomly set to zero. By doing so, some elements of the characteristics of certain data are absent. The loss of data features makes the original data "impure". This kind of data "impurity" is similar to adding noise. Repeatedly changing the encoder to eliminating the noise interference, the encoder eventually manages to accomplish the objective of recreated the original data, or in other words, denoising. To get the matrix, in particular, the vectorized representation of the text in matrix form was destroyed. The torn-up matrix was fed into the encoder to get the encoding, and reconstruction matrix was then retrieved through the reconstruction of the decoder. The reconstructed matrix was also not compared to the original matrix to get the reconstruction error. The encoder and decoder parameters were then varied to get the minimum reconstruction error in order to get the final encoding. Taking the encoding features obtained from the upper layer as the input of the lower layer, the coding of the lower layer was obtained using the same method. The above were repeated until a given number of layers was encoded. Concisely, with the creation of a deep network and training of a low-dimensional feature code layer, layer after

layer, the low-dimensional feature that best represented the text was extracted to guarantee a feature dimensionality reduction to high-dimensional data of a text. Figure 2 shows the structure of the denoising autoencoder.

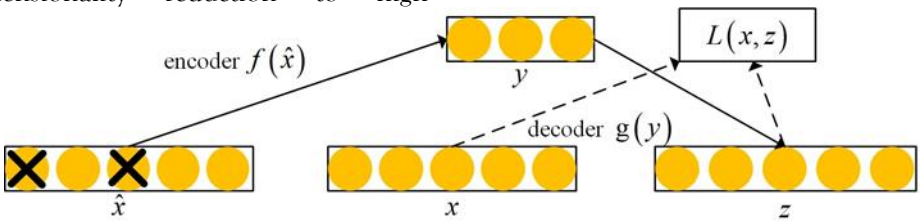

Figure 2. Structure diagram of the denoising autoencoder.

The encoder f (xˆ) was used for dimensionality reduction of high-dimensional data. First, the input vector x was destructed to obtain xˆ, and then it was input into the encoder f (xˆ). After the linear transformation and activation function, the implicit encoding result is finally obtained. Decoder g (y) was used for the reconstruction process of low-dimensional coding, that is, the hidden layer data were mapped back to the reconstruction, which is expressed as the following functions:

$$y = f(\hat{x}) = S_f \quad W\hat{x} + b_y \quad , \tag{1}$$

$$z = g(y) = S_g \quad W'y + b_z \quad , \tag{2}$$

where Sf is a nonlinear activation function, and its expression is:

$$S_f = sigmoid(y) = \frac{1}{1 + e^{-}}, \tag{3}$$

where Sg is the activation function of the decoder, W′ = WT is the transposition of W. Therefore, we only need to train W. Moreover, by and bz are bias vectors. We used the sigmoid function in this paper. The training process for a DAE is to find the minimum reconstruction error of parameter θ = W, by, bz} in the training sample set. The expression of the reconstruction error is as follows:

$$H_{DAE} = \sum_{x \in D} L(x, g(f(\hat{x}))), \tag{4}$$

with L being the reconstruction error function. In the experiment, cross-entropy loss functional is always better than the square deviation loss functional. So, we used the cross- entropy loss function and the expression will be the following:

$$L(x, z) = -\frac{1}{n} \sum_{i=1}^{d_x} [x_i \ln z_i + (1 - x_i) \ln (1 - z_i)],$$

Image-Feature Extraction Module

We have employed a better attention-based variational auto-encoder (VAE-ATT) to distill image features. We aim to use the data to learn how to map the raw data to the data representation, both in an automatic fashion. Deep neural network VAE is made of an encoder and decoder. As in Figure 3 the VAE is all about extracting the hidden characteristics of the information and constructing a model based on the hidden characteristics to the generated targets. The encoder identifies the potential reasonable

variables of the original data, and conditions the coding results with Gaussian noise in such a way that encourages them into hidden features that follow the Gaussian distribution. The decoder built model projects the hidden features to the reconstructed probability distribution that is as close as possible to the original distribution.

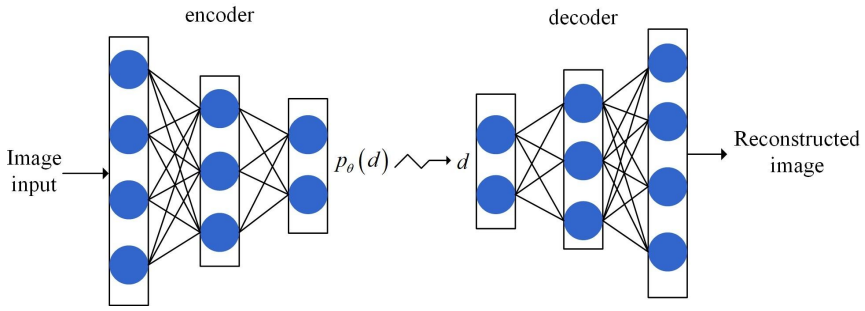

Figure 3. How VAE works.

The network VAE has two components: the encoder network E with parameter φ and decoder D with parameter μ. The loss function is as follows:

$$L(\varphi, \mu, a) = E_{q\varphi(d|a)}[\log p_\mu(a|d)] - D_{KL} \quad q_\varphi(d|a) \,||\, p_\mu(d) \quad ,$$

where $q_\varphi(d|a)$ represents the encoder from the data space to the hidden space, and $p_\mu(a|d)$ stands for the decoder from the hidden space to data space.

The loss function consists of two parts. The first term of Formula is the reconstruction error, which drives the reconstructed distribution $p_\theta(a|d)$ closer to the input distribution $p_u(a)$. The second term aims to reduce the KL divergence and drive $q_\varphi(d|a)$ closer to the prior distribution. To achieve this reconstruction, VAE captures the most important characteristic factors that can represent the original input data.

In particular, we experimented with the variant β-*VAE* of VAE β-*VAE* introduces the understanding of the entanglement prior, assuming that the data are generated based on mutually independent factors. Therefore, various independent variables can be used to represent these factors in the representation. The unentangled one previously can facilitate the encoder to learn the abstract representation of the data in a compact form that can be applied in other downstream tasks and enhance the efficiency of the samples.

$$L(\varphi, \mu, a) = E_{q\varphi(d|a)}[\log p_\mu(a|d)] - \beta D_{KL} \quad q_\varphi(d|a) \,||\, p_\mu(d) \quad .$$

According to Formula 2: The dimension of hidden variables and reconstruction error can be balanced by introducing a super parameter 2 to thank Formula 2 -VAE. Further, the isotropic quality of the Gaussian prior also raises invisible restrictions to the posteriori of learning. Modulation of β may also affect the amount of learning in training hence promoting different representations to learn. When carrying out the experiment, its value should be varied in such a way that it facilitates unwrapping of the characterization.

Bengio et al. suggested that the representation which fits certain tasks and data domains can dramatically enhance the learning success rate and the model robustness of the training task. Thus, we built attention mechanism of the high-level representation of VAE extraction. The same applies to the self-attention and human visual attention mechanisms which filter certain important information out of a significant amount of information and direct attention to the important information. Figure 4 shows the inner architecture of the attention model. This module breaks down the overall properties of the input data, captures the relationship between channels, and forecasts the significance of the channels in order to selectively accentuate some features.

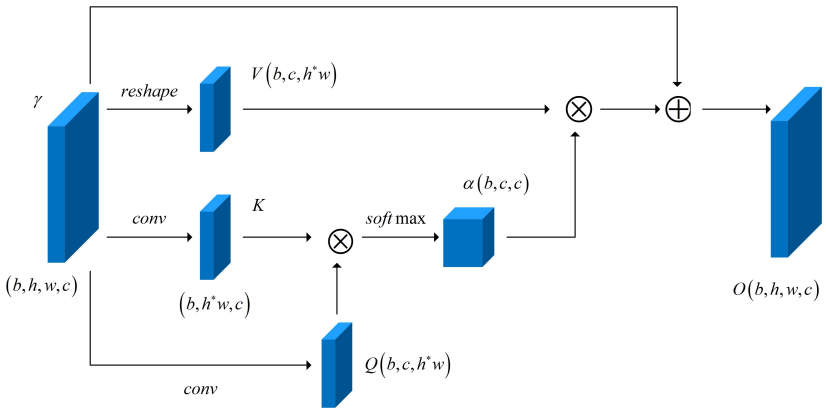

Figure 4. Network structure of the attention model and its corresponding feature dimensions.

The input of the attention module was constructed according to the hidden feature γ produced by the pretraining encoder, γ ∈ Rb×h×w×c, where b is the batch size, h and w are the length and width of the feature map respectively, and c is the number of channels. As indicated in Formulas Q and K are new feature graphs obtained by integrating input feature γ through cross-channel information of a 1 × 1 convolution kernel, and the dimensions were transformed into Ra×c, where a = h × w. Then, matrix multiplication was performed between Q and K transposed, and finally, the Softmax function was used for normalization to obtain the probability distribution αij of the attention mechanism with the dimension c × c. The significance of this design is to calculate the influence weight between each channel number of γ, which can highlight the role of the key feature graph and reduce the influence of redundant features on the overall classification performance.

$$Q = reshape\left(F_{CNN}\left(\gamma;\mu_1\right)\right), \quad (11)$$

$$K = reshape\left(F_{CNN}\left(\gamma;\mu_2\right)\right), \quad (12)$$

$$V = reshape\left(\gamma\right), \quad (13)$$

$$\alpha_{ij} = \frac{\exp\left(Q_i \cdot K_j\right)}{\sum_{i=1}^{C} \exp\left(Q_i \cdot K_j\right)}. \quad (14)$$

Finally, the weight coefficient αij and the original feature were weighted and summed, and then the scale coefficient β was adjusted to obtain the highly discriminative feature expression Oj, as follows:

$$O_j = \beta \sum_{i=1}^{c} \alpha_{ij} \; V_i + \gamma_j, \quad (15)$$

where β was initialized to zero, and gradually assigned to a large weight in the learning process.

Feature-Fusion Module

For the features extracted by the text-feature extraction module and image-feature extraction module, we designed a cross-feature-fusion module based on the attention mechanism, CFF-ATT, to fuse them. The details of the feature-fusion module are presented in Figure 5. The output characteristics of the two previously mentioned modules were of the input of the feature fusion module where one of them acts as the primary input, and the other as the secondary. The two modes of input were combined in order to produce.

target mode output. Setting the main input as E = {E1, E2, . . . , En} ∈ Rde ×n, and secondary input as

G = {G1, G2, . . . , Gl} ∈ Rdt ×l , we project the main input E and the secondary input G into the same shared vector space:

$$E_{emb_i} = \tanh\left(P_{E_{emb}} E_i + C_{E_{emb}}\right),$$

$$G_{emb_j} = \tanh\left(P_{G_{emb}} G_i + C_{G_{emb}}\right),$$

where PEemb ∈ Rdv ×de , PGemb ∈ Rdv ×dt , CEemb ∈ Rdv , and CGemb ∈ Rdv in the above formula are the training parameters, and dv is the dimension of the shared vector space. We used Eemb and Gemb to calculate the attention matrix M ∈ Rn×l , where Mij represents the correlation between the i-th content of the main input and the i-th content of the secondary input. The attention matrix can be expressed as follows:

$$M_{ij} = E_{emb_i}{}^{T} G_{emb_j}. \quad (18)$$

Then, we obtained the secondary input J based on the attention mechanism:

$$J = G \cdot MT. \quad (19)$$

Finally, the main input E and the attention-based secondary input J were spliced at the full connection layer to obtain the fusion feature U = {U1, U2, . . . , Un} :

$$U_i = \tanh(P_u [E_i : J_i] + C_u), \quad (20)$$

where Ui ∈ Rde , Pu ∈ Rde ×(de +dt).

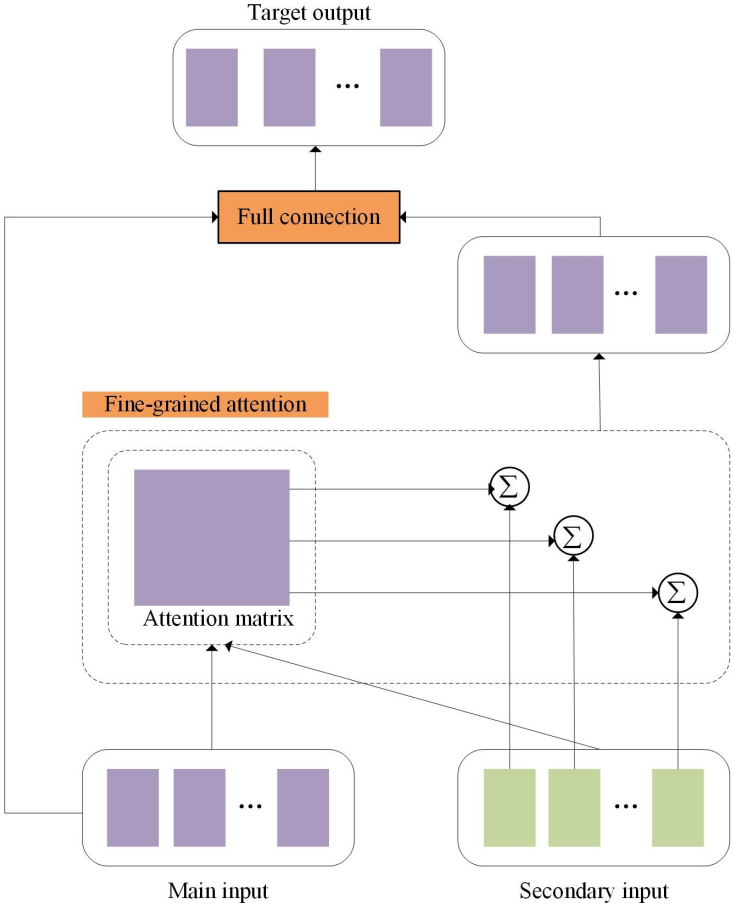

Figure 5. Feature fusion module.

Output Layer

We got the fused features through the above modules. One is the feature where the text forms the main feature and the image is the auxiliary feature and the other one is where the image forms the main feature and the text is the auxiliary

feature. Linking them, we feed the last output layer and applied the Softmax classifier to the sentiment classifier.

Experiments and Results

Here, we will explain experimental results of our model on two Twitter datasets, and the comparative experiments with other methods of the bases. Finally, we make a visual verification of the effect of the attention mechanism on image feature extraction.

Datasets and Setup

We have used two popular multimodal sentiment datasets in this experiment, that is MVSA-Single and MVSA-Multiple [26]. The MVSA-Single dataset consists of 5129 text-image pairs that are Twitter-based. Each text and image were annotated by an annotator using one of three sentiments (positive, negative or neutral). The MVSA-Multiple datasets are a collection of 19,600 pairs of text and images (Twitter). Unlike the previous dataset, each text–image pair was annotated with one of the three sentiments by three different annotators. The judgment of each annotator on the text or image was independent of the judgment of others.

In the case of the MVSA-multiple datasets, since each text or image is given three sentimental tags, the same sentimental tags might appear or they might be different and this will influence our judgement of the sentimental polarity of the text or image. Thus, we think that the text or image is meaningful and can only be used if two or more of the three sentimental tags of text or image are the same. Moreover, the sentimental polarities of a text-image pair reflected in the sentimental tag were the same, i.e., were both sentimental. Same, it is easy to classify the sentiment. But in other tweets in these two datasets, the sentiment polarity of the text and images did not always match with one another, which impacted our further sentiment classification task. To gain more precise data, we removed the tweets having a single polarity of positive sentiment, and two polariities of negative sentiment. In case of tweets having one polarity of neutral sentiment, and the other polarity having a positive (or negative) sentiment, we treated the sentimental polarity as positive (or negative), in other words we disregarded the neutral sentimental label. We had received the MVSA-Single dataset containing 4511 textimage pairs, and MVSA-Multiple datasets containing 17, 024 textimage pairs after processing the dataset using the above technique. We randomly subdivided the data set into training.

(80%), verification (10%) and testing sets (80%) according to the ratio of 8:1:1.

Experimental Parameter Setting

We scaled each image in the dataset to 224 × 224 pixels. The enhanced variational auto-encoder is trained on the training dataset. During the β-VAE training phase, Adam optimizer was utilized, and the fixed learning rate was 0.001. CNN is utilized in the encoder model. The output of each image is 2048 10 × 10 area feature maps. The deconvolutional neural network is a part of the decoder. Loss function is used to offer the reconstruction error of the image via cross-entropy and the discrepancy between the distribution of hidden variables and the unit Gaussian distribution via KL divergence. The convergence property of the loss function stated that in order to achieve randomness and to not get into local optimization, it was necessary to choose the batch size used in this paper, namely 128. During the training, we also employed dropout and early-stopping in order to prevent overfitting of the models used. We chose the common space of feature-fusion module as 100 and its dropout was 0.3.

Baselines

The following are two graphs that we compared our model with independently on two MVSA datasets with the baseline approaches.

- SentiBank + SentiStrength were extracted 1200 pairs of adjectives and nouns to determine the characteristics of the image and estimate the sentimental score of the part of the text.
- CBOW + DA + LR were based on an unsupervised and semi-supervised learning of the internal properties of both the text and image with the skip-gram and denoising autoencoders and related them to sentiment classification.
- CNN-Multi: CNN-Multi comprises two independent CNNs to acquire text and image features, and two other CNNs aided in sentimental classification by feeding the features into a further CNN.
- DNN-LR used the neural network to classify text and image which subsequently was used to extract their respective features, relate them and feed them into a logistic regression that was used to classify sentiment.
- MultiSentiNet employed the deep semantic features of the image, such as the visual, object and scene features and suggested an attention LSTM model based on visual features to assimilate these significant text words that are relevant to sentiment analysis.
- Mrs. CoMn took advantage of the interaction between the text and the image and suggested a stacked co-memory network to represent the interaction of the visual and text information environmentally to perform sentimental analysis

Experimental Results

Our indicators of the experimental evaluation were the accuracy rate and F1-score, and the calculation is as follows:

$$Precision = \frac{TP}{TP+FP}$$

$$Recall = \frac{TP}{TP+FN}$$

$$Accuracy = \frac{Precision \cdot Recall}{TP+FN+TN+FP}$$

TP = the number of samples that were originally positively labeled and that have been labeled in this manner, FP = the number of samples that have been incorrectly labeled as positive, TN = the number of samples that were negatively originally labeled and that are now incorrectly in the negative category, FN = the number of samples that have been incorrectly labeled as negative but actually positive.

The results of the comparison of our model with other baseline methods in two MVSA datasets are shown in Table 1. Published papers were used to extract the experimental findings of the baseline technique in Table 1. CoMN(6) is CoMN that has six memory jumps. It is the CoMN model whose classification effect is the best in the corresponding paper.

SentiBank + SentiStrength does the worst as per the results in Table 1. This model only used the traditional statistical features to represent the sentimental information of text and images, and failed to extract the most useful internal features of the text and images. It therefore cannot infer well the polarity of sentiments of the tweets in terms of polarity.

Then, CBOW + DA + LR were a denoising autoencoder that created images. CBOW + DA + LR can learn more accurate features of the images

by decreasing the image error, and use them together with word embedding to accomplish the sentiment classification task. That is why, its effect is superior to that of SentiBank + SentiStrength. Moreover, CNN-Multi and DNN-LR used the deep learning neural network to extract text and image features. They performed poorly compared with CBOW + DA + LR when the amount of data was small as is the case in MVSA-single dataset but used compare to CBOW + DA + LR when the amount of data was very large as is the case in MVSA-multiple datasets. This shows that these two deep learning models need additional training to enhance their learning effect.

Table 1: *Experimental results of baseline method and our model on two MVSA datasets*

		MVSA-Single		MVSA-Muliiple	
	Model	ACC	F1	ACC	F1
Baselines	SentiBank + SentiStrength	0.5205	0.5008	0.6562	0.5536
	CBOW + DA + LR	0.6386	0.6352	0.6422	0.6373
	CNN-Multi	0.6120	0.5837	0.6630	0.6419
	DNN-LR	0.6142	0.6103	0.6786	0.6633
	MultiSentiNet	0.6984	0.6963	0.6886	0.6811
	CoMN(6)	0.7051	0.7001	0.6892	0.6883
	Proposed Model	0.7144	0.7106	0.6962	0.6935

MultiSentiNet, when applying both text and image, took into account the visual feature semantic information and accentuated situating the effect of the image information on the text information. Thus, it obtained good experimental results.

Moreover, CoMN took into account two-way interaction and simultaneous promotion of text information and image informatio; hence, its experimental results were superior to MultiSentiNet that works only with the impact of the image information on the text information. Nonetheless, CoMN attention mechanism involved one attention vector to engage in various content types, potentially resulting into information loss. The noise interference in the textual information was removed in our model and the more significant image features were extracted. We applied the feature-fusion module which relies on a fine-grained attention-based mechanism to learn modal fusion features of the text and image information on an interactive basis. Thus, we got superior results of the experiment compared to CoMN.

Attention Visualization

Here, we visualize the impact of the attention mechanism to the extraction of image features. The algorithm that we used in the experiment was the t-SNE to further decrease the dimensionality of the eigenvalues output of the network and project it on a two-dimensional plane. The feature visualization diagram of part of the testing set in the MVSA-Single dataset experiment is figure 6. Figure a represents the initial characteristic of the image feature prior to it joining the attention module and Figure 6b represents the image feature following the manner in which it has been improved by the attention mechanism. In order to achieve a clearer image expression, three types of images positive, neutral and negative were sampled in the t-SNE experiment, and the size of the three types of images was simultaneously decreased. Figure 6 has three marking symbols that depict three categories. Class1 (positive) and class2 (negative) and class3 (neutral) are some of them. The difference between the distribution of various categories of images becomes more evident after the enhancement of the features of the attention module. The intra class distance standard

deviation is lowered and the standard deviation of the spacing between classes is increased. The experimental results reveal that the attention mechanism can capture key features in advanced features, which is helpful to improve the accuracy of the sentiment classification task.

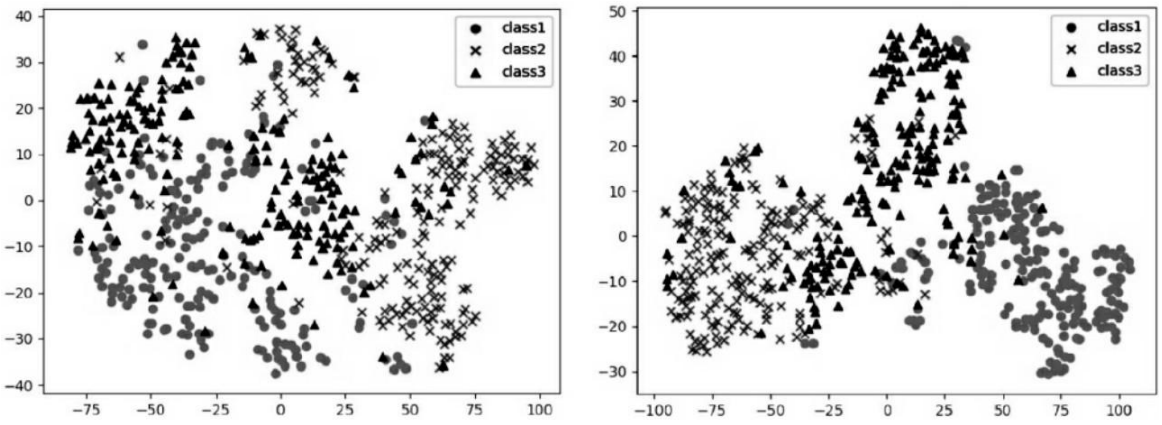

Figure 6. Gaussian visualization outputs of the mapped two-dimensional space with image features using t-SNE. (a) Before entering the attention module, the early feature of the image feature. (b) The image feature after it is enhanced by the attention mechanism.

5. Conclusions and Future Work

Sentiment analysis Multimodal sentiment analysis of social media is not an easy one. To achieve this, this paper will propose a multimodal sentiment analysis model on the attention mechanism. The model can be successfully used to remove noise interference in the textual information of social media and acquire more precise text characteristics. The image features that are more significant to classifying sentiment are retrieved, when combined with the attention mechanism. On features fusion, attention mechanism is reintroduced to combine features in other modes successfully that learns the interactive data between text and images. The model used modal internal information and modal interaction information to effectively obtain the sentimental feature representation of multimodal data, accurately judged the sentimental polarity of users' tweets, better revealed users' real feelings, and helped us understand people's attitudes and views towards certain events on social media. Our proposed model is feasible and more effective as has been shown by the experimental results on the two open datasets.

We plan to refine the current models and techniques in the future and explore other modalities of working, such as audio, video, etc. We will seek to determine more efficient ways of extracting features and fusing the features in order to offer more useful information to sentiment analysis.

Author Contributions: Conceptualization, K.Z.; methodology, K.Z.; software, J.L.; validation, W.L.; formal analysis, Y.G. and K.Z.; investigation, J.L. and W.L.; resources, Y.G. and J.Z.; data curation, J.Z.; writing—original draft preparation, K.Z.; writing—review and editing, Y.G.; visualization, J.L. and W.L.; supervision, J.Z.; project administration, Y.G. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Key R&D Program of China (No. 2019YFB1404700).

Acknowledgments: The authors are very thankful to the editor and referees for their valuable comments and suggestions for improving the paper.

Conflicts of Interest: The authors declare no conflict of interest.

REFERENCES

1. Liu, B. Sentiment analysis and opinion mining. Synth. Lect. Hum. Lang. Technol. 2012, 5, 1–167.
2. Wilson, T.; Wiebe, J.; Hoffmann, P. Recognizing contextual polarity in phrase-level sentiment analysis. In Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Vancouver, BC, Canada, 6–8 October 2005; pp. 347–354.
3. Rodger, J.; Murrar, A.; Chaudhary, P.; Foley, B.; Balmakhtar, M.; Piper, J. Assessing American Presidential Candidates Using Principles of Ontological Engineering, Word Sense Disambiguation, and Data Envelope Analysis. Management 2020, 20, 22.
4. Fan, F.; Feng, Y.; Zhao, D. Multi-grained attention network for aspect-level sentiment classification. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, 31 October–4 November 2018; pp. 3433–3442.
5. Li, Z.; Wei, Y.; Zhang, Y.; Yang, Q. Hierarchical attention transfer network for cross-domain sentiment classification. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, Hilton, New Orleans Riverside, New Orleans, LA, USA, 2–7 February 2018.
6. Woo, S.; Park, J.; Lee, J.Y.; So Kweon, I. Cbam: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 3–19.
7. Li, X.; Xie, H.; Chen, L.; Wang, J.; Deng, X. News impact on stock price return via sentiment analysis. Knowl.-Based Syst. 2014, 69, 14–23.
8. 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.
9. Xu, N.; Mao, W.; Chen, G. A co-memory network for multimodal sentiment analysis. In Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, Ann Arbor, MI, USA, 8–12 July 2018; pp. 929–932.
10. Liu, B.; Zhang, L. A survey of opinion mining and sentiment analysis. In Mining Text Data; Springer: Boston, MA, USA, 2012; pp. 415–463.
11. Borth, D.; Ji, R.; Chen, T.; Breuel, T.; Chang, S.F. Large-scale visual sentiment ontology and detectors using adjective noun pairs. In Proceedings of the 21st ACM International Conference on Multimedia, Barcelona, Spain, 21–25 October 2013; pp. 223–232.
12. Cao, D.; Ji, R.; Lin, D.; Li, S. A cross-media public sentiment analysis system for microblog. Multimed. Syst. 2016, 22, 479–486.
13. Poria, S.; Cambria, E.; Hussain, A.; Huang, G.B. Towards an intelligent framework for multimodal affective data analysis. Neural Netw. 2015, 63, 104–116.
14. Wang, M.; Cao, D.; Li, L.; Li, S.; Ji, R. Microblog sentiment analysis based on cross-media bag-of-words model. In Proceedings of the International Conference on Internet Multimedia Computing and Service, Xiamen, China, 10–12 July 2014; pp. 76–80.

15. Baecchi, C.; Uricchio, T.; Bertini, M.; Del Bimbo, A. A multimodal feature learning approach for sentiment analysis of social network multimedia. Multimed. Tools Appl. 2016, 75, 2507–2525.
16. 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 on Web Search and Data Mining, San Francisco, CA, USA, 22–25 February 2016; pp. 13–22.
17. 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.
18. Chen, Y.; Zhang, Z. Research on text sentiment analysis based on CNNs and SVM. In Proceedings of the 2018 13th IEEE Conference on Industrial Electronics and Applications (ICIEA), Wuhan, China, 31 May–2 June 2018; pp. 2731–2734.
19. 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.
20. Cai, G.; Xia, B. Convolutional neural networks for multimedia sentiment analysis. In Natural Language Processing and Chinese Computing; Springer: New York, NY, USA, 2015; pp. 159–167.
21. Poria, S.; Peng, H.; Hussain, A.; Howard, N.; Cambria, E. Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis. Neurocomputing 2017, 261, 217–230.
22. Zadeh, A.; Chen, M.; Poria, S.; Cambria, E.; Morency, L.P. Tensor fusion network for multimodal sentiment analysis. arXiv 2017.
23. Vincent, P.; Larochelle, H.; Bengio, Y.; Manzagol, P.A. Extracting and composing robust features with denoising autoencoders. In Proceedings of the 25th International Conference on Machine Learning, Helsinki, Finland, 5–9 July 2008; pp. 1096–1103.
24. Higgins, I.; Matthey, L.; Pal, A.; Burgess, C.; Glorot, X.; Botvinick, M.; Mohamed, S.; Lerchner, A. Beta-Vae: Learning Basic Visual Concepts with a Constrained Variational Framework. 2016.
25. Bengio, Y.; Courville, A.; Vincent, P. Representation learning: A review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. 2013, 35, 1798–1828.
26. Niu, T.; Zhu, S.; Pang, L.; El Saddik, A. Sentiment analysis on multi-view social data. In Proceedings of the International Conference on Multimedia Modeling, Miami, FL, USA, 4–6 January 2016; pp. 15–27.
27. Maaten, L.v.d.; Hinton, G. Visualizing data using t-SNE. J. Mach. Learn. Res. 2008, 9, 2579–2605.