# MULTIMODAL SENTIMENTAL ANALYSIS BASED ON DEEP LEARNING

**Jiao BianBian[1], Leelavathi Rajamanickam[2*], N. Lohgheswary[3] & Z. M. Nopiah[4]**

**Abstract**
Multimodal sentiment analysis is automatically detecting, analysing and extracting emotions and opinions in multimodal data. Deep learning is becoming very popular for multimodal sentiment analysis because it can automatically extract meaningful and abstract semantic features. The main objective of this study was to introduce an efficient model for multimodal sentiment analysis using deep learning method. The system is divided into four parts namely data layer, single-modality feature extraction layer, multimodal features fusion layer and sentiment analysis layer, it adopted the public dataset: CMU-MOSI and CMU-MOSEI. It can be concluded that the system can improve and surpass the traditional textual sentiment analysis.

**Keywords:** CMU-MOSI; deep learning; multimodal sentiment analysis

[1,2*]Centre for Software Engineering, Faculty of Engineering, Built Environment & Information Technology, SEGi University, Kota Damansara, 47810 Selangor Darul Ehsan, Malaysia,
[3]Department of Electrical and Electronics Engineering, Xiamen University Malaysia, Sepang, 43900 Selangor Darul Ehsan,
[c]Department of Engineering Education, Faculty of Engineering & Built Environment, University Kebangsaan Malaysia

***Corresponding Author:-** Leelavathi Rajamanickam
[*]Centre for Software Engineering, Faculty of Engineering, Built Environment & Information Technology, SEGi University, Kota Damansara, 47810 Selangor Darul Ehsan, Malaysia, leelavathiraj@segi.edu.my

**DOI:** - 10.48047/ecb/2023.12.si5a.0249

## I. Introduction

Sentiment analysis is automatically detecting, analysing and extracting emotions and opinions expressed by people. Every day, Twitter, Facebook, Sina Weibo and other social media platforms generate a great number of comments which can be used to analyse peoples' opinions and emotions. At present, text sentiment analysis is widely used for customer satisfaction assessment and public opinion monitoring.

However, With the rapid development of social multimedia, the arrival of complementary data streams such as audio, video and pictures will help to improve and surpass the previous text-based sentiment analysis. Especially, people tend to post short text in social media, which increases the difficulty of sentiment analysis.

On the basis of text, multimodal information which adding facial expression, voice and tone, can provide more intuitive expression, convey more accurate and rich emotional information and help people to reveal the emotional information that may be hidden in the text. The sentiment analysis is evolving from text-based sentiment analysis to multimodal sentiment analysis which integrates text, video and audio. Multimodal sentiment analysis has great potential and has many issues to be addressed.

Multimodal signals include textual, visual and audio information, E.g., vlogs, spoken reviews, text messages with accompanying images or video.The extraction methods of text emotional features are well developed, however the extraction methods of image emotional features, video emotional features and audio emotional features are still on the way.

Especially, how to measure the weight of text, video and audio in determining sentiment classification and how to deal with the correlation of three single-modality are still in its infancy and more research are needed to demonstrate its full potential. Lin et al.

proposed that sentiment dictionaries, machine learning methods and deep learning methods are the three main existing methods for multimodal sentiment analysis[1]. Deep learning is proved efficient in multimodal sentiment analysis practice.

This research aims to explore the key problems, including the extraction of unimodal emotional features, the fusion of multimodal emotional features based on deep learning methods and improving the accuracy of sentiment prediction.

Among many issues that need to be addressed in multimodal sentiment analysis, the extraction of unimodal features and the design of a multimodal sentiment analysis model are the core issues. The research questions include:

RQ1: How to explore the emotional features of multimodality information on YouTube?

RQ2: How to fuse the multimodal emotional features?

RQ3: How to effectively predict sentiment using multimodality information on YouTube?

This research focuses on the key issues in multimodal sentiment analysis and proposes some novel solutions. The research objectives include:

RO1: To identify a model on the emotional features of multimodality on YouTube using deep learning.

RO2: To develop a multimodal information fusion mechanism.

RO3: To analyses sentimental based multimodal mechanism on YouTube.

The multimodal sentiment analysis is more complicated than the previous text-based sentiment analysis. It combines the sentiment analysis based on text, audio and video.

In term of text sentiment analysis, the mainstream methods include sentiment dictionary and machine learning. The popular methods for image sentiment analysis include low-level features-based method, mid-level features-based method, ANP and DBN. The mainstream methods for audio sentiment include SVM, HMM, KNN and GMM.

The current multimodal fusion modes include early fusion, late fusion and mixed fusion. There is a special multimodal sentiment analysis named sentiment analysis based on image and text fusion which integrate text and image information.

According to the function, Multimodal sentiment analysis is divided into four layers: data layer, coding layer, fusion layer and classification layer. In data layer, multimodal information is preprocessed and transferred into the system. In the coding layer, the information of three modes are firstly encoded to obtain the raw vectors of each mode which is the vector representation set of all elements of each mode.

After averaging or weighted summation of the raw vectors, a single vector representation of each mode is obtained. After the raw vectors and single vector representation pass through the representation fusion layer, the reconstructed feature vector of

each mode fused with other modal information is obtained. Finally, the reconstructed feature vectors of the three modes are processed by the mode fusion layer to obtain the final fusion vector as the input of the classification layer. Figure 1 describes the conceptual framework for multimodal sentiment analysis.

With the development of social multimedia, the multimodal sentiment analysis become very important for individual, corporations and governments. It can widely apply in e-commerce, public opinion monitoring, political election. It also can used for suicide detection, Intelligent customer service, human-computer interaction and other aspects. This research introduces an efficient model for multimodal sentiment analysis

which can improve and surpass the traditional textual sentiment analysis.

We propose a model to process multimodal sentiment analysis based on transformer. It can extract multimodal information features using transformer encoder and fuse features based on transformer attention mechanism. This model is more sensitive to the target in the video and more effective in improving sentiment predict accuracy.

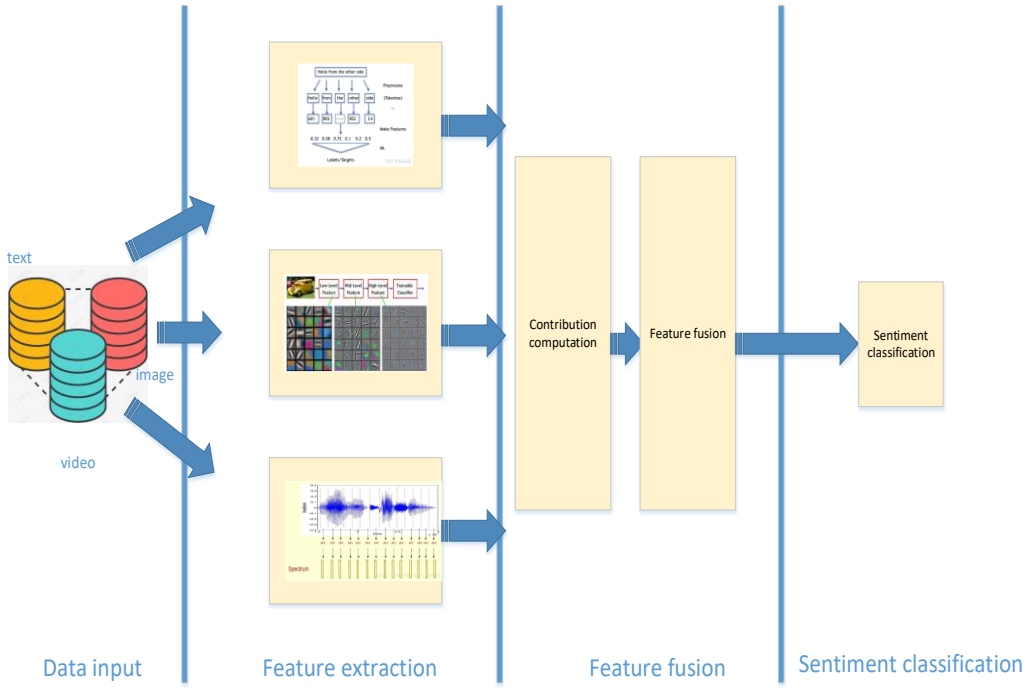


**Figure 1** Conceptual framework for multimodal sentiment analysis

## II. Literature Review

Because the multimodal sentiment analysis is a rather new research topic, there are only a few notable researches focusing on it. Multimodal sentiment analysis was introduced by Morency et al., It proposed analysing audio and visual content in addition to text for sentiment analysis.

Machine learning and deep learning are two main existing methods for multimodal sentiment analysis. Li et al. adopted Machine learning to fuse the multimodalities features, it uses text emotion score to calculate the image emotion score, and the image score is input into the Logistic Regression as a feature to obtain the emotion result[2].

However, Machine learning method separates feature extraction from decision-making process, so it is necessary to select the best features carefully and transfer it to machine learning algorithm. Feature selection process is very expensive and

difficult, which requires a lot of time and expertise. Moreover, machine learning methods require many labelled data. Although self-supervised learning

can solve the problem of lacking training data to some extent. However, deep learning can do better. Deep learning can avoid tedious process of feature extraction, and automatically carry out high-dimensional abstract the features through neural networks, reducing the composition of feature engineering and saving a lot of time.

Therefore, deep learning methods are becoming very popular because it can automatically extract meaningful and abstract semantic features for sentiment analysis. Deep learning methods have

been proved to outperform the traditional methods in multimodal sentiment analysis.

"Deep Learning" is a special kind of machine learning, which was first introduced in 1986. LeCun et al. stated that Deep learning approaches are composed of multiple layers to learn features of data with multiple levels of abstraction. Goodfellow et al. stated that Deep learning methods can automatically combine simple features into more complex features.

Young et al. introduced Deep learning models and architectures, mainly used in Natural Language Processing (NLP) domain[3]. They introduced and compared various Deep learning applications and models in various NLP fields, and discussed possible future trends. Zhang et al. presented state-of-the-art deep learning methods used in recognition systems.

Deep learning promoted the development of artificial intelligence. In recent years, deep learning has been widely used in image processing, speech recognition, natural language processing and other fields.

Deep learning models highly complex data through multi-level nonlinear transformation, which enables it to automatically and effectively learn deeper information from high-dimensional data. moreover, deep learning is outstanding for processing large-scale data and heterogeneous data. Therefore, Deep learning has become an important method for multimodal sentiment analysis.

Xu et al. proposed a method for aspect based multimodal sentiment analysis to capture the impact of aspects on texts, images, and their interactions[4]. The method consists of two interactive memory networks to supervise the textual and visual information for a given aspect.

Another important work of this paper is to construct a new public dataset: Multi-ZOL. The model can capture the correlations on aspect-based multimodal sentiment data. However, the Accuracy is only 61.59% and MACRO F1 Score is 60.51%. The current multimodal information fusion modes are divided into two types: feature-level fusion (early fusion) and decision level fusion (late fusion).

Based on the characteristics of deep learning algorithm, most deep learning methods use feature-level fusion, whereas some paper adopt decision-level fusion. MIAO et al. introduced a model in which CNN model is pre-trained on large-scale image data set[5], and then transferred to fine-tuned CNN to train the text emotion classification model by inputting the trained word vector into BILSTM, the final emotional result is obtained by decision fusion.

With the recent trend of attention mechanism, much research on multimodal sentiment analysis began to introduce different attention-based neural network to model the multimodalities fusion, the attention mechanism performs well because it concerns the correlation between different modalities. Harish A. and Sadat F. used Attention-based DNN on CMU-MOSI dataset[6].

Chen et al. presented some new methods to extract sentimental features from text, audio and video, further use these sentimental features to verify the multimodal analysis model based on multi-head attention mechanism[7].

Ashish Vaswani et al. published the landmark paper: attention is all you need and firstly proposed the Transform architecture which was used in neural machine translation tasks[8].

In the following years, Transform became the mainstream architecture in the field of natural language processing, and successfully crossed over to many fields such as Computer Vision.

Because of its characteristics, it is also suitable for multimodal information processing. Tsai Yao-Hung Hubert et al. introduced the Multimodal Transformer (MulT) to process the multimodal data in an end-to-end manner without explicitly alignment[9].

In this model, there are three unimodal transformers and 6 bimodal transformers, the center of the model is the directional pairwise cross-modal attention, which enables one modality to receive information from another modality.

But this paper focused on multimodal information processing instead of processing multimodal sentiment analysis. Yu and Jiang proposed a multimodal sentiment classification method based on adapted BERT[10].

However, compared to the good performance on textual sentiment analysis, Multimodal BERT is not sensitive to the target in the image representation. the comparison of mentioned methods was listed in Table 1.

**Table 1**. the comparison of mentioned methods

| Title | Author and Published year | Modality | Data set | Fusion mode | algorithm | perfomance |
|---|---|---|---|---|---|---|
| **Image sentiment prediction based on textual descriptions with Adjective Noun Pairs** | LI et al. 2018 | Text, Image | Twitter 1269 | Decision - level fusion | Logistic Regression | Accuracy: 71.1%; F1 Score: 77.2%. |
| **Multi-interactive memory network for aspect based multimodal sentiment analysis** | XU et al. 2019 | Text, Image | Multi-ZOL | Feature-level fusion | Multi-Interactive Memory Network | Accuracy: 61.59%; MACOR F1 Score: 60.51%. |
| **Joint visual-textual approach for microblog sentiment analysis** | MIAO et al. 2019 | Text, Image | data crawled from WEIBO | Decision - level fusion | FCNN and WBLSTM | Accuracy: 88.2%; F1 Score: 88.8%. |
| **Multimodal transformer for unaligned Multimodal language sequences** | YAO. et al. 2019 | Text, audio and video | CMU-MOSI, MOSEI, IEMOCAP | Feature-level fusion | Multimodal Transformer | Accuracy: 81.6%; F1 Score: 81.6%. |
| **Trimodal attention module for multimodal sentiment analysis** | A. Harish and F. Sadat, 2020 | Text, audio and video | CMU-MOSI | Feature-level Fusion and Decision - level fusion | Attention-based DNN | Accuracy: 79.3%; F1 Score: 83.8% |
| **Multimodal sentiment analysis based on multi-head attention mechanism** | X. et al. 2020. | Text, audio and video | MOUD, CMU-MOSI | Feature-level fusion | Multi-head attention mechanism | MOUD: accuracy 90.43%; CMU-MOSI: accuracy 82.71% |
| **Adapting bert for target-oriented multimodal sentiment classification** | J.-F. Yu and J. Jiang,2020 | Text, Image | Twitter15 and Twitter17 | Feature - level fusion | multi-layer attention mechanism | Twitter15：Accuracy 77.15%；Twitter17：Accuracy 70.50% |

In recent years, researchers have established various types of multimodal datasets to provide experimental data for multimodal sentiment analysis models. The relatively well-known datasets in China include the seed emotion EEG dataset, the CHEAVD2.0 multimodal sentiment analysis dataset and the multi-ZOL multimodal dataset. The international well-known datasets include YELP, IEMOCAP, CMU-MOSEI and so on. Table 2 lists the available public multimodal datasets.

**Table 2.** The public multimodal data

| Name | data |
|---|---|
| Multi-ZOL | Text and image |
| Flickr | Video |
| Yelp | Text and image |
| CMU-MOSI | Text, audio and video |
| CMU-MOSEI | Text, audio and video |
| MOSEI | Text, audio and video |
| Twitter15 and Twitter17 | Text and image |
| MELD | Text, image and video |

## III. Methodology

Our research proposes a system to do multimodal sentiment analysis based on deep learning methods. The system is divided into four parts namely data layer, single-modality feature extraction layer, multimodal features fusion layer and sentiment analysis layer. Figure 2 shows the overall architecture for multimodal sentiment analysis.

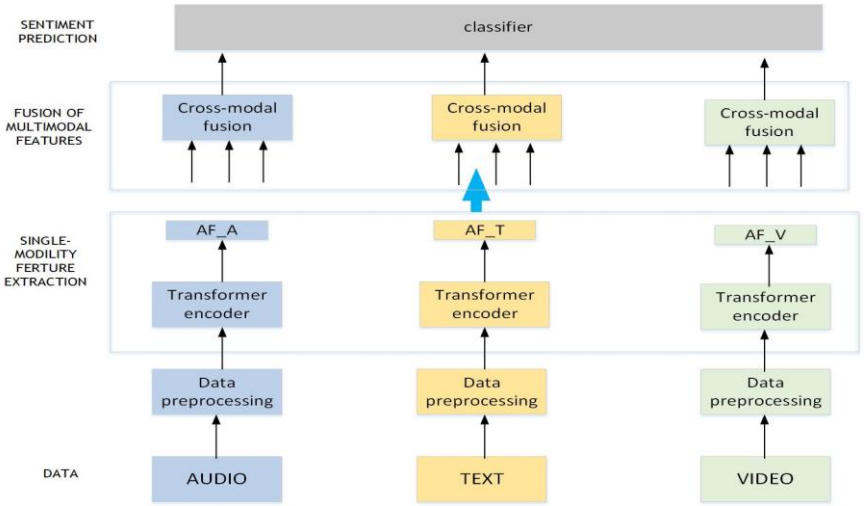


**Figure 2** Overall architecture for Multimodal sentiment analysis

The research adopted the public datasets: CMU-MOSI and CMU-MOSEI, they contain numerous video blogs(vlog) of social media platform named YouTube, the vlogs are movie comment expression videos. The samples in these two datasets are marked by human annotators. CMU-MOSI consists of 2,199 short video clips which are labeled with sentiment scores from -3 to 3. CMU-MOSEI consists of 23,453 annotated video clips which are annotated with sentiment scores from -3 to 3.

In data layer, three kinds of information: audio, text and video are preprocessed firstly. Text, audio and video have different process strategy. In term of text, glove word embeddings were used to represent the textual features, the dimension of each word embedding is 300. In term of video, Facet was used to represent a set of visual features including facial action units. In term of audio, COVAREP was used to represent t acoustic features, the dimension of the acoustic feature is 74.The features were then transferred to the transformer encoder. Three transformers based on self-attention can extract the features of single-modality to AF_A, AF_T and AF_V respectively. In multimodal features fusion layer, a transformer based on cross-attention fuses the different features using feature-level fusion strategy. In general, the current multimodal information fusion modes are divided into two types: feature-level fusion (early fusion) and decision-level fusion (late fusion). The mechanism of feature-level fusion is shown in figure 3. The basic principle of feature-level fusion is to map the features from different sources and different modalities to a high-dimensional vector space, then put these vectors together, later use them in the classical classification model.

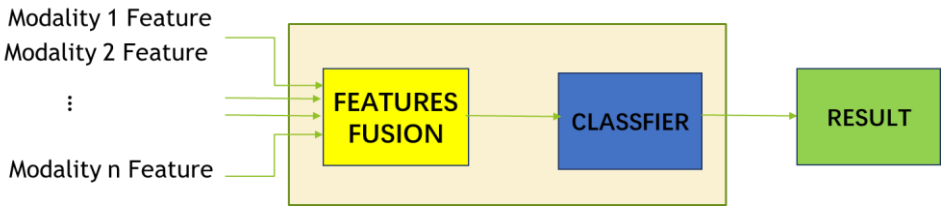


**Figure 3** the feature-level fusion

The mechanism of decision-level fusion is shown in figure 4, Firstly, different models are selected for each mode for training, and then these models are combined by voting decision-making method and so on. The decision-level fusion is more difficult to calculate and time-consuming.

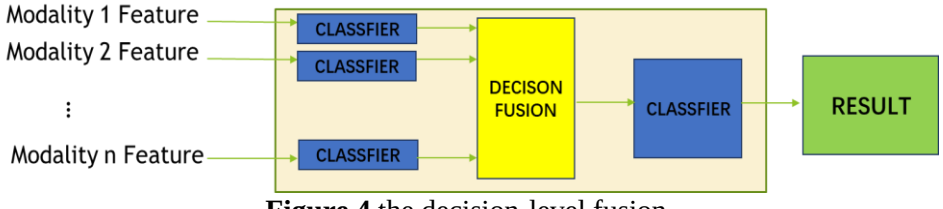


**Figure 4** the decision-level fusion

At last, the fusion features are transferred to classifier to predict the sentiment of multimodal information.

**ACKNOWLEDGEMENTS**
The authors wish to express gratitude towards SEGi University (SEGiIRF/2020-8/FoEBEIT-37/103) for supporting the research.

**References**

1. P. Lin, X. Luo and Y. Fan: "A Survey of Sentiment Analysis Based on Deep Learning", World Academy of Science, Engineering and Technology International Journal of Computer and Information Engineering,2020, Vol.14, No.12, pp.473-485
2. Z. Li, , Y. Fan : "Image sentiment prediction based on textual descriptions with adjective noun pairs", Multimedia Tools and Applications, 2018, Vol.77, No.1, pp. 1115-1132
3. T. Young, D. Hazarika, S. Poria and E. Cambria, "Recent Trends in Deep Learning Based Natural Language Processing" , IEEE Computational Intelligence Magazine, 2018,vol. 13, No. 3, pp. 55-75.
4. N. Xu, W. Mao and G. Chen, "Multi-Interactive Memory Network for Aspect Based Multimodal Sentiment Analysis", Proceedings of the AAAI Conference on Artificial Intelligence, 2019, Vol. 33, No. 1, pp.371-378
5. Miao et al., "Joint visual-textual approach for microblog sentiment analysis", Computer Engineering and Design, 2019, Vol. 40, No.4, pp.1099-1105
6. Harish A. and Sadat F., "Trimodal attention module for multimodal sentiment analysis," in Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, 2020,pp. 13803-13804
7. X. Chen, G.-M. Lu and J. Yan, and Yan J., "Multimodal sentiment analysis based on multi-head attention mechanism", Proceedings of the 4th International Conference on Machine Learning and Soft Computing, 2020, pp. 34–39
8. A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin, "Attention is All you Need", Proceedings of NeurIPS, 2017, pp.5998–6008
9. Tsai Yao-Hung Hubert, Bai Shaojie, Pu Liang Paul, Kolter J Zico, Morency Louis-Philippe, Salakhutdinov Ruslan., "Multimodal Transformer for Unaligned Multimodal Language Sequences", Proceedings of the conference Association for Computational Linguistics, 2019, pp.6558-6569
10. J.-F. Yu and J. Jiang, "Adapting bert for target-oriented multimodal sentiment classification", Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence, 2020, pp. 5408–5414.

**About the Author (30 words)**


Photo

Jiao Bianbian(1981-), female, Han nationality, China, Computer science major, master's degree, associate professor, the main research areas include Natural Language Processing and data mining.

ASSOC. PROF. TS. DR. LEELAVATHI R received her Doctoral in Philosophy (PhD) in Computer Science & Engineering, her area of research is in Software Engineering, Big Data, Data Science, Image Processing, Machine Learning, HCI and revolutionary projects dealing with advanced computing.