Proceedings of MSS 2025 Symposium: Automation and Smart Technologies in Petroleum Engineering
DOI: 10.54254/2755-2721/142/2025.KL22289

Integrating Visual and Textual Cues for Sentiment Analysis: Multi-modal Approach

Hanqi Bai

School of Computer Science and Engineering, University of New South Wales, Sydney, Australia
z5286125@zmail.unsw.edu.au

Abstract: With the steady growth of social media and online platforms, sentiment analysis has become a critical task to understand public opinion, customer feedback, and social trends. This study investigates multi-modal sentiment analysis by exploring state-of-the-art models such as Bidirectional encoder representations from transformers (BERT), Bootstrapping Language-Image Pretraining (BLIP), Generative Image-to-text Transformer (GIT), and Contrastive Language-Image Pretraining (CLIP) for sentiment classification tasks using both image and text data. The effectiveness of these models on sentiment analysis tasks is evaluated under different configurations such as text-only BERT, image-to-text augmented BERT, and CLIP-based classification. The results show that while BERT achieves 76% accuracy on text-only sentiment analysis, combining text with image-generated descriptions does not significantly improve performance, with accuracy remaining around 74%. On the other hand, CLIP achieves a moderate 62% accuracy using image-text embeddings. While CLIP performs well on multi-modal mapping, it demonstrates challenges in deep semantic understanding compared to BERT's 0.76 F1 score on the text-only task. These findings highlight the challenges of effectively merging different modalities and point out future directions for improving sentiment analysis in multi-modal settings, enhancing the ability of models to fully understand the semantic content in both images and text.

Keywords: Multi-modal Sentiment Analysis, Sentiment Analysis, CLIP, BERT.

1. Introduction

With the rapid growth of multimedia content on social media platforms, users publish content through various channels, methods, and carriers, which makes analysis, feedback, recommendation, interaction, and other applications based on user emotional state particularly important. However, traditional sentiment analysis mainly focuses on text data and ignores data in other modalities such as images and audio, while human emotions have multi-modal characteristics. To cope with this situation, multi-modal sentiment analysis has become a potential solution, integrating multiple data types to understand content better and perform sentiment classification.

The latest progress in single-modal and multi-modal models has shown potential in sentiment analysis tasks. Among them, the pre-trained model BERT has been proven to be effective in extracting features from text data [1-3]. Multi-modal models such as BLIP and GIT utilize pre-trained image and language models to enhance vision-language tasks. BLIP integrates encoder-decoder architecture, which is efficient for both image-grounded text generation and image-text retrieval [4] and BLIP-2 further leveraging frozen models for better computational efficiency [5], bridging the

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/142/2025.KL22289

modality gap with a Querying Transformer that aids in vision-to-language generation, these advances contribute to state-of-the-art results while reducing the computational load [6]. GIT simplifies vision-language tasks with unified architecture, combining image encoder and text decoder trained using a language modeling objective. This approach eliminates the need for complex structures, such as object detectors, by directly mapping raw image inputs to text descriptions, these advances contribute to state-of-the-art results while reducing the computational load [7].

Inspired by the concept of aligning image and text representations, Linearly Mapping Between Representation spaces (LiMBeR) focuses on transferring image representations into the text space through a simple linear transformation [8]. Building on this idea, the CLIP model [9] and its variant Alpha-CLIP [10] offer direct potential solutions for multi-modal analysis. These advances highlight the necessity of exploring how different modalities interact and their integration in sentiment analysis tasks. Despite these advances, there are still some challenges in multi-modal sentiment analysis. One problem is the alignment and fusion of heterogeneous modalities, which usually have different feature distributions. How to combine these models efficiently and conveniently is a challenge. At the same time, there is a lack of large-scale, well-annotated multi-modal sentiment datasets, and the existing data is relatively lacking in the quality and quantity of images and texts. Solving these challenges can not only promote the development of natural language processing and computer vision in sentiment analysis.

This study aims to explore and compare various state-of-the-art multi-modal sentiment analysis methods. By leveraging pre-trained models such as BERT, CLIP, and BLIP, their ability to capture individual and joint modality features to accurately predict sentiment was investigated. This work seeks to identify existing gaps and propose potential directions for future research in this area.

2. Methodology

2.1. Data Source

Using the Eriched MVSA dataset, which is developed based on twitter data the data consists of image, text, and label pairs, where the text is an English twitter post with random special symbols and the corresponding image is attached [11]. There are 4869 data in total, that is, 4869 pairs of text, pictures, and labels.

As shown in Figure 1, There are three types of labels that include positive, negative, and neutral, and positive labels account for 34% of the total data, negative labels account for 30%, and 36% for neutral labels.

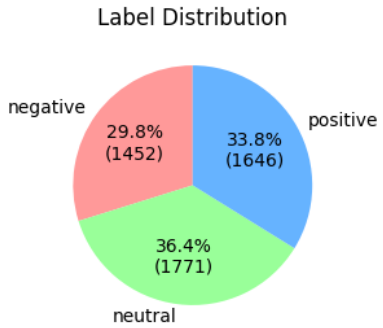


Figure 1: Data label distribution.

The overall average length of text data after removing special characters is 84.9, which is 91.8 positive, 87.5 negative and 76.3 of neutral.

DOI: 10.54254/2755-2721/142/2025.KL22289

2.2. Methods and Models

2.2.1. Text-only BERT Baseline

BERT is a transformer-based model that has revolutionized natural language processing (NLP) tasks due to its ability to capture context from both directions (left-to-right and right-to-left) in a sentence, this bidirectional understanding allows BERT to classify text data effectively. Therefore, it is reasonable to use Bert to perform sentiment analysis on text only as a baseline for text data. Use the pre-trained Bert basic model and add three classification heads corresponding to positive, negative and neutral. Then train and verify on the dataset.

The results will be regarded as the emotional level carried by the text in the dataset, providing a reference for other methods of sentiment analysis.

2.2.2. Image-to-Text Augmented BERT

The BLIP and GIT models were selected as the trial solution for converting images to text. They generated text descriptions of the images, combined them with text data in different ways, and finally sent them to BERT for sentiment analysis. BLIP is a new type of image-to-text pre-trained model released by Salesforce in 2022. It introduces Query Queue into the model and combines CNN and Transformer. It has a wide range of applications in tasks such as image description generation and image-text matching. GIT is a generative model based on the Transformer architecture, which converts images into text representations, compared with BLIP, it focuses more on generation tasks.

Once obtained the text description of the image, the description is combined with the text data. There are two methods worth trying, the first approach shown in Figure 2, to directly connect the text and description and then give it to BERT for classification.

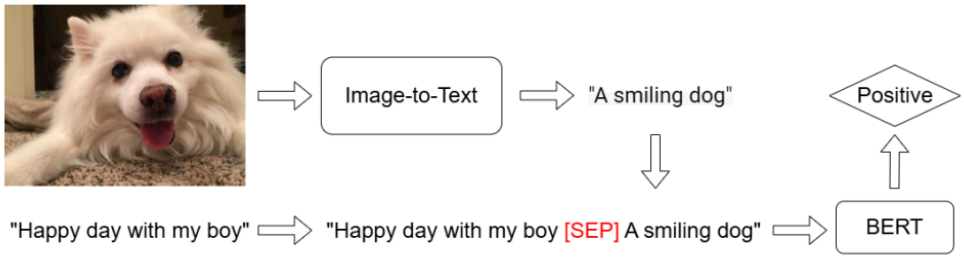


Figure 2: Merging of description and text data.

The second method uses the weighted summation method described in Figure 3, leveraging BERT to obtain the embedding of the text and description, and combine the two embeddings with different weights, and then input them into the BERT classification head for three classifications.

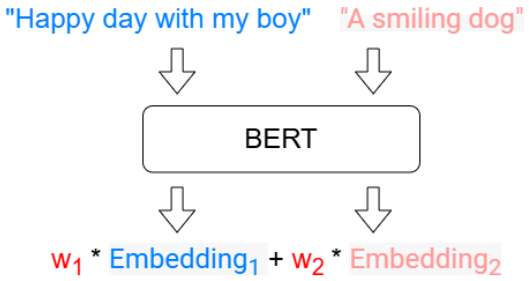


Figure 3: Weighted sum of description and text embeddings.

DOI: 10.54254/2755-2721/142/2025.KL22289

Based on the results of the early verification, the embedding weighted summation has significantly reduced accuracy, the experiment will not conduct in-depth research on this method. One possible reason is that the weighted summation method destroys the emotional factors in the text data, where the embedding of the descriptive text exacerbates this effect.

2.2.3. CLIP-based Classification

CLIP is trained on a large number of image and corresponding text pairs, using CNN to encode images and Transformer to encode text, ultimately mapping images and text into the same embedding space. With the help of zero-shot classification capability, it is introduced into the task of multi-modal sentiment analysis. The experiment uses CLIP to obtain the embedding of text and image, and the embeddings of the two are input into a fully connected classifier by merging or weighted sum.

2.3. Evaluation

To evaluate the performance of each method, the dataset is divided into training, validation, and test sets. The primary evaluation metrics include accuracy, precision, recall, and F1-score, which provide a comprehensive assessment of the model's performance in sentiment classification across different categories.

For the generated descriptions by the model, perform manual evaluation to randomly evaluate the quality of descriptions, evaluating the accuracy and sentiment consistency of the generated text with the corresponding sentiment by describing the length, whether there is repetition, and the alignment of the specific content with the image content.

3. Experiments

3.1. BERT Configuration

Pre-trained model version is BERT-base-uncased, the parameters of the tokenizer use padding and truncation, the padding strategy is a maximum length of 128. The trainer in the Transformer library is used for training, with a learning rate of 5E-5, a weight decay of 0.01 and AdamW as optimizer.

3.2. Image-to-Text models

The pre-trained model GIT used is microsoft/git-base-coco. AutoProcessor is used for image and text preprocessing with padding set on the left. The maximum length of the captions generated by this model is 75 tokens. The BLIP used is Salesforce/blip-image-captioning-base with padding and skip special token when decode from model output.

3.3. CLIP Classification

Freeze all parameters in CLIP, encode the image and text to obtain the embedding, and then pass the embedding into the fully connected classification head according to the size.

Figure 4 provides the detailed structure of the classification header, that is, the merged size is 1024, the weight sum is 512, and the specific size is 1025/512 in the first layer, 128 in the second layer, 32 in the third layer, and three neurons in the output layer represent the corresponding labels. The classification head uses the ReLU activation function, a dropout of 0.1 and a learning rate of 1E-3 during training.

DOI: 10.54254/2755-2721/142/2025.KL22289

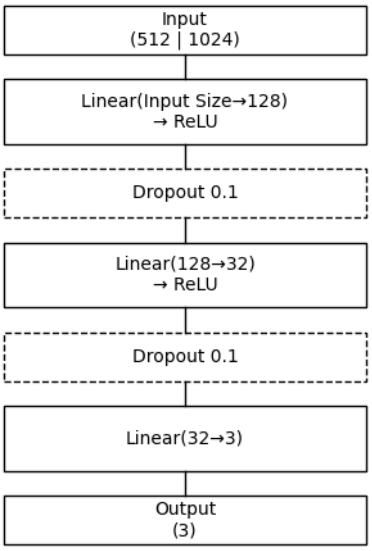


Figure 4: Fully connected classification head with CLIP.

4. Results & Discussion

BERT was trained on plain text data for 7 epochs and achieved an overall F1-score of 0.75 on the testing set, this indicates a relatively balanced performance in sentiment classification, where the model is able to correctly identify sentiments in most instances.

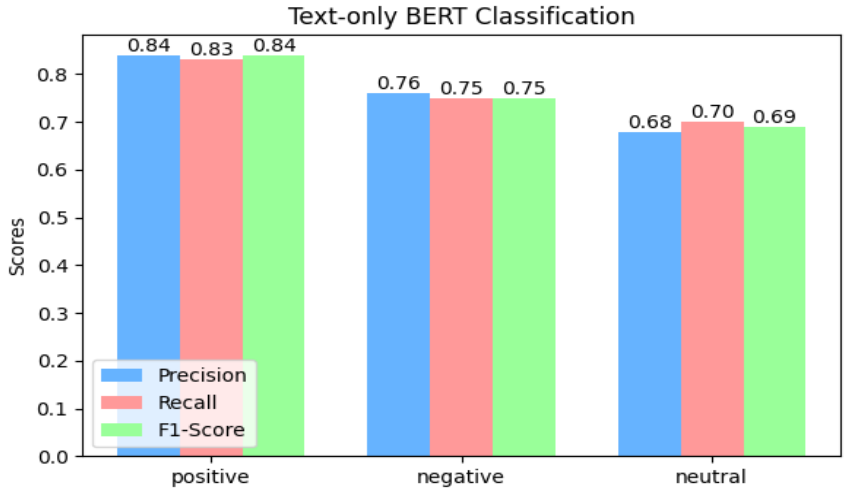


Figure 5: Fully connected classification head with CLIP.

The results of the BERT model in three classes are shown in Figure 5. The F1-score for the positive class was 0.84, which suggests that the model performed well in detecting positive sentiments, with fewer false positives and false negatives. For the negative class, the F1-score was 0.75, indicating moderate performance. While the model correctly identified negative sentiments, there was a higher rate of misclassification between negative and neutral instances. The neutral class showed a lower F1-score of 0.69, which can be attributed to the difficulty in distinguishing between neutral and other sentiment categories, especially since neutral data often lacks strong emotional cues, making it more challenging to classify accurately. Overall, these results suggest that while BERT excels in identifying

DOI: 10.54254/2755-2721/142/2025.KL22289

clear positive sentiments, there are challenges in distinguishing between negative and neutral sentiments, which requires further model refinement or enhanced feature engineering.

For the descriptions generated by BLIP and GIT, the average length of the descriptions generated by the BLIP model is 44, and the average length of the descriptions generated by the GIT model is 41. Upon conducting a preliminary manual evaluation, the descriptions generated by GIT and BLIP were found to have similar quality, However, the output of BLIP shows slightly greater instability compared with GIT as shown in Figure 6, which is manifested as meaningless special symbols and repeated words, which can significantly affect downstream tasks.

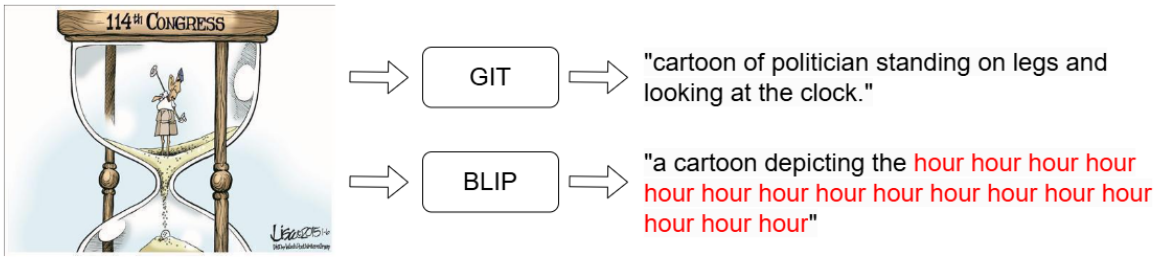


Figure 6: Duplicate content in BLIP model output.

In natural language processing, this might lead to a loss of coherence and clarity. For sentiment analysis and other NLP tasks, these inconsistencies can result in incorrect interpretations, as models may fail to properly process or understand the text and undermining the model's ability to accurately classify sentiments or generate coherent responses. From the results, there are some differences between BLIP and GIT models in sentiment analysis tasks. Although the description text generated by GIT is higher quality to a certain extent, it has a relatively lower accuracy rate. One possible reason is that the description text is more neutral and has less emotional color.

Table 1: BLIP and GIT model results.

	BLIP			**GIT**		
	precision	recall	f1-score	precision	recall	f1-score
positive	0.83	0.81	0.82	0.79	0.80	0.80
negative	0.74	0.76	0.75	0.72	0.76	0.74
neutral	0.66	0.66	0.66	0.64	0.59	0.61

However, it can be observed from Table 1, both description data provide a minor decreased f1-score than plain text with BERT, the negative class has the smallest gap, which is less than 1%, the positive class decreased by about 3% and the neutral class decreased by about 6%. The method of using string merging does not bring obvious improvement to the testing set. Compared with pure text data, the scheme of adding image data seems to show slight interference, which causes a decrease in the final performance.

in short, the method of using text descriptions does not bring significant results. The possible reason is that both BLIP and GIT generated descriptions are declarative, which is different from traditional spoken text. Descriptive language may have longer dependencies. Another possibility is that the descriptive text itself has less emotional color, the attempt to train and analyze BERT alone on descriptions gives 43% accuracy on BLIP data and 40% on GIT data, which shows there is a lack of data for emotional descriptions during the training process of BLIP and GIT.

The CLIP with classification combined with both concatenated embedding and weighted sum was trained for 15 epochs on the training set. It can be seen from Figure 7 and Figure 8, embedding

DOI: 10.54254/2755-2721/142/2025.KL22289

concatenate achieved an accuracy of about 60% on the test set, while the weighted summation achieved 62%, the weights of the weighted sum are 0.8 for text embedding and 0.2 for description embedding. Overall, more complex classification head structures tend to slightly improve accuracy in both methods. Compared with models such as BERT, CLIP's ability to capture complex language semantics may not be as strong, one possible reason is that it does not focus on pure text understanding like BERT.

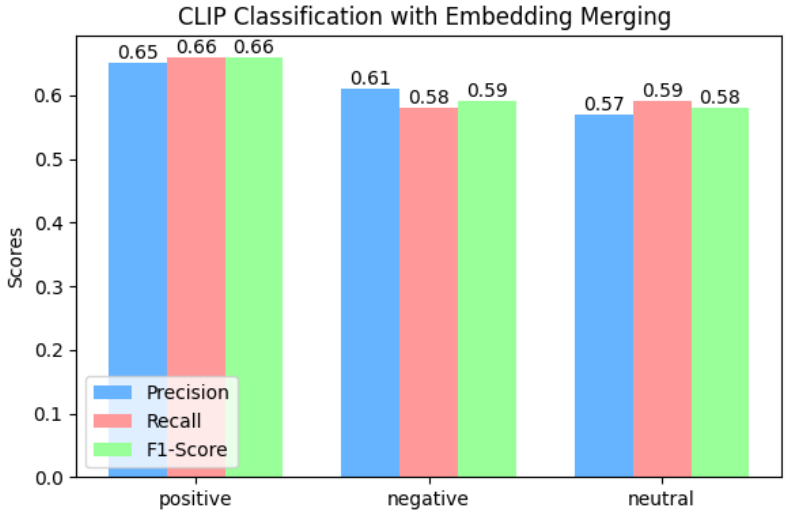


Figure 7: Results of CLIP classification head with embedding concatenation.

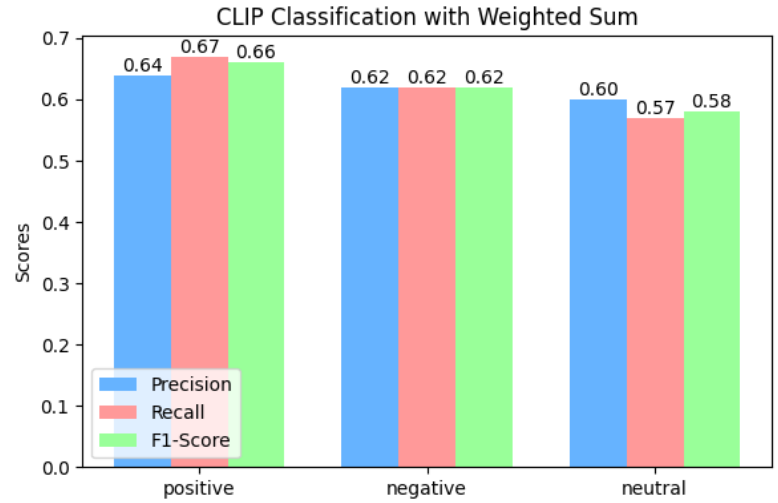


Figure 8: Results of CLIP classification head with weighted summation.

5. Conclusion

This study explored several multi-modal sentiment analysis approaches, leveraging state-of-the-art models such as BERT, BLIP, GIT, and CLIP to evaluate their feasibility for multi-modal sentiment classification.

The results highlight several important observations, that BERT performs well in text-based sentiment analysis, achieving solid results with an F1 score of 0.76. It demonstrates a solid ability to classify three sentiments when analyzing plain text, with high accuracy in predicting positive sentiment labels.

DOI: 10.54254/2755-2721/142/2025.KL22289

On the image-2-text side, while both BLIP and GIT generate text descriptions from images, merging these descriptions with text does not significantly improve the results. In fact, it seems to slightly hinder the results, likely due to the lack of sentiment in the generated text. There is still much room for improvement in the fusion of text-image data in the future.

On the other hand, the CLIP model using image-text embedding also shows good potential compared to BERT but falls short in capturing complex semantics. The concatenated embedding and weighted sum embedding methods have an accuracy of about 60-62%, which suggests that CLIP may not have an advantage in deeply understanding text content.

Overall, while multi-modal sentiment analysis is an exciting and growing field, it faces challenges in integrating different modalities. Future research can focus on improving the quality of image-generated text and its ability to express emotions, exploring better fusion techniques and filling gaps in current multi-modal datasets. This work lays the foundation for further exploration of the complex interactions of language, images, and emotions, highlighting the potential for more refined methods in the future.

References

[1] *Bello, A., Ng, S.C., & Leung, M.-F. (2023). A BERT Framework to Sentiment Analysis of Tweets. Sensors, 23(1), 506.*

[2] *Mickel, H.A., Bihorac, O.A. and Rouces, J. (2019) Aspect-Based Sentiment Analysis Using BERT. Proceedings of the 22nd Nordic Conference on Computational Linguistics, 187–196.*

[3] *Zhao, A.P. and Yu, Y. (2021) Knowledge-Enabled BERT for Aspect-Based Sentiment Analysis. Knowledge-Based Systems, 227, 107220.*

[4] *Li, J., Li, D., Xiong, C. and Hoi, S. (2022) BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. Proceedings of the 39th International Conference on Machine Learning, Proceedings of Machine Learning Research, 162, 12888-12900.*

[5] *Li, J., Li, D., Savarese, S. and Hoi, S. (2023) BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. Proceedings of the 40th International Conference on Machine Learning, Proceedings of Machine Learning Research, 202, 19730-19742.*

[6] *Srinivas, P. (2023) BLIP-NLP Model for Sentiment Analysis. 2023 2nd International Conference on Edge Comput ing and Applications (ICECAA), Namakkal, India, pp. 468-475. doi: 10.1109/ICECAA58104.2023.10212253.*

[7] *Wang, J., Yang, Z., Hu, X., Li, L., Lin, K., Gan, Z., Liu, Z., Liu, C. and Wang, L. (2022) GIT: A Generative Image-to-Text Transformer for Vision and Language. arXiv.*

[8] *Bustos, C., Civit, C., Du, B., Solé-Ribalta, A. and Lapedriza, A. (2023) On the Use of Vision-Language Models for Visual Sentiment Analysis: A Study on CLIP. 2023 11th International Conference on Affective Computing and Intelligent Interaction (ACII), Cambridge, MA, USA, pp. 1-8. doi: 10.1109/ACII59096.2023.10388075.*

[9] *Sun, Z., Fang, Y., Wu, T., Zhang, P., Zang, Y., Kong, S., Xiong, Y., Lin, D. and Wang, J. (2023) Alpha-CLIP: A CLIP Model Focusing on Wherever You Want. arXiv.*

[10] *Mohammed, D.J. and Aleqabie, H.J. (2022) The Enrichment of MVSA Twitter Data via Caption-Generated Label Using Sentiment Analysis. 2022 Iraqi International Conference on Communication and Information Technologies (IICCIT), Basrah, Iraq, pp. 322-327. doi: 10.1109/IICCIT55816.2022.10010435.*

[11] *Niu, T., Zhu, S.A., Pang, L. and El Saddik, A. (2016) Sentiment Analysis on Multi-View Social Data. MultiMedia Modeling (MMM), 15-27..*