

PMIG-Net: A Confidence-Aware and Prior-Guided Multimodal Fusion Framework for Weakly Supervised News Sentiment Analysis

XUYANG GONG*

*Student, Department of Big Data Analytics, Kyung Hee University, Seoul, Korea

[Abstract]

The rapid growth of online news and social media has increased the importance of multimodal sentiment analysis that jointly leverages textual and visual information. However, existing studies have mainly focused on social media, online reviews, and video data, and have not sufficiently addressed modality-level sentiment inconsistency, image noise, and uncertainty in weakly supervised labels in news articles. To address these issues, this study constructs CNN-MMSA, a weakly supervised multimodal news sentiment dataset consisting of 8,273 CNN news samples, and proposes a PMIG-Net-based sentiment analysis framework. PMIG-Net utilizes BERT-based text representations and CLIP-based image representations, selectively adjusts the contribution of visual information through prior-guided attention fusion and image gating, and jointly learns modality-specific and final multimodal representations through hierarchical multi-task learning. On CNN-MMSA, PMIG-Net achieved an Accuracy of 0.7164, a Macro-F1 of 0.6732, and a Macro-Recall of 0.6828. Compared with adapted baselines that reimplement the core fusion concepts of TFN, MulT, MAG-BERT, and Self-MM, the proposed model showed relatively stronger performance, particularly in Macro-F1 and Macro-Recall. Ablation experiments showed clear decreases in Macro-F1 and Macro-Recall when the image gate was removed, while the other components exhibited different trade-offs across evaluation metrics. In an additional benchmark evaluation on MVSA-Single, PMIG-Single-Pass also outperformed CLIP-only, BERT-only, and BERT+CLIP Concat under a controlled experimental setting, indicating its applicability to image-text sentiment analysis in social media data. By integrating weak-label construction and selective multimodal fusion within a unified framework, this study provides a practical multimodal sentiment analysis approach for news opinion analysis and intelligent media content analysis.

► **Key words:** Multimodal Sentiment Analysis, News Sentiment Analysis, Weakly Supervised Learning, Prior-Guided Attention, Image Confidence Gate, Hierarchical Multi-Task Learning

-
- First Author: XUYANG GONG, Corresponding Author: XUYANG GONG
 - *XUYANG GONG (gongxuyang@khu.ac.kr), Department of Big Data Analytics, Kyung Hee University
 - Received: 2026. 07. 23, Revised: 2026. 08. 31, Accepted: 2026. 09. 02.

[요 약]

온라인 뉴스와 소셜 미디어의 확산으로 텍스트와 이미지를 함께 활용하는 멀티모달 감성 분석의 중요성이 높아지고 있다. 그러나 기존 연구는 주로 소셜 미디어, 온라인 리뷰 및 동영상 데이터를 중심으로 수행되어 뉴스 기사에서 나타나는 모달리티 간 감성 불일치, 이미지 노이즈 및 약지도 라벨의 불확실성을 충분히 반영하지 못하였다. 본 연구에서는 이러한 문제를 다루기 위해 8,273개의 CNN 뉴스 샘플로 구성된 약지도 기반 멀티모달 감성 데이터셋 CNN-MMSA를 구축하고, PMIG-Net 기반 감성 분석 프레임워크를 제안한다. PMIG-Net은 BERT 기반 텍스트 표현과 CLIP 기반 이미지 표현을 활용하고, 사전 지식 기반 어텐션 융합과 이미지 게이트를 통해 이미지 정보의 반영 정도를 조절하며, 계층적 멀티태스크 학습을 통해 모달리티별 표현과 최종 멀티모달 표현을 함께 학습한다. CNN-MMSA 실험에서 PMIG-Net은 Accuracy 0.7164, Macro-F1 0.6732, Macro-Recall 0.6828을 기록하였다. TFN, MulT, MAG-BERT 및 Self-MM의 핵심 융합 개념을 재구현한 adapted baselines와의 비교에서 특히 Macro-F1과 Macro-Recall에서 상대적으로 높은 성능을 보였다. 어블레이션 실험에서는 이미지 게이트 제거 시 Macro-F1과 Macro-Recall이 뚜렷하게 감소하였으며, 다른 구성요소들은 지표에 따라 상이한 trade-off를 보였다. 또한 MVSA-Single 추가 벤치마크 평가의 통제 비교에서도 PMIG-Single-Pass가 CLIP-only, BERT-only 및 BERT+CLIP Concat보다 높은 성능을 보여 이미지-텍스트 기반 소셜 미디어 감성 데이터에 대한 적용 가능성을 확인하였다. 본 연구는 약지도 라벨 구축과 선택적 멀티모달 융합을 하나의 프레임워크로 통합함으로써, 뉴스 여론 분석과 지능형 미디어 콘텐츠 분석을 위한 실용적인 멀티모달 감성 분석 방법을 제시한다.

▶ **주제어:** 멀티모달 감성 분석, 뉴스 감성 분석, 약지도 학습, 사전 지식 기반 어텐션, 이미지 신뢰도 게이트, 계층적 멀티태스크

I. Introduction

온라인 뉴스 플랫폼과 소셜 미디어의 확산에 따라 멀티모달 감성 분석(Multimodal Sentiment Analysis)은 자연어 처리(NLP), 컴퓨터 비전(CV), 감성 컴퓨팅 분야에서 중요한 연구 주제로 다루어지고 있다. 멀티모달 감성 분석은 텍스트와 이미지 등 서로 다른 모달리티의 정보를 통합함으로써 텍스트 단일모달 기반 방법보다 다양한 감성 단서를 활용할 수 있으며, 여론 분석, 뉴스 분석 및 추천 시스템 등 여러 응용 분야에 활용되고 있다[1][2]. 또한 사전학습 언어모델, 비전-언어 표현 학습 및 크로스모달(Cross-modal) 융합 기술의 발전으로 멀티모달 감성 분류 성능도 지속적으로 향상되고 있다[3][4-5].

그러나 뉴스를 대상으로 한 멀티모달 감성 분석 연구는 상대적으로 제한적이다. 대표적인 이미지-텍스트 기반 멀티모달 감성 데이터셋인 MVSA는 Twitter에서 수집된 이미지-텍스트 쌍을 기반으로 구축되어 소셜 미디어 환경의 감성 분석을 주된 대상으로 한다[6]. 뉴스 도메인에서도 아삼어(Assamese) 뉴스 기사를 대상으로 이미지와 텍스트를 결합한 감성 분석 연구가 제시되었으나[7], 제목, 본문

및 이미지의 감성 기여도 차이와 모달리티 간 불일치를 세분화하여 다루는 연구는 여전히 제한적이다.

뉴스 감성 분석은 소셜 미디어 기반 멀티모달 감성 분석과 구별되는 특성을 갖추고 있다. 소셜 미디어에서는 텍스트와 이미지가 감정이나 의견을 직접적으로 표현하는 경우가 많지만[8], 뉴스 이미지는 인물, 장소, 사건 등 기사 맥락을 보완하는 역할을 하는 경우가 많다[9]. 따라서 뉴스의 제목, 본문 및 이미지가 항상 동일한 감성 경향을 보이지 않으며, 일부 이미지는 사실 전달에 더 큰 비중을 둘 수 있다. 기존의 어텐션, 크로스 모달 Transformer 및 게이트 기반 융합 방법은 모달리티 간 상호작용을 학습하지만[10], 뉴스 환경의 약한 시각적 감성 단서와 텍스트-이미지 간 불일치를 충분히 반영하지 못할 수 있다. 이 경우 시각 정보를 일률적으로 결합하면 불필요한 노이즈가 포함될 가능성이 있다.

또한 대규모 멀티모달 데이터에 대한 수작업 감성 라벨링에는 많이 시간과 비용이 요구된다. 최근의 연구에서는 약지도 학습과 의사 라벨을 활용하여 학습 데이터를 구축

하는 방법이 활용되고 있으나, 자동으로 생성된 라벨에는 노이즈와 불확실성이 포함될 수도 있다[11][12].

본 연구에서는 이러한 문제를 다루고자 약지도 라벨 구축과 PMIG-Net 기반 감성 분류를 결합한 멀티모달 뉴스 감성 분석 프레임워크를 제안한다. 주요 기여는 다음과 같다.

첫째, 제목, 본문 및 이미지를 포함하는 약지도 기반 멀티모달 뉴스 감성 데이터셋 CNN-MMSA를 구축하고 복수의 약한 분류기와 인간 참조 라벨을 활용하여 약지도 라벨을 생성하였다.

둘째, 뉴스 감성 분류를 위한 PMIG-Net을 제안하였다. 제안 모델은 사전 지식 기반 어텐션, 이미지 게이트 및 계층적 멀티태스크 학습을 결합하여 모달리티 정보를 통합하고 이미지 특징의 반영 정도를 조절한다.

셋째, CNN-MMSA에서 비교 및 어블레이션 실험을 수행하고, MVSA-Single에서 추가 벤치마크 평가를 통해 제안 방법의 성능 특성과 적용 가능성을 분석하였다.

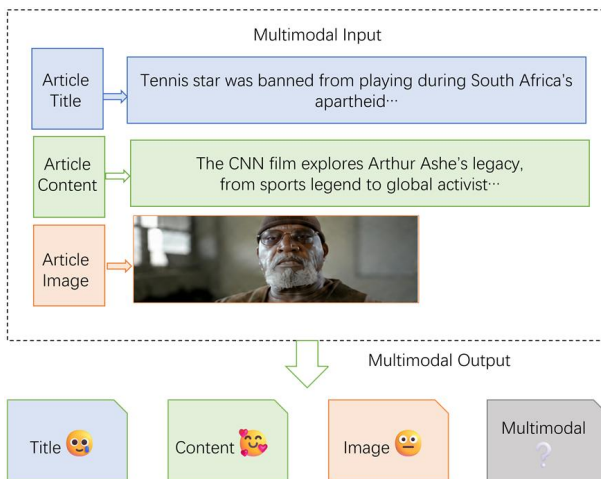


Fig. 1. Examples of Modality Sentiment Inconsistency in News Articles

II. Related Work

2.1 Sentiment Analysis Based on Text

텍스트 기반 감성 분석은 텍스트에 포함된 감성 극성이나 정서적 성향을 식별하는 것을 목적으로 하며, 초기에는 감성 사전과 수작업 특징을 활용한 기계학습 방법이 주로 사용되었다. 최근에는 사전학습 언어모델을 활용한 약지도 텍스트 분류[13], 감성 분석 방법론에 대한 체계적 연구[14], LLM 기반 감성 분석[15] 등으로 연구 범위가 확대되고 있다.

뉴스 텍스트의 경우 감성 단서가 명시적으로 드러나지 않는 경우가 많아 문맥을 고려한 분석이 중요하다. 기존 연구에서도 뉴스 감성 분석을 위해 정교한 라벨링과 문맥 적응의 필요성이 제기되었다[16]. 그러므로 뉴스 감성 분석연구에서는 개별 감성 어휘에 더해 제목과 본문의 문맥 정보를 함께 고려할 필요가 있으며, 이는 이후 멀티모달 분석에서 텍스트 표현을 구성하는 중요한 기반이 된다.

2.2 Multimodal Sentiment Analysis and Fusion Techniques

멀티모달 감성 분석은 서로 다른 모달리티의 정보를 결합하여 감성 표현을 학습하며, 최근에는 어텐션, 그래프 신경망, 대조 학습 및 Transformer 기반 융합 방법이 활발히 연구되고 있다. 이미지-텍스트 감성 분석에서는 그래프 기반 상호작용을 활용한 MGNNS[17], 대조 학습과 다층 융합을 결합한 CLMLF[18], 의미 강화와 크로스 모달 상호작용을 적용한 SECIF[19], 이미지 정보와 어텐션 기반 융합을 활용한 MSACA[20] 등이 제안되었다.

그럼에도 이러한 연구들은 주로 일반적인 이미지-텍스트 쌍의 상호작용과 융합에 초점을 두고 있으며, 제목-본문-이미지로 구성된 뉴스 구조와 모달리티별 사전 중요도를 함께 고려하는 연구는 상대적으로 제한적이다.

2.3 Weakly Supervised Sentiment Labeling and Multimodal Learning

약지도 학습은 휴리스틱 규칙, 외부 모델 및 의사 라벨링 등을 활용하여 대규모 수작업 라벨링에 대한 의존도를 줄이는 방법이다. 감성 분석에서는 높은 라벨링 비용과 라벨 노이즈 문제를 해결하기 위하여 자기 지도적 라벨 생성과 멀티태스크 학습을 결합한 방법이 제안되었으며[21], 노이즈 라벨 환경에서는 메타 학습 기반 라벨 보정을 통해 분류 강건성을 향상시키는 연구도 이루어졌다[22].

다만 기존 연구들은 주로 멀티모달 표현 융합, 모달 불균형, 감성 불일치, 또는 라벨 노이즈와 같은 개별 문제에 집중해 왔다. 뉴스의 제목-본문-이미지 구조를 고려하면서 약지도 라벨 구축과 멀티모달 융합을 하나의 프레임워크에서 함께 다룬 연구는 상대적으로 한정적이다. 본 연구는 복수의 약한 분류기와 인간 참조 라벨을 활용한 약지도 라벨 구축과, 사전 지식 기반 어텐션, 학습형 이미지 게이트 및 계층적 멀티태스크 학습으로 구성된 PMIG-Net을 하나의 프레임워크에서 결합한다는 점에서 기존 연구와 차별화된다.

III. The Proposed Scheme

3.1 Overview of the Proposed Framework

본 연구는 뉴스 기사의 특수성, 모달리티 간 불일치 문제, 또는 약지도 라벨의 노이즈와 신뢰도 문제를 고려하여 약지도 기반 뉴스 멀티모달 감성 분석 프레임워크를 설계한다.

그 프레임워크는 (1) 약지도 기반 감성 라벨 생성 단계와 (2) PMIG-Net 기반 감성 분류 단계로 구성된다. 첫 번째 단계에서는 복수의 약한 감성 분류기, 가중치 학습 및 저신뢰 표본 검토를 통해 멀티모달 감성 라벨을 구성한다. 두 번째 단계에서는 PMIG-Net이 사전 지식 기반 어텐션, 이미지 신뢰도 게이트 및 계층적 멀티태스크 학습을 활용하여 최종 감성을 분류한다. 각 단계의 세부 구성은 이후 절에서 설명한다.

3.2 Weakly Supervised Multimodal Sentiment Label Generation

본 연구는 약지도 기반 CNN-MMSA 데이터셋을 목적으로 뉴스 기사의 제목, 본문 및 이미지에 대한 감성 점수와 라벨을 생성한 후, 이를 텍스트 및 멀티모달 수준으로 단계적으로 통합하였다. 전체 과정은 그림 2에 제시하였다.

먼저 제목과 본문의 감성 정보를 추출하기 위해 VADER[23], TextBlob[24], NLTK Opinion Lexicon[25], 사전학습 Twitter-RoBERTa[26]의 네 가지 약한 감성 분류기를 사용하였다. VADER는 compound score를 기준으로 0.05 이상을 긍정, -0.05 이하를 부정, 그 사이를 중립으로 분류하였다. NLTK Opinion Lexicon은 긍정 및 부정 어휘의 출현 빈도 차이를 기반으로 감성 점수를 산출하였으며, TextBlob과 함께 ± 0.02 의 임계값을 적용하였다. Twitter-RoBERTa는 부정, 중립 및 긍정 클래스의 예측 확률 중 가장 높은 클래스를 감성 라벨로 선택하였다.

둘째, 약한 감성 주석기의 신뢰성을 평가하고 결합 가중치를 학습하기 위해 전체 데이터 D 에서 300개 샘플을 무작위로 추출하여 참조 집합 D_h (Step 1). 연구자 A와 B는 D_h 의 제목과 본문에 대해 독립적으로 감성 주석을 수행하였으며, Cohen's Kappa를 이용하여 주석자 간 일치도를 평가하였다. 그 결과 제목과 본문은 각각 0.9191과 0.8881로 나타나 높은 수준의 일치도를 보였다. 이후 불일치 사례는 두 연구자가 공동으로 검토하고 논의하여 최종 합의 라벨 Y_h 을 구성하였다(Step 2).

셋째, Y_h 을 종속변수로 하고, 네 가지 약한 감성 분류기의 감성 점수를 입력변수로 하여 Ridge 회귀를 학습하였다. Ridge 회귀는 식 (1)의 목적함수를 최소화하도록 최적화하였다(Step 3). 이를 통해 제목과 본문에 대한 분류기별 결합 가중치를 각각 추정하였다. Twitter-RoBERTa, VADER, TextBlob 및 NLTK Opinion Lexicon 순으로 본문 가중치(w_c)는 0.3153, 0.0810, 0.3624, 0.2414였으며, 제목 가중치(w_t)는 0.2334, 0.0643, 0.4645, 0.2378로 나타났다. 이후 전체 데이터의 제목과 본문 감성 점수는 네 분류기의 단순 평균이 아니라, 인간 참조 라벨을 기반으로 학습된 가중합으로 계산하였다(Step 5). 이 과정에서 제목의 최종 감성 점수와 감성 라벨을 각각 S_t 와 y_t , 본문의 최종 감성 점수와 감성 라벨을 각각 S_c 와 y_c 로 정의하였다.

$$\min_{w,b} \sum_{i=1}^n (y_i - (X_i w + b))^2 + \lambda \|w\|_2^2 \quad (1)$$

넷째, 제목과 본문의 최종 감성 점수 절댓값이 0.1 미만인 샘플을 저신뢰 텍스트로 정의하였다. 이 중 네 개의 약한 분류기 가운데 최소 세 개가 동일한 라벨을 제시한 경우 다수결로 보정하고, 그렇지 않은 경우 수작업 검토를 수행하였다(Step 6). 또는 이미지 라벨은 CLIP ViT-B/32 기반 제로샷 분류를 통해 생성하였으며[29]. 이미지 감성 점수 S_i , 이미지 라벨 y_i , 예측 신뢰도 q_i 를 산출하였다(Step 7). 이후 $q_i < 0.5$ 인 샘플은 저신뢰 이미지로 분류하고 수작업 검토를 수행하였다(Step 8).

다섯째, 최종 약지도 라벨은 단계적 가중 결합을 통해 생성하였다. 먼저 제목과 본문 감성 점수를 0.3:0.7의 비율로 결합하여 텍스트 감성 점수 S_{text} 을 산출하고(Step 9), 이후 텍스트와 이미지 감성 점수를 0.8:0.2의 비율로 결합하여 최종 멀티모달 감성 점수 S_{multi} 을 생성하였다(Step 10). 이때의 가중치는 뉴스에서 본문이 제목보다 더 많은 감성 맥락을 포함하고, 이미지는 텍스트를 보조하는 경우가 많다는 구조적 특성을 반영한 사전 결합 가중치이다. 해당 값은 학습을 통해 추정된 최적 파라미터가 아니며, 가중치 설정에 대한 민감도는 5.5절에서 별도로 분석하였다. 또한 이 가중치는 PMIG-Net 내부의 특징 융합 가중치와는 구분된다.

$$\begin{aligned} S_{text} &= 0.3S_t + 0.7S_c \\ S_{multi} &= 0.8S_{text} + 0.2S_i \end{aligned} \quad (2)$$

Algorithm 1 Weakly Supervised Multimodal Sentiment Label Generation

Input : News dataset $D=\{X_t, X_c, X_i\}$
Output : $\{y_t, y_c, y_i, y_{text}, y_{multi}\}$

1. $D_h \leftarrow \text{RandomSample}(D; 300)$
2. $Y_h \leftarrow \text{ConsensusAnnotation}(D_h)$
3. $W_t, W_c \leftarrow \text{LearnWeakAnnotatorWeights}(Y_h)$
4. for $n = 1$ to N do
5. $(S_t, y_t), (S_c, y_c) \leftarrow \text{WeakTextAnnotation}(X_t, X_c, W_t, W_c)$
6. Correct low-confidence text labels
7. $(S_i, y_i, q_i) \leftarrow \text{WeakImageAnnotation}(X_i)$
8. Review low-confidence image labels
9. $S_{text} \leftarrow \text{TextLabelFusion}(S_t, S_c)$
10. $S_{multi} \leftarrow \text{MultimodalLabelFusion}(S_{text}, S_i)$
11. end for
12. return $\{y_t, y_c, y_i, y_{text}, y_{multi}\}$

Fig. 2. Weakly Supervised Sentiment Label Construction Procedure

마지막으로 제목, 본문, 텍스트 및 최종 멀티모달 감성 점수에 기준 임계값을 적용하여 각각 제목 라벨 y_t , 본문 라벨 y_c , 텍스트 라벨 y_{text} , 이미지 라벨 y_i 및 최종 멀티모달 라벨 y_{multi} 을 생성하였다(Step 12). 이상의 절차를 통해 CNN-MMSA의 다섯 종류의 약지도 감성 라벨을 최종적으로 구축하였으며, 생성된 라벨은 PMIG-Net의 계층적 멀티태스크 학습에 사용하였다.

3.3 Overall Architecture of PMIG-Net

그림 3은 PMIG-Net의 전체 구조를 나타낸다. 뉴스 제목 (X_t), 본문(X_c), 이미지(I)를 입력으로 하여 BERT와 CLIP 인코더를 거쳐 제목 특징 h_t , 본문 특징 h_c , 이미지 특징 h_i 을 추출한다. 이후 제목과 본문 특징을 사전 지식 기반 어텐션으로 융합하여 텍스트 표현 h_{text} 을 생성하고, 이미지 신뢰도 게이트를 통해 이미지 특징을 샘플별로 조절한다. 최종적으로 제목, 본문 및 조절된 이미지 특징을 사전 지식 기반 멀티모달 어텐션으로 융합한 후 h_{multi} 을 생성한다.

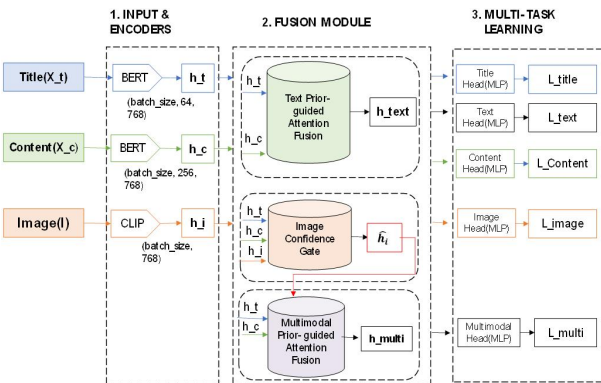


Fig. 3. Structure of the Proposed PMIG-Net Model

마지막으로 계층적 멀티태스크 학습을 통해 각 모달리티 출력과 최종 멀티모달 출력을 함께 최적화한다. $F_{PMIG}(\cdot)$ 는 제안 모델의 전체 처리 함수를 의미한다.

$$(h_t, h_c, h_i, h_{text}, h_{multi}) = F_{PMIG}(X_t, X_c, I) \quad (3)$$

3.4 Multimodal Feature Representation

이 단계에서 뉴스는 제목, 본문, 이미지로부터 멀티모달 특징을 추출하고자 함이다. 텍스트 특징 추출을 목적으로 뉴스 제목과 본문에 사전 학습된 BERT 인코더를 각각 적용하였고 마지막 히든 레이어에서 [CLS] 토큰 표현을 추출하여 Dropout을 활용하여 제목과 본문의 특징을 도출하였다.

$$\begin{aligned}
 h_t &= \text{Drop}(BERT_{title}(X_{t_{[CLS]}})) \\
 h_c &= \text{Drop}(BERT_{content}(X_{c_{[CLS]}}))
 \end{aligned} \quad (4)$$

여기서는 각각 뉴스 제목과 본문의 의미 정보를 나타내는 텍스트 특징 벡터이다. $BERT_{title}$ 과 $BERT_{content}$ 는 동일한 사전학습 가중치로 초기화되지만, 학습 과정에서는 서로 독립적으로 최적화되어 제목과 본문의 상이한 언어적 특성을 반영하도록 설계하였다.

시각 특징 추출을 위해 CLIP 기반의 이미지 인코더를 사용하였다. 이미지는 CLIP Vision Encoder에 입력되며, 인코더의 최종 출력으로부터 이미지 전체의 의미 정보를 포함하는 전역 시각 특징을 추출한다. 이후 과적합을 방지하기 위해 Dropout을 적용하였다.

$$h_i = \text{Drop}(CLIP_{vision}(I)_{pool}) \quad (5)$$

$CLIP_{vision}(I)_{pool}$ 은 CLIP 이미지 인코더의 pooled output을 의미하며, $\text{Drop}(\cdot)$ 은 과적합을 완화하기 위한 dropout 연산을 나타낸다. h_i 는 이미지 전체의 의미 정보를 집약적으로 표현하는 전역 시각 특징 표현을 나타낸다.

3.5 Prior-Guided Attention Fusion and Image Confidence Gate

본 절에서는 PMIG-Net의 핵심 융합 구조인 사전 지식 기반 어텐션 융합과 이미지 신뢰도 게이트를 설명한다. 그림 4는 제목, 본문 및 이미지 특징이 두 모듈을 거쳐 최종 멀티모달 표현으로 융합되는 과정을 나타낸다.

모달리티별 사전 가중치는 뉴스의 구조적 특성을 반영하여 제목, 본문 및 이미지의 전역적 중요도를 설정한다. 그러나 이미지가 제공하는 감성 정보의 정도는 샘플마다 다를 수 있으므로, 고정된 사전 가중치만으로는 이러한 차이를 충분히 반영하기 어렵다. 이에 이미지 신뢰도 게이트를 적용하여 샘플별 이미지 특징의 반영 정도를 조절한다.

따라서 사전 가중치와 이미지 신뢰도 게이트는 각각 모달리티 수준의 전역적 중요도와 샘플 수준의 이미지 특징 조절을 담당한다. 이때 이미지 신뢰도 게이트의 출력값은 이미지 신뢰도를 직접 측정하는 지표로 해석하지 않으며, 최종 이미지 기여도는 게이트 값과 어텐션 가중치의 결합을 통해 결정된다. 본 연구에서 Image Confidence Gate의 ‘confidence’는 이미지의 인간 수준 신뢰도나 유용성을 직접 측정하는 확률적 지표를 의미하지 않으며, 학습 과정에서 샘플별 이미지 특징의 반영 정도를 적응적으로 조절하는 게이팅 메커니즘을 의미한다.

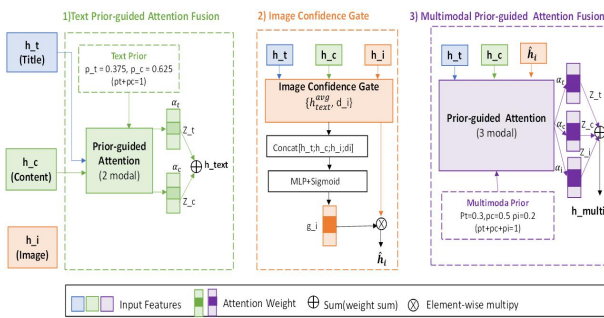


Fig. 4. Prior-Guided Attention Fusion and Image Confidence Gate

3.5.1 Prior-Guided Attention Fusion

그림 4의 첫 번째 단계에서는 제목 특징 h_t 와 본문 특징 h_c 을 입력으로 받아 텍스트 수준의 사전 지식 기반 어텐션을 수행한다. 일반적인 데이터 기반 어텐션은 학습된 점수에 따라 각 특징의 중요도를 결정하지만, 본 연구에서는 뉴스의 제목과 본문이 제공하는 정보의 차이를 반영하기 위해 모달리티별 사전 가중치를 함께 사용한다. 이를 통해 학습된 어텐션과 사전 정보를 결합하여 텍스트 융합 표현 h_{text} 을 생성한다.

입력 특징 집합을 $H = \{h_1, h_2, \dots, h_M\}$ 이라고 할 때, 각 모달리티 특징 h_m 은 먼저 학습 가능한 선형 투영을 통해 공통 융합 공간의 특징 Z_m 로 변환된다. 이후 MLP 기반의 어텐션 스코어링 함수 $fatt_{(.)}$ 에 입력되며, 각 모달리티에 대한 정규화 이전의 학습된 어텐션 점수 S_m 은 다음과 같이 계산된다.

$$Z_m = W_p h_m \quad (6)$$

$$S_m = fatt(Z_m) = w_a^T \tanh(W_a z_m + b_a) + b_s$$

이후 사전 가중치는 학습된 어텐션 점수에 로그 형태로 결합되며, 이를 통해 최종 어텐션 가중치를 계산한다. 식 (7)에서 p_m 은 모달리티별 사전 가중치이며, ϵ 은 수치적 안정성을 위한 작은 상수이다.

$$\alpha_m = \frac{\exp(s_m + \log(p_m + \epsilon))}{\sum_{k=1}^M \exp(s_k + \log(p_k + \epsilon))} \quad (7)$$

계산된 최종 어텐션 가중치를 이용하여 가중합을 수행하고 최종 텍스트 융합 표현 h_{text} 을 생성한다.

$$h_{text} = \sum_{m=1}^M \alpha_m z_m \quad (8)$$

그림 4의 마지막 단계에서는 제목 특징 h_t , 본문 특징 h_c 및 조절된 이미지 특징 $\hat{h}_i$ 을 이용하여 멀티모달 수준의 사전 지식 기반 어텐션을 수행한다.

$$h_{multi} = \sum_{m=1}^M \alpha_m z_m \quad (9)$$

최종적으로 계산된 어텐션 가중치를 이용하여 멀티모달 융합 표현 h_{multi} 을 생성하며, 이는 이후 멀티태스크 분류기의 입력으로 사용된다.

3.5.2 Image Confidence Gate

이미지 신뢰도 게이트는 개별 샘플에서 이미지 특징의 반영 정도를 동적으로 조절하기 위한 모듈이다. 먼저 제목 특징 h_t 와 본문 특징 h_c 의 평균을 이용하여 텍스트 기준 표현 h_{text}^{avg} 을 구성하고, 이를 이미지 특징 h_i 와 비교하여 텍스트-이미지 간 차이 특징을 계산한다.

$$h_{text}^{avg} = \frac{h_t + h_c}{2} \quad d_i = |h_{text}^{avg} - h_i| \quad (10)$$

이후 제목 특징 h_t , 본문 특징 h_c , 이미지 특징 h_i 및 차이 특징 d_i 를 MLP에 입력하여 이미지 게이트 계수를 산출한다.

$$g_i = \sigma(MLP([h_t, h_c, h_i, d_i])) \quad (11)$$

여기서 식 (11)에 $\sigma(\cdot)$ 는 sigmoid 함수이며, g_i 는 샘플별 이미지 특징의 반영 정도를 조절하는 게이트 계수이다. 게이트를 통해 조절된 이미지 특징은 이후 멀티모달 사전 지식 기반 어텐션에 입력되므로, 최종 융합에서 이미지의 상대적 기여도는 게이트 값뿐만 아니라 멀티모달 어텐션 가중치에 의해 함께 결정된다.

$$\hat{h}_i = g_i \cdot h_i \quad (12)$$

3.6 Hierarchical Multi-Task Learning and Optimization

본 연구는 모달리티별 감성 단서와 최종 멀티모달 표현을 함께 학습하기 위해 계층적 멀티태스크 학습 전략을 적용하였다. 이를 위해 제목, 본문, 이미지, 텍스트 융합 표현 및 최종 멀티모달 표현에 각각 독립적인 분류 헤드를 구성하고, 각 수준의 감성 표현을 동시에 최적화하도록 설계하였다. 이러한 보조 태스크의 공동 학습은 특정 모달리

티에 대한 과도한 의존을 완화하고, 표현 학습의 안정성과 일반화 성능을 높이는 데 목적이 있다.

각 분류 헤드에는 교차엔트로피 손실을 적용하였으며, 전체 멀티태스크 손실은 다음과 같이 다섯 개 태스크 손실의 가중합으로 정의하였다.

$$L_{MT} = \lambda_l L_l + \lambda_c L_c + \lambda_i L_i + \lambda_{text} L_{text} + \lambda_m L_{multi} \quad (13)$$

여기서 $L_l, L_c, L_i, L_{text}, L_{multi}$ 는 각각 제목, 본문, 이미지, 텍스트 융합 및 최종 멀티모달 분류 태스크의 교차엔트로피 손실을 의미한다. 손실 가중치는 $\lambda_l=0.15$, $\lambda_c=0.20$, $\lambda_i=0.15$, $\lambda_{text}=0.3$, $\lambda_m=0.45$ 로 설정하였다. 본문, 텍스트 융합 및 최종 멀티모달 태스크에 상대적으로 높은 가중치를 부여한 것은 해당 표현들이 제목이나 이미지 단독 표현보다 더 풍부한 문맥 및 감성 정보를 포함하기 때문이다.

또한 이미지 신뢰도 게이트의 결정이 지나치게 모호해지는 것을 방지하기 위해 게이트 정규화 항을 추가하였다. 식 앞에서 정의한 이미지 게이트 값 g_i 에 대해 정규화 항은 다음과 같이 계산한다.

$$L_{gate} = \frac{1}{B} \sum_{n=1}^B g_i^{(n)} (1 - g_i^{(n)}) \quad (14)$$

해당 항은 $g_i=0.5$ 일 때 최대가 되고, g_i 가 0 또는 1에 가까워질수록 감소한다. 따라서 이를 최소화함으로써 이미지 특징을 억제하거나 유지하는 게이트 결정이 보다 명확해지도록 유도하였다. 최종 학습 목표 함수는 다음과 같다.

$$L_{total} = L_{MT} + \lambda_g L_{gate} \quad (15)$$

여기서 $\lambda_g=0.05$ 로 설정하였다. 이를 통해 PMIG-Net은 모달리티별 보조 태스크와 최종 멀티모달 태스크를 함께 최적화하는 동시에, 이미지 게이트 값이 불확실한 중간 영역에 머무르는 경향을 완화하도록 학습된다.

IV. Experimental Design

4.1 Dataset and Label Construction

본 연구에서는 CNN 뉴스 웹사이트에서 수집한 기사 데이터를 바탕으로 약지도 학습 기법을 적용하여 구축한 멀티모달 감성 분석 데이터셋을 활용하였다. 그림 5는 데이터에 관한 주제별 분포를 나타낸다.

라벨 구축 절차는 3.2절에서 설명한 바와 같으며, 표 1은 Ridge 회귀 기반 가중치 학습과 주석 일치도 검증을 위해 무작위로 선정한 300개 표본의 인간 참조 주석 결과와 저신뢰 표본의 보정 현황을 요약한 것이다. 저신뢰 표본 중 제목 726개와 본문 923개는 약한 감성 분류기 간 다수결을 통해 보정하였으며, 나머지 제목 718개와 본문 521

개는 수작업 검토를 통해 최종 라벨을 확정하였다. 또한 CLIP 예측 신뢰도가 0.5 미만인 이미지 1,194개에 대해서도 수작업 검토를 진행하였다.

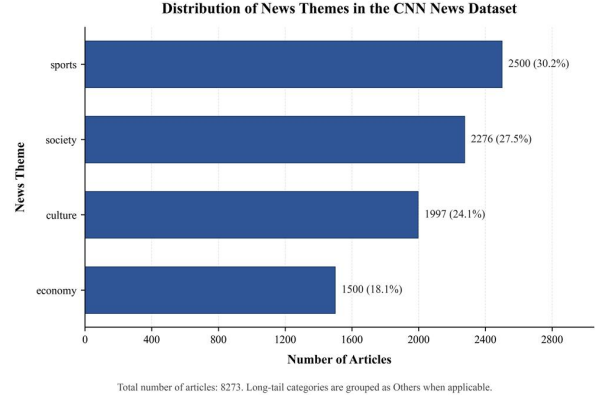


Fig. 5. Distribution of News Topics in the CNN News Dataset

Table 1. Manual Annotation Agreement and Label Correction Results

Metric	Title	Content	Image
Human sample	300	300	-
Cohen's Kappa	0.919	0.888	-
Reviewed samples (low-confidence)	1,444	1,444	1,194
Majority-vote corrected	726	923	-
Manual-review corrected	718	521	1,194

또한 데이터 누수를 방지하기 위해 동일 article_id뿐 아니라 article_link 중복의 샘플을 제거하였다. 다음으로 전체 멀티모달 감성 라벨을 기준으로 stratified sampling을 적용하여 학습, 검증 및 테스트 세트를 구성하였다. 데이터 분포는 표 2에 제시되었다.

Table 2. Sentiment Label Distribution of the CNN-MMSA Dataset

Dataset	Negative (-1)	Neutral (0)	Positive (1)	Total
Train	642	2,277	2,872	5,791
Validation	137	488	616	1,241
Test	138	488	615	1,241
Total	917	3,253	4,103	8,273

4.2 Experimental Settings and Evaluation Metrics

본 연구의 기본 실험 환경과 주요 하이퍼파라미터는 표 3에 제시하였다. CNN-MMSA의 단일·멀티모달 입력 비교, 주요 모델 비교 및 어블레이션 실험에는 동일한 데이터 분할, 전처리, 최대 입력 길이, batch size, optimizer, learning rate, weight decay 및 학습 epoch를 적용하였

다. 입력 모달리티에 따라 BERT-base-uncased와 CLIP ViT-B/32 인코더를 사용하였으며, random seed는 42로 설정하고 validation Macro-F1이 가장 높은 checkpoint를 최종 평가에 사용하였다.

감성 클래스의 분포가 불균형한 환경에서는 Accuracy만으로 모델의 균형적인 분류 성능을 충분히 평가하기 어렵다. 이에 본 연구에서는 Accuracy와 함께 Macro-F1 및 Macro-Recall을 주요 평가 지표로 사용하여 전체 분류 정확도와 클래스별 예측 성능을 함께 분석하였다.

Table 3. Experimental Settings and Hyperparameters

Parameter	Setting	Parameter	Setting
Framework	PyTorch	GPU	RTX 3070 Laptop GPU
Text Encoder	BERT-base-uncased	Image Encoder	CLIP ViT-B/32
Optimizer	AdamW	Epochs	5
Batch Size	4	Dropout	0.1
Learning Rate	2×10^{-5}	Weight Decay	1×10^{-4}
Title Max Length	64	Content Max Length	256
Image Size	224×224	Number of Classes	3
Random Seed	42	Selection Metric	Validation Macro-F1

4.3 Baseline Models

제안 모델의 성능을 검증하기 위해 단일모달리티 기준 모델과 멀티모달 기준 모델을 구성하였다. 단일모달리티 모델로는 BERT-base-uncased 기반의 제목 단독 모델과 본문 단독 모델, CLIP ViT-B/32 기반의 이미지 단독 모델을 사용하였다. 멀티모달 기준 모델로는 TFN[4], MulT[5], MAG-BERT[27] 및 Self-MM[28]을 사용하였다. 다만 본 연구에서는 각 방법의 공식 코드를 직접 사용한 것이 아니라, 각 모델의 핵심 융합 개념을 CNN-MMSA의 제목-본문-이미지 입력 구조에 맞게 재구현하였다. 따라서 이하에서 사용되는 TFN, MulT, MAG-BERT 및 Self-MM은 원 모델의 엄밀한 재현이라기보다 본 연구의 공통 실험 환경에 맞게 적용한 adapted implementation을 의미한다. TFN은 제목, 본문 및 이미지 특징과 모달리티 간 요소별 상호작용 특징을 결합하여 멀티모달 표현을 구성하였으며, MulT는 텍스트와 이미지 특징 간 크로스 모달 어텐션을 적용하였다. MAG-BERT는 텍스트와 이미지 특징으로부터 산출한 게이트를 이용하여 이미지 정보를 텍스트 표현에 결합하였고, Self-MM은 텍스트 및 이미지 보조 분류 태스크와 최종 멀티모달 분류 태스크를 함께 최적화하도록 구성하였다.

비교 모델의 복잡도를 파악하기 위해 전체 파라미터 수, 학습 가능한 파라미터 수 및 융합-분류 모듈의 파라미터 수를 비교하였다. 모든 모델은 동일한 BERT-base-uncased와 CLIP ViT-B/32 인코더를 사용하였으며, 주요 비교 실험에서는 모든 인코더와 융합 모듈을 함께 미세 조정하였으며 전체 파라미터가 모두 학습 가능 파라미터에 해당한다. 세부 결과는 표 4에 제시하였으며, PMIG-Net의 파라미터 규모는 비교 모델과 유사한 수준으로 나타나 성능 차이를 단순한 모델 규모 증가만으로 설명하기는 어렵다. 또는 주요 비교 실험에서는 모든 인코더와 융합 모듈을 함께 미세 조정하였으므로 전체 파라미터가 모두 학습 가능 파라미터에 해당한다.

Table 4. Parameter Counts and Model Complexity of Adapted Baseline Models

Model	Total Params	Trainable Params	Fusion/Head Params
TFN	308.91M	308.91M	2.49M
MulT	310.88M	310.88M	4.46M
MAG-BERT	308.98M	308.98M	2.56M
Self-MM	309.38M	309.38M	2.96M
PMIG-Net	309.41M	309.41M	2.99M

(Note. All models share the same BERT-base-uncased text encoders (218.96M parameters) and CLIP ViT-B/32 image encoder (87.46M parameters).)

V. Experiments and Analysis

5.1 Performance Comparison of Unimodal and Multimodal Models

본 절에서는 입력 모달리티에 따른 최종 감성 분류 성능과 제안 모델의 멀티모달 융합 효과를 대조하고자 Title, Content, Image, Text 및 PMIG-Net을 평가하였다. Title과 Content는 각각 BERT-base-uncased를 이용하여 제목 또는 본문만을 입력으로 사용하는 단일 텍스트 모델이며, Image는 CLIP ViT-B/32를 이용한 단일 이미지 모델이다. Text는 제목과 본문을 각각 독립적인 BERT로 인코딩한 후 사전 지식 기반 어텐션을 통해 융합하는 텍스트 모델이다. 이들 비교 모델은 서로 다른 입력 조건에서 동일한 최종 감성 예측 능력을 비교하기 위해 모두 최종 멀티모달 라벨을 감독 신호로 사용하여 독립적으로 학습 평가하였다. 따라서 표 5의 결과는 PMIG-Net 내부의 모달리티별 보조 헤드 성능과 구분된다.

표 5에서 Title과 Content는 Image보다 높은 성능을 보여 뉴스 감성 분석에서 텍스트 정보가 최종 감성 판단의

주요 단서로 작용함을 보여주었다. 한편 제목과 본문을 함께 융합한 Text는 Content 대비 뚜렷한 성능 향상을 보이지 않았다. PMIG-Net은 Accuracy 0.7164, Macro-F1 0.6732, Macro-Recall 0.6828로 가장 높은 성능을 기록하였다. 이는 텍스트 중심의 감성 정보에 이미지 정보를 선택적으로 결합하고 이를 멀티모달 수준에서 융합하는 전체 구조가 단일 입력 및 텍스트 전용 모델보다 효과적인 성능을 보였음을 나타낸다.

Table 5. Comparison of Sentiment Classification Performance of Unimodal and Multimodal Model

Model	Accuracy	Macro-F1	Macro-Recall
Title	0.6753	0.6254	0.6455
Content	0.6962	0.6182	0.6005
Image	0.4956	0.2209	0.3333
Text	0.6938	0.6103	0.5967
PMIG-Net	0.7164	0.6732	0.6828

5.2 Performance Analysis of Hierarchical Prediction Heads in PMIG-Net

표 6은 PMIG-Net의 계층적 멀티태스크 학습에서 Title, Content, Image, Text 및 Multimodal 예측 head의 성능을 분석한 결과이다. 각 head는 해당 계층의 개별 감성 라벨을 감독 신호로 사용하므로, 표 6은 동일한 분류 과제의 직접적인 성능 비교보다는 각 계층에서 감성 정보가 어떻게 학습되는지를 보여준다.

Title head는 Macro-F1 0.6745와 Macro-Recall 0.6907로 비교적 균형적인 성능을 보였으며, Content head는 상대적으로 낮은 Macro-F1과 Macro-Recall을 기록하였다. Image head는 Macro-F1 0.2716, Macro-Recall 0.3333으로 이미지 수준의 감성 분류가 상대적으로 어려움을 보여주었다. Text head는 Macro-F1 0.6538과 Macro-Recall 0.6825를 기록하였고, 최종 Multimodal head는 Accuracy 0.7164, Macro-F1 0.6732, Macro-Recall 0.6828을 나타냈다.

이러한 결과는 PMIG-Net이 제목, 본문, 이미지, 텍스트 융합 및 최종 멀티모달 수준에서 각각 별도의 감독 신호를 활용하여 계층별 감성 표현을 학습함을 보여준다.

Table 6. Performance Analysis of Prediction Heads in PMIG-Net's Hierarchical Multi-Task Learning

Model	Accuracy	Macro-F1	Macro-Recall
Title head	0.6890	0.6745	0.6907
Content head	0.6865	0.5139	0.5905
Image head	0.6873	0.2716	0.3333
Text head	0.6890	0.6538	0.6825
Multimodal head	0.7164	0.6732	0.6828

5.3 Analysis of Fusion Strategies and Comparison with Existing Multimodal Models

본 절에서는 융합 방식 자체의 영향을 비교하기 위해 동일한 계층적 멀티태스크 구조를 유지한 상태에서 BERT 및 CLIP 인코더를 고정하고, 융합 모듈과 분류 헤드만 학습하였다. Mean Fusion은 각 모달리티를 동일하게 반영하고, Prior-Weighted Fusion은 사전에 설정된 모달리티 가중치를 적용한다. Attention Fusion은 샘플별 특징에 따라 융합 가중치를 학습하며, Prior-guided Attention Fusion은 학습된 어텐션에 사전 가중치를 함께 반영한다. 모든 방법은 동일한 데이터 분할과 학습 조건에서 최종 멀티모달 출력의 성능을 기준으로 평가하였다.

표 7에서 Mean Fusion은 Accuracy 0.6832, Macro-F1 0.5015, Macro-Recall 0.5196으로 가장 낮은 성능을 보였다. Prior-Weighted Fusion은 세 지표 모두에서 Mean Fusion보다 향상되었으며, Attention Fusion은 Accuracy 0.6890, Macro-F1 0.5486, Macro-Recall 0.5465로 가장 높은 성능을 기록하였다. Prior-guided Attention Fusion은 Macro-F1 0.5482로 Attention Fusion과 유사한 수준을 보였으나, Accuracy와 Macro-Recall에서는 소폭 낮았다. 이는 고정 인코더 조건에서 학습 가능한 어텐션 기반 융합이 단순 평균 및 고정 가중 융합보다 효과적임을 보여주는 반면, 사전 가중치를 추가한 효과는 뚜렷하게 나타나지 않았음을 의미한다. 따라서 본 실험은 동일한 구조와 학습 조건에서 융합 방식 자체의 영향을 비교하기 위한 통제 실험으로 활용되며, 이후에는 인코더까지 함께 학습하는 종단 간 조건에서 PMIG-Net과 기존 멀티모달 모델의 성능을 비교한다.

Table 7. Comparison of Multimodal Fusion Strategies with Frozen Encoders

Fusion Method	Accuracy	Macro-F1	Macro-Recall
Mean Fusion	0.6832	0.5015	0.5196
Prior-Weighted Fusion	0.6865	0.5371	0.5382
Attention Fusion	0.6890	0.5486	0.5465
Prior-guided Attention Fusion	0.6849	0.5482	0.5426

PMIG-Net의 전체적인 성능을 검증하기 위해 인코더를 포함한 전체 구조를 학습 가능한 상태로 설정하고 TFN, MulT, MAG-BERT 및 Self-MM과 비교하였다. 비교 모델에는 앞서 설명한 동일한 데이터 분할, 공통 인코더 및 주요 학습 조건을 적용하였다.

표 8에서 TFN은 Accuracy 0.7212로 가장 높은 값을 보였으며, PMIG-Net은 0.7164로 이에 근접한 성능을 기

록하였다. 반면 PMIG-Net은 Macro-F1 0.6732와 Macro-Recall 0.6828로 모든 비교 모델 중 가장 높은 값을 보였다. 이는 PMIG-Net이 전체 정확도를 경쟁력 있는 수준으로 유지하면서 클래스별 균형 성능에서 상대적인 강점을 보였음을 나타낸다.

Table 8. Comparison with Multimodal Baseline Models on CNN-MMSA

Method	Accuracy	Macro-F1	Macro-Recall
MAG-BERT	0.7075	0.6628	0.6730
TFN	0.7212	0.6462	0.6249
Self-MM	0.7010	0.6369	0.6323
MuT	0.7139	0.6481	0.6336
PMIG-Net	0.7164	0.6732	0.6828

5.4 Ablation Study and Balanced Performance Analysis

본 절에서는 PMIG-Net의 각 구성 요소가 최종 감성 분류 성능에 미치는 영향을 분석하기 위해 어블레이션 실험을 실시하였다. 실험은 두 단계로 구성되어 있다. 첫째, 전체 모델에서 하나의 모듈만 제거하여 각 구성 요소의 개별 기여도를 분석하였다. 둘째, 이미지 신뢰도 게이트를 제거한 구조를 기준으로 다른 모듈을 함께 제거하여 모듈 간 결합 효과를 분석하였다. 각각 실험의 구성은 다음과 같다.

- w/o P : 사전 가중치를 제거하고 각 모듈리티에 동일한 가중치를 적용한 모델이다.
- w/o A : 사전 가중치는 유지하되, 샘플별 적응적 어텐션 가중치 학습만 제거한 모델이다.
- w/o G : 이미지 신뢰도 게이트를 제거하고 나머지 융합 및 멀티태스크 구조를 유지한 모델이다.
- w/o M : 계층적 보조 태스크의 감독을 제거하고 최종 멀티모달 태스크만 학습한 모델이다.

표 9는 PMIG-Net의 각 구성 요소를 개별적으로 제거한 단일 모듈 어블레이션 결과이다. 각 모듈의 제거는 모든 지표에서 일관된 성능 저하를 보이지 않았으며, Accuracy와 클래스 균형 지표 간에 서로 다른 trade-off가 나타났다. 특히 w/o G는 Accuracy가 0.7268로 증가했지만 Macro-F1과 Macro-Recall은 각각 0.6433과 0.6230으로 크게 감소하여, 이는 이미지 게이트가 클래스별 균형 성능 유지에 기여할 수 있음을 시사한다. w/o A와 w/o M 역시 Accuracy는 증가했으나 Macro-Recall은 감소하였고, w/o P는 Accuracy가 소폭 감소한 반면 Macro-F1과 Macro-Recall은 다소 증가하였다. 따라서 각 구성 요소의 효과는 단일 지표보다 Accuracy, Macro-F1 및 Macro-Recall을 함께 고려하여 해석할 필요가 있다.

Table 9. Single-Module Ablation Study Results of PMIG-Net

Model	Accuracy	Macro-F1	Macro-Recall
w/o P	0.7123	0.6748	0.6910
w/o A	0.7349	0.6753	0.6665
w/o G	0.7268	0.6433	0.6230
w/o M	0.7212	0.6733	0.6696
PMIG-Net	0.7164	0.6732	0.6828

표 10에서 P, A, G, M은 각각 사전 가중치, 샘플별 적응적 어텐션 가중치, 이미지 신뢰도 게이트, 계층적 멀티태스크 학습을 의미하며, '+'는 해당 모듈을 동시에 제거한 조합을 나타낸다. w/o P+G는 Accuracy 0.7349로 높아졌고 Macro-F1 PMIG-Net보다 소폭 높은 값을 보였으나, Macro-Recall은 0.6616으로 현저히 감소하였다. w/o A+G와 w/o M+G에서는 Accuracy가 PMIG-Net보다 높았으나, Macro-F1과 Macro-Recall은 모두 감소하였으며, 특히 w/o M+G에서는 Macro-F1과 Macro-Recall이 각각 0.6183과 0.6007로 가장 큰 성능 저하를 보였다. 결론적으로 PMIG-Net은 Accuracy 0.7164를 유지하면서 Macro-F1 0.6732와 가장 높은 Macro-Recall 0.6828을 기록하여, 단순히 전체 정확도를 높이기보다는 클래스별 예측 성능을 보다 균형적으로 유지하는 특성을 보였다.

Table 10. Module Combination Ablation Results without the Image Confidence Gate

Model	Accuracy	Macro-F1	Macro-Recall
w/o P + G	0.7349	0.6745	0.6616
w/o A + G	0.7292	0.6543	0.6350
w/o M + G	0.7252	0.6183	0.6007
PMIG-Net(Ours)	0.7164	0.6732	0.6828

5.5 Validation of Weak-Supervision Labels

약지도 라벨의 구성 타당성과 평가 신뢰성을 확인하기 위해, 라벨 생성 가중치에 대한 민감도 분석과 독립적인 인간 참조 라벨을 이용한 추가 평가를 수행하였다. 먼저 약지도 라벨 생성 과정에서 사용한 결합 가중치의 영향을 분석하는 점을 고려할 때 데이터 분할, 모델 구조, 학습 하 이퍼파라미터 및 평가 절차를 동일하게 유지하고 라벨 생성 단계의 가중치만 변경하였다. 제목-본문 가중치는 기존 0.3:0.7을 기준으로 0.4:0.6 및 0.2:0.8로 변경하였으며, 텍스트-이미지 가중치는 기존 0.8:0.2를 기준으로 0.9:0.1 및 0.7:0.3으로 변경하여 비교하였다.

표 11의 결과에서 기존 설정인 제목 0.3, 본문 0.7, 텍스트 0.8, 이미지 0.2는 Accuracy 0.7164, Macro-F1 0.6732, Macro-Recall 0.6828로 검토한 설정 중 가장 우수한 전반적 성능을 보였다. 제목-본문 비율을 변경한 경

우 세 평가 지표가 모두 감소하였으며, 이미지 가중치를 0.1로 낮춘 경우에는 기존 설정과 유사한 성능을 유지하였다. 반면 이미지 가중치를 0.3으로 증가시킨 경우 성능이 크게 감소하였다. 이는 검토한 범위에서 텍스트 중심, 이미지 보조의 가중치 설정이 상대적으로 적절하며, 기존 설정이 비교적 균형적인 성능을 제공함을 보여준다.

Table 11. Sensitivity Analysis of Weak-Supervision Label Fusion Weights

Title:Content / Text:Image	Accuracy	Macro-F1	Macro-Recall
0.3:0.7 / 0.8:0.2	0.7164	0.6732	0.6828
0.4:0.6 / 0.8:0.2	0.6857	0.6354	0.6368
0.2:0.8 / 0.8:0.2	0.7091	0.6549	0.6414
0.3:0.7 / 0.9:0.1	0.7147	0.6726	0.6704
0.3:0.7 / 0.7:0.3	0.6181	0.5814	0.6171

또한 약지도 라벨에 대한 의존성을 추가로 확인하기 위해 기존 라벨 구축에 사용된 참조 표본과 중복되지 않는 200개 테스트 샘플에 대해 독립적인 인간 참조 라벨을 구축하고 동일한 PMIG-Net 예측 결과를 평가하였다. 인간 참조 라벨은 두 명의 평가자가 약지도 라벨 및 PMIG-Net의 예측 결과를 확인하지 않은 상태에서 각 샘플을 독립적으로 주석하여 구축하였다. 두 평가자의 판단이 일치하지 않은 경우에는 해당 사례를 함께 재검토하고 논의하여 최종 합의 라벨을 확정하였다. 표 12에 대한 인간 참조 라벨을 기준으로 평가했을 때 Accuracy는 0.6650에서 0.6800으로, Macro-F1은 0.6312에서 0.6480으로, Weighted-F1은 0.6651에서 0.6861로 소폭 증가하였으며, Macro-Recall은 유사한 수준을 유지하였다. 이러한 결과는 제안 모델이 약지도 라벨에만 의존하지 않고 인간 판단과 일치하는 감성 패턴도 학습하고 있음을 보여주는 보조적 근거를 제공한다.

Table 12. PMIG-Net Performance on the Same 200 Samples According to Ground Truth

Evaluation Ground Truth	Accuracy	Macro-F1	Macro-Recall	Weighted-F1
Weak labels	0.6650	0.6312	0.6494	0.6651
Independent Human labels	0.6800	0.6480	0.6396	0.6861

5.6 Robustness and Random Seed Stability Analysis

표 13은 5개의 랜덤 시드에 대한 반복 실험을 활용하여 PMIG-Net의 성능 변동과 MAG-BERT와의 차이를 비교한

결과이다. PMIG-Net은 Accuracy와 Macro-F1에서 각각 4개 시드, Macro-Recall에서 3개 시드에서 MAG-BERT보다 높은 값을 보였으며, 세 지표의 평균 성능도 모두 높게 나타났다. 이는 제한된 반복 조건에서 제안 모델이 랜덤 초기화와 데이터 샘플링 변화에도 비교적 일관된 성능 경향을 보였음을 나타낸다.

표 14는 5개의 랜덤 시드 반복 실험을 바탕으로 PMIG-Net과 MAG-BERT 간 성능 차이를 통계적으로 분석한 결과이다. PMIG-Net은 세 지표에서 평균적으로 높은 값을 보였으며, Accuracy는 대응표본 t-검정에서 유의한 차이를 보였으나($p=0.0207$), Wilcoxon 검정에서는 유의수준에 도달하지 않았다($p=0.0625$). Macro-F1과 Macro-Recall에서는 통계적으로 유의한 차이가 확인되지 않았다. 따라서 PMIG-Net은 5개의 랜덤 시드에서 평균적으로 더 높은 성능을 보였으나, 통계적 유의성이 모든 지표와 검정 방법에서 일관되게 확인되지는 않았다.

Table 13. Random Seed-Based Performance Stability and Model Comparison Results

Seed	PMIG Acc	MAG Acc	PMIG F1	MAG F1	PMIG Rec	MAG Rec
42	0.7164	0.7075	0.6732	0.6628	0.6828	0.6730
52	0.7421	0.7196	0.6889	0.6527	0.6741	0.6383
62	0.7204	0.7091	0.6505	0.6589	0.6328	0.6549
72	0.7357	0.7115	0.6854	0.6465	0.6841	0.6287
82	0.7357	0.6938	0.6587	0.6431	0.6382	0.6475
Mean	0.7301	0.7083	0.6713	0.6528	0.6624	0.6485
Std	0.0111	0.0093	0.0166	0.0082	0.0249	0.0169

(Note. PMIG and MAG refer to PMIG-Net and MAG-BERT, respectively, while Acc, F1, and Rec denote Accuracy, Macro-F1, and Macro-Recall, respectively.)

Table 14. Statistical Significance Analysis Results

Metric	Mean Difference	Paired t-test (p)	Wilcoxon (p)	Cohen's d
Accuracy	+0.0218	0.0207	0.0625	1.66
Macro-F1	+0.0185	0.1008	0.1250	0.95
Macro-Recall	+0.0139	0.3839	0.4375	0.44

5.7 Additional Benchmark Evaluation on MVSA-Single

5.7.1 Dataset and Evaluation Protocol

제안 방법의 추가적인 적용 가능성을 평가하기 위해 공개 벤치마크인 MVSA-Single을 사용하였다. 전처리 후 4,511개의 이미지-텍스트 쌍을 학습 3,611개, 검증 450개, 테스트 450개로 분할하였다. 모든 모델은 동일한 데이터 분할에서

MVSA-Single로 다시 학습하였으며, CNN-MMSA에서 학습된 모델을 직접 전이하지 않았다. 따라서 본 실험은 교차 데이터셋 일반화가 아닌 추가 벤치마크 평가로 해석한다. 모든 실험은 seed=42, 5 epochs 조건에서 수행하고 validation Macro-F1을 기준으로 최적 모델을 선택하였다.

5.7.2 Effect of Duplicated Text Input

MVSA-Single은 하나의 텍스트만 제공하므로, 기존 PMIG-Net을 그대로 적용할 경우 동일한 텍스트가 제목과 본문 분기에 반복 입력된다. 이에 입력 복제의 영향을 검증하기 위해 Duplicated, Shared, Single-Pass 및 Single 구조를 비교하였다. Duplicated는 두 개의 독립적인 BERT를 사용하고, Shared는 하나의 BERT를 공유하여 두 번 인코딩하며, Single-Pass는 텍스트를 한 번만 인코딩한 후 동일한 표현을 두 분기에 재사용한다. Single은 제목-본문 구분을 제거하고 하나의 텍스트 분기만 사용하는 구조이다.

Table 15. Effect of Text-Input Duplication on MVSA-Single

Variant	#Enc.	#Pass	Acc.	Macro-F1	Macro-Recall
Duplicated	2	2	0.6889	0.5922	0.5946
Shared	1	2	0.7000	0.5819	0.5742
Single-Pass	1	1	0.7244	0.6146	0.6017
Single	1	1	0.7000	0.6061	0.5972

(Note. #Enc. denotes the number of BERT encoders, and #Pass denotes the number of BERT forward passes per sample.)

표 15에서 Single-Pass는 Accuracy 0.7244, Macro-F1 0.6146, Macro-Recall 0.6017로 가장 높은 성능을 보였으며, Single 역시 Duplicated에 비교하여 높은 Macro-F1을 기록하였다. 이는 단일 텍스트 환경에서 동일 입력을 여러 인코더에 반복 입력하는 것이 필수적이지 않으며, 한 번 생성한 텍스트 표현을 재사용하는 방식이 MVSA-Single에 보다 적합할 수 있음을 보여준다.

5.7.3 Controlled Comparison on MVSA-Single

제안 모델의 멀티모달 적용 효과를 확인하고자 동일한 실험 조건에서 BERT-only, CLIP-only 및 BERT+CLIP Concat과 비교하였다. BERT-only는 텍스트만을, CLIP-only는 이미지만을 사용하며, BERT+CLIP Concat은 두 모달리티의 특징을 단순 연결하여 분류한다. 모든 모델은 동일한 데이터 분할과 학습 조건에서 평가하였다.

표 16에서 CLIP-only는 가장 낮은 성능을 보였고, BERT-only는 상대적으로 높은 성능을 기록하여 MVSA-Single에서 텍스트 정보의 중요성을 확인하였다. 반면 BERT+CLIP Concat은 BERT-only보다 뚜렷한 향상을 보이지 않았다. PMIG-Single-Pass는 Accuracy 0.7244, Macro-F1 0.6146, Macro-Recall 0.6017, Weighted-F1 0.7174로 모든 지표에서 가장 높은 성능을 보였다. 이는 단순한 특징 연결보다 제안된 융합 구조가 텍스트와 이미지 정보를 보다 효과적으로 활용할 가능성을 보여준다.

Table 16. Controlled Comparison on MVSA-Single under an Identical Experimental Protocol

Model	Acc.	Macro-F1	Macro-Recall	Weighted-F1
CLIP-only	0.5000	0.3345	0.3579	0.4822
BERT-only	0.7156	0.5866	0.5623	0.6975
BERT+CLIP Concat	0.6889	0.5775	0.5710	0.6843
PMIG-Single-Pass	0.7244	0.6146	0.6017	0.7174

5.7.4 Comparison with Literature-Reported

MVSA-Single Results

MVSA-Single에서 제안 모델의 상대적 성능을 확인하기 위해 기존 연구의 보고 결과와 비교하였다. 다만 연구마다 데이터 전처리, 분할 및 평가 방식이 다른 바, 표 17은 엄밀한 직접 비교가 아닌 참고적 비교로 해석한다.

표 17에서 PMIG-Single-Pass는 일부 MVSA-Single 특허 모델보다 낮은 절대 성능을 보였다. 그러나 동일한 조건에서 수행한 표 16의 통제 비교에서는 BERT-only, CLIP-only 및 BERT+CLIP Concat보다 높은 성능을 기록하였다. 따라서 표 17은 제안 모델의 공개 벤치마크 내 상대적 위치를 확인하기 위한 보조적 결과로 활용하였다.

Table 17. Literature-Reported Performance on MVSA-Single

Method	Year	Acc.	F1	F1 Type
MGNNS[17]	2021	0.7377	0.7270	W-F1
CLMLF[18]	2022	0.7533	0.7346	W-F1
SECIF[19]	2025	0.7910	0.7874	Reported F1
MSACA[20]	2026	0.7578	0.7462	W-F1
PMIG-Single-Pass	This study	0.7244	0.7174	W-F1

(Note. W-F1 denotes Weighted-F1. Results of previous studies are reproduced from the corresponding publications. Because experimental protocols and F1 definitions may differ across studies, these values are provided for contextual comparison rather than strict head-to-head evaluation.)

5.8 Qualitative and Visualization Analysis

5.8.1 Representative Case-Based Analysis of Image Contribution

테스트 셋에서 텍스트와 이미지의 감성 단서가 불일치한 사례, 일치한 사례 및 중립적 사례를 선정하여 이미지 융합 동작을 정성적으로 분석하였다. 그림 6에서 세 사례의 Image Gate 값은 모두 0.97 이상으로 나타나, Gate 값만으로 이미지의 유용성이나 신뢰도를 직접 구분하기에는 한계가 있었다. 반면 Image Attention Weight는 사례별로 0.0110~0.0236의 차이를 보였다.

이에 Gate와 Attention의 결합 효과를 살펴보기 위한 보조적 지표로 두 값의 곱을 Effective Image Weight로 정의하였다. 불일치 사례에서는 0.0108로 가장 낮았고, 일치 사례에서는 0.0234로 약 2.2배 높았으며, 중립 사례에서는 0.0121로 나타났다. 이러한 결과는 Image Gate를 이미지 신뢰도의 직접적인 측정치라기보다 이미지 특징의 반영 정도를 조절하는 학습형 게이트 계수로 해석하는 것이 적절하며, 최종 이미지 반영 정도는 Gate와 멀티모달 어텐션의 결합에 의해 달라질 수 있음을 시사한다.

Case Type	Text	Text Label	Image Label	Image Gate	Image Attention	Effective Image Weight	Prediction	Result
Text-Image Conflict	Michael Jordan's high school basketball coach, ...	Neutral	Positive	0.9804	0.0110	0.0108	Positive	Correct
Text-Image Consistent	The Fed's huge challenge: Taming inflation...	Positive	Positive	0.9886	0.0236	0.0234	Positive	Correct
Weak/Neutral Image	Facebook knew it was being used to incite...	Neutral	Neutral	0.9734	0.0124	0.0121	Neutral	Correct

Fig. 6. Representative Cases for Qualitative Analysis

5.8.2 Visualization Analysis

그림 7은 테스트 세트 전체에서 산출된 Effective Image Weight의 분포를 제시한다. Effective Image Weight는 Image Gate와 Image Attention Weight의 결합 효과를 나타내는 보조적 지표이며, 전체 평균은 0.016으로 나타났다. 대부분의 값이 비교적 낮은 범위에 분포하므로, 본 데이터에서 이미지 특징이 최종 융합 과정에서 제한적인 상대적 가중치로 반영되는 경향을 보여준다. 앞서 그림 6에서 분석한 대표 사례들도 이러한 분포의 서로 다른 구간에 위치하였으며, 이는 샘플에 따라 이미지 반영 정도가 달라질 수 있음을 정성적으로 보여준다.

불균형한 클래스 분포에서의 분류 성능을 보다 구체적으로 분석하기 위해 대표 비교 모델인 MAG-BERT와 혼동 행렬을 비교하였다. 그림 8과 같이 PMIG-Net은 Accuracy

를 0.7075에서 0.7164로, Macro-F1을 0.6628에서 0.6732로 향상시켰으며, Neutral 클래스의 정분류 비율도 63.1%에서 66.0%로 증가하였다. 이러한 결과는 제안 모델이 감성 경계가 상대적으로 모호한 Neutral 클래스에서도 분류 성능을 개선하고, 클래스 불균형 환경에서 보다 균형적인 감성 예측을 수행할 수 있음을 보여준다.

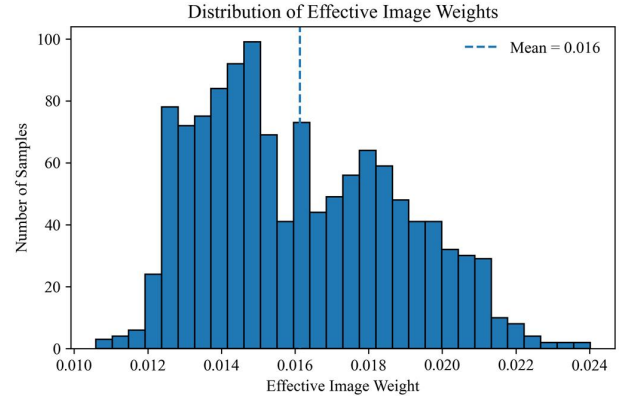


Fig. 7. Distribution of Effective Image Weights

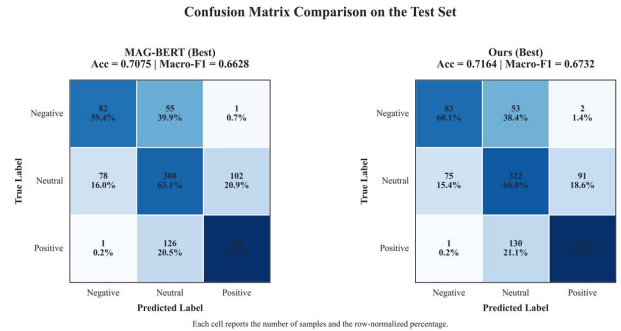


Fig. 8. Confusion Matrix Comparison of MAG-BERT and PMIG-Net on the Test Set

그림 9는 학습 epoch에 따른 대표 멀티모달 모델의 검증 성능 변화를 보여준다. 그림 9(a)는 Accuracy, 그림 9(b)는 Macro-F1, 그림 9(c)는 Macro-Recall의 변화를 각각 나타낸다. Accuracy에서는 모델별로 epoch에 따른 일정한 변동이 관찰되었으며, 일부 비교 모델은 학습 후반부에서 상대적으로 높은 값을 기록하였다. 반면 PMIG-Net은 Macro-F1과 Macro-Recall에서 전반적으로 안정적인 성능을 유지하였으며, 특히 일부 비교 모델이 특정 epoch에서 큰 성능 변동을 보인 것과 달리 비교적 작은 변동 폭을 나타냈다. 이러한 결과는 제안 모델이 클래스 불균형과 약지도 기반 라벨 노이즈가 존재하는 환경에서도 보다 안정적으로 수렴하며, 균형적인 감성 분류 성능을 유지할 수 있음을 시사한다. 다만 본 결과는 최종 테스트 성능을 비교하기 위한 것이 아닌, 모델의 학습 과정

과 수렴 특성을 분석하기 위한 보조적 결과로 해석하였다.

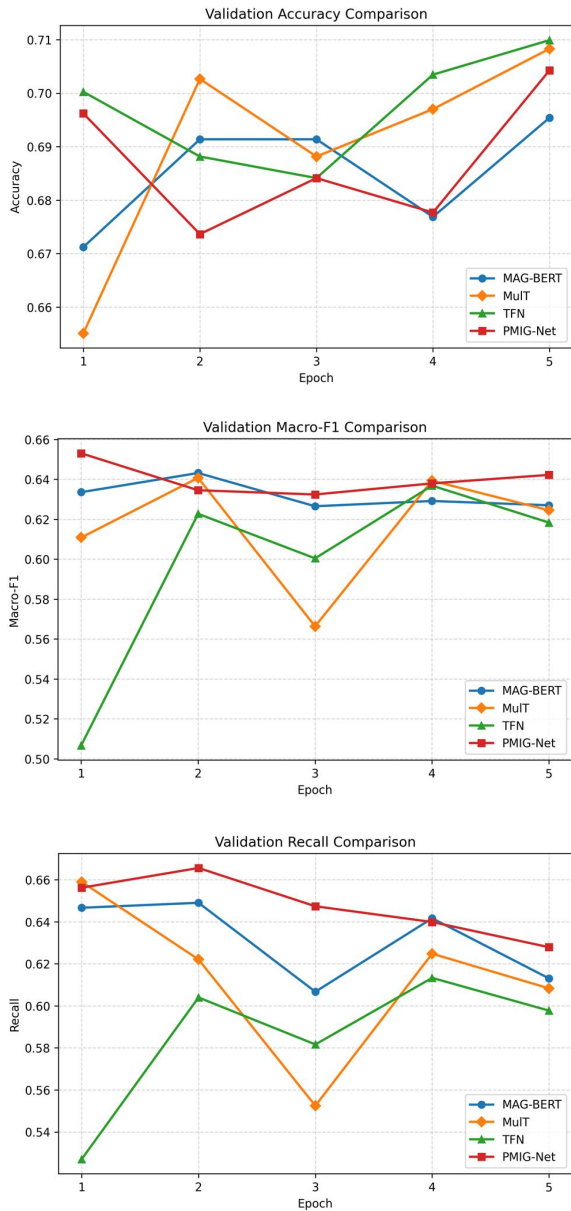


Fig. 9. Validation performance trends of representative multimodal models: (a) Accuracy; (b) Macro-F1; (c) Macro-Recall.

VI. Discussion and Conclusion

6.1 Discussion of the Research Findings

본 연구에서는 약지도 기반 뉴스 멀티모달 감성 분석에 기여하고자 CNN-MMSA 데이터셋을 구축하고, 뉴스 데이터의 특성을 반영한 PMIG-Net을 제안하였다. 제안 모델은 텍스트와 이미지가 항상 동일한 감성 정보를 제공하지 않는다는 점을 고려하여, 텍스트를 중심으로 한 감성 표현

을 유지하면서 이미지 특징의 반영 정도를 조절하도록 설계하였다.

실험 결과에 따르면 PMIG-Net은 대표 비교 모델과 비교하여 경쟁력 있는 Accuracy를 유지하는 동시에 Macro-F1과 Macro-Recall에서 상대적으로 높은 성능을 보였다. 융합 전략 비교와 어블레이션 실험에서는 각 구성 요소의 제거가 모든 지표에서 일관된 성능 저하로 이어지는 않았으며, Accuracy와 클래스 균형 지표 간에 서로 다른 trade-off가 나타났다. 특히 이미지 게이트의 제거는 Macro-F1과 Macro-Recall의 감소와 관련되어 나타났다. 이러한 결과는 각 구성 요소가 모든 성능 지표를 독립적으로 향상시키기보다는, 사전 지식 기반 융합, 적응적 어텐션, 이미지 게이트 및 멀티태스크 학습이 서로 보완적으로 작용할 가능성이 있음을 시사한다.

또한 MVSA-Single에서 수행한 추가 벤치마크 평가에서는 PMIG-Single-Pass가 동일한 실험 조건의 BERT-only, CLIP-only 및 BERT+CLIP Concat에 비하여 높은 성능을 보였다. 이는 제안 프레임워크가 CNN-MMSA 이외의 이미지-텍스트 감성 데이터에도 적용될 가능성을 보여준다. 다만 MVSA-Single은 뉴스 데이터가 아니며 제목과 본문이 구분되지 않는 바람에 뉴스 도메인에서의 직접적인 일반화 성능으로 해석하는 데에는 한계가 있다.

6.2 Limitations and Future Work

본 연구에는 몇 가지 한계가 존재한다. 첫째, 실험이 제한된 규모의 영어 뉴스 데이터를 중심으로 수행되었으므로, 다양한 언어와 도메인에 대한 일반화 가능성은 추가적인 검증이 필요하다. 둘째, 약지도 기반 라벨을 사용하였기 때문에 일부 샘플에서는 라벨 노이즈가 존재할 수 있으며, 특히 감성 경계가 모호한 사례에서 평가 결과에 영향을 줄 가능성이 있다. 또한 초기 약지도 라벨 구축에 사용된 300개의 인간 참조 표본 중 일부(42개)가 이후 테스트 세트에 포함되어 데이터 독립성 측면의 한계가 존재한다. 이를 보완하기 위해 기존 300개 참조 표본과 중복되지 않는 별도의 200개 테스트 샘플에 대해 독립적인 인간 참조 라벨을 구축하여 추가 평가를 수행하였다. 다만 해당 평가의 규모 역시 제한적이므로, 향후 보다 대규모의 수작업 라벨을 활용한 검증이 필요하다. 셋째, 본 연구는 전역적 이미지 특징을 중심으로 활용하였기 때문에 객체 및 이벤트 수준의 세밀한 시각 정보는 충분히 반영하지 못하였다. 마지막으로, 통계 분석은 5개의 랜덤 시드에 기반하므로 검정력에 한계가 있으며, 지표별 유의성도 일관되게 확인되지 않았다.

향후 연구에서는 다양한 언어와 뉴스 도메인으로 실험 범위를 확장하고, 더욱 대규모의 인간 주석 데이터를 활용하여 약지도 라벨의 신뢰성을 추가로 검증할 필요가 있다. 또한 객체 및 장면 수준의 시각 표현과 최신 멀티모달 파운데이션 모델을 결합하여 모델의 일반화 가능성과 해석 가능성을 확장하고, 반복 실험 횟수를 늘려 통계적 신뢰성을 추가로 확인할 필요가 있다.

REFERENCES

- [1] S. Lai, X. Hu, H. Xu, Z. Ren, and Z. Liu, "Multimodal sentiment analysis: A survey," *Displays*, Vol. 80, Article 102563, December 2023. DOI: 10.1016/j.displa.2023.102563.
- [2] A. Gandhi, K. Adharyu, S. Poria, E. Cambria, and A. Hussain, "Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions," *Information Fusion*, Vol. 91, pp. 424-444, March 2023. DOI: 10.1016/j.inffus.2022.09.025.
- [3] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *Proceedings of NAACL-HLT 2019*, pp. 4171-4186, Minneapolis, MN, USA, June 2019. DOI: 10.18653/v1/N19-1423.
- [4] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor Fusion Network for multimodal sentiment analysis," *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1103-1114, Copenhagen, Denmark, September 2017. DOI: 10.18653/v1/D17-1115.
- [5] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal Transformer for unaligned multimodal language sequences," *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 6558-6569, Florence, Italy, July 2019. DOI: 10.18653/v1/P19-1656.
- [6] Niu, S. Zhu, L. Pang, and A. El Saddik, "Sentiment analysis on multi-view social data," *Proceedings of the 22nd International Conference on MultiMedia Modeling (MMM 2016)*, pp. 15-27, Miami, FL, USA, January 2016. DOI: 10.1007/978-3-319-27674-8_2.
- [7] R. Das and T. D. Singh, "Image-text multimodal sentiment analysis framework of Assamese news articles using late fusion," *ACM Transactions on Asian and Low-Resource Language Information Processing*, Vol. 22, No. 6, Article 161, pp. 1-30, June 2023. DOI: 10.1145/3584861.
- [8] L. Wang, J. Peng, C. Zheng, T. Zhao, and L. Zhu, "A cross modal hierarchical fusion multimodal sentiment analysis method based on multi-task learning," *Information Processing & Management*, Vol. 61, No. 3, Article 103675, May 2024. DOI: 10.1016/j.ipm.2024.103675.
- [9] T. Qu, T. Tuytelaars, and M.-F. Moens, "Visually-aware context modeling for news image captioning," *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 2927-2943, Mexico City, Mexico, June 2024. DOI: 10.18653/v1/2024.naacl-long.162.
- [10] H. Zhang, W. Wang, and T. Yu, "Towards robust multimodal sentiment analysis with incomplete data," *Advances in Neural Information Processing Systems*, Vol. 37, pp. 55943-55974, Vancouver, Canada, December 2024. DOI: 10.52202/079017-1779.
- [11] E. Arazo, D. Ortego, P. Albert, N. E. O'Connor, and K. McGuinness, "Pseudo-labeling and confirmation bias in deep semi-supervised learning," *Proceedings of the 2020 International Joint Conference on Neural Networks (IJCNN)*, pp. 1-8, Glasgow, UK, July 2020. DOI: 10.1109/IJCNN48605.2020.9207304.
- [12] W. Zhang, Y. Zhou, S. Liu, Y. Zhang, and X. Shang, "Weakly supervised text classification framework for noisy-labeled imbalanced samples," *Neurocomputing*, Vol. 610, Article 128617, December 2024. DOI: 10.1016/j.neucom.2024.128617.
- [13] L. Xiao, "Extremely Weakly-supervised Text Classification with Wordsets Mining and Sync-Denoising," *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7167-7179, Mexico City, Mexico, June 2024. DOI: 10.18653/v1/2024.naacl-long.397.
- [14] Y. Mao, Q. Liu, and Y. Zhang, "Sentiment analysis methods, applications, and challenges: A systematic literature review," *Journal of King Saud University - Computer and Information Sciences*, Vol. 36, No. 4, Article 102048, April 2024. DOI: 10.1016/j.jksuci.2024.102048.
- [15] W. Zhang, Y. Deng, B. Liu, S. Pan, and L. Bing, "Sentiment analysis in the era of large language models: A reality check," *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 3881-3906, Mexico City, Mexico, June 2024. DOI: 10.18653/v1/2024.findings-naacl.246.
- [16] O. Ring, M. K. Szabó, C. Guba, B. Váradi, and I. Üveges, "Approaches to sentiment analysis of Hungarian political news at the sentence level," *Language Resources and Evaluation*, Vol. 58, No. 4, pp. 1233-1261, December 2024.
- [17] X. Yang, S. Feng, Y. Zhang, and D. Wang, "Multimodal Sentiment Detection Based on Multi-channel Graph Neural Networks," *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*

- (Volume 1: Long Papers), pp. 328-339, Online, August 2021. DOI: 10.18653/v1/2021.acl-long.28
- [18] Z. Li, B. Xu, C. Zhu, and T. Zhao, "CLMLF: A Contrastive Learning and Multi-Layer Fusion Method for Multimodal Sentiment Detection," Findings of the Association for Computational Linguistics: NAACL 2022, pp. 2282-2294, Seattle, United States, July 2022. DOI: 10.18653/v1/2022.findings-naacl.175
- [19] G. Mu, Y. Chen, X. Li, L. Dai, and J. Dai, "Semantic Enhancement and Cross-Modal Interaction Fusion for Sentiment Analysis in Social Media," PLOS ONE, Vol. 20, No. 4, e0321011, April 2025. DOI: 10.1371/journal.pone.0321011
- [20] Y. Chang, Z. Li, Y. Ruan, and G. Yin, "Image-Text Multimodal Sentiment Analysis Algorithm Based on Curriculum Learning and Attention Mechanisms," Big Data and Cognitive Computing, Vol. 10, No. 1, Article 23, January 2026. DOI: 10.3390/bdcc10010023
- [21] Y. Hwang and J.-H. Kim, "Self-supervised unimodal label generation strategy using recalibrated modality representations for multimodal sentiment analysis," Findings of the Association for Computational Linguistics: EACL 2023, pp. 35-46, Dubrovnik, Croatia, May 2023. DOI: 10.18653/v1/2023.findings-eacl.2.
- [22] K. Jiang, B. Cao, and J. Fan, "A robust framework for multimodal sentiment analysis with noisy labels generated from distributed data annotation," Computer Modeling in Engineering & Sciences, Vol. 139, No. 3, pp. 2965-2984, March 2024. DOI: 10.32604/cmcs.2023.046348.
- [23] C. J. Hutto and E. Gilbert, "VADER: A parsimonious rule-based model for sentiment analysis of social media text," Proceedings of the International AAAI Conference on Web and Social Media, Vol. 8, No. 1, pp. 216-225, Ann Arbor, MI, USA, June 2014. DOI: 10.1609/icwsm.v8i1.14550.
- [24] S. Loria, "TextBlob: Simplified text processing," TextBlob Documentation, <https://textblob.readthedocs.io/>, accessed June 4, 2026.
- [25] M. Hu, and B. Liu, "Mining and Summarizing Customer Reviews," Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 168-177, Seattle, WA, USA, August 2004. DOI: 10.1145/1014052.1014073
- [26] F. Barbieri, J. Camacho-Collados, L. Espinosa-Anke, and L. Neves, "TweetEval: Unified Benchmark and Comparative Evaluation for Tweet Classification," Findings of the Association for Computational Linguistics: EMNLP 2020, pp. 1644-1650, Online, November 2020. DOI: 10.18653/v1/2020.findings-emnlp.148
- [27] W. Rahman, M. K. Hasan, S. Lee, A. B. Zadeh, C. Mao, L.-P. Morency, and E. Hoque, "Integrating multimodal information in large pretrained transformers," Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pp. 2359-2369, Online, July 2020. DOI: 10.18653/v1/2020.acl-main.214.
- [28] W. Yu, H. Xu, Z. Yuan, and J. Wu, "Learning modality-specific representations with self-supervised multi-task learning for multimodal sentiment analysis," Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, No. 12, pp. 10790-10797, May 2021. DOI: 10.1609/aaai.v35i12.17289.
- [29] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning Transferable Visual Models from Natural Language Supervision," Proceedings of the 38th International Conference on Machine Learning, Vol. 139, pp. 8748-8763, Virtual Conference, July 2021. URL: <https://proceedings.mlr.press/v139/radford21a.html>

Authors



XUYANG GONG is currently pursuing the M.S. degree in the Department of Big Data Analytics at Kyung Hee University, Seoul, Republic of Korea. His research interests include large language models (LLMs),

multimodal learning, and generative artificial intelligence (GenAI).