한국어 부정적 밈 탐지를 위한 컴퓨터 비전 기반 접근법

성아영[1], 박은일[2]
[1]성균관대학교 인공지능융학학과 석사과정
[2]성균관대학교 인공지능융학학과 교수
achieve00@skku.edu, eunilpark@skku.edu

A Computer Vision-Based Approach for Detecting Negative Korean Memes

Ayeong Seong[1], Eunil Park[1]
[1]Dept. of Applied Artificial Intelligence, Sungkyunkwan University

요 약

최근 온라인 환경에서 밈(meme)이 중요한 소통 수단으로 자리 잡으면서, 혐오 발언이나 모욕적 표현을 담은 부정적 밈에 대한 분석의 필요성이 커지고 있다. 기존 연구들은 대부분 영어권에 집중되었고, 텍스트 분석에 의존하거나 특정 이슈에만 국한되어 다양한 유형의 콘텐츠를 포괄적으로 다루는 데 한계가 있었다. 본 연구는 이러한 한계를 극복하고자 한국의 독특한 밈인 짤방을 대상으로 컴퓨터 비전 기반 접근법을 제안한다. 이를 위해 폭력성, 선정성, 비하/비방 등 다양한 카테고리로 라벨링한 한국어 밈 데이터셋을 구축하고, 제안 모델은 객체 탐지 모델인 Mask R-CNN과 특징 추출 모델인 EfficientNetV2를 결합한 하이브리드 아키텍처로, 이미지 내 시각적 정보를 효과적으로 분석한다. 실험 결과, 제안 모델은 이진 및 다중 분류에서 기존 모델 대비 우수한 성능을 보였다. 본 연구는 한국어 밈 연구의 새로운 방향을 제시하고, 온라인 유해 콘텐츠 분류에 기여할 것으로 기대된다.

1. 서론

밈(meme)이란 리처드 도킨스의 저서 이기적 유전자에서 처음 사용된 단어로 인터넷 유저들 사이에서 널리 퍼지고 파생되는 유머러스한 이미지, 비디오, 텍스트 등의 창작물을 의미한다 [1]. 최근 인터넷 밈은 사용자가 원하는 메시지를 표현하는 수단으로 온라인 커뮤니케이션의 핵심 요소로 자리 잡았다. 유머러스한 표현으로 자신의 감정, 생각 등을 전할 뿐만 아니라 타인을 향한 풍자, 비방, 혐오 표현 등의 메시지가 되기도 한다 [2].

이러한 문제로 인해 최근 밈 콘텐츠 분석 연구도 활발하게 진행되고 있지만, 이러한 연구는 영어권에 한정되어 있다는 한계점이 존재한다. 해외 커뮤니티(Reddit, 4chan)에서 유행하는 이미지 매크로(image macro) 밈은 일반적으로 텍스트와 이미지가 결합된 형태로 구성된다 [3]. 반면 한국어 이미지 밈 중에서도 짤방은 텍스트보다 이미지에 의존하는 경향이 있으며, 그림 1처럼 텍스트 없이 사용되는 경우도 많다. 짤 혹은 짤방은 원래 '짤림방지'의 줄임말로, 게시물 삭제를 방지하기 위해 흥미로운 이미지를 첨부하는 온라인 커뮤니티의 관행에서 유래했다. 현재 '짤'과 '짤방'은 인터넷 밈 중 이미지 기반 형식을 통칭하는 용어로 사용되고 있다. 그림 2는 텍스트와 이미지가 결합된 한국어 밈의 예시이다. 하지만 그림 1과 같이 텍스트 없이 이미지 자체로 의미를 전달하는 짤방은 기존 NLP 기반 접근법만으로 탐지하는 데 한계가 존재한다.

따라서 본 연구에서는 한국 밈을 이용하여 텍스트를 직접 분석하지 않고, 이미지 내 시각적 요소를 활용하여 부정적 밈을 탐지하는 컴퓨터 비전 기반 접근법을 제안한다.

(그림 1) 텍스트가 없는 밈.

(그림 2) 텍스트가 있는 밈.

2. 관련 연구

최근 밈 콘텐츠 분석 연구가 활발히 진행되면서 다양한 데이터셋이 구축되었다. 대표적으로 Hateful Memes Dataset은 Facebook AI에서 공개한 데이터셋으로, 텍스트와 이미지의 조합으로 구성된 혐오성 밈을 탐지하는 데 활용된다 [4]. 10,000개 이상의 이미지로 구성되어 있으며, 혐오와 비혐오 두가지로 구분하여 라벨링되었다.

HarMeme과 Harm-P는 특정 사회적 이슈와 관련된 밈을 다룬 데이터셋이다 [5]. HarMeme은 COVID-19 관련 밈을, Harm-P는 미국 정치 관련 밈을 포함하며, 밈의 유해성을 '매우 유해', '부분적 유해', '무해' 세 가지 카테고리로 세분화하여 라벨링했다. 이는 단순한 혐오 표현을 넘어 밈의 유해성 정도를 정량적으로 분석하려는 시도이다.

PrideMM 데이터셋은 LGBTQ+ 커뮤니티에 대한 혐오 발언을 탐지하고 분석하기 위해 다양한 플랫폼에서 수집되었다 [6]. 이는 단순히 혐오 여부를 넘어, 혐오 대상 식별, 주제적 입장 분류, 유머 탐지 등 여러 세부 태스크를 포함하는 것이 특징이다.

하지만 이러한 연구에는 다음과 같은 한계점이 존재한다. 첫째, 대부분 영어권 콘텐츠에 집중되어 있어 한국어 밈과 같은 비영어권 밈을 다루지 못한다. 둘째, 혐오 발언이나 특정 이슈에 한정되어 있어, 모욕적 표현, 폭력, 자해 유도 등 다양한 유형의 부정적 콘텐츠를 포괄적으로 다루지 못한다.

따라서 본 연구는 이러한 한계를 극복하고자 혐오 발언 뿐만 아니라 폭력성, 선정성, 비하/비방 등 다양한 카테고리로 세분화하여 부정적 밈을 탐지하는 연구로 확장하고자 한다.

3. 제안 방법

3.1 데이터셋

본 연구에서 사용한 데이터는 '오늘의 짤방' 아카이브 사이트에서 2015년 3월 10일부터 2024년 5월 20일까지 업로드된 짤방을 웹 크롤링을 통해 수집하였다. 해당 사이트는 X, 페이스북, 인스타그램, 다음 카페 등 다양한 소셜 미디어와 국내 주요 커뮤니티에서 유통된 짤방을 아카이브하여 한국 문화권의 밈을 수집하기에 적합하다고 판단했다. 초기 수집된 16,282개의 이미지 중 파일 손상, 해상도 저하 등으로 인해 내용을 식별하기 어려운 이미지를 필터링하였다. 최종적으로 라벨링에 사용된 데이터는 총 10,590개의 이미지를 사용하였다.

데이터 라벨링은 4명의 인공지능 관련 전공의 한국어 사용자를 대상으로 진행되었으며, 데이터의 신뢰성과 일관성을 확보하기 위해 세 단계를 거쳐 진행되었다. 첫째, 사전 정의된 라벨링 지침을 제공하여 각 라벨에 대한 명확한 기준을 확립하였다. 다양한 짤방 예시와 카테고리별 라벨링 샘플을 제공하였다. 둘째, 두 그룹으로 나뉜 라벨러들이 독립적으로 라벨링을 수행하였다. 각 그룹은 전체 데이터 세트와 함께 이전에 제공한 라벨링 지침의 기준에 따라 데이터 라벨링을 진행하였다. 마지막으로 교차 검토를 통해 의견이 맞지 않는 라벨을 조정하고, 일관성을 평가하였다.

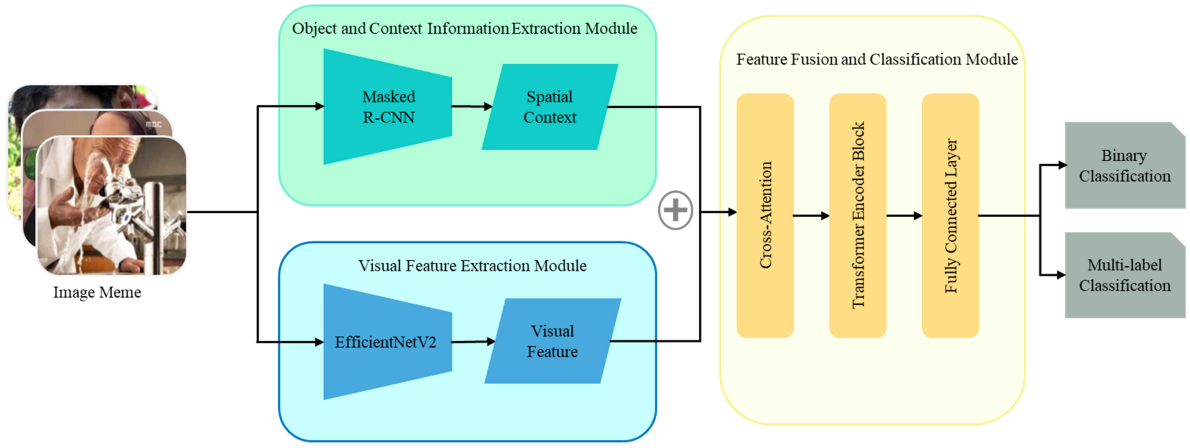
본 연구에서는 밈 콘텐츠의 복합적인 특성을 포착하기 위해 각 이미지에 대해 이미지 파일명, 이미지 형식, 텍스트 포함 여부, 부정적 콘텐츠 여부, 부정적 카테고리 유형의 5가지 항목을 정의하였다. 특히 '부정적 카테고리 유형'를 표 1과 같은 사전 정의된 분류 체계를 기반으로 세분화하여 라벨링을 진행하였다.

<표 1> 부정적 카테고리 유형

카테고리명	설명
욕설	욕설이나 비속어 포함됨
폭력성	폭력적인 장면을 묘사하거나 폭력적인 행동을 부추김
선정성	성적 암시나 노골적인 성적 이미지를 포함됨
비하/비장	특정 개인이나 집단을 비하하거나 조롱함
혐오	특정 사회적 집단에 대한 증오를 조장하거나 편견을 부추기는 표현이 포함됨
자해/자살조장/자기비하	자해를 묘사하거나 자살을 부추기는 내용, 자기 비하와 같은 자신을 해하는 표현이 포함됨
기타	허위사실과 같은 위 카테고리에 포함되지 않는 부정적 내용이 포함됨

3.2 비전 기반 부정적 밈 탐지 모델

한국어 이미지 밈에 존재하는 텍스트, 객체, 배경 등 시각적인 맥락적 정보를 이용하고자 본 연구는 이미지의 시각적 특징과 맥락적 정보를 통합적으로 처리하는 새로운 모델을 제안한다. 제안 모델은 객체 탐지 모델인 Mask R-CNN과 특징 추출 모델인 EfficientNetV2를 결합하여 설계되었다. 제안하는 모델의 구조는 그림 3과 같다.



(그림 3) 제안하는 모델 구조.

제안하는 모델은 세 가지 주요 모듈로 구성된다. '객체 및 맥락 정보 추출 모듈'은 Mask R-CNN을 활용하여 이미지 내 텍스트, 객체, 배경 정보를 탐지하고 이를 컨텍스트 벡터로 변환한다. '시각적 특징 추출 모듈'은 EfficientNetV2를 통해 원본 이미지의 고차원적 시각적 특징을 특징 벡터로 추출한다. '특징 융합 및 분류 모듈'은 이 두 벡터를 결합한다. 이 과정에서 크로스 어텐션(Cross-Attention)과 트랜스포머 인코더 블록을 활용하여 시각적 특징과 맥락적 정보 간의 상호작용을 극대화하고, 최종적으로 밈 콘텐츠의 부정적 여부와 해당 카테고리를 분류한다.

4. 실험 및 결과

4.1 실험 설정

본 실험에서는 구축한 데이터셋을 활용하여 이미지와 GIF를 대상으로 실험을 진행하였다. 사용한 한국어 밈 데이터셋의 부정적 데이터 분포와 부정적 카테고리 유형 분표는 각각 표 2, 표 3과 같다.

<표 2> 부정적 데이터 분포

부정적 콘텐츠 여부	이미지 개수	gif 개수
비부정적	5,073	1,560
부정적	3,063	894

<표 3> 부정적 카테고리 유형 분포

카테고리명	이미지 개수	gif 개수
욕설	722	92
폭력성	758	278
선정성	157	58
비하/비장	1,645	311
혐오	407	128
자해/자살조장/자기비하	362	23
기타	72	4

데이터는 학습, 검증, 평가 세트를 6:2:2의 비율로 나누었으며, 이미지와 GIF 파일은 독립적으로 평가하여 파일 유형에 따른 성능 차이를 분석했다. 모델 성능 평가 지표로는 정확도, F1-score, macro F1-score, weighted F1-score를 사용했다.

본 실험은 Google Colab 환경에서 NVIDIA A100 GPU를 이용해 PyTorch 프레임워크로 구현되었다. 모델은 AdamW Optimizer와 BCE 손실 함수를 사용하여 30 epoch 동안 학습되었으며, 과적합을 방지하기 위해 Early Stopping 기법을 적용했다.

4.2 비교 모델

분류 성능 평가를 위해 기존 혐오 밈 탐지 연구에서 사용된 아키텍처들을 선정하여 이용하였다.

총 4가지 모델을 실험에 이용했으며, 이미지 실험에는 ResNet50, EfficientNetV2, Visformer를, GIF 실험에는 3D-CNN, EfficientNetV2, Visformer를 비교 모델로 활용하였다.

4.3 이진 분류 실험 결과

본 실험에서는 밈 콘텐츠가 부정적인지 여부를 이진 분류를 통해 예측하고, 다양한 모델의 성능을 비교하여 제안된 모델의 효과를 검증하였다. 표 4는 이미지 데이터에 대한 이진 분류 실험 결과를, 표 5

<표 4> 이미지 데이터 이진 분류 실험 결과

모델	정확도	F1-score	macro F1-score
ResNet50	0.5714	0.5344	0.5687
EfficientNetV2	0.6239	0.5357	0.6098
Visformer	0.5847	0.5089	0.5746
Proposed model	**0.6432**	**0.5529**	**0.6280**

<표 5> GIF 데이터 이진 분류 실험 결과

모델	정확도	F1-score	macro F1-score
3D-CNN	**0.7500**	**0.5848**	**0.7030**
EfficientNetV2	0.5609	0.3916	0.5240
Visformer	0.3283	0.3953	0.3199
Proposed model	0.6696	0.5250	0.6358

는 GIF 데이터 실험 결과를 나타낸다.

이미지 데이터의 실험 결과, 모든 평가 지표에서 제안하는 모델이 가장 높은 성능을 보였다. GIF 데이터 실험에서는 3D CNN이 가장 높은 성능을 기록하였지만 제안하는 모델도 그에 준하는 성능을 보이는 것을 확인할 수 있었다.

4.4 다중 분류 실험 결과

부정적 밈 콘텐츠가 포함된 경우, 해당 콘텐츠를 카테고리별로 분류하는 실험도 진행하였다. 부정적 밈 콘텐츠는 욕설, 폭력성, 혐오 등의 다양한 카테고리로 구분될 수 있으며, 하나의 콘텐츠가 여러 카테고리에 속할 수 있기에 멀티 레이블 분류 방식으로 실험을 진행하였다. 이를 통해 제안된 모델이 기존 모델 대비 각 부정적 카테고리를 얼마나 효과적으로 분류할 수 있는지를 검증하였다. 표 6는 이미지 데이터에 대한 다중 분류 실험 결과를, 표 7는 GIF 데이터 실험 결과를 나타낸다.

<표 6> 이미지 데이터 다중 분류 실험 결과

모델	정확도	macro F1-score	weighted F1-Score
ResNet50	0.3587	0.3885	0.5182
EfficientNetV2	0.3354	0.3798	0.4968
Visformer	0.3183	0.3902	0.4959
Proposed model	**0.3789**	**0.4340**	**0.5578**

<표 7> GIF 데이터 다중 분류 실험 결과

모델	정확도	macro F1-score	weighted F1-score
3D-CNN	0.2838	0.2748	0.4400
EfficientNetV2	0.3243	0.3254	0.4923
Visformer	0.2027	0.2195	0.3338
Proposed model	**0.3784**	**0.3556**	**0.5327**

3D-CNN은 이진 분류에서는 강점을 보였지만, 다중 분류에서는 정확도, macro F1-score, weighted F1-score 모두에서 성능이 떨어졌다. 이는 3D CNN이 개별 클래스를 예측하는 데 효과적이지만, 여러 개의 라벨이 동시에 존재하는 상황에서의 일반화 능력이 상대적으로 부족하기 때문으로 해석된다. 반면 제안하는 모델은 다중 분류에서 가장 높은 성능을 보여주며 안정적인 성능을 보여주었다.

5. 결론 및 향후 연구

본 연구는 텍스트가 없는 한국어 밈인 '짤방'의 부정적 콘텐츠를 탐지하기 위해 컴퓨터 비전 기반 모델을 제안했다. Mask R-CNN과 EfficientNetV2를 결합한 하이브리드 아키텍처를 사용하여 이미지 내의 시각적 특징과 맥락 정보를 효과적으로 분석했으며, 이를 통해 혐오 발언뿐만 아니라 폭력성, 선정성 등 다양한 유형의 부정적 밈을 포괄적으로 탐지하는 능력을 입증했다. 제안하는 모델은 이미지 데이터에 대한 이진 및 다중 분류 실험에서 가장 뛰어난 성능을 보였으며, 이는 한국어 밈 연구의 새로운 가능성을 제시하고 실제 온라인 환경의 유해 콘텐츠를 분류하는 데 기여할 것으로 기대된다.

사사(Acknowledgement)

이 논문은 정부(과학기술정보통신부)의 재원으로 정보통신기획평가원-대학ICT연구센터(ITRC; RS-2024-00436936), 학석사연계ICT핵심인재양성사업(RS-2023-00259497), 인간지향적 차세대 도전형 AI 기술 개발 (RS-2025-25440264, 인간의 감각 인지 프로세스 기반 범용인공지능 개발)의 지원을 받아 수행된 연구임.

참고문헌

[1] Oxford Languages. (n.d.). meme. Retrieved from https://www.oxfordlearnersdictionaries.com/definition/english/meme

[2] Chen, C. (2012). The creation and meaning of Internet memes in 4chan: Popular Internet culture in the age of online digital reproduction. Habitus, 3(1), 6-19.

[3] Dancygier, B., & Vandelanotte, L. (2017). Internet memes as multimodal constructions. Cognitive Linguistics, 28(3), 565-598.

[4] Kiela, D. et al. (2020). The hateful memes challenge: Detecting hate speech in multimodal memes. Advances in Neural Information Processing Systems, 33, 2611-2624.

[5] Pramanick, S. et al. (2021). MOMENTA: A Multimodal Framework for Detecting Harmful Memes and Their Targets. In Findings of the Association for Computational Linguistics: EMNLP 2021, pp. 4439-4455.

[6] Cooksey, S. et al. (2020). PrideMM: Second Order Model Checking for Memory Consistency Models. In Formal Methods. FM 2019 International Workshops, pp. 507-525. Springer International Publishing.