멀티모달 데이터와 머신러닝 기법을 활용한 유튜브 영상 조회수 예측 연구

박현철[1], 문태건[1], 박건희[1], 최영락[1], 이택[2], 김정동[2], 이정빈[2]
[1]선문대학교 컴퓨터공학부 학부생
[2]선문대학교 컴퓨터공학과 교수
{byp006531, mtg57919141, hold0316, jeus0216, comtaek, kjd4u, jungbini}@sunmoon.ac.kr

Predicting YouTube Video Views Using Multimodal Data and Machine Learning Techniques

Hyun-Cheul Park[1], Tae-Geon Moon,[1]Gun-Hee Park[1],Yeong-Rak Choi[1],
Taek Lee[2], Jeong-Dong Kim[2], Jung-Been Lee[2]
[1,2]Div. of Computer Science and Engineering, Sun Moon University

요 약

유튜브는 개인부터 기업·공공기관까지 활용이 급증하며 방대한 영상 데이터와 치열한 경쟁 환경을 형성하고 있다. 따라서 본 연구에서는 메타데이터, 자막, 썸네일을 포함한 데이터를 수집·분석하여 영상 조회수를 예측하는 모델을 제안하였다. 제안된 모델은 Baseline 대비 향상된 성능을 보였으며, 주요 영향 요인으로 제목 토픽, URL, 업로드 규칙성, 썸네일 요소 등이 확인되었다. 결과적으로 본 연구 결과를 활용하여 데이터 기반 콘텐츠 전략을 수립할 수 있으며, 이는 향후 추천 시스템 개선, 콘텐츠 기획 자동화 연구의 기초 자료로 활용될 수 있다.

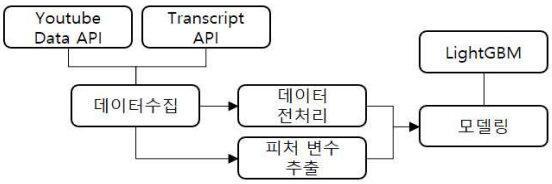
1. 서론

유튜브는 개인 크리에이터부터 기업·공공기관까지 폭넓게 활용되며, 업로드 영상 수와 시청 시간이 매년 급격히 증가하고 있다[1-3]. 또한, 한국인의 유튜브 1인당 일일 사용 시간은 지속적으로 증가하여 방대한 시청 데이터를 형성하고 있으며, 매분 수백 시간 이상의 영상이 업로드되고 있다. 이에 따라 콘텐츠 경쟁이 심화되고 성과 측정과 전략적 대응이 중요해지고 있다[1,4]. 따라서, 데이터 기반의 조회수 예측 및 분석 도구 개발이 필요하다.

본 연구에서는 YouTube Data API, Transcript API와 썸네일 이미지 분석을 통해 수집한 멀티모달 데이터(메타데이터, 자막, 이미지 피처)를 결합하여 유튜브 영상 조회수를 예측할 수 있는 회귀모델을 제안한다. Linear Regression을 Baseline으로 설정하여 LightGBM 모델과의 성능 차이를 RMSE, R^2, F1-score(회귀 결과를 중앙값 기준으로 이진화하여 계산)으로 비교 평가하고, SHAP 분석으로 설명 가능성을 확보하였다.

2. 멀티모달 데이터 기반 영상 조회수 예측 모델

멀티모달 데이터 기반의 유튜브 영상 조회수 예측 모델의 개발 절차는 그림1과 같다.



(그림 1) 모델의 개발 절차

분석 대상으로 2024년 5월 13일 기준 국내 PC방 점유율이 높은 상위 5종의 게임(리그 오브 레전드, 배틀그라운드, 서든어택, FC온라인, 발로란트)을 선정하였다. PC방 점유율은 게임의 이용자 선호도를 반영하는 대표적 지표로, 점유율이 높은 게임일수록 관련 영상 업로드 흐름을 잘 보여줄 것으로 판단하였다. 이후 영상 29,289개와 자막 18,715개를 각각 Youtube Data API와 Transcript API로 수집하여 CSV·TXT 파일로 저장, 관리하였다. 데이터 전처리에서 결측치·중복을 제거하고 조회수 0인 영상을 제외했으며, 예측 정확도 향상을 위해 썸네일·자막·제목·설명·업로드 시간에서 피처를 생성하였다. 썸네일에서 얼굴 수·면적 비율·텍스트 비율·주요 감정 및 세부 감정을 추출하고, 자막에서 주제 특성 피처를 도출하였다. 제목은 인기 토픽 사용 여부와 길이에 따른 조회수 경향, 설명은 챕터 존재 여부와 URL 포함 여부, 업로드 시간은 게시 주기·요일·시간대로

피처화하였다. 이후 생성한 피처들을 통합하여 LightGBM 모델로 학습하였으며 Randomized SearchCV를 통해 최적 파라미터를 선정하였다.

3. 유튜브 조회수 예측 모델 성능 평가 결과

그림 2는 실제 조회수와 학습된 예측 모델의 분포를 나타낸 그림이다.

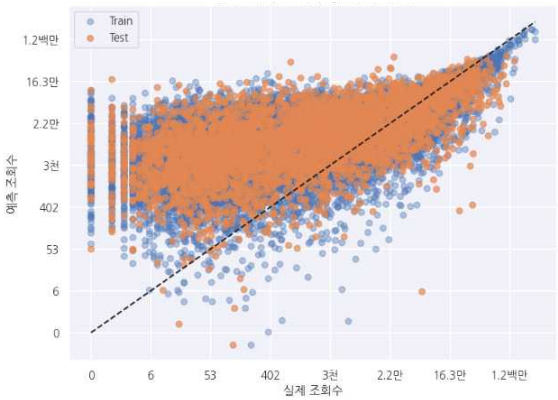
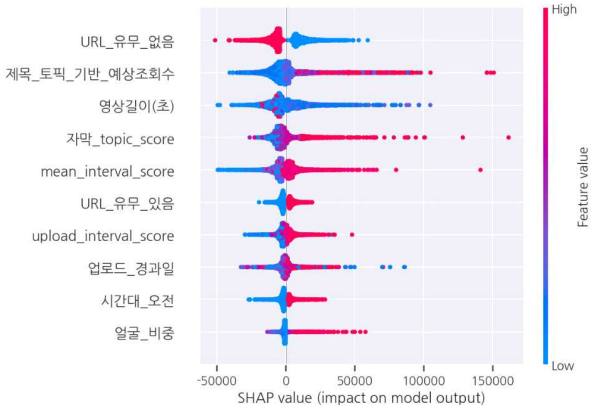


그림 2) 조회수 예측 모델과 실제 조회수 분포도

Baseline 모델인 Linear Regression 모델의 RMSE는 73,543이었으며, 멀티모달 데이터 기반 LightGBM 모델은 61,679으로 본 연구의 모델이 Baseline 모델보다 더 적은 차이를 보였다. R^2가 0.394, F1-score가 0.732로 나타나 예측값과 실제 조회수는 평균적으로 약 6만 회 이상의 오차가 있지만, 전체 변동성의 약 39.4%를 설명하며, 이진 분류 성능 또한 우수하였다. 다음 그림 3은 Shapley value를 통한 모델 요약 분석 결과이다.



(그림 3) 학습된 모델의 SHAP 분석 결과

조회수 예측에 영향을 미친 상위 10개 변수는 URL 유무, 제목 토픽 기반 예상 조회수, 영상 길이, 자막_topic_score(자막 주제 인기도 점수), mean_interval_score(채널 업로드 규칙성), upload_interval_score(이전 영상과의 업로드 간격), 업로드 경과일, 시간대, 썸네일 얼굴 비중(썸네일 내 얼굴이 차지하는 비율) 등이었다. 특히, 인기 토픽 활용, 설명란에 URL 포함, 짧은 영상 길이, 일정한 업로드 패턴, 인기 주제 반영, 썸네일 내 얼굴 노출, 오전 업로드가 조회수 증가에 긍정적인 영향을 주는 요인으로 확인되었다.

4. 결론

본 연구는 멀티모달 데이터 기반 유튜브 조회수 예측 모델을 제안하여 Baseline 대비 성능을 향상시켰으며, SHAP 분석을 통해 제목 토픽, URL, 업로드 규칙성, 썸네일 텍스트·얼굴 등 주요 예측 변수를 확인하였다. 이를 통해 게임 트렌드 분석, 최적 업로드 전략, 썸네일·키워드 설계 등 데이터 기반 콘텐츠 기획 가능성을 제시하였다. 본 연구는 채널 운영의 수익 안정화와 성장 전략 수립에 기여할 수 있으며, 향후 다양한 장르와 사용자 반응 데이터를 반영한 연구가 필요하다.

감사의 글

이 논문은 2025년도 과학기술정통신부 및 정보통신기획평가원의 SW중심대학지원사업(2024-0-00023) 및 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(RS-2023-00243114).

참고문헌

[1] 박지민, “유튜브에 빠진 한국… 1인당 매달 40시간씩 봐”, 조선일보, 2024.03.05., https://www.chosun.com/economy/tech_it/2024/03/05/WSC7JI7AIVFLFO57IV5C3FQ3AI

[2] GMI Research Team, “YOUTUBE STATISTICS 2025 (DEMOGRAPHICS, USERS BY COUNTRY & MORE)”, Global Media Insight (GMI), 2025.06.,https://www.globalmediainsight.com/blog/youtube-users-statistics

[3] SEO.AI’s Content Team, How Many Videos Are on YouTube? Statistics & Facts, SEO.AI, 2025.02.15., https://seo.ai/blog/how-many-videos-are-on-youtube

[4] Robin Geuens, “How many hours of video are uploaded to YouTube each minute?, SOAX, 2024.10.30., https://soax.com/research/how-many-hours-of-video-are-uploaded-to-youtube-every-minute