영상 선호도 예측을 위한 유튜브 영상에 대한 토픽어와 비공감 유형 매칭

정지민* · 김승진 · 이동윤 · 김교태

한국외국어대학교

Matching of Topic Words and Non-Sympathetic Types on YouTube Videos for Predicting Video Preference

Jimin Jung* · Seungjin Kim · Dongyun Lee · Gyotae Kim

Hankuk University of Foreign Studies

E-mail : stopmin02@hufs.ac.kr

요 약

전 세계 최대 규모의 동영상 공유 플랫폼인 유튜브는 수많은 영상을 제공하며 원하는 정보를 손쉽게 얻을 수 있다는 점에서 많은 사람의 사랑을 받고 있다. 그렇지만 영상마다 공감 비율(싫어요/좋아요)은 동일 채널의 영상일지라도 주제나 업로드 시기 등에 따라 많은 편차를 보여, 기존 연구들은 공감 비율과 영상 조회 수와 같은 수치를 통해 그 원인을 유추하거나 해석하려 한다. 이러한 방식은 공감 현황을 파악하는 데는 도움을 주지만, 특정 영상의 선호도 원인을 파악하는 데는 한계가 있다. 따라서 본 연구는 영상별로 수집한 댓글들로부터 추출한 토픽어(명사 상당 어구)와 사전에 분류한 비공감 유형 간 매칭을 통해 그 원인을 파악하고자 한다. 공감 비율에서 있어서 아웃라이어가 많이 발생하는 반려동물과 요리 분야의 상위 10개 채널에서 제작한 영상 중 비공감 지수(비공감 수/공감 수)가 가장 높은 상위 10개 영상(반려동물 임계값: 4.000, 요리 임계값: 0.723)에 대해 유튜브 API를 통해 수집한 11,110개 댓글들을 수집하고 토픽어를 추출하여 사전에 정의한 비공감 유형과 매칭시켰다. 이를 통해 댓글 분석만으로도 비공감 비율이 높을 것인지, 어떠한 비공감 유형일지 예측 가능함을 확인하였다. 향후 유튜브 채널 운영자를 위한 비공감 영상 예측 및 제작 기준을 구축하는 후속 연구를 통해 사용자에게 긍정적인 영상을 제공할 수 있는 유튜브 환경을 개선할 수 있을 것으로 기대한다.

ABSTRACT

YouTube, the world's largest video sharing platform, is loved by many users in that it provides numerous videos and makes it easy to get helpful information. However, the ratio of like/hate for each video varies according to the subject or upload time, even though they are in the same channel; thus, previous studies try to understand the reason by inspecting some numerical statistics such as the ratio and view count. They can help know how each video is preferred, but there is an explicit limitation to identifying the cause of such preference. Therefore, this study aims to determine the reason that affects the preference through matching between topic words extracted from comments in each video and non-sympathetic types defined in advance. Among the top 10 channels in the field of ‘pets’ and ‘cooking’, where outliers occur a lot, the top 10 videos (the threshold of pet: 4.000, the threshold of cooking: 0.723) with the highest ratio were selected. 11,110 comments collected totally, and topics were extracted and matched with non-sympathetic types. The experimental results confirmed that it is possible to predict whether the rate of like/hate would be high or which non-sympathetic type would be by analyzing the comments.

키워드

Youtube, Non-sympathetic Video, Topic Word, Non-sympathetic Type, the Ratio of Like/Hate

* corresponding author

Ⅰ. 서 론

전 세계 최대 규모의 동영상 공유 플랫폼 유튜브는 수많은 영상을 제공하며 원하는 정보를 손쉽게 얻을 수 있다는 점에서 많은 사람이 찾고 있다. 하지만 영상마다 공감 비율(싫어요/좋아요)에 있어서 많은 편차를 보인다. 채널 운영자의 관점에서 영상 선호도가 극명하게 갈리는 이유를 미리, 그리고 정확히 알 수 있다면 채널 운영에 도움이 되겠지만, 이러한 시스템을 갖추고 있지는 못하다. 영상 선호도에 관심을 가진 기존 연구들은 공감 비율, 영상 조회 수와 같은 단순 수치 위주의 분석에 초점을 맞추고 있어 특정 영상에서 선호도 양상을 파악하는 데는 한계가 있다[1-3]. 본 연구는 이러한 이슈를 해결하기 위해, 댓글들로부터 토픽어를 추출하여 사전에 분류된 비공감 유형과 매칭함으로써 선호도 양상을 이해하는 도움을 주고자 한다.

Ⅱ. 영상 선호도 분석

본 연구에서는 그림 1과 같은 절차로 본 연구를 수행하고자 한다. 먼저, 조회 수가 높은 분야(실험에서 10)의 영상 데이터를 유튜브 키워드 영상 수집 API*를 이용하여 수집한다(그림 1의 첫 번째 단계)[4]. 수집된 각 분야 영상들의 공감 비율 분포를 확인하고 아웃라이어가 상대적으로 많은 분야(실험에서 2)를 선정한다(그림 1의 두 번째 단계). 선정된 분야에 대해 구독자 수 기준 상위 N개(실험에서 10)의 채널들을 선택하여, 해당 채널이 제작한 영상들을 유튜브 채널 영상 수집 API**를 이용하여 수집한다. 영상들의 비공감 지수(비공감 수/공감 수) 상위 M개(실험에서 10)에 대해 유튜브 댓글 수집 API***를 이용하여 댓글들을 수집한다(그림 1의 세 번째 단계)[4]. 댓글로부터 명사 추출기를 이용하여 명사 상당 어구들을 추출하고, 토픽어를 수작업으로 선정한다(그림 1의 네 번째 단계). 마지막으로 토픽어들을 사전에 정의한 비공감 유형(실험에서 반려동물에 대해 6가지, 요리에 대해 4가지)과 매칭하고 본 방법의 유효성을 확인한다.

* youtube.search().list(q=keyword, type="video", pageToken=token, order=order, part="id, snippet", maxResults=max_results, location=location, locationRadius=location_radius).execute()

** youtube.channels().list(id=channel_id, part='contentDetails').execute()

*** youtube.commentThreads().list(part='snippet, replies', videoId=video_ID, maxResults=N).execute()

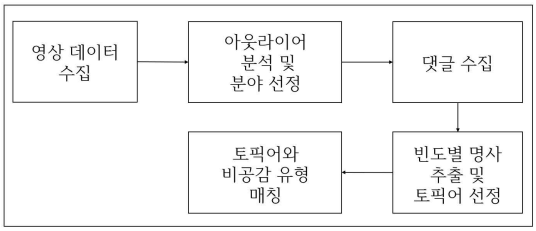

그림 1. 영상 선호도 분석 절차

Ⅲ. 실험 결과

영상 데이터 수집을 위해 조회 수가 높고, 저자들에게 공통으로 추천되는 음악, 피트니스, 액션어드벤처게임, 만화영화, 수학, 축구, 요리, 반려동물, 드라마, 공예 등 10개 분야를 선정하였다. 각 분야에 대해 대표적인 20개 키워드를 선정하고 유튜브 키워드 영상 수집 API를 이용하여 88,186개 영상을 수집하였다(표 1 참조).

표 1. 10개 분야에서 수집한 영상 현황

번호	분야	#수집 영상
1	음악	9,417개
2	피트니스	10,628개
3	액션어드벤처게임	6,820개
4	만화영화	9,933개
5	수학	9,254개
6	축구	7,367개
7	요리	8,697개
8	반려동물	6,771개
9	드라마	9,862개
10	공예	9,437개
합계		88,186개

분야별로 수집된 영상들을 상호 비교하기 쉽도록 '좋아요' 수를 기준으로 정규화하였는데, 분야에 따라 아웃라이어(추세선을 상당히 이탈하는 영상)의 빈도가 차이가 남을 확인하였다(그림 2 참조). 예를 들어, 음악이나 수학 등 교양과 교육 분야들은 추세선을 일정하게 따라가는 양상을 보였지만, 반려동물과 요리에서는 공감 지수(공감 수/비공감 수), 비공감 지수(비공감 수/공감 수)에 있어서 극단적인 아웃라이어들이 상대적으로 많이 분포함을 확인할 수 있다.

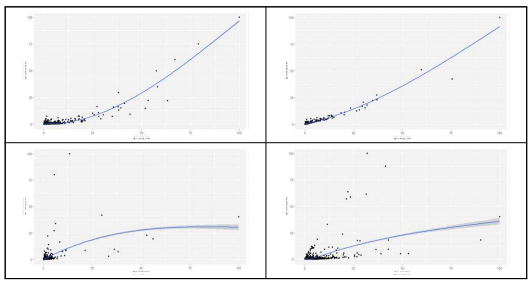

그림 2. 분야별 공감 비율 분포 (좌측 위에서부터 시계 방향으로 음악, 수학, 반려동물, 요리. X축: 정규화된 '좋아요', Y축: 정규화된 '싫어요', 파란 선: 추세선)

본 연구의 목적이 비공감 원인을 파악하고 예측하는 데 있으므로, 이러한 아웃라이어가 많이 발생하는 반려동물과 요리 분야를 대상으로 구독자 수 기준 상위 10개 채널을 선정하고, 해당 채널들에서 제작한 영상들을 수집하였다[5].

그림 3은 두 분야에 대해 수집된 상위 10개 채널 영상들의 선호 양상을 살펴본 예이다. 반려동물 분야는 사용자에 의해 클릭 되는 전체 공감/비공감 수치가 증가하더라도 공감 지수와 비공감 지수가 극단적으로 유지되는 동시에 비공감 지수 값도 상당히 높은 경우가 많이 보임을 확인할 수 있다. 즉, 전체 수치와 상관없이 선호하는 영상과 선호하지 않는 영상이 극명하게 나뉜다는 의미이다. 반면에, 요리 영상은 전체 수치가 증가하는 데 어느 정도 비례하여 공감 비율이 일정하게 유지되는 것을 확인할 수 있다. 흥미로운 점은 일부 영상들은 극단적인 비공감 지수를 보이는 반해, 극단적인 공감 지수를 보이는 영상의 비율이 상대적으로 낮음을 알 수 있다.

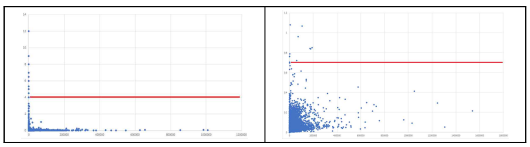

그림 3. 반려동물, 요리 분야 비공감 상위 10개 임계값 (반려동물: 4.000, 요리: 0.723, X축: 공감 수+비공감 수, Y축: 비공감 수/공감 수)

두 분야에 대해 댓글들을 수집하여 명사 추출기를 통해 명사 상당 어구들을 추출하고 빈도순으로 정리하였다.*. 이들 중에서 비공감 유형에 영향을 미치지 않는 불용어들(예. I, They 등)을 제거한 후 토픽어들을 선정하여 사전에 정의한 비공감 유형과 매칭을 시도하였다(표 2와 표 3 참조).

표 2. 비공감 유형과 매칭되는 토픽어(반려동물)

번호	유형	토픽어 예시
1	시청 불가 영상(국가락)	
2	논쟁 유발 영상(예. 인종차별, 남녀혐오)	사회의식, 동남아, 갈라치기, 남자들
3	어그로성 영상	어그로
4	논란 채널 영상	계정, 구독자, 소유권, 유튜브
5	제작 태도 논란 영상	스트레스, 장난, 웃음거리, 편집
6	동물 학대 영상	스트레스, 편집, 학대

표 3. 비공감 유형과 매칭되는 토픽어(요리)

번호	유형	토픽어 예시**
1	혐오 요리 영상	Alive, Bad, Fucking
2	재료 논란 요리 영상	Batman, China, Bats, Corona
3	논란 채널 영상	Back, Support, Welcome, 악플
4	동물 학대 영상	Cruelty, Dead, Live, Hell

* 반려동물 분야 2개 영상은 해당 분야와 상관없는 영상이라 분석에서 제외

반려동물 분야는 채널명이나 동물명이 많아 영상별 댓글 간에 중복되는 토픽어가 작게 나타나지만, 요리 분야는 채널명보다 요리에 대한 의견과 감성어가 많이 등장함을 확인할 수 있었다. 즉, 본 연구에서는 감성 분석을 수행하지 않았지만, 토픽어만으로 비공감 유형을 결정하는 데 제약이 있을 수 있다는 것을 의미하는 것으로, 본 연구가 분야에 따라 토픽어와 감성 분석의 중요도가 다르다는 것을 확인하였다는 의의도 있다. 연구 환경과 시간적 제약으로 본 연구에서는 최대 30개의 토픽어만 사용하였지만, 향후 자동화를 통해 군집화와 분류 기법을 적용한다면 더 광범위한 확인이 가능할 것으로 본다.

반려동물 분야의 유형1(시청 불가 영상)은 비공감 지수가 매우 높았지만, 댓글을 허용하지 않아 직접 확인한 결과, 내셔널 지오그래픽 코리아 채널에서 한국을 국가락에 포함했음을 확인할 수 있었고, 유형5와 유형6과 같이 한 영상이 두 가지 유형을 동시에 가지는 경우도 있었다. 요리 분야는 외국어 댓글들의 비중이 높았으며, 감성어들을 많이 포함하는 경향이 있었다. 특히, 'Indonesian Food - FRUIT BAT MEAT Cooked Two Ways Manado Indonesia'와 같은 영상은 최근 코로나 시국이라는 특수성으로 다양한 감성어와 높은 비공감 지수를 보임을 알 수 있었다.

Ⅳ. 결론 및 향후 연구

본 연구는 사용자들이 어떠한 이유로 비공감을 표시하는 성향이 있는지를 분석하기 위해, 10개 분야의 8만 건 이상의 영상들을 수집하고 여러 단계를 거쳐 토픽어의 중요성과 감성 분석의 필요성을 확인할 수 있었다. 또한, 분야별로 공감 비율 분포가 상이하고, 공감 지수와 비공감 지수의 분포도 많은 차이가 있음을 확인할 수 있었다는 점도 중요한 기여가 될 수 있다고 본다.

향후 자동으로 토픽어를 추출하고 유형별로 분류하는 기법을 도입한다면, 채널 운영자들이 비공감 영상을 이른 시점에서 예측하고 그 원인을 파악하는 데 도움을 주는 동시에 사용자에게 긍정적인 영상을 제공할 수 있는 유튜브 환경을 개선할 수 있을 것으로 기대한다.

References

[1] W. Hoiles, "Engagement and Popularity Dynamics of YouTube Videos and Sensitivity to Meta-Data," IEEE Transactions on Knowledge and Data Engineering 29(7), 2017.

[2] J. Seol, "An Exploratory Study on Most

** 영어는 형용사 포함

Popular YouTube Channel Genres and Their Popularity," Media Economics & Culture 13(1), 2021.
[3] J. Cho, "A Study for Development of Indicators for Evaluation of Local Governments' YouTube PR Activities," Journal of Digital Contents Society, 21(9), 2020.
[4] https://365kim.tistory.com/93 (It was accessed on 2021.9.13.)
[5] https://youtube-rank.com/board/?mid=home (It was accessed on 2021.9.13.)