유튜브 데이터 마이닝을 통한 영상 소비자들의 관심주제 분석 : 유아동 콘텐츠의 사례

Analysing the subjects of online video consumer's interest using Youtube data mining : the case of kids channels

박영준, 송용백, 이승윤, 이준송, 윤장혁 교수
건국대학교 산업공학과
{pharos.veritatis, ybsong55, ecosy34, tlagksshl4, janghyoon}@gmail.com

Keyword: Social data, Sentiment Analysis, Youtube, Text mining, Deep Neural Network

Abstract.

Customer analysis has had a significant impact on a variety of business areas. Especially, it plays a key role in the Digital Media field. According to recent studies, various analysis based on social data from popular Digital Media, such as YouTube, is in progress. However, few studies use a customer's actual response, "user's sentiment". Thus, we study how to understand their reactions by using Deep-Learning Sentimental Analysis and Time Series Analysis to catch the megatrends of Digital Media and predict intellectual trends soon. In this study, our theme only focuses on Youtube kids' channel contents. We extract megatrends which we define as a 'Subjects of user's interest' using Weighted Term-Frequency Matrix from related video title and description. As the following step, we use both Sentimental Analysis and Time Series Analysis to discover potential trends. We expect this study to contribute to Multi-Channel Network(MCN) business and help creators to make popular content considering real consumer interests.

1. 서론

소셜 미디어의 등장은 미디어를 소비하는 소비자의 양상을 기존의 일방향적이고 폐쇄적이던 방식에서, 생산자와 소비자 간의 활발한 상호작용이 이뤄지는 방식으로 변화시켰다. (Kim et al., 2014; Cho and Kim, 2011; Lee et al., 2009). 소비행태가 바뀌면서 새로운 콘텐츠 플랫폼이 각광을 받고 있는데, 유튜브(Youtube)가 좋은 예시이다. 유튜브는 온라인 오픈 플랫폼이자 구글에 이은 세계 제 2위의 정보검색 서비스로써 2018년 기준 매월 19 억 명에 달하는 이용자가 유튜브를 사용하고 있다. 유튜브에서 활동하는 크리에이터는 생산자와 소비자의 역할을 동시에 갖는 참여형 소비자(프로슈머 prosumer)이기 때문에, 유튜브의 영상은 소비자의 생각과 니즈를 빠르고 민감하게 반영한다. 그래서 유튜브 영상 콘텐츠를 분석하는 것은 시시각각 변화하는 소비자의 관심과 니즈를 파악하는데 도움이 될 수 있다.

2018년 기준으로 유아, 아동 관련 유튜브 콘텐츠 시장은 4조 억원에 달하는 규모로 성장하였다. 유튜브에서 제작되어 선풍적인 인기를 끈 콘텐츠로 '핑크퐁'이 있다. 핑크퐁은 아시아, 유럽, 미주 지역을 아울러 세계적인 인기를 누리며 등장한지 2년 만에 수많은 영상의 조회 수가 억 단위를 넘겼다. 또 다른 킬링 콘텐츠(Killing contents)로는 전문 MC가 나와 장난감을 가지고 노는 '캐리와 장난감 친구들'이 있으며, 이 역시 수백만의 구독자를 거느리고 있다. 그 밖에도 참신하고 창의적인 아동 콘텐츠가 물밀 듯이 나오며 소비자의 선택을 받고 있다. 콘텐츠를 이용자가 소비했다는 것은 영상이 그들의 필요를 이해하고, 충족시켰다는 것을 뜻한다. 이는 유아, 아동 콘텐츠 시장에 많은 기회가 있음을 의미하기도 한다. 그래서 본 연구에서는 유튜브 유아, 아동 콘텐츠를 대상으로 소비자의 관심과 니즈를 분석해보았다.

유튜브와 관련된 연구를 살펴보면, Chatzopoulou et al.(2010)의 연구는 댓글 수, 조회 수 등 정량적인 요소를 통해 상관관계를 분석하였다. 하지만 이 방식으로는 불명확하고 간접적인 결론 밖에 도출할 수 없었다. 그 후에도 유튜브를 활용한 연구가 있었으나 콘텐츠 자체가 아닌 콘텐츠를 이용하는 구독자와 친구 중심의 인적 네트워크 연구가 대부분이었다.(Susarla, 2012; Yoganarasimhan, 2012). 최근의 한 연구에서는 콘텐츠를 중심으로하여 일반요인의 상관관계와 네트워크 분석을 활용한 통합모델을 제시한 바 있으나(Park and Lim, 2015) 실제 유튜브 콘텐츠를 이용하는 소비자의 관심과 니즈를 효과적으로 파악하기에는 한계가 있었다. 이처럼 콘텐츠에 대한 소

비자의 실질적인 반응을 고려한 연구는 아직 미흡한 실정이다.

영상 콘텐츠에 대한 소비자의 적극적인 반응은 댓글을 통해 확인할 수 있다. 소비자는 영상의 제목과 썸네일에 혹해 영상을 클릭할 수 있기 때문에 조회 수는 영상에 대한 단순한 호기심과 관련이 된다. 하지만 댓글은 직접 의지를 들여 의견을 표명하는 과정이 필요하기 때문에 소비자의 적극적인 반응 그 자체라 할 수 있다. 실제로 댓글분석은 마케팅적 관점에서 소비자들의 인식을 적극적으로 반영하는 것으로 알려져 있으며, 그 활용이 크게 확대되고 있다(김성태, 2014). 따라서 본 연구는 기존에 집중적으로 다루지 않은 '댓글'이라는 요인을 심층적으로 활용하였다.

본 연구에서는 콘텐츠에 대한 실질적인 관심을 분석하기 위해 댓글의 수와 게시날짜, 작성내용을 활용하였다. 또한 영상의 제목과 Description에는 콘텐츠의 주제가 함축적으로 담겨있다는 관점 하에, 제목과 Description에서 추출한 여러 키워드 중 소비자의 관심을 나타내는 핵심 단어를 '관심주제'로 정의한다. 본 연구에서는 콘텐츠의 주제를 의미하는 수많은 키워드 중에서 소비자의 적극적 관심을 받으며 미래 수요를 충족시킬 수 있는 관심주제를 도출하고자 한다. 단순히 영상의 조회 수나 댓글 수 등 널리 알려진 지표뿐만 아니라 댓글의 내용에 대한 감성분석을 실시하여 각 관심주제의 시계열적 동향을 파악함으로써 각 관심주제의 near-future를 확인하고, 가까운 미래에 영향력 있는 콘텐츠를 제작하고자 하는 유튜버 및 다중 채널 네트워크(Multi Channel Network, MCN) 비즈니스에 의사결정을 위한 정보를 제공하고자 한다.

2. 선행연구

대표적인 소셜 미디어로써 유튜브에 대한 많은 연구가 진행되어왔다. 이들은 공통적으로 조회 수를 종속 변수로 상정하고 그에 영향을 끼치는 요인에 대한 심도있는 방법론과 모델을 제안하였다.

Chatzopoulou et al.(2010)는 조회 수에 영향을 끼치는 여러 요인들을 탐구하였는데, 그 요인들 중, 댓글 수, 즐겨찾기 등록 수, Like 수가 조회 수와 상당히 높은 상관관계를 가지는 것으로 분석이 되었다. 그리고 콘텐츠를 게시한 시간을 기준으로 요인들의 변화를 확인했으며 이것이 조회 수와 큰 연관이 없음을 확인했다. 또한 네트워크 분석을 통해 콘텐츠 간에 많은 네트워킹이 형성될수록 더 잦은 추천 상위권에 속해 있음을 확인했지만 이 역시 조회 수와 큰 연관은 없음을 확인하였다.

Susarla(2012)는 이전의 연구에서 조명하지 않은 구독자 시스템 등 유튜브 내 인적 네트워크 기반의 콘텐츠 분석을 진행하였다. 이 연구를 통해 댓글 수, 평점, 좋아요 수와 같은 일반적인 요인 뿐만 아니라 네트워크 요인을 친구 네트워크, 구독자 네트워크 방식으로 나누어서 분석을 하였다. 그 결과 콘텐츠 게시 이후 초기단계에서 구독자 네트워크 방식이 영향력이 있었고, 시간이 흐른 뒤 콘텐츠가 어느정도 만연한 상태에서는 친구 네트워크 방식이 더 영향을 끼침을 확인하였다.

Park and Lim(2015)은 위 논문들과 분석의 흐름은 동일하지만 실제적인 콘텐츠 중심의 분석을 통해 기획의 관점에서 콘텐츠 확산을 위한 가이드라인을 제시해주었다. 그 중에는 다수의 콘텐츠와 유사도를 높이고, 검색빈도를 줄이는 방법과, 구독자 수를 늘리는 전략, 콘텐츠 등록 시간을 오래 유지할수록 유리하다는 방안이 있다. 추가적으로, 네트워크 요인 중 연결정도와 고유벡터 중심성이 조회 수에 negative한 영향력을 준다는 점을 찾아냈다. 즉, 많은 콘텐츠와 관련될수록 유사한 콘텐츠 군에 속하여 조회 수를 효과적으로 올리지 못하고 분산된다는 뜻이다. 또한, 고유벡터 중심성이 높을수록 이웃한 콘텐츠가 인기 있을 확률이 높으므로 이러한 콘텐츠와는 연결을 피하는 편이 좋았다. 이 연구에서 제시하는 내용이 기존의 연구와 비교해 현실적인 측면을 담았다는 점에서 의미가 있지만, 한계점으로는 키워드를 기반으로 하여 콘텐츠에 대한 일반 법칙을 확인하는 과정에서 간접적인 콘텐츠 네트워크만을 고려했다는 점이다. 실제적인 콘텐츠의 내용과 그 이용자의 반응까지 고려한 현실적인 연구가 필요한 시점이다.

따라서 본 연구에서는 유튜브 유아, 아동 분야를 대상으로 인기 영상 콘텐츠에 대한 제목과 Description, 그리고 이용자의 실질적 반응을 볼 수 있는 댓글 내용의 감성, 댓글 개수, 게시날짜 및 조회 수를 활용하여 유튜브 영상 콘텐츠를 대표하는 관심주제의 주요 동향을 파악하고자 한다.

3. 연구 방법

본 연구는 유튜브 영상 데이터를 활용하여 특정분야의 주요 관심주제의 Near-future를 확인하기 위한 방법론을 제시한다. 분석 대상이 되는 분야는 앞서 서론에서 언급한 유아, 아동 분야이며, 이와 관련된 데이터를 수집 및 활용하였다.

우선 유튜브 유아, 아동 분야 상위 100개 채널에 대하여 파이썬 기반의 웹 스크래퍼를 제작 및 활용하여 모든 영상정보를 추출한다. 이후 콘텐츠가 가지고 있는 {제목,Description} 문자열로부터 파이썬 Konlpy 패키지를 사용하여 명사단위 키워드를 추출하고 정제하여 관심주제들을 도출한다. 도출된 관심주제에 대하여 이를 포함하는 영상들의 모든 댓글을 웹 스크래핑하여 모으고, 딥러닝으로 설계한 감성

분석 모델을 활용하여 댓글의 감성을 판단한다. 그 결과 각 관심주제의 댓글들에 기재된 작성시기를 바탕으로 해당 관심주제의 시기별 총 댓글 수, 긍정 댓글 수, 부정 댓글 수 3가지 정보로 시각화를 진행한다. 시각화된 댓글 수 및 감성정보를 통하여 해당 관심주제의 시계열적 인기 변화 및 소비자의 감성적 반응추이를 직관적으로 해석한다. 이를 통해 유튜브 유아, 아동 콘텐츠에 대한 소비자 관심주제의 시기별 동향과 Near-Future를 예측할 수 있다.

<Figure 1> Total Process of this Study

3.1 데이터 수집 및 전처리

유아, 아동 분야에 대한 기회를 도출하기 위하여 유튜브 유아, 아동 관련 채널 top100 안의 모든 영상들에 대한 정보를 수집한다. 본 연구는 유튜브 콘텐츠의 제목과 Desription, 영상에 달린 댓글 내용, 게시날짜, 댓글 수를 기반으로 진행되므로 분석과정에서 필요한 정보를 모두 수집하여야 한다. 파이썬 기반의 웹 스크래퍼를 제작하였고 top100 채널에 포함되는 모든 유튜브 영상의 URL를 활용하여 위에서 언급한 정보를 모두 추출한다. 수집 결과, 유아, 아동 분야 상위 100개의 채널에 대한 영상이 약 8만개, 영상에 대한 댓글은 약 830만개가 수집되었다.

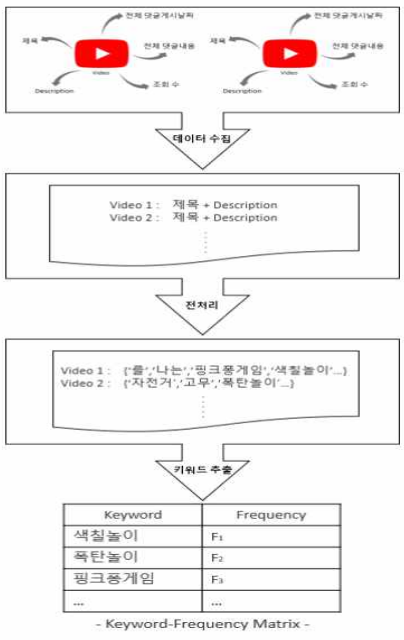
3.2 키워드 추출

이전 단계에서 도출한 영상 콘텐츠 정보에서 관심주제 중심의 분석을 진행하기 위하여 영상별 {제목, Description} 문자열로부터 명사 단위의 정보를 추출한다. 해당 과정은 파이썬 Konlpy을 활용한다. 다음으로 실제적인 분석에서의 관심주제로 선정할 유의미한 단어를 필터링 하기 위해 단어의 글자 수가 3글자 이상인 단어만 추출한다. 키워드 추출을 통해 추출된 단어들을 관심주제로 정하기 위해서는 그 단어가 콘텐츠를 대표하는 주제를 의미해야 하지만 일반적으로 단일, 2글자 단어의 경우 완성된 단어에서 모호한 의미를 갖는 경우가 많기 때문에 3글자 이상의 단어만을 추출하였다. 예를 들어, '나', '놀이', '색칠'은 너무 일반적인 의미를 갖고 있거나 다른 명사와의 관계를 유추하기 쉽지 않다. 하지만 3글자 이상의 단어는 '색칠공부', '영화리뷰'처럼 대부분 복합명사이거나 실제적인 의미를 가진 단어일 가능성이 높기에 3글자 이상으로 2차 필터링 과정을 거친다.

3글자 이상으로 추출된 단어들 중에서 핵심이 되는 단어를 선택하기 위해 단어를 포함하는 영상의 개수와 영상의 조회 수를 곱한 값을 가중치로 주었다. 추출된 3글자 이상의 단어 각각은 TF(Term-Frequency)값을 갖는데, 이는 해당 단어를 포함하는 영상이 몇 개인지를 나타낸다. 다음으로 특정 키워드를 포함하는 영상마다 서로 다른 조회 수를 가지고 있다는 점을 고려하여, 추출된 단어의 TF값에 각 영상의 조회 수를 곱한다. 이 과정은 조회 수가 높은 영상에 등장한 단어일수록 높은 가중치를 부여함을 의미한다. 단, 조회 수가 1억이 넘어가는 영상과 그에 비해 아주 낮은 영상이 있기 때문에 조회 수에 Log scale 취하여 적용했다. <eq 1>은 추출된 3글자이상 단어의 중요도를 계산하는 식이다.

$$Applied\ TF_{keyword\beta} = \log(View\ Count) \times TF_{keyword\beta}$$

<eq 1>



<Figure 2> Total Process of Web Scraping, Preprocessing, Extracting Major Keywords

3.3 댓글에 대한 감성분석

감성분석(Sentiment Analysis)이란 텍스트에 드러난 감정, 의견과 같은 주관적인 정보를 추출해내는 분석방법론이다. 본 연구에서는 댓글의 감성분석을 통해 해당 댓글을 긍정 혹은 부정 중 하나로 이진분류(binary classification) 하였다. 일반적으로 감성분석은 데이터 수집(Data Collection), 주관성 탐지(Subjectivity Detection), 극성 탐지(Polarity Detection)의 3 단계를 거쳐 진행된다.

3.3.1 데이터 수집(Data Collection)

본 연구에서는 지도학습(Supervised Learning)과 딥러닝(Deep Neural Network, DNN)을 이용하여 감성분석 모델을 설계하였다. 지도학습에 필요한 학습 데이터(training data set)와, 모델 테스트를 위한 테스트 데이터(test data set)는 네이버에서 제공하는 '네이버 영화평 리뷰 데이터(naver sentiment movie corpus)'를 사용하였다.

'네이버 영화평 리뷰 데이터'는 네이버 영화 서비스에 작성된 리뷰를 대상으로, 영화당 100개의 리뷰를 모아 총 20만개의 리뷰(train data set: 15만개, test data set: 5만개)로 구성된다. 리뷰는 10점 척도로 평점이 매겨져있다. 중립이라고 판단되는 5~8점 리뷰는 제외하고 9~10점은 긍정으로, 1~4점은 부정으로 판단하여 리뷰의 감정이 labelling(긍정: 1, 부정: 0) 되어있다. 긍정과 부정의 리뷰를 각각 50%의 비율로 데이터에 포함시켜 학습 데이터(15만개)와 테스트 데이터(5만개)를 제공한다.

유튜브의 댓글과 영화의 리뷰는 공통적으로 영상을 대상으로 하고 있으며, 영상에 대한 소비자의 감정을 특정 분야에서만 사용되는 단어가 아닌, 감정을 나타내는 일반적인 단어로 나타내고 있다. 그래서 본 연구의 감성분석 모델을 학습시킬 데이터로써 타당하다고 할 수 있다.

Id	Document	Label
7156791	액션이 없는데도 재미 있는 몇안되는 영화	1
9045019	교도소 이야기구먼 ..솔직히 재미는 없다..평점 조정	0
2718894	보면서 웃지 않는 건 불가능하다	1
8480268	주제는 좋은데 중반부터 지루하다	0
164908	재밌는데 별점이 왜이리 낮은고	1
2009382	뭐냐..시작하고 3분만에 나왔다. 리플릿 사진 보며 불안하더니만..	0

<Table 1> Example of Training Data Set

3.3.2 주관성 탐지(Subjectivity Detection)

수집한 데이터에서 감정, 의견과 같은 작성자의 주관을 나타내는 부분을 추출하여 텍스트 데이터를 특성벡터(feature vector)로 변환해주는 단계가 주관성 탐지이다.

본 연구에서는 문맥(context)을 이용하여 댓글의 감성을 분석하였다. 단순히 감성을 나타내는 단어의 유무만으로는 댓글의 감성을 판단하기 어렵기 때문이다. 예를 들어, '재미가 있다.'와 '재밌지 않다.'는 둘 다 '재미'라는 단어를 포함하지만 감성은 정반대이다. 하지만 문맥을 이용하여 감성을 판단하면 더 정확하게 댓글의 감성을 판단할 수 있다.

문맥은 정해진 구간 내의 단어들로 알 수 있는데, 보통 3~17개의 단어로 문맥을 파악한다고 한다(Jurafsky et al, 2015). 댓글의 경우 한 문장이 대부분이며, 한 문장은 보통 3~17개의 단어로 이루어지므로, 댓글에 등장한 단어들의 조합(Co-occurrence)이 문맥을 만들어내며 감성을 판단할 수 있는 근거가 된다.

댓글을 특성벡터(feature vector)로 변환하기 위해서는 특정 단어들로 구성된 Bag of words가 필요하다. 정보성이 높은 단어들로 Bag of words를 구성하기 위해 학습 데이터(15만개)의 댓글을 형태소 단위로 분리하고, 각 형태소의 TF-IDF(Term Frequency - Inverse Document Frequency)를 계산하여 상위 10000개의 단어를 특성(feature)으로 선택하였다. 학습 데이터의 각 댓글은 구성된 Bag of words을 기준으로 TF(Term Frequency)을 계산하여 특성벡터(feature vector)로 바꿔준다.

단어수 (중복)	1,223,753
단어수 (미중복)	49,845

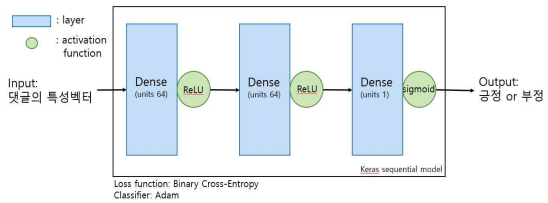
<Table 2> Result of Tokenizing Training Data

3.3.3 극성 탐지(Polarity Detection)

극성 탐지 단계에서는 주어진 데이터를 긍정 혹은 부정으로 판단한다. 감성분석 모델은 딥러닝(DNN)을 활용하여 설계하였다. 최근 딥러닝을 이용한 감성분석이 좋은 성과를 보여주고 있기 때문에 본 연구에서도 감성모델 설계에 적용하였다(서상현 김준태, 2016). 모델을 구성하고 학습 데이터의 벡터화 된 댓글을 이용하여 모델을 학습시켰고, 테스트 데이터를 이용하여 모델의 성능을 평가하였다. 모델의 성능은 정확도(accuracy)가 83.86%, 정밀도(precision)가 84.03%, 재현율(recall)이 83.88%로 측정되었다.

	word	count		word	count
1	영화/noun	14092	2	./punctuation	14054
3	보다/verb	10658	4	../punctuation	9550
5	.../punctuation	8718	6	하다/verb	8000
7	재밌다/adj	6458	8	좋다/adj	5444
9	없다/adj	5404	10	너무/adverb	4944
11	정말/noun	4417	12	!/punctuation	4402
13	?/punctuation	3954	14	최고/noun	3954
15	점/noun	3815	16	있다/adj	3484
17	~/punctuation	3390	18	ㅋㅋㅋ/korean Particle	3067
19	재미있다/adj	3062	20	않다/verb	2867

<Table 3> Top 20th Keywords except postposition



<Figure 3> Deep Learning Model used in Sentimental Analysis

3.4 관심주제의 시계열 분석

각 관심주제와 관련된 모든 댓글의 감성을 분석하여 관심주제별로 댓글의 게시 날짜를 기준으로 총 댓글 수, 긍정 댓글 수, 부정 댓글 수 3가지에 대해 시계열 그래프를 작성한다.

Keyword : '로보카폴리'		
Comment	Time	Sentiment
"너무너무 축하드려요"	11개월 전	1
"저도 엄마한테 사달라 해야쥐"	11개월 전	1
"딸아이가 너무 좋아하겠어요"	10개월 전	1
"난다재미없어"	10개월 전	0
"우와재밌겠다"	9개월 전	1
"윽 궁금"	9개월 전	1
"데미안은 인성쓰레기입니다"	6개월 전	0
"데미안 나쁨"	6개월 전	0
"재밌내"	5개월 전	1

<Table 4> Example data format of keyword, '로보카폴리' after Sentimental Analysis

최종적으로 본 연구에서는 시기에 따라 관심주제에 대한 소비자의 감성 변화를 분석하여 시각화 한다.

4. 연구결과

(4.1) Weighted Term Frequency Matrix

유아, 아동 분야 조회 수 상위 100개 채널의 82,007개 영상에 대하여 제목과 영상 설명의 텍스트를 대상으로 영상의 목적과 영상의 내용을 대표하며 소비자에게 선택받는 관심주제를 찾는 키워드 분석을 실시하였다. 한글 텍스트에서 의미 있는 정보를 분석하고 추출하는 여러 가지 자연어 처리 방법들 중 본 연구는 한글자연어 처리에 가장 대중적으로 사용되는 KONLPY를 사용하여 한글의 품사를 구분하였다. 영상의 텍스트에 포함된 대명사를 제외한 명사들이 영상이 다루고자 하는 내용을 대표하는 것으로 핵심단어는 명사만을 수집한다.

3글자 이상의 단어로 구성된 핵심단어들은 총 36,059개가 도출되었고 앞서 설명한 바와 같이 영상의 조회 수를 가중치로 설정하여 weighted-Term-Frequency 매트릭스를 생성하였고 키워드 당 가중 출현빈도 값이 1,000 이하인 키워드를 제거하였다. 그 결과 421개의 관심주제를 도출하였으며 유튜브 영상의 내용을 설명하지 못하는 명사 '구독하기', '유튜브', '인스타그램' 등 49개를 분석대상으로 적합하지 않다고 판단하여 정성적(定性的)으로 제거하였다. 최종적으로 남은 372개의 관심주제들에 대해 분석을 실시하였다. <Table 5>은 분석대상 키워드 상위 25개와 가중 출현빈도 값을 나타낸 표이다.

순위	키워드	값	순위	키워드	값
1	장난감	1341484	13	미니특공대	72654
2	뽀로로	520435	14	꼬꼬스토이	67879
3	터닝메카드	268452	15	코코몽	61098
4	애니메이션	209425	16	인기동요	59428
5	헬로카봇	198868	17	슈퍼윙스	58520
6	장난감놀이	189577	18	다이노코어	58382
7	꼬마버스 타요	132119	19	토이 스페이스	57338
8	로보카폴리	111557	20	장난감 애니메이션	48047
9	파워레인저	103156	21	영어동요	44096
10	깨비키즈	99917	22	토이팩토리	43890
11	상상놀이터	81678	23	다이노포스	41938
12	장난감친구	81419	24	액체괴물	40483

<Table 5> Top 24 Keyword – Weighted Term Frequency value

상위 24개의 관심주제에 뽀로로, 헬로카봇, 꼬마버스타요 등 유명 캐릭터가 다수 위치하였으며 인기동요, 영어동요, 액체괴물 등의 영상이 다루는 내용에 관련된 단어들도 도출되었다.

(4.2) 관심주제별 댓글의 감성분석 결과

키워드-영상 매트릭스를 통해 각 키워드를 포함하는 영상의 집단을 만들고 각 영상마다 가지는 고유 url code를 통하여 영상에 속하는 댓글을 모아 keyword-댓글 set을 구축한다. 키워드별 댓글 set에 대하여 본 연구에서 개발한 감성평가 딥러닝 모델을 적용하여 댓글을 긍정, 부정으로 분류하였다. 감성분석이 완료된 댓글들을 2019.05을 기준으로 월별로 집계하였다. <Table 6>은 '다이노포스', '인형놀이' 두 관심주제에 대한 월별 댓글 집계 결과이다.

다이노포스			인형놀이		
time	긍정	부정	time	긍정	부정
18.05	21087	8025	18.05	10427	5421
18.06	25591	8714	18.06	9275	4029
18.07	27439	10491	18.07	9041	4123
18.08	23004	7756	18.08	9839	4054
18.09	16374	5649	18.09	10010	5175
18.10	12518	3265	18.1	10495	5450
18.11	10683	2852	18.11	11347	6381
18.12	12749	3346	18.12	14972	7667
19.01	8617	2581	19.01	19172	10169
19.02	4885	983	19.02	18055	9480
19.03	3967	905	19.03	19993	10980
19.04	3639	742	19.04	20103	12190

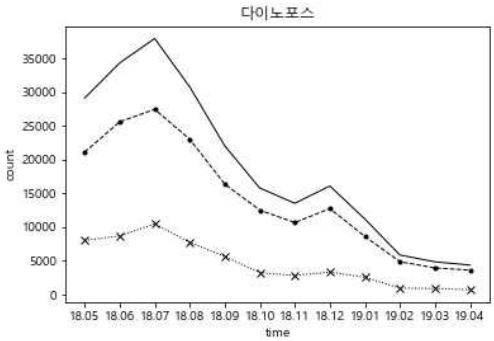
<Table 6> Example of Sentimental Analysis per keywords, '다이노포스', '인형놀이'

(4.3) 관심주제별 댓글의 시계열 분석 결과

총 372개의 선별된 관심주제에 대하여 댓글들의 감성 및 시계열 분석을 한 결과, 크게 5가지 패턴을 확인할 수 있었다. 우선 전체적으로 상향적 동향을 보이는 16개의 관심주제 집단, 하향적 추세를 보이는 99개의 관심주제 집단이 있다. 또한 시계열적으로 106개의 관심주제가 계절적 추세를 보이고 있으며, 이와 반대로 큰 변동을 보이지 않는, 41개의 안정적 관심주제 집단이 있다. 이외에 댓글 수가 적어서 유의성을 내포하지 않거나, 매우 불규칙적 패턴을 보이는 76개의 관심주제가 있다.

(4.3.1) 하향적 추세의 관심주제

2018년 5월부터 2019년 4월까지 전체적으로 하향적인 추세를 보이는 관심주제는 총 99개로 다이노포스, 터닝메카드, 파워레인저 등이 포함되어 있다. 또한 공통적으로 <Figure 4>과 같은 하향적 추세를 보이며, 전체 집단을 최대 댓글 수로 분류를 하였을 때 <Table 7>과 같이 나뉜다.



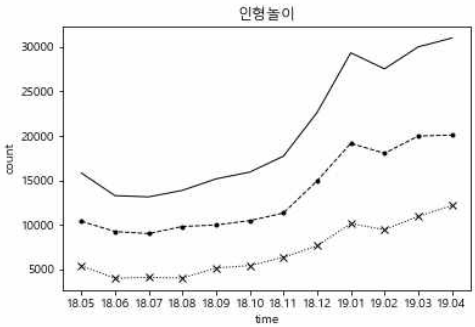
<Figure 4> Downward treand of Keyword '다이노포스'

Max number of Comment (N)	Representative Keywords
N > 10000	다이노포스, 뽀로로, 시크릿쥬쥬, 터닝메카드, 파워레인저
N > 5000	변신로봇, 베이블레이드, 캐리앤송, 토이박스, 디즈니애니메이션
N > 1000	인기동화, 인형드라마, 스타워즈, 옵티머스, 마인크래프트
N > 100	로봇트레인, 교통안전이야기, 토이스토리, 율동동요, 미니피규어

<Table 7> Representative Keywords among Downward Trend Set classified by Maximum number of comments

(4.3.2) 상향적 추세의 관심주제

2018년 5월부터 2019년 4월까지 전체적으로 상향적 동향을 보이는 관심주제는 총 16개로 인형놀이, 상황극, 상상이야기 등이 포함되어 있었다. 또한 <Figure 5>과 같은 공통적으로 상향적 추세를 보이며, 관심주제를 최대 댓글 수로 분류를 하였을 때 <Table 8>과 같이 정리된다.



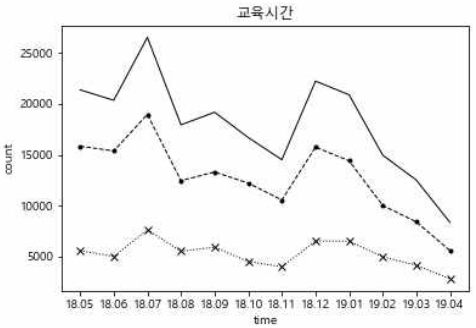
<Figure 5> Upward treand of Keyword '인형놀이'

Max number of Comment (N)	Representative Keywords
N > 10000	인형놀이, 상황극, 상상이야기, 밀착중계, 사우루스
N > 5000	병원놀이, 바비인형, 구급차
N > 1000	과학도감, 낱말카드, 영어동요, 캐비키즈, 추천영상
N > 100	꿀잼 동영상, 라이센스, 래빗미디어, 상상더하기, 세계공룡싸움

<Table 8> Representative Keywords among Upward Trend Set classified by Maximum number of comments

(4.3.3) 계절적 추세의 관심주제

2018년 5월부터 2019년 4월까지 전체적으로 계절적 동향을 보이는 관심주제는 총 106개로 겨울왕국, 교육시간, 바다탐험대 등이 포함되어 있었다. 또한 <Figure 6>과 같이 계절적으로 증감하는 동향을 보이며, 관심주제를 최대 댓글 수로 분류를 하였을 때 <Table 9>과 같이 정리된다.



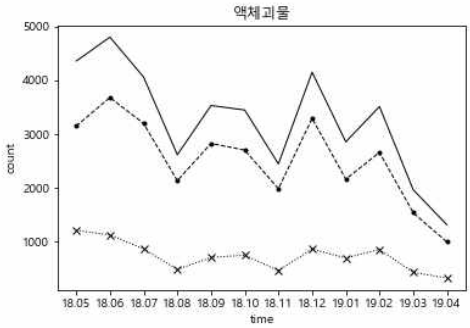
<Figure 6> Seasonal trend of Keyword '교육시간'

Max number of Comment (N)	Representative Keywords
N > 10000	겨울왕국, 교육시간, 꼬바머스, 인스타그램, 디즈니
N > 5000	대문밖장난감, 영어공부, 색칠공부, 신제품, 포켓몬스터
N > 1000	생활습관, 상허가족, 영어그림사전, 동화이야기, 놀이공원
N > 100	과학호기심, 바다탐험대, 직업놀이, 시사상식, 미니언즈

<Table 9> Representative Keywords among Seasonal Trend Set classified by Maximum number of comments

(4.3.4) 안정적 추세의 관심주제

2018년 5월부터 2019년 4월까지 안정적 동향을 보이는 관심주제는 교육채널, 장난감놀이, 안전수칙 등 총 41개가 나왔다. 대체적으로 <Figure 7>과 같이 안정적 추세를 보이며, 관심주제를 최대 댓글 수로 분류를 하였을 때 <Table 10>과 같이 정리된다.



<Figure 7> Seasonal trend of Keyword '액체괴물'

Max number of Comment (N)	Representative Keywords
N > 10000	게임소개, 교육채널, 장난감놀이, 쥬라기월드, 웹드라마
N > 5000	로보카폴리, 액체괴물, 주방놀이, 티라노사우루스, 만들기놀이
N > 1000	우주여행, 공룡송, 프테라노돈, 메가비스트
N > 100	가면라이더, 일상생활, 안전수칙, 괴짜발명소녀, 긴급출동센터

<Table 10> Representative Keywords among Steady Trend Set classified by Maximum number of comments

종합적으로 도출된 관심주제를 보았을 때 하향적 추세의 관심주제에는 기존의 유명 애니메이션인 '뽀로로', '터닝메카드', '파워레인저'가 도출되었다. 상향적 시계열 특성의 관심주제 집단에는 '상황극', '상상이야기', '인형놀이' 등 다양한 놀이와 관련된 단어들이 나왔다. 한편, 계절적으로 증감하는 추세를 보이는 관심주제로는 '겨울왕국', '영어공부', '색칠공부' 등 계절적 콘텐츠나 교육 관련된 단어가 나왔다. 마지막으로 시기별로 안정적인 댓글 수를 갖는 관심주제는 '게임소개', '교육채널', '안전수칙' 등 게임, 안전 및 교육에 관련된 내용이 나왔다.

5. 토의 및 결론

한국소비자원의 2017년 보고서에 의하면 2017년 동영상 광고비는 5,137억 원을 기록하였다. 1인 미디어 시장의 발달로 동영상 광고비는 2015년부터 매년 60%가 넘는 성장률을 보이고 있다(한국소비자원, 2017). 특히, 유튜브는 국내 1인 미디어 플랫폼의 89.2%를 점유하며 디지털 미디어 시장을 선도하고 있는 추세이다(김청용, 2018). 이러한 유튜브 플랫폼은 양방향 매체이기 때문에 소비자의 니즈를 파

악하는 것이 무엇보다 중요하다. 하지만 단순 영상 조회 수, 댓글 개수만으로 소비자의 니즈를 파악하는 데에 한계가 있다. 본 연구에서는 이용자의 주요 관심사를 파악하기 위해 영상 댓글에 남겨진 소비자의 감성적 반응을 이용하였다. 딥러닝 모델을 사용하여 댓글의 감성을 분석하였고 이를 시계열적으로 나타냄으로서 소비자 지적 관심사의 변화추이를 살펴보고자 하였다. 결과적으로 18년 4조억 원의 시장규모를 갖는 유아, 아동 콘텐츠를 중심으로 이용자의 관심주제와 그 동향을 파악하였다.

키워드 분석을 통해 도출된 가장 영향력 있는 단어는 '장난감'으로써 유아, 아동분야 유튜브에서 다양한 장난감을 활용한 콘텐츠가 활발히 제작됨을 알 수 있다. 또한 '뽀로로'와 '헬로카봇', '로보카폴리', '파워레인져', '코코몽' 등 유명 만화영화와 캐릭터들이 주요 키워드임을 확인할 수 있었다. 이는 EBS와 투니버스 등 TV 채널과 병행하여 유튜브도 애니메이션과 캐릭터가 소비되는 주요 영상 플랫폼이라는 것을 의미한다. '칠교놀이', '색칠놀이', '공룡탐험', '과학도감', '변신로봇' 등의 관심주제는 과거 컬러북과 그림책으로 소비 되었던 각종 유아 대상 놀이들이 유튜브 영상으로 제작되어 아이들에게 제공되는 것을 알 수 있다.

영상에 대한 직접적인 관심과 참여를 나타내는 댓글에 대한 시계열 분석 결과를 통해 관심주제를 5가지 패턴으로 분류하였다. 안정적 추세를 보이는 관심주제는 '액체괴물', '장난감놀이', '주방놀이', '우주여행', '공룡송'과 관련된 단어가 포함되어 있고, 이들은 시간에 따라 다소 변동을 보이지만 넓은 관점에서 꾸준한 수요를 보이고 있다. 이러한 관심주제를 포함하는 콘텐츠는 가까운 미래에도 유아, 아동 영상의 주요 소비자인 아동과 그 부모에게 관심의 대상이 될 것임을 알 수 있다. 한편 상향적 추세를 보이는 '인형놀이', '병원놀이', '영어동요', '과학도감' 관심주제는 앞서 결과에서 제시한 <Figure 4>와 같은 추세를 보이기 때문에 가까운 미래에도 영향력 있을 것을 예상할 수 있다. 그와 반대로 하향적 추세를 보이는 다이노포스, 변신로봇, 디즈니애니메이션 등의 관심주제는 지난 1년간 급격한 관심의 하향세에 있기 때문에 미래 유튜브 콘텐츠로의 제작은 경쟁력이 없음을 예측할 수 있다.

본 연구의 댓글 감성분석을 이용한 소비자 니즈의 시계열적 분석은 다중 채널 네트워크(Multi Channel Network, MCN) 비즈니스에 도움을 줄 수 있다. 여러 개의 유튜브 채널과 제휴하여 기술지원, 자금지원, 파트너 관리, 디지털 저작권 관리 등을 제공하는 조직을 MCN이라 한다. MCN은 전반적으로 유튜브 크리에이터를 지원해 줄 뿐 아니라, 콘텐츠 제작에 도움을 줄 수 있는 소비자 데이터도 분석 및 제공하고 있다. MCN에서 분석시 누적 조회 수 및 댓글 수를 활용하는 경우가 있는데 이는 서론에서 언급한 것과 같은 한계점을 갖는다. 하지만 영상의 누적 조회 수를 사용하는 것은 시간의 개념이 빠져 있기 때문에 시시각각 변화하는 소비자의 관심도를 파악할 수 없다. 분석에 댓글수를 이용하는 것은 소비자의 관심도를 파악할 수는 있지만 영상 콘텐츠의 선호도를 파악할 수는 없다. 그렇기 때문에 영상 댓글의 감정을 분석하고 시계열적으로 살펴본 본 연구의 결과는 유튜브 크리에이터와 MCN 비즈니스에 도움을 줄 것으로 전망한다.

본 연구에서는 1년간의 댓글을 대상으로 동향을 분석하였다. 유튜브 영상의 경우 댓글의 작성일시가 절대적 시점이 아닌 상대적 시점으로 기록된다. 즉, 1년 안에 작성된 댓글은 '몇 개월 전'으로 표시되지만 1년이 넘은 댓글은 '몇 년 전'으로 표시가 된다. 분석을 월 단위로 맞추기 위해서는 1년이라는 제한된 기간을 사용할 수밖에 없었던 이유이다. 분석기간이 1년으로 제한되었기 때문에 계절적 추이를 보이는 관심주제에 대한 소비자 니즈 패턴을 찾는 것이 원활하지 못했다는 점이 본 연구의 한계점으로 남는다. 추후 3~5년 사이의 댓글을 월 단위로 수집할 수 있게 된다면 더 정확하게 관심사의 동향을 파악할 수 있을 것으로 기대된다.

참고문헌

Chatzopoulou, G., C. Sheng and M. Faloutsos, "A first step towards understanding popularity in youtube" INFOCOM IEEE Conference on Computer Communication Workshops,(2013),1~5

Cho I. D AND N. G. Kim, "Recommending Core and Connecting Keywords of Research Area Using Social Network and Data Mining Techniques," Journal of Intelligence and Information Systems, Vol.17 No.1(2011), 127~138

Lee, S., S. Park, G, G. Lim, and S. Baek, "A Roadmap for Developing Digital Content Distribution Infrastructure," Journal of Korea Society of IT Services, Vol. 8, No.4(2009), 75~86

Yoganarasimhan, H., "Impact of social network structure on contect propagation : A study using YouTube data," Quantitative Marketing and Economics, Vol. 10. No. 1(2012), 111~150

김선우, 안희웅, 장유나, 홍민예, 서민지, 김성태, " 유투브 댓글 빅데이터 분석을 통한 제품별 소비자 구매 여정 연구 ", 영상문화콘텐츠연구, 16( ), 57 - 89, 2019