유튜브 낚시성 콘텐츠의 주요 구성요소 분석

이서우[1], 조미정[1], 채은비[1], 김해인[1]
[1]덕성여자대학교 컴퓨터공학과
leesw2366@gmail.com, ing06164@gmail.com , chaeeunbi981127@gmail.com,
snowman8775@gmail.com

Analysis of major components of YouTube fishing content

Seo-Woo Lee[1], Mi-jeong Jo[1], Eun-bi Chae[1], Hae-in Kim[1]
[1]Dept. of Computer Science, Duksung University

요 약

본 연구에서는 낚시성 콘텐츠의 주요 구성 요소인 썸네일과 제목을 MLKit와 TF-IDF를 이용하여 분석하고 이를 딥러닝 Sentence BERT 모델에 적용하였다. 이를 활용하여 추후 낚시성 콘텐츠를 걸러내는 알고리즘을 개발 예정이다.

1. 서론

세계 최대 규모의 동영상 플랫폼인 유튜브의 영향은 유저의 수가 나날이 증가함에 따라 그의 영향력도 커지고 있다. 빅데이터 플랫폼 기업 아이지에이 웍스의 모바일인덱스에 따르면[1][2] 유튜브 사용자는 4천164명으로 앱별 사용자 1위를 차지했고 정보 검색에서도 절대 강자에 오른 네이버를 유튜브가 추격하여 네이버(86.7%)와 유튜브(60.3%)의 격차는 크게 줄었다. 하지만 유튜브에 있는 모든 정보가 사실만을 담고 있지는 않다. 콘텐츠 크리에이터들은 사람들의 인기를 얻어야 한다는 생각에서 자극적이고 허위 정보를 담은 콘텐츠를 올리기도 한다. 현재 유튜브 내에서 이에 대하여 제재하고 있지만, 제목과 썸네일을 통해서 사용자가 동영상을 시청하게 하고 사실은 전혀 관련된 내용을 담고 있지 않은 일명 '낚시성' 콘텐츠들은 아직도 유효하다.

따라서 본 연구에서는 낚시성 콘텐츠의 구성 요소를 분석하고 이를 바탕으로 추후 낚시성 콘텐츠를 걸러내는 알고리즘을 개발 예정이다.

2. 본론

2-1) 낚시성 콘텐츠의 구성 요소

사용자가 유튜브 영상을 선택하기 위해 고려할 수 있는 요소로는 썸네일, 영상 제목 크게 2가지로 정의할 수 있다. 그중 영상을 한눈에 표현해 주는 썸네일은 긴 영상을 단 한 장의 이미지로 표현하기 때문에 조회수 향상에 가장 주요한 요인으로 꼽힌다.

두 개의 매력적인 썸네일을 가진 영상 중 하나를 골라야 한다면 판단 기준이 되는 것은 영상의 제목이다. 영상에 대한 텍스트 중 가장 큰 크기인 제목은 사람들이 영상을 시청할 때 참고하게 되는 두 번째 요인이다.

사람들은 유튜브에서 동영상의 검색 순위를 높이기 위해 구글 키워드플래너와 같은 도구를 활용하여 관련 키워드의 정보를 찾아 이를 영상의 썸네일, 제목 등에 활용한다. 경쟁력이 높은 키워드들로 구성된 콘텐츠는 다른 영상에 비해 높은 클릭률을 가져오기도 하지만 경쟁력을 높이기만을 위한 키워드는 영상의 내용과 항상 일치하지는 않는다.

2-2) 영상 정보 추출에 사용된 기술

낚시성 콘텐츠를 걸러내는 필터링을 구현하기 위해서는 콘텐츠를 구성하는 요소들에 대한 분석이 요구된다. 제목, 썸네일, 설명, 내용 등 영상의 구성요소 분석을 위해 다양한 기술을 적용해 볼 수 있다.

썸네일 이미지에서 텍스트 정보 추출을 목적으로 광학문자인식(OCR) 기술인 MLKit를 사용하였다. 썸네일을 비트맵으로 변환하는 과정을 통해 이미지를 전처리 후 텍스트만을 추출하여 영상 분석의 데이터로 활용하였다.

그림 1 MLkit 실행 결과 예시

영상의 내용 분석은 영상의 음성데이터를 활용하는 방법을 선택하였다. Amazon Transcribe 기술을 적용하여 음성언어를 문자 데이터로 전환(STT)을 통해 스크립트를 추출한다. 이 같은 방법으로 추출된 데이터는 낚시성 콘텐츠 판단 알고리즘에 사용된다.

2-3) 주요 분석 기법

앞선 내용을 바탕으로 본 연구에서는 STT(Speech-To-Text)를 통해 유튜브 콘텐츠의 음성을 자막화하고 이를 바탕으로 TF-IDF(Term Frequency Inverse Document Frequency)를 이용한 키워드 추출 및 Sentence BERT(Bidirectional Encoder Representations from Transformers)를 통한 제목과 STT 문장 간 유사성 계산을 수행하여 낚시성 콘텐츠의 내용 분석을 수행하고자 한다.

2-4) TF-IDF(Term Frequency Inverse Document Frequency)

TF-IDF는 텍스트 마이닝에서 사용되는 가중치로 문서 군 내의 각 문서에 대해 특정 단어의 상대적 중요도를 나타내는 통계적 모델이다. 이는 단어의 등장 빈도를 이용해 중요도를 나타내며, 문서의 키워드를 추출하기 위해 사용되었다. TF-IDF는 다음과 같은 수식을 통해 계산된다.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}}$$

* $n_{i,j}$: 특정 단어 t가 특정 문서 d에 등장한 횟수
* $\sum_k n_{k,j}$: 문서 d에서 모든 단어의 등장 횟수

TF-IDF는 [3]의 정의에 따르면 TF(Term Frequency)와 IDF(Inverse Document Frequency)를 곱한 값으로 TF는 1개의 문서에서 특정 단어가 출현한 빈도수를, IDF는 전체 문서 수를 특정 단어가 포함된 문서의 수로 나눈 값을 나타낸다.

$$idf_i = \log \frac{|D|}{|d_j | t_j \in d_j|}$$

* $|D|$: 문서군에 포함된 총 문서의 수
* $|d_j | t_j \in d_j|$: 단어 t가 등장하는 문서의 수

$$TF-IDF_{i,j} = tf_{i,j} \times idf_i$$

본 연구에서는 위 수식을 그대로 이용하지 않고 파이썬의 라이브러리인 TfidfVectorizer를 이용해 STT에서 TF-IDF 행렬을 만들고 이를 워드 클라우드 형태로 나타내었다.

그림 2 샘플영상의 워드 클라우드 결과

이는 'A양 인터뷰'라는 검색어로 검색한 결과 중 한 영상에 대한 워드클라우드 결과이다.

제목	TF-IDF
A양, 외모까지 바꿔놓은 희귀병 충격정체...A양의 현재상태가 심상치않은 이유ㄷㄷ	['A양', '서류', '인정'] ['0.52', '0.29', '0.17']

<표 1> TF-IDF값이 가장 높은 주요 키워드 목록

결과를 통해 알 수 있듯, 해당 영상은 제목에서 '희귀병', '충격', '심상치않은' 등의 단어와 썸네일에서 해당 연예인의 눈물 흘리는 모습 등을 통해 해당 연예인의 건강 상태가 악화하였음을 암시하지만 실제 영상의 내용에는 이를 짐작하는 내용들뿐이며 그림 2에서 알 수 있듯이 그보다는 다른 내용들이 주를 이루는 콘텐츠이다. 또한 <표1>을 통해 해당 영상은 주요 키워드와 낮은 연관성을 가지는 낚시성 영상이라는 것을 확인할 수 있었다.

2-5) Sentence BERT(Bidirectional Encoder Representations from Transformers)

BERT는 2018년 구글이 공개한 사전 훈련된 언어 모델이다. 본 연구에서는 파이썬의 sentence_transformers 라이브러리를 이용해 BERT 모델을 불러와 klue sts, korSTS 데이터셋으로 fine tunning 하여 사용한다. Sentence BERT는 문장 간 유사도를 계산

하기 때문에 콘텐츠의 내용을 (초반, 중반, 후반)으로 나누어 중반 4개 문장을 택해 썸네일과 제목의 유사도를 계산하기로 하였다. 초반이나 후반에는 영상과 크게 상관없는 대부분 인사말, 끝을 맺는말로 이루어져 있었다.

유튜브 영상 선택의 가장 높은 원인을 차지하는 썸네일의 경우 텍스트를 추출하여 일치율을 확인해 본 결과 아래와 같은 결과가 나왔다.

STT	score
A양은 평소에 멤버 중 D양과 친한 것으로 유명했다.	0.245
그런데 탈퇴한 E양이 2022년 홍보를 시작했다	0.193
지난 1월 배우 B양와 C군이 결혼했는데, 이 결혼식에서 두 사람이 마주쳤다는 소식도 관심을 받았다.	0.136
직원은 C군과 절친으로 알려졌으며, A양 역시 C 드라마에서 함께 호흡을 맞췄기 때문에 결혼식에 참석한 것으로 전해졌다.	0.033

<표 2> 썸네일과 STT 중반 4개 문장 간 유사도

<표 2>는 앞선 낚시성 콘텐츠에 대해 해당 영상의 썸네일과 STT 문장 간의 유사도를 Sentence BERT를 이용해 계산한 결과이다. Sentence BERT를 이용한 STS(Semantic Textual Similarity) 태스크의 경우, 유사도(score)를 0에서 5 사이로 계산하지만 fine tunning 과정에서 유사도를 0에서 1로 정규화했다. 따라서, A양에 관한 영상 제목인 [A양, 외모까지 바꿔놓은 희귀병 충격정체...A양의 현재상태가 심상치않은 이유ㄷㄷ]과 중반 문장 간의 유사도는 최대 24%로 낮은 유사도를 가지고 있다고 판단할 수 있었다.

위와 같은 방법으로 Sentence Bert를 이용하여 영상의 제목과 중반 4개 문장의 유사도를 계산해 본 결과 아래의 <표3>과 같은 결과가 나왔다.

STT	score
A양은 평소에 멤버 중 D양과 친한 것으로 유명했다.	0.348
그런데 탈퇴한 E양이 2022년 홍보를 시작했다	0.273
지난 1월 배우 B양와 C군이 결혼했는데, 이 결혼식에서 두 사람이 마주쳤다는 소식도 관심을 받았다.	0.230
직원은 C군과 절친으로 알려졌으며, A양 역시 C 드라마에서 함께 호흡을 맞췄기 때문에 결혼식에 참석한 것으로 전해졌다.	

<표 3> 제목과 STT 중반 4개 문장 간 유사도

이렇게 TF-IDF를 통해 낚시성 콘텐츠의 제목, 썸네일에 사용되는 단어와 실제 영상의 키워드가 다르다는 것을 시각화할 수 있었고, Sentence BERT를 통해 제목과 STT 문장 간의 유사도를 계산하는 방법을 도출함으로써 낚시성 콘텐츠, 즉 내용과 낮은 유사도를 가지는 정보들은 제목에서 주로 나타난다는 것을 확인할 수 있었다.

결론

본 연구는 기존에 이루어지지 않던 낚시성 콘텐츠의 필터링을 시도했다는 점에 의의가 있다. 본 연구에서는 딥러닝을 활용한 문맥 비교와 텍스트 마이닝을 활용하는 차별화된 새로운 필터링 방식을 제시한다. 이를 통해 사용자 검색어와 콘텐츠의 상관관계를 수치화하는 일치율을 도출할 수 있다.

최종적으로 낚시성 콘텐츠의 무분별한 게재에 제재를 가하고, 건전한 유튜브 문화를 형성함으로써 사용자들에게 원하는 콘텐츠만을 보여주는 편의성을 제공하는 것을 목표로 한다. 본 연구는 더 나아가 낚시성 콘텐츠 판단 알고리즘의 정확도에 초점을 맞추어 성능을 고도화할 계획이다. 유튜브뿐만이 아니라 다양한 미디어 플랫폼으로 적용 범위를 확장한다면 신뢰도 높은 서비스 플랫폼 환경을 조성하고 가짜 뉴스 확산 감소를 도와 사회의 긍정적 변화에 기여할 수 있을 것으로 기대한다.

사사

본 논문은 과학기술정보통신부 정보통신창의인재양성사업의 지원을 통해 수행한 ICT멘토링 프로젝트 결과물입니다.

참고문헌

[1]임은진. (2022.7.20.). 모바일인덱스 "6월 유튜브앱 국내 이용자 4천164만명…엔터 1위". 연합뉴스

[2]김종우. (2022.4.26.). '정보검색 1위' 네이버, 유튜브가 맹추격…나스미디어 '인터넷 이용자 조사'. 부산일보

[3] 이성직, 김한준.(2009). TF-IDF의 변형을 이용한 전자뉴스에서의 키워드 추출 기법. 한국전자거래학회지 (pp. 59 - 73).