k-평균 군집화 기법을 활용한 SNS의 부적절한 광고성 콘텐츠 탐지

이동환*, 임희석**
고려대학교 컴퓨터정보통신대학원 빅데이터융합학과
*tarje2@korea.ac.kr, **limhseok@korea.ac.kr

Detection of inappropriate advertising content on SNS using k-means clustering technique

Dong-Hwan Lee*, Heui-Seok Lim**
*Dept. of Bigdata Convergence, Graduate School of Computer and Information Technology, Korea University

요 약

오늘날 SNS를 사용하는 사람들이 증가함에 따라, 생성되는 데이터도 많아지고 종류도 매우 다양해졌다. 하지만 유익한 정보만 존재하는 것이 아니라, 부정적, 반사회적, 사행성 등의 부적절한 콘텐츠가 공존한다. 때문에 사용자에 따라 적절한 콘텐츠를 필터링 할 필요성이 증가하고 있다. 따라서 본 연구에서는 SNS Instagram을 대상으로 콘텐츠의 해시태그를 수집하여 데이터화 했다. 또한 k-평균 군집화 기법을 적용하여, 유사한 특성의 콘텐츠들을 군집화하고, 각 군집은 실루엣 계수(Silhouette Coefficient)와 키워드 다양성(Keyword Diversity)을 계산하여 콘텐츠의 적절성을 판단하였다.

1. 서론

오늘날의 SNS(Social Network Service)는 단순히 개인의 생각이나 관심을 표현하는 수단을 초월하여 정보 공유, 불특정 다수와의 의사소통, 그리고 기업들의 마케팅 수단으로도 활용되고 있다.[1] 이러한 SNS 상에서의 다양하고 많은 콘텐츠를 누구나 쉽게 접할 수 있지만, 모든 콘텐츠가 사람들에게 유용한 정보를 포함하고 있는 것은 아니다.[2]

(그림 1) 해시태그('코로나일상', '코로나') 검색 결과

특히 (그림 1)과 같이 '코로나일상', '코로나'라는 해시태그가 사용된 콘텐츠 검색 결과에서 부정적, 반사회적, 사행성 광고 등의 부적절한 콘텐츠들을 확인할 수 있다. 이와 같은 콘텐츠들은 많은 대중에게 노출되기 위해, 연관성 없는 해시태그를 남발하고, 반복적인 업로드를 통해 사용자들에게 불편한 감정을 느끼게 한다. 때문에 사용자에 따라 적절한 콘텐츠를 필터링할 필요성이 증가하고 있다.[3] 본 연구에서는 군집화 기법을 활용하여 부적절한 광고성 콘텐츠를 분류할 수 있는 방안을 제시하였다.

2. 관련 연구

부적절한 광고를 필터링하는 연구들은 이메일, 문자 메시지, 블로그, 유튜브 등 사용자가 많이 사용하는 시대적 흐름에 따라 영역이 확장되고 있다. 그 예로, BERT 언어 모델을 이용하여 글의 문맥을 파악하고 광고 유무를 판단하는 연구와[4] 순환신경망 RNN(Recurrent Neural Network)의 LSTM(Long Short-Term Memory)기법을 활용한 유튜브 댓글 스팸 분류의 연구 등이 있었다.[5] 이상의 연구들과 같이 대부분의 광고성 스팸 분류 모델은 언어 모델이나 인공 신경망을 활용하고 콘텐츠의 본문, 내용, 댓글 등에 초점을 맞추고 있다. 하지만 이러한 연구들은 일반적인 단어들로 구성된 콘텐츠들을 분류하는 데에는 어려움이 있다. 또한 분석 결과를 얻기 위해 많은 자원과 시간이 소비된다는 단점이 있다. 따라서 본 연구에서는 콘텐츠의 주제를 의미하는 해시태그를 활용하여 상대적으로 빠르게 분석된 결과를 얻을 수 있는 k-평균 군집화 기법을 적용하였다. 분석

결과는 실루엣 계수와 키워드 다양성을 측정하여 콘텐츠의 적절성을 판단할 수 있었다.

3. 데이터

본 연구에서는 SNS Instagram 콘텐츠 정보를 수집하기위해 Python을 활용한 수집 모듈을 개발하였다. 콘텐츠의 정보는 <표 1>과 같이 '콘텐츠 고유 ID', '콘텐츠 본문' 만을 수집하여 사용 목적에 맞는 최소한의 정보만을 활용하였다.

<표 1> Instagram 콘텐츠 수집 항목

수집 항목	수집 목적
ID	중복 수집된 콘텐츠 처리
본문	콘텐츠 본문 내 해시태그 추출

수집 대상은 특정 해시태그를 통해 검색 결과에 해당하는 콘텐츠들을 수집하였다. 아래 <표 2>는 콘텐츠를 수집하기 위해 사용한 해시태그와 그에 따른 수집된 건수를 보여준다.

<표 2> Instagram 콘텐츠 수집에 사용된 해시태그

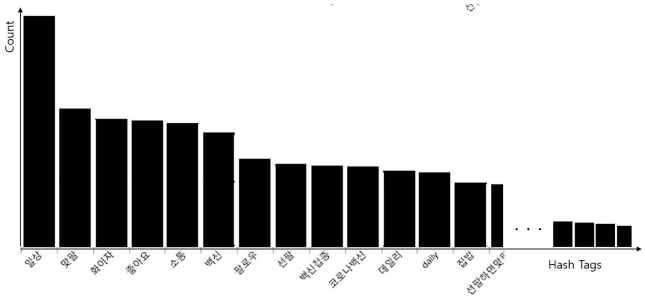
해시태그	수집 콘텐츠 수
코로나	606
코로나19	606
코로나시대	606
코로나일상	606

다양한 콘텐츠를 확보하기 위해서 최근 게시물을 기준으로 6일간 총 2,424개의 콘텐츠를 수집하였다. 콘텐츠의 본문 내용이 없거나, '광고'라는 해시태그가 사용된 콘텐츠는 제거하여 아래 <표 3>과 같은 1,448개의 데이터 세트를 생성하였다.

<표 3> 데이터 세트

ID	본문
ID_0001	#선셋 #해변 #서피비치 #코로나 #아름다운 #느낌 #영화 #데일리룩 #데일리 #소통합시다 #소통그램 #일상룩 #일상 #인생샷 #핫플 #신기 #예쁜 #하늘 #핫플레이스
ID_0002	시원해? #귀여움 #습하다 #멍스타그램 #코로나 #집콕 #가족
:	:
ID_1448	#2차백신완료 #코로나 #백신 #아스트라제네카 | 2차백신완료,백신,아스트라제네카,코로나

또한 수집에 사용된 '코로나', '코로나19', '코로나시대', '코로나일상'이라는 4개의 해시태그는 동일한 의미를 나타내는 단어로써, '코로나'라는 해시태그로 통일시켰다. 수집된 각 콘텐츠의 해시태그는 최대 38개, 평균 13개가 사용되었다. 또한 사용된 전체 해시태그는 10,260개였으며, 해시태그에 따른 등장 빈도는 아래 (그림 2)와 같은 분포를 보였다.



(그림 2) 해시태그별 빈도수 분포

해시태그는 정형화된 단어가 아닌 사용자가 원하는 텍스트로 사용할 수 있기 때문에, 등장 빈도수가 적은 것들은 자율성이 있고 반복적으로 사용되지 않는 해시태그로 판단할 수 있다. 따라서 등장 빈도수에 따른 분포의 상위 90%에 해당하는 해시태그 217개를 키워드로 정의하였다.

4. 연구 방법 및 결론

1,448개의 콘텐츠에 대해서 217개의 키워드 등장 여부에 따라 Document-Term Matrix를 생성하였다.

<표 4> Document-Term Matrix

콘텐츠	가족	먹방	불금	…	코로나
ID_0001	0	0	0	…	1
ID_0002	1	1	0	…	1
:	:	:	:	:	:
ID_1448	0	0	0	…	1

이를 통해 k-평균 군집화를 수행하고 군집별 콘텐츠 수, 실루엣 점수를 도출할 수 있었다.

k-평균 군집화 기법은 전체 데이터의 각 관측치의 최소평균거리를 계산하여 k개의 군집으로 분할하는 비지도 학습 방법의 알고리즘이다.[6] k-평균 군집화 기법의 과정은 초기 k개를 임의로 설정하여 관측 값들과의 거리를 계산하고, 계산한 결과를 바탕으로 관측치를 가장 가까운 군집으로 형성시킨다. 이 과정을 통해 k개의 군집이 형성되면, 각 군집간

의 중심점을 계산하여 새로운 군집을 형성한다. 이를 반복하면서 중심점의 위치를 이동시켜 각 군집의 데이터와 중심점의 거리가 최소가 되도록 하고, 더 이상 중심점이 바뀌지 않으면 최종 군집이 결정된다.

또한 군집화 평가 방법으로는 실루엣(Silhouette) 분석 방법이 있는데, 실루엣 계수(Coefficient)는 군집이 얼마나 잘 분류되어 있는지 평가할 수 있는 지표이다.[7] 실루엣 계수는 (-1, 1)의 값을 갖는데, 값이 1에 가까울수록 군집화가 잘 되었음을 의미한다. 군집화가 잘 되었다는 것은 다른 군집들과는 거리가 멀고, 동일한 군집내의 데이터끼리는 서로 가깝게 잘 뭉쳐있다는 것이다.

그 외에도 콘텐츠에 얼마나 많은 키워드가 등장했는지 측정하기 위하여 아래 (1)과 같이 키워드 다양성(Keyword Diversity) 값을 산출하였다.

$$\text{Keyword Diversity} = \frac{\text{콘텐츠에 등장한 키워드 수}}{\text{전체 키워드 수}} \quad (1)$$

아래 <표 5>는 임의의 군집 수 k=5 일 때, k-평균 군집화를 수행한 결과이다.

<표 5> k=5 일 때 군집화 결과

군집	콘텐츠 수	실루엣 계수(S)	키워드 다양성(K)
C1	1,396	0.6181	0.8295
C2	17	0.7500	0.1336
C3	11	0.7257	0.0461
C4	14	0.6734	0.1336
C5	10	0.7500	0.1751
계	1,448	0.6219	0.2636

군집 C_1은 1,396개의 콘텐츠가 포함되어 있고 실루엣 계수는 전체 군집의 실루엣 계수의 평균보다 작고, 키워드 다양성은 매우 큰 값을 갖는다. 이는 군집 내 콘텐츠들의 응집도는 낮고 사용된 키워드가 규칙적이지 않다는 것을 의미한다. 반면에, 군집 C_2의 실루엣 계수는 전체 군집의 실루엣 계수 평균보다 크고 키워드 다양성은 작은 결과를 보였다. 실루엣 계수는 크고, 키워드 다양성이 작다는 것은 적은 키워드를 규칙적으로 사용했다는 것이다.

이와 같이 대중에게 많이 노출되기 위한 목적의 부적절한 광고성 콘텐츠들은 관련 없는 키워드를 남발하여 사용하고, 동일한 콘텐츠를 반복적으로 업로드 하는 특징이 있다. 때문에 부적절한 광고성 콘텐츠들이 포함된 군집들은 상대적으로 실루엣 계수의 값이 크고, 키워드 다양성 값은 작게 도출 되었다. 더 많은 실험을 위해 군집 수를 5, 10, 15, 20으로 변화시키며 k-평균 군집화를 수행하였으며, 그 결과의 요약은 아래 <표 6>와 같다.

<표 6> 군집 수에 따른 군집화 결과 요약

K	평균 실루엣 계수	평균 키워드 다양성
5	0.6220	0.2636
10	0.5516	0.2198
15	0.5591	0.2375
20	0.5563	0.1956

이와 같은 군집화 결과를 바탕으로 분류된 군집의 콘텐츠 적절성은 아래 (2)의 의사코드(Pseudo Code)를 기준으로 판단하였다.

if : (2)

$S(C_i) > Average(S(C))$ *and*

$K(C_i) < Average(K(C))$

then, ***"Inappropriate Advertising Content"***

else :

"General Content"

C는 군집을 의미하고 *i*를 통해 각 군집을 구분하였다. 또한 S는 군집의 실루엣 계수, K는 군집의 키워드 다양성 값을 의미한다. 즉, 군집 *i*의 실루엣 계수 값이 전체 군집의 실루엣 계수 평균값보다 크고, 군집 *i*의 키워드 다양성 값이 전체 군집의 키워드 다양성 값의 평균보다 작으면 해당 군집*i*는 부적절한 광고성 콘텐츠들의 군집이라 판단하였다. 위의 군집에 따른 콘텐츠 적절성 평가 결과는 아래 <표 7>과 같다.

<표 7> 군집 수에 따른 군집화 분류 결과

k	일반 콘텐츠 군집 수	부적절한 콘텐츠 군집 수
5	1	4
10	3	7
15	4	11
20	6	14

또한 분류된 군집들의 평가는 서로 다른 분야에 종사하는 임의의 5명이 <표 8>과 같이 군집에 따라 등장 빈도수가 높은 키워드 상위 5개를 통해 판단하였으며, 아래 (3)에 따라 분류 결과를 점수로 환산하였다.

<표 8> k=5, 군집별 등장 빈도수 높은 키워드 상위 5개

군집	키워드
C1	코로나, 코로나19, 거리두기, 마스크, 백신
C2	Voyagers, 강철부대, 도지코인, 별자리, 에버랜드
C3	UnitedStates, coronavirus19, covid19, dinein, stayhome
C4	개웃, 개잼, 꿀잼, 유머짤, 웃긴짤
C5	fff, 나들이룩, 데일리룩, 빈티지, 원피스

$$\text{Human Score} = \frac{\text{올바르게 분류된 군집 수}}{\text{전체 군집 수}} \quad (3)$$

임의의 5명이 평가한 군집화 분류 결과 점수는 아래 <표 9>와 같다.

<표 9> 군집 수에 따른 군집화 분류 결과 점수

k	Human Score
5	1.0
10	0.9
15	0.8
20	0.8

군집의 수가 증가함에 따라 군집화 분류 결과 점수가 약간 감소하였으나, 군집화 기법을 통해 분류된 콘텐츠의 적절성은 신뢰할 수 있다고 판단된다.

하지만 각 군집의 실루엣 계수가 군집들의 평균 실루엣 계수 보다 작고, 키워드의 다양성이 군집들의 평균 키워드 다양성 보다 크더라도 그 차이가 미세하면 부적절한 광고성 콘텐츠들의 군집이라 판단할 수 있었다. 또한 군집화 결과에서 어떤 군집에 포함된 콘텐츠의 수가 상대적으로 많을수록 해당 군집에는 일반적인 콘텐츠와 부적절한 광고성 콘텐츠가 섞여 있는 것을 확인할 수 있었다.

6. 결론

본 연구에서는 SNS Instagram 콘텐츠의 해시태그 키워드를 중심으로 k-평균 군집화 기법을 적용하였다. 특히 콘텐츠들을 군집화하고 실루엣 계수와 키워드 다양성의 평가 지표를 통해 콘텐츠들의 적절성을 판단하였다.

지금까지 이메일, 문자, 메시지, 댓글 등의 광고성 스팸 분류에 대한 연구는 많이 있었지만, 이번 연구에서와 같이 군집화 기법을 적용하여 부적절한 광고성 콘텐츠를 분류하는 실험은 찾을 수 없었다.

이번 연구 결과를 통해 군집화 기법을 활용한 광고성 스팸 분류 가능성을 보여주었다. 단순히 광고를 필터링하는 것이 아니라 반사회적, 사행성, 음란성 등의 반복적으로 업로드 되는 부적절한 광고성 콘텐츠들을 실루엣 계수와 키워드 다양성을 통해 빠르게 군집으로 분류해 낼 수 있었다.

향후 연구에서는 데이터를 수집할 때, 다양하고 더 많은 해시태그를 사용하여 데이터 세트의 크기와 키워드의 수를 극대화 할 계획이다. 또한 군집 수에 따른 변화폭도 감소시켜 군집 수에 따른 세분화된 결과를 도출할 것이다.

참고문헌

[1] 김병희, 한상필, “기업 커뮤니케이션에서 소셜미디어의 활용 가능성 : 의세설정과 소셜 프레즌스를 중심으로”, 한국광고학회 광고학연구, 제22권 4호, 91-113, 2011

[2] 장정현, 나스리디노프아지즈, “SNS기반 유해사이트 판단 및 수집 시스템”, 한국정보처리학회 춘계학술대회, 2017, 812-815

[3] 김지아, 이금분, “SNS 내 효과적인 불법 스포츠 도박 광고 차단 방법”, 한국정보과학회 논문지, 제24권 제 12호(통권 제189호), 201-207, 2019

[4] 박래근, 윤혁진, 신의철, 안영진, 정승도, “광고글 필터링 모델 적용 및 성능 향상 방안”, 한국산학기술학회 논문지, 제21권 제11호, 1-8, 2020

[5] 이동석, 오하영, “LSTM을 이용한 유투브 스팸 댓글 분류 - 문장 토근화 및 인공신경망 중심으로”, 한국정보통신학회 추계종합학술대회, 2020, 241-244

[6] MacQueen, J, “Some methods for classification and analysis of multivariate observations”, Proceedings of the 5th Berkerly symposium on Mathematical Statistics and Probability, Vol. 1, pp. 281-297, 1967

[7] Reter J.Rousseeuw, “Silhouettes: A graphical aid to the interpretation and validation of cluster analysis”, Journal of Computational and Applied Mathematics, Vol. 20, pp.53-65, 1987