JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 23, No. 8, pp. 1419-1427, Aug. 2022

Check for updates

토픽 모델링과 회귀분석을 활용한 모바일 푸시 메시지의 OTT 서비스 콘텐츠 소비에 미치는 영향 분석

김 상 연[1] · 이 희 대[2] · 박 학 용[3] · 황 동 욱[4*]
[1,4*]광운대학교 미디어커뮤니케이션학부 교수
[2]경희대학교 저널리즘학과 겸임교수
[3]광운대학교 커뮤니케이션 대학원 박사과정

The Effect of Push Notifications on OTT Content Consumption

Sang-Yeon Kim[1] · Hee-Dae Lee[2] · Hak-Yong Park[3] · Dongwook Hwang[4*]
[1,4*]Professor, School of Media and Communication, Kwangwoon University, Seoul 01890, Korea
[2]Adjunct Professor, Graduate School of Media and Communication, Kyunghee University, Seoul 02447, Korea
[3]Doctoral Student, Graduate School of Communication, Kwangwoon University, Seoul 01890, Korea

[요 약]

본 연구는 모바일 앱 기반 OTT 서비스로 요리 콘텐츠 분야 한국 내 1위를 기록 중인 '만개의 레시피'가 이용자와의 상호작용 활성화를 위해 실시한 4년여간 푸시 알림글을 분석하였다. Latent Dirichlet Allocation (LDA)기반 토픽 모델링을 활용하여 모바일 앱에서 이용자에게 제공하는 푸시 알림 메시지의 잠재 토픽들을 추출하고, 엘라스틱넷 회귀분석을 활용하여 메시지에 포함된 단어들 중 콘텐츠의 조회수에 긍정/부정적 영향을 미치는 단어들을 파악하였다. 연구 결과, 푸시 메시지는 '계절음식', '판촉', '식재료 손질', '일품요리', 그리고 '간단식'의 다섯 가지 토픽으로 분류가 가능하였으며, 대체로 판촉 관련 단어를 사용하였을 때 특히 이용자 조회수가 하락하는 것으로 나타났다.

[Abstract]

This study analyzes texts used in push notifications from an OTT (Over-the-Top) service, "10,000 recipes," to explore for topic categories and identify words that most/least attract viewers. In this study, representative latent topics and words of push notification messages provided by mobile apps were extracted using Latent Dirichlet Allocation(LDA)-based topic modeling. In addition, the Elastic Net Regression was used to identify the representative words from the push notification message that either positively or negatively affect the number of views of mobile OTT service content. Five topics emerged from the analysis: 'seasonal food,' 'promotion,' 'preparing ingredients,' 'a la carte,' and 'easy meals.' Use of promotion-oriented words tended to dampen viewer attraction.

색인어 : 마케팅, OTT 서비스, 예측 모델링, 푸시 메시지, 토픽 모델링

Keyword : Marketing, OTT service, Predictive modeling, Push notifications, Topic modeling

http://dx.doi.org/10.9728/dcs.2022.23.8.1419

Received 24 May 2022; **Revised** 16 August 2022
Accepted 24 August 2022

***Corresponding Author; Dongwook Hwang**
Tel: +82-2-940-8453
E-mail: dongwookkr@kw.ac.kr

Ⅰ. 서 론

전 세계적으로 미디어 시장은 웹(WEB)과 모바일(Mobile)을 기반으로 비디오 스트리밍 서비스를 제공하는 플랫폼인 OTT (Over-The-Top)의 회광으로 콘텐츠의 유통 창구가 다양해지면서 PC·모바일 인터넷을 통한 동영상 이용이 급속하게 증대하고 있다. 이러한 OTT 서비스는 넷플릭스, 아마존 프라임비디오, 왓챠, 쿠팡플레이 등 유료 구독 전용 서비스를 비롯해 구글의 유튜브, 네이버의 네이버NOW, 카카오의 카카오TV 등 포털 사업자 또는 티빙(tving)과 시즌(Seezn), 웨이브(WAVVE) 등 방송사 또는 통신사들, 그리고 메타(META), 인스타그램, 트위터, 틱톡과 같은 소셜 미디어 기업들의 동영상 서비스까지 유선과 무선 모두를 포함하는 동영상 콘텐츠 서비스 전반을 통칭하는 광의의 개념으로 확대되었다. 현재 OTT 서비스 콘텐츠 종류로는 크게 두 가지로 요약해볼 수 있는데, 먼저 넷플릭스, 왓챠와 같은 콘텐츠 중개상, 즉 콘텐츠 애그리게이터(CA; Contents Aggregator) 플랫폼을 이용한 기성 제작 콘텐츠(RMC; Ready Made Content)와 유튜브, 틱톡과 같이 누구나 참여 가능한 SNS(Social Networking Service) 등을 기반으로 하는 1인 미디어 형태의 개방형 플랫폼을 통해 콘텐츠 이용자가 직접 콘텐츠 창작자로 생산에 나서는 사용자 제작 콘텐츠(UGC; User Generated Content)로 분류해볼 수 있다. 두 가지 OTT 서비스 콘텐츠의 종류와 현재 미디어를 소비하는 이용자의 양상을 고려해볼 때, 과거 TV 방송사 홈페이지의 프로그램 시청자 의견란에 피드백을 남기는 정도의 단방향적, 수동적이던 유형에서, 생산자와 소비자 간에 웹 또는 모바일 앱(Application)상에서 각 콘텐츠 별로 댓글 및 추천, 공유와 같은 쌍방향 의견 게재는 물론 소비자가 직접 콘텐츠를 제작해서 업로드하는 등의 보다 적극적인 참여에 이르기까지 상호작용이 활발하게 이뤄지는 양태로 전환하게 되었다.

이처럼 활성화된 OTT 서비스 환경에서 이용자들의 관심과 니즈가 반영된 주요 댓글, 좋아요 등의 추천, 시청 조회 수와 같은 상호작용 기록들을 통해 다종다양한 이용자들의 반응을 신속하게 습득하는 것이 가능하며, 이와 같은 이용자들의 온라인상에서 수집된 데이터를 분석 및 파악하고 반영함으로써 현 콘텐츠의 개선 또는 차기작의 신규 기획과 제작의 공정에 매우 중요하게 작용할 수 있다.

현재 OTT 서비스의 콘텐츠는 주로 스마트폰의 앱을 통한 시청 형태가 주를 이루고 있다. 이때 앱을 통해 사전에 신규 콘텐츠의 정보를 노출시킴으로써 시청 조회를 적극적으로 유도하는 푸시 알림(push notification) 기능은 OTT 서비스 제공자가 이용자와의 상호작용 활성화를 위해 운용할 수 있는 주요 요소 중 하나다. 이와 관련하여 스마트폰 어플리케이션의 푸시 알림을 통한 사용자의 서비스 이용행태 상호작용 효과를 조사한 기존 연구가 다수 존재한다[1]-[6]. 하지만 대부분의 기존 연구에서는 포커스 그룹 인터뷰(FGI; Focus Group Interview)나 전문가 심층 인터뷰(In-Depth Interview), 그리고 설문조사(Survey) 등 간접 조사 방법이 주로 사용되었으며, 알림 메시지의 텍스트 내용 자체를 분석한 연구는 찾아보기 어렵다. 또한, 동영상 기반의 OTT 서비스 어플리케이션을 분석 대상으로 활용한 연구 사례가 부족하며, 해당 OTT 서비스 관련 데이터를 통해 사용자의 행태를 예측한 연구는 없는 상황으로 관련한 추가 연구의 필요성을 제기하고자 한다.

본 연구에서는 텍스트 마이닝 기법을 활용하여 OTT 서비스 앱의 알림 메시지 제목과 본문, 조회 수 등을 데이터화하여 메시지의 키워드를 중심으로 한 토픽들과 실제 사용자의 조회수와의 연관성을 살펴보고, 궁극적으로 회귀분석을 통해 메시지에 사용된 단어들로 해당 알림 메시지의 조회수를 예측할 수 있는지 규명해 보고자 한다.

이를 위해 본 연구에서는 LDA(Latent Dirichlet Allocation)기반 토픽 모델링을 활용하여 거대 텍스트 데이터에서 일반화 가능한 소수의 이슈를 추출하고, 엘라스틱넷 회귀분석(Elastic Net Regression)을 통해 조회수의 상승이나 하락에 영향을 주는 단어들을 탐색한다. 이는 텍스트 데이터양의 폭발적인 증가로, 과거 연구자가 직접 텍스트를 코딩하던 내용분석(Content Analysis)이 머신러닝 기법에 기반한 분석으로 이행하는 방법론적 트렌드 변화를 반영한다고 하겠다. 두 분석 방법 모두 세밀한 데이터 분석에는 한계가 있으나 빅데이터를 설명하는 심플한 모델 개발에 효과적이다.

Ⅱ. 문헌연구

2-1 푸시알림 관련 기존 연구 동향

모바일에서 제공하는 메시지 서비스를 노출 형식에 따라 분류하면, 서비스 제공자가 이용자들을 대상으로 적극적으로 메시지를 전달하는 푸시(push) 형식과 이용자들이 스스로 능동적으로 메시지를 찾아서 살펴보는 풀(pull) 형식이 있다. 푸시 알림 메시지는 다시 옵트인(opt-in)과 옵트아웃(opt-out)으로 나누어 볼 수 있다. 옵트인 방식은 광고 메시지의 푸시를 받겠다는 의사를 이용자에게 사전 확인한 후 전송하는 방식이고, 옵트아웃 방식은 국내에서 주로 채택하고 있는 방식으로 광고 메시지를 전달받은 후 이용자가 이에 대한 거부 의사를 통보하는 방식이다. 특히, 옵트인 방식은 이용자 타깃팅을 통해 선별된 이용자나 서비스 신청자들에게 광고 메시지를 전송하는 형태를 의미하며, 이용자의 정보를 사전에 수집하여 특정 시간대 또는 특정 지역에 있는 지정된 고객에게 타깃형 방식이 가능한 장점이 있는 대신 일방적으로 제공자 측의 메시지를 전달하기 때문에 고객들의 거부감으로 인해 광고 효과가 낮아질 가능성이 있는 것이 단점이다[2].

모바일 앱 이용자들을 대상으로 사전에 새로운 콘텐츠 정보를 노출시키고 안내하는 서비스 이용을 유도하는 푸시 알림 기능과 이에 따른 이용자 반응을 분석한 선행 연구들이 다

수 존재한다. 기존의 관련 연구에 따르면 푸시 알림 메시지가 수신자들의 의사결정에 실제적으로 영향을 줄 수 있음을 밝혔다[7], [8]. 예를 들어, 스마트폰 날씨 앱을 활용해 푸시 알림 메시지의 수신 여부에 따라 실험을 통해 푸시 알림의 효과를 파악한 연구결과가 있다 [9]. 해당 연구 결과에서는 메시지를 수신한 그룹이 더 자주 앱에 접속하고 푸시 알림이 왔을 때는 이를 클릭한 사용자들이 앱에 더 빈번하게 방문했다는 결과를 도출하였다. 또한, 푸시 알림 메시지는 사람들의 행동을 유발하는 촉매 역할을 하기도 한다. 김정현[10]에 따르면, 모바일 앱의 푸시 메시지를 포함한 설득 메시지의 영향은 정보원에 대한 소비자의 신뢰도에 영향을 주는 것으로 파악되어, 신뢰도가 높을수록 태도 변화의 유발 가능성이 높은 것으로 나타났다. 또한 소비자들 사이에서 신뢰도가 높은 매체를 통해 광고 메시지를 전파할 때 소비자의 구매의도가 상승하는 것을 확인하였다.

이 밖에 스마트폰 앱의 푸시 알림에 관한 기존의 연구는 발신 횟수, 메시지의 글자 수 등에 의한 사용자 선호 의향을 분석한 이미향 외 [3] 연구와 푸시 알림 메시지의 효과적인 UX 디자인 유형화를 연구한 김소현과 권혜수 [4] 연구, 푸시 알림 메시지가 스마트폰 게임의 이용자 수와 매출에 영향을 미치는 지를 조사한 이진우 외 [5] 연구, 푸시 알림 메시지가 이용자의 모바일 구매의도에 주는 영향력을 관찰한 연구(심선영[6]) 등 다양한 시도가 있었으며, 푸시 메시지와 이용자들의 행동 변화 간 상관관계를 밝혀냈다.

2-2 OTT 서비스 관련 데이터 활용 사례 연구 동향

기존 연구들에 따르면, OTT 서비스의 사용자 이용행태 데이터를 수집하여 분석한 사례 연구들을 다수 찾아볼 수 있다. 예를 들어, 김정희와 백지원[11]의 연구에서는 다년간의 한국 미디어 패널조사의 다이어리 자료를 이용하여 방송서비스 이용 시간 증가 대비 OTT 서비스 이용 시간이 얼마나 증가하였는지를 나타내는 탄력성을 측정하고 추정하는 연구를 진행하였다. 또한, 박연진과 신현문[12]도 한국 미디어 패널 자료를 활용, OTT 서비스 이용률과 빈도, 회당 이용 시간 등 다양한 데이터를 분석하여 국내 소비자의 이용 행태를 탐색하였다. 김주현[13]은 닐슨 코리안클릭이 수집, 공개하는 패널 데이터를 이용, 모바일 OTT 동영상 서비스 이용 시간과 TV 시청시간 사이의 관계를 분석하였다. 같은 맥락으로 최민재[14]의 연구에서는 스마트폰 이용자 1,000여 명을 대상으로 스마트폰을 통한 방송서비스 이용행태에 따른 TV를 통한 방송매체 시청 시간의 변화에 대하여 분석하였다. 이처럼 OTT 서비스 이용행태와 관련한 데이터를 바탕으로 다수의 선행 연구들이 진행되었지만, 주로 OTT 서비스 이용행태에 따라 방송매체에 어떠한 영향을 미치는지를 보고자 하는 OTT와 방송매체의 경쟁 관계를 다루는 연구들이 주를 이루었다. 이외에 임진솔 외[15]의 연구에서는 소셜 미디어인 트위터의 텍스트 데이터를 분석하여 서비스 제공자 관점에서 어떠한 이유로 OTT 플랫폼을 구독하고 콘텐츠를 이용하는지 요인들을 파악한 연구도 있지만, 특정 텍스트에 대한 상호작용 효과 등을 파악하기 어렵고, 모바일 푸시알림 메시지와 같은 데이터 분석은 고려되지 않고 있다.

2-3 토픽 모델링 기법과 관련한 선행 연구

토픽 모델링(Topic Modeling) 분석은 비정형화된 텍스트 데이터로부터 의미 있는 정보를 추출하는 텍스트 마이닝과 자동화된 방식을 통해 잠재적 확률모델에 의해 공통된 주제를 추출하는 방법으로 알려져 있다. 현재 다양한 알고리즘이 존재하지만, Blei을 필두로한 연구진(2003)이 제시한 '잠재 디리클레 할당'(LDA; Latent Dirichlet Allocation)이 일반적으로 가장 널리 활용되고 있다 [16]. LDA는 문서와 단어 간 상관관계를 기반으로 다수의 문서들에 잠재하는 토픽을 추출해내는 확률분포 모형이다. 디리클레 분포(Dirichlet distribution)를 활용하는 LDA는 문서의 말뭉치(corpus), 문서 내의 단어 수, 문서의 양으로 계산되는 잠재된 파라미터 값을 통해 각 주제별 핵심 단어들을 골라낸다. LDA의 전제조건으로는 모든 문서들이 토픽들의 혼합으로 구성되어 있고, 토픽들은 확률분포를 기반으로 하여 단어들을 생성하여 구성하는 것이다. LDA 알고리즘의 메카니즘은 그림 1과 같이 도식화할 수 있다.

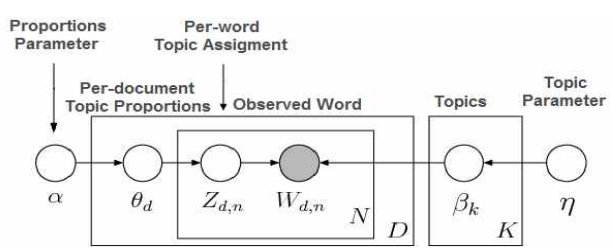


그림 1. LDA 시각적 모델 예시
Fig. 1. Visual representation of LDA[17]

K: 전체 토픽의 개수
D: 말뭉치 전체 문서의 개수
N: 문서에 속한 단어의 개수
α: 문서별 토픽 분포의 하이퍼파라미터
η: 토픽별 단어 분포의 하이퍼파라미터
β_k: k번째 토픽에 해당하는 벡터
Θ_d: d번째 문서에 포함된 토픽의 비중
$Z_{d,n}$: d번째 문서내 n번째 단어가 속하는 토픽
$W_{d,n}$: 실제 관측 가능한 d번째 문서의 n번째 단어

LDA를 활용한 토픽 모델링은 다양한 분야에서 연구되어 오고 있다. 이중 특히 소셜 커머스 및 온라인 플랫폼을 통해 수집된 텍스트 데이터를 수집하여 트렌드를 분석하는 연구가 다수 진행되고 있다. 배정환과 그의 동료들의 연구[18]에서는 트위터에서 생성된 메시지 데이터를 대상으로 주요 이슈들을

추출하고 이를 웹에 시각화시키는 '트위터 이슈 트래킹 시스템'을 구축하였다. 채승훈 외의 연구[19]에서는 모바일 소셜커머스와 오픈 마켓의 이용자 내용을 집계하고 이를 토픽 모델링을 활용하여 토픽을 유용성 및 편리성으로 분류하여 구분된 토픽에서 감성분석을 실시하였다. 최정균 외의 연구[20]에서는 네이버의 뉴스 포털 서비스에서 생성된 메시지 데이터를 분석하여 각 언론사의 보도 행태를 탐사하였다. 정원준[21]은 연구를 통해 한국 내 사드 배치 이후 한·중간에 생성된 대립과 갈등의 상황에서 갈등 쟁점을 도출해 갈등 주기 모형에 맞춰 분석하였다. 강창완 외의 연구[22]는 한국 자료분석 학회지 논문들을 대상으로 한 토픽모델링을 수행하여 해당 학회지의 연구 트렌드 및 학회 본연의 목적을 잘 수행하고 있는지 분석하였다. 이외에도 이영준[23]은 금융통화위원회 의사록 전체에 대한 토픽모델링과 감성분석을 통해 미래의 금리 의사결정에 대한 추가적인 정보를 미리 예측 및 검증하고자 하였다.

종합해보면, 온·오프라인의 다종다양한 메시지 분석을 통해 실시간 여론 트렌드 파악 등 재빠른 이슈 추출을 위해 토픽 모델링이 활용되고 있음을 알 수 있다. 하지만 현재까지 모바일 앱을 기반으로 한 OTT 서비스의 푸시 알림 메시지를 대상으로 한 토픽 모델링 기법을 활용한 사례는 미비한 실정이다.

2-4 엘라스틱넷 회귀분석(Elastic Net Regression)

엘라스틱넷 회귀분석(Elastic Net Regression)는 정규화 선형회귀분석 방법의 일종으로 선형회귀 계수(weight)의 절대값의 합과 제곱합에 대한 제약조건을 추가하여 예측모형이 과도하게 최적화되는 현상인 과적합(overfitting) 현상을 막기 위해 제한된 방법이다. 엘라스틱넷 회귀모델은 조율 모수인 λ와 규제항(L1, L2)의 비중을 조절하는 α를 하이퍼파라미터로 갖고, 특히 α값을 0과 1 사이로 조절하여 L1(lasso)과 L2(ridge)를 절충하는 정규화(regularizaion) 모델을 탐색한다.

본 연구에서는 메시지 내의 텍스트가 콘텐츠의 조회 수에 미치는 영향을 예측할 수 있는지를 탐구하고자 엘라스틱넷 회귀분석을 통해 분석하고자 한다. 이를 통해 궁극적으로 각 단어들의 가중치(weight)들을 축소 또는 무력화하여, 설명력을 크게 희생하지 않으면서도 최대한 작은 수의 단어들로 조회수에 미치는 영향을 파악하고자 한다.

Ⅲ. 연구방법

본 연구에서는 '만개의 레시피'라는 요리 콘텐츠 분야 국내 1위 OTT 서비스를 제공하는 모바일 어플리케이션을 바탕으로 해당 OTT 서비스의 데이터베이스에 등록된 푸시 알림 메시지를 분석하여 메시지의 핵심 토픽(키워드)를 추출하고 해당 토픽과 관련하여 실제 OTT 서비스 이용자의 조회수와 어떤 관계를 형성하는지 보고자 엘라스틱넷 회귀모델을 활용하여 예측하고자 한다. 전반적인 연구방법의 과정은 그림 2와 같다.

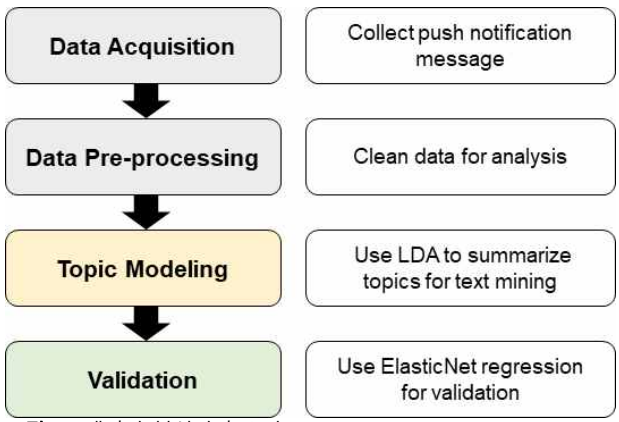


그림 2. 데이터 분석방법 요약
Fig. 2. The overview of data analysis

3-1 데이터 수집 및 키워드 추출

2016년 2월 17일부터 2020년 6월 22일까지 '만개의 레시피'의 데이터베이스에 등록된 푸시 알림 메시지 총 3,354개를 분석하였다. 해당 데이터베이스는 푸시 알림 메시지의 전송일시, 제목, 본문, 조회 수 등의 정보를 담고 있다. 본 연구에서는 텍스트 마이닝 분석을 위해 R 3.6.2를 활용하였고, 비정형 텍스트 데이터를 정형화된 데이터 구조로 변환하기 위한 전처리 작업을 진행하였다. 한글 자연어 분석 패키지(Korean Natural Language Processing; KoNLP) 0.80.1에 탑재된 형태소 사전(NIADic)을 사용하였다. 본 연구에서는 사전에 등록되어 있는 총 983,012개의 단어와 매칭되는 명사들을 추출한 뒤 불용어들을 제거하였고, 불용어 리스트에 포함된 구두점, 기호, 각종 이모티콘, 날자, 숫자와 데이터 내에서 자주 등장하나 분석 상 큰 의미가 없는 것으로 판단되는 '만개의 레시피', '레시피', '광고', '음식' 등을 제거하였다.

앞서 전처리 된 텍스트 데이터를 메타 데이터를 포함한 말뭉치로 변환하고, 이를 다시 단어문서행렬(Term-Document Matrix, TDM)으로 변환 처리하였다. 대부분의 푸시 메시지에 걸쳐 자주 등장하는 의미 없는 단어들의 영향력을 제한하기 위해 각 단어의 출현 빈도는 TF-IDF(term frequency-inverse document frequency)로 계산하였다. 최종적으로 2,598 단어 × 3,325 푸시 메시지의 희소행렬(sparse matrix)을 생성하였고, 이를 바탕으로 분석을 진행하였다.

3-2 토픽 모델링 분석

데이터 전체에 나타난 단어 2,600여 개를 소수의 잠재 토픽(latent topic)으로 축소가 가능한지 탐색하고자 qdap 패키지에 탑재된 LDA 분석을 수행하였다.

토픽 모델링 분석에는 두 개의 하이퍼파라미터가 활용되는

데, 그림 1에서 제시된 알파(α; topic distribution-per-document prior)와 에타(η; term distribution-per-topic prior)를 활용한다. 이때, 잠재 토픽의 수(k)를 임의로 조정해가며 토픽 모델링 분석을 반복 수행하고, 각각의 토픽에 포함되는 단어들의 구성을 나타내는 결과값들을 바탕으로 실제 단어들 간의 관계와 컨텍스트를 파악한다. 본 연구에서는 토픽의 수, k를 설정하기 위한 방법으로는 연구자의 주관적 판단과 임의 설정을 최소화하기 위하여 엘보우법(elbow method [24,25,26,27])을 활용하였다.

3-3 엘라스틱넷 회귀분석

푸시 알림 메시지에 포함된 단어들 중 조회수를 예측하는데 가장 많은 기여를 하게 되는 단어를 예측하기 위하여 데이터 전처리 후 선형회귀 분석을 수행하였다. 먼저, 전처리가 완료된 텍스트 데이터 3,325개의 푸시 알림 메시지 중 80% (n = 2,660)를 학습데이터 셋(training set)으로, 나머지 20%(n = 665)를 평가 데이터(test set)로 나누었다. 단순 랜덤이 아닌, 두 샘플에 나타난 조회 수의 분포가 서로 일치되도록하는 방식을 선택하였다.

희소행렬의 크기를 축소하여 조회 수 예측에 실질적인 정보를 제공할 수 있는 단어들만 데이터에 남도록 하였다. 이때, 전체 데이터 내 출현 빈도가 10%를 넘지 않는 단어들을 희소단어로 정의하였고, 이와 반대로 50%를 초과하여 빈번하게 쓰이는 단어들은 정보로서의 활용 가치가 낮다고 판단하여 데이터셋에서 삭제하였다. 결과적으로, 2,661(푸시 알림 메시지) × 307(단어)의 조회수 예측에 유효한 정보를 지닌 행렬을 얻게 되었다.

이후 데이터에 남아있는 단어들로 푸시 알림 메시지의 조회수를 예측하는 선형회귀 모델을 구축하였고, 특히 모델의 과적합을 방지하기 위하여 엘라스틱넷 회귀분석을 사용하였다. 교차검증(cross-validation)을 위해 학습데이터 셋을 다시 다섯(k-fold = 5)으로 나누어, 임의의 네 개 학습데이터 셋에서 생성된 모델을 나머지 하나의 학습데이터 셋에 적용, 조회 수 예측 정확도(R^2)를 확인하였다. 이 과정을 반복하여 획득한 결과값들의 평균으로 모델의 전반적인 정확도를 가늠하였다. 끝으로, 학습데이터 셋에서 얻은 최종 모델로 평가 데이터 셋을 예측하였다.

Ⅳ. 연구 결과

4-1 토픽 모델링 분석 결과

본 연구 결과, LDA 토픽 모델링에서 두 가지 하이퍼파라미터(α & η)를 각각 .05이고, 토픽의 수를 5와 7 사이일 때 최적의 모델을 도출하였다. 이때, 토픽의 수를 설정하기 위해 앞서 언급한 4가지의 엘보우법을 활용하였다. 그 결과, 그림 3과 같이, 토픽이 5개일 때 'Juan[24]'의 평균 토픽간 거리(average topic distance) 인덱스가 리턴감소 시작점이었으며, 토픽이 7개일 때 'Griffiths[25]'의 log$P$(w|T), 즉 주어진 토픽 수에 따른 데이터의 최우도가 더 이상 큰 폭 상승하지 않는 것으로 나타났다. 나머지 두 개의 인덱스('Arun[26]'과 'Deveaud[27]')는 눈에 띄는 엘보우를 형성하지 않아 토픽 수를 결정하는데 활용되지 않았다. 결과적으로, 토픽의 수를 5, 6, 7로 늘려가며 분석을 반복 시행한 결과, 토픽의 수가 5일 때, 해석이 가장 용이한 모델을 발견하였다.

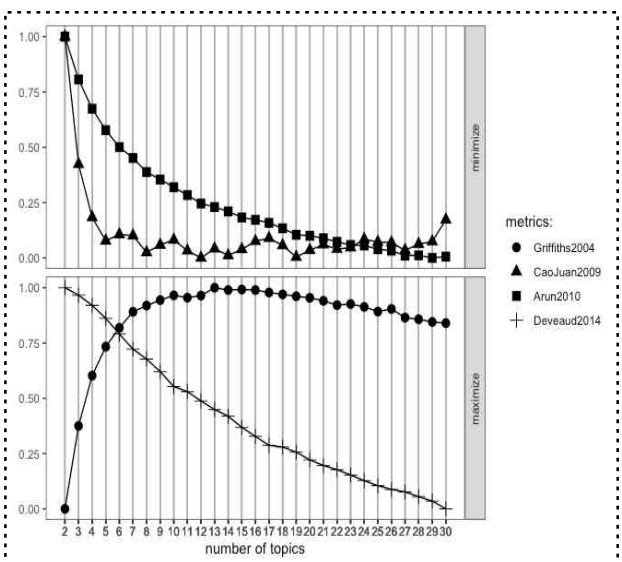


그림 3. 엘보우법(elbow methods)을 활용한 토픽 수 결정
Fig. 3. Results of different elbow methods for identifying the appropriate number of topics

다섯 개 각각의 토픽을 가장 잘 대표하는 최상위 단어 10개를 열거한 뒤 주관적 해석을 더해 추출하였다. 첫 번째 토픽은 '계절음식'으로 이를 대표하는 최상위 단어는 '제철,' '시원,' '여름,' '사과,' '겨울' 등과 같은 계절과 제철에 관련한 단어들과 '달콤,' '쫄깃,' '치즈,' '고소,' '효능' 등 특정 음식에 대한 형용사와 연관 단어들이 포함되었다. 두 번째 토픽은 '판촉'으로 '특가,' '가격,' '할인,' '상품,' '주방,' '세트' 등과 같이 판촉물과 관련된 단어들의 빈도가 높았으며, '행복,' '우리집,' '우리,' '디자인' 등의 연관 단어들이 도출되었다. 세 번째 토픽은 '식재료 손질법'으로 최상위 단어로는 '보관,' '손질,' '방법,' '요리,' '껍질,' '꿀팁', '하면,' '공모전,' '사용,' '해보다' 등의 연관 단어들이 도출되었다. 네 번째 토픽은 '일품요리 팁'으로 '요리,' '메뉴,' '주말' 등의 단어들과 '국물,' '날씨,' '간단,' '뜨끈,' '오늘,' '호로록,' '그릇' 등 국물 요리와 관련된 단어들이 다수 나타났다. 마지막으로 다섯째 토픽으로는 '간단식'으로 최상위 단어로는 '반찬,' '간단,' '에어,' '프라이,' '밥도둑,' '요리,' '다이어트,' '밑반찬,' '아침,' '술안주' 등 간단하게 요리 가능한 메뉴 또는 그와 관련된 요리 기구 및 간

단식 요리의 주된 목적 등이 도출되었다.

각각의 토픽들을 대표하는 최상위 두 개의 푸시 알림 메시지들은 우리의 해석과 의미상 일맥상통함을 알 수 있다. 먼저, '계절음식'의 토픽인 경우 "더울 때 생각나는 디저트! 시원해서 소름 돋아! 후르츠 칵테일이 가득! 옛날 팥빙수"와 "추울 때 생각나는 화끈한 맛! 쭈꾸미 먹고 볶음밥은 필수인거 아시죠? 맥주 안주로도 최고!"과 같은 푸시 알림 메시지를 추출할 수 있었고, '판촉'의 토픽인 경우는 "주방이 깔끔해지는 마법 같은 비법! 정리 안되는 주방공간! 이거 하나면 깔끔하게 해결가능!"과 "이 가격 이 구성 최고다! 소문날만해! 거기다 무료 배송까지! 소문 날대로 난 곱창, 막창, 대창세트! 저렴하게 즐겨보세요"와 같은 푸시 알림 메시지를 추출하였다. '식재료 손질법'과 같은 토픽에서는 "수박껍질 여태껏 버렸다규? 수박껍질 활용 BEST5. 수박껍질의 대변신! 수박껍질 요리는 여기~!"와 "양파 넌 어디까지 가능해? 양파요리 BEST5. 양파로 이렇게 다양한 요리가! 양파요리 TOP5" 등의 대표 푸시 알림 메시지를, '일품요리 팁'의 토픽으로는 "당신의 요리를 간편하게! 꽁치부대찌게! 요정들도 반한 간편 캠핑요리, 꽁치부대찌게!"와 "밀가루만 있다면 OK! 수제비 OK! 뜨끈한 국물이 필요한데 면이 없다구? 칼국수 & 수제비!" 푸시 알림을 추출할 수 있었다. 마지막으로, '간단식' 토픽으로는 "에어프라이로 구운 생선과 어울리는 오늘의 식단. 오늘 저녁 반찬 걱정할 필요없는 맞춤 식단 레시피!"와 "초 신박템! 이거 하나만 있으면 요리가 뚝딱! 너무 간편! 후라이펜, 에어프라이어, 오븐 어디에나 칙칙! 요리가 이렇게 쉽다니..."가 대표성이 가장 높은 두 개의 푸시 알림 메시지로 추출되었다.

4-2 엘라스틱넷 회귀분석 결과

그림 4에서 확인할 수 있듯이, 데이터에 남아있는 총 307개의 단어 중 정규화 과정을 통과한 119개의 단어로만 구성된 모델의 경우 과적합을 피하면서도 조회수를 예측함에 있어 오류(MSE)가 가장 작은 것으로 나타났다. 본 모델의 설명력은 학습데이터 셋(n = 2,660)에 적용했을 때 R^2 = .22, 평가 데이터 셋(n = 665)에 대입했을 때 R^2 = .06으로, 사회과학적 기준으로 볼 때 각각 중효과(moderate effect)와 최소효과(recommended minimum effect)에 해당된다[23].

푸시 알림 메시지 조회수 상승에 가장 크게 기여하는 단어들 중 '인기', '밑반찬', '성공', '만점', '양념장', '국민', '일품', '별미', '밥상', '백종원', '혼밥' 등이 눈에 띈다. 조회수 상승에 기여한 단어들은 주로 '인기', '백종원'과 같이 인기 레시피와 관련된 토픽들, '밑반찬', '양념장', '혼밥'과 같은 일상에서 자주 접하게 되는 음식들 또는 '일품', '별미' 등과 같이 계절에 따른 변화에 민감한 요리들로 요약해볼 수 있다.

반대로 조회수 하락에 영향을 주는 단어들은 '모음', '상품', '사은품', '특가', '구성', '최저가' 등 주로 판촉행사 홍보와 관련되어 있음을 짐작해 볼 수 있다(그림 5 참조).

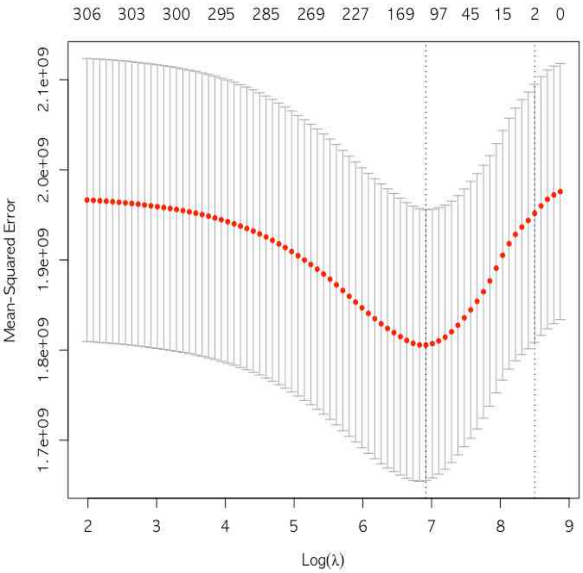
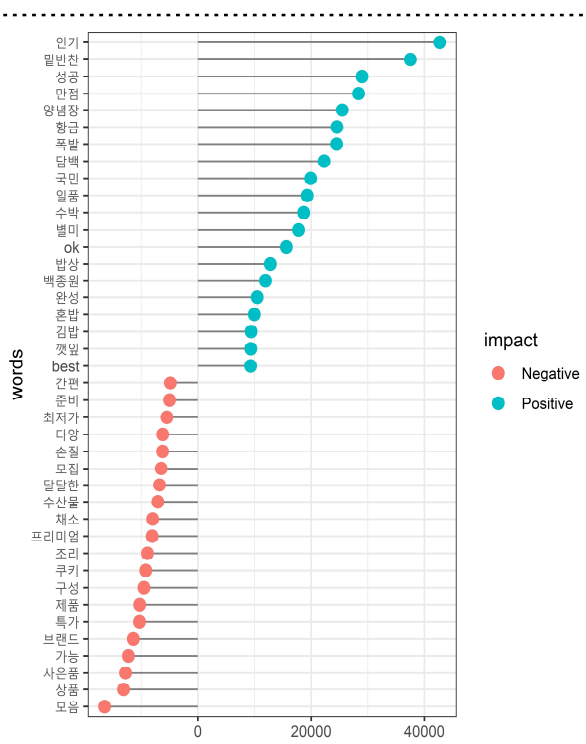


그림 4. 엘라스틱넷 회귀분석을 활용한 모델 정규화 결과
Fig. 4. Results of Elastic Net regression



* The image above summarizes results from analyzing the original Korean text data to fully deliver the nuance of each word. Interested readers can obtain the results translated in English from the corresponding author.

그림 5. 푸시 메시지 조회수에 긍/부정적 영향을 주는 주요 단어들
Fig. 5. Top 20 words with positive/negative impact for the number of views

V. 결 론

본 연구에서는 OTT 서비스에서 이용자들을 대상으로 전파한 알림 메시지의 패턴을 토픽 모델링 기법을 활용해 분석하고, 메시지에 포함된 단어들을 기반으로 시청 조회수를 예측할 수 있는지 관찰하였다. 알림 메시지는 크게 '계절음식', '판촉', '식재료 손질', '일품요리', 그리고 '간단식' 다섯 개의 잠재 토픽으로 분류가 가능했으며, 대체로 판촉과 연관된 단어들이 포함되었을 때 이용자 조회수가 떨어질 수 있음을 관찰하였다.

본 연구는 대형 글로벌 OTT 서비스들이 한국 시장으로 대거 진출하고 있는 상황에서 국내 OTT 서비스의 대응 콘텐츠 개발을 위한 탐사적 연구로서 의미가 있다고 하겠다. 특히 기존 연구에서 자주 시도되지 않았던 텍스트 마이닝 기법을 적용하여 알림 메시지들의 유형화가 가능하고 이용자들의 (비)선호 토픽이 실재하고 있음을 확인할 수 있었다.

본 연구는 OTT 마케팅 실무자들로 하여 서비스를 통해 생성되는 데이터 분석을 통해 현재의 마케팅 트랜드를 파악하고, 이를 신규 콘텐츠 기획에 활용할 수 있음을 시사한다. 인기있는 콘텐츠 제작을 위해서는 어떠한 토픽들에 시청자들이 반응하는지에 대한 철저한 분석이 필요할 것이다. 물론 데이터 분석 결과를 차기 콘텐츠의 기획 혹은 제작 방식 결정에 구체적으로 어떻게 적용할지에 대한 실무자들의 고민이 뒷받침되어야 할 것이다.

본 연구는 다음과 같은 한계를 가진다. 첫째, 본 연구는 서비스 제공자가 생산한 텍스트 데이터 분석에 초점을 두고 있으며 이용자에 대한 이론적 접근과 논의는 포함하지 않고 있다. 후속 연구에서는 소비자 중심 데이터(예: 댓글) 분석 및 소비자 심리에 기반한 가설수립, 결과 해석 등을 통해 균형을 맞출 필요가 있다. 둘째, OTT 서비스 제공자, 즉 메시지 발송자는 다양한 목적을 갖고 메시지를 기획하고 발송한다. 이러한 목적들이 토픽모델링 결과와 어떻게 연관되는지 현장의 실무자 인터뷰 등을 통해 추가적으로 보완된다면 보다 실질적인 함의를 도출할 수 있을 것으로 판단된다. 셋째, 본 연구는 요리 콘텐츠라는 특정 영역에 전문화된 OTT 서비스를 대상으로 진행하였으므로, 이후 다른 콘텐츠 장르의 서비스까지 분석 대상을 넓혀 현재 결과의 일반화 가능성을 테스트할 필요가 있다. 마지막으로, 본 연구에서 조회 수 하락을 예측하는 것으로 나타난 판촉 관련 단어들이 다른 영역의 광고 메시지에서도 역효과를 내는지에 대한 추가 검증이 요구된다. 현재의 결과가 계속된다면 '광고 언어를 사용하지 않는 광고' 메시지 작성이 가능한지, 그렇다면 실무적으로 어떤 접근 방식이 효과적인지에 등에 대한 실질적 논의가 뒤따라야 할 것이다.

참고문헌

[1] J. Lee and H. Kim, "A study of the mobile advertising effect according to Diverse Phrases," *The Korean Society of Food Preservation*, Vol. 5, pp. 257-275, 2004.

[2] J. Lee, S. Lee, and C. Lee, "A Study on the advertising effect of push-type mobile advertising," *Korean Academic Society of Business Administration*, pp. 1-22, 2009.

[3] M. Lee, D. Kim, and Y. Lim, "Research on the properties that affect the users' reaction to the smartphone-based push services," *The Korean Journal of Art and Media*, Vol. 12, No. 1, pp. 87-95, 2013, https://doi.org/10.36726/cammp.2013.12.1.87

[4] S. Kim and H. Kwon, "Study on formalization of push notification UX design," in *Proceedings of the Human Computer Interaction Korea Conference*, pp. 323-330, 2014.

[5] J. Lee, H, Lim and J. Jung, "Study on results of push notification for smartphone game: Focusing on the number of users and revenue," *Entrue Journal of Information Technology*, Vol. 14, No. 3, pp. 101-113, 2015, G704-001673.2015.14.3.002

[6] S. Shim and Y. Kim, "An empirical study on the effect of smartphone push notification and SNS information on the mobile purchasing," *Journal of Information Technology Applications & Management*, Vol. 22, No. 4, pp. 105-126, 2015, https://doi.org/10.21219/jitam.2015.22.4.105

[7] J. E. Fischer, N. Yee, V. Bellotti, N. Good, S. Benford, and C. Greenhalgh, "Effects of content and time of delivery on receptivity to mobile interruptions," in Proceedings of the 12th international conference on Human computer interaction with mobile devices and services, pp. 103-112, 2010, https://dl.acm.org/doi/10.1145/1851600.1851620

[8] S. A. Grandhi and Q. Jones, "Conceptualizing interpersonal interruption management. A theoretical framework and research program," in *Proceedings of the 42nd Hawaii International Conference on System Sciences*, pp. 1-10, 2009, https://dl.acm.org/doi/10.1109/HICSS.2009.124

[9] M. Kim, "The effect of push notification alerts on mobile application usage habit," *Korean Society for Journalism and Communication Studies*, Vol. 59, No. 5, pp. 358-387, 2015.

[10] J. Kim, "The antecedents of influence on the attitude toward mobile advertising," *Advertising Research*, Vol. 75, pp. 35-59, 2007.

[11] J. Kim, and J. Baek, "An Empirical Analysis of the Effects of OTT Services on Changes in the Media Use Pattern" *International Telecommunications Policy Review*, Vol. 26, No. 1, pp. 47-79, 2019, https://doi.org/10.7236/IJIBC.2021.13.4.55.

[12] Y. Park, & H. Shin, (2021). "An analysis of OTT users' behaviors based on Korean Media Panel Data" in

Proceedings of the Korean Institute of Communication Sciences Conference, pp. 68-69, 2021.

[13] J. Kim, "A Study on the Displacement of Mobile OTT Video Services on Home TV," *The Journal of the Korea Contents Association*, Vol. 18, No. 8, pp. 434-445, 2018, https://doi.org/10.5392/JKCA.2018.18.08.434

[14] M. Choi, "A Study on the Displace Effect of Smartphone Broadcasting and Video Service for Watching TV," *Korean Journal of Broadcasting and Telecommunication Studies*, Vol. 27, No. 3, pp. 172-205, 2013, G704-000045.2013.27.3.001.

[15] J. Lim, H. So, and H. Oh, "OTT Platform User Interest Analysis through Buzz Analysis: Based on Twitter Data," *Journal of Digital Contents Society,* Vol. 23, No. 5, pp. 837-845, 2022, https://doi.org/10.9728/dcs.2022.23.5.837.

[16] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research*, Vol. 3, pp. 993-1022, 2003, https://dl.acm.org/doi/10.5555/944919.944937

[17] D. M. Blei, "Probabilistic topic models", *Communications of the ACM*, Vol. 55, No. 4, pp. 77-84, 2012, https://dl.acm.org/doi/10.1145/2133806.2133826

[18] J. H. Bae, N. G. Han, and M. Song, "Twitter issue tracking system by topic modeling techniques," *Journal of Intelligence and Information Systems*, Vol. 20, No. 2, pp. 109-122, 2014, https://doi.org/10.13088/jiis.2014.20.2.109

[19] S. H. Chae, J. I. Lim, and J. Kang, "A comparative analysis of social commerce and open market using user reviews in Korean mobile commerce," *Journal of Intelligence and Information Systems*, Vol. 21, No. 4, pp. 53-77, 2015, https://doi.org/10.13088/jiis.2015.21.4.053

[20] J. G. Choi, S. H. Jin, and J. H. Choi, "A study on differences of aspect of report by news media using text mining analysis" *Journal of the Korean Data Analysis Society*, Vol. 19, No. 5, pp. 2509-2522, 2017, https://doi.org/10.37727/jkdas.2017.19.5.2509

[21] W. J. Chung, "Keyword and topic analysis on the THAAD conflict between south Korea and China : based on a time-series topic modeling and a semantic network analysis," *The Korean Journal of Advertising and Public Relations*, Vol. 20, No. 3, pp. 143-196, 2018, https://doi.org/10.16914/kjapr.2018.20.3.143

[22] C. W. Kang, K. K. Kim, and S. B. Choi, "A topic analysis of abstracts in Journal of the Korean Data Analysis Society," *Journal of the Korean Data Analysis Society*, Vol. 20, No. 6, pp. 2907-2915, 2018.

[23] Y. J. Lee, and B. H. Yoon, "Analyzing the minutes of the monetary policy board through topic modeling and sentiment analysis," *Journal of the Korean Data Analysis Society*, Vol. 21, No. 2, pp. 889-900, 2019, https://doi.org/10.37727/jkdas.2019.21.2.889

[24] C. Juan, X. Tian, L. Jintao, Z. Yongdong, and T. Sheng, "A density-based method for adaptive lda model selection," in *Neurocomputing 16th European Symposium on Artificial Neural Networks 2008*, Vol. 72, No. 7-9, pp. 1775−1781, 2009, http://doi.org/10.1016/j.neucom.2008.06.011

[25] T. L. Griffiths and M. Steyvers, "Finding scientific topics," in *Proceedings of the National Academy of Sciences,* Vol. 101, pp. 5228−5235, 2004, https://doi.org/10.1073/pnas.0307752101

[26] R. Arun, V. Suresh, C. E. V. Madhavan, and M. N. N. Murthy, "On finding the natural number of topics with latent dirichlet allocation: Some observations," In *Advances in Knowledge Discovery and Data Mining*, Berlin, Heidelberg: Springer, pp. 391−402, 2010, https://doi.org/10.1007/978-3-642-13657-3_43

[27] R. Deveaud, É. S. Juan, and P. Bellot, "Accurate and effective latent concept modeling for ad hoc information retrieval," *Document Numérique*, Vol. 17, No. 1, pp. 61−84, 2014, https://doi.org/10.3166/dn.17.1.61-84

[28] C. Ferguson, "An effect size primer: A guide for clinicians and researchers," *Professional Psychology: Research and Practice*, Vol. 40, No. 5, pp. 532-538, https://doi:10.1037/a0015808

김상연(Sang-Yeon Kim)

2005년 : 위스콘신-밀워키 대학교 (석사)
2009년 : 미시건주립대학교 (박사)

2009년~2015년: 조교수, 위스콘신-밀워키 대학교
2016년~2020년: 부교수, 위스콘신-밀워키 대학교
2020년~2021년: 부교수, 광운대학교
2022년~현 재: 교수, 광운대학교
※관심분야 : 데이터 사이언스, 스토리텔링, 문화비교, 전략 커뮤니케이션

이희대(Hee-Dae Lee)

2013년 : 연세대학교 언론홍보대학원 (석사)
2020년 : 광운대학교 (박사)

1997년~2003년: TV프로듀서, 삼성전자㈜ 영상사업단
2003년~2019년: 편성국장, 한국DMB㈜ 지상파DMB
2020년~현 재: OTT 미디어본부장, ㈜옴니더이앤엠
2017년~현 재: 겸임교수, 광운대
2019년~현 재: 겸임교수, 명지대
2021년~현 재: 겸임교수, 경희대 대학원
※관심분야 : OTT 미디어, 1인 미디어, 인공지능, 문화예술, 스토리텔링, 방송영상

박학용(Hak-Yong Park)

1988년 : 한국외국어대학교 대학원 (석사)
2019년 : 광운대학교 (박사수료)

2008년~2012년: 편집국장, 문화일보
2012년~2018년: 논설위원, 문화일보
2018년~현 재: 대표이사, ㈜디지털타임스
※관심분야 : 디지털 저널리즘, 미디어 커뮤니케이션

황동욱(Dongwook Hwang)

2011년 : 한국과학기술원 학사
2019년 : 서울대학교 대학원 (공학박사 -산업공학과)

2021년~현 재: 광운대학교 미디어커뮤니케이션학부 조교수
※관심분야 : 인간공학, HCI, VR/AR, 3D프린팅, UI/UX