An Analysis on the Relationship Between Content Topics of Social Media and Customer Engagement Using Machine Learning Methodology
소셜미디어 콘텐츠 주제와 고객 인게이지먼트 간의 관계분석: 머신러닝 방법론을 중심으로

Jungwon Lee(First Author)
Ph.D. Candidate, Dept. of Corporate Management, Korea University
(*d2ljw510@naver.com*)
Cheol Park(Corresponding Author)
Professor of Global Business at Korea University Sejong
(*cpark@korea.ac.kr*)

Customer engagement is regarded as a performance indicator of social media marketing, and previous studies have reported that the characteristics of content to increase customer engagement. However, the topic of content has not been sufficiently studied. This study analyzes the relationship between the topic of social media content and customer engagement and suggests an analysis procedure that can apply a machine learning model, a key tool for recent digital transformation. For empirical analysis, 154,705 social media data of 51 global brands were collected, and content topics were classified using a topic modeling method. And the relationship between content topic and customer engagement was analyzed using zero-inflated negative binomial regression analysis and machine learning model. As a result of the analysis, contents of 51 brands were classified into 18 contents topics, and there was a difference in the impact on customer engagement according to the topic. In addition, using a machine learning model, it was possible to predict the customer engagement performance of the content with an accuracy of about 90%. This study contributed to the marketing literature by analyzing the relationship between social media content topics and customer engagement through machine learning methodology.

Key Words: Social media, Customer engagement, Contents marketing, Machine learning, Topic modeling, Zero-inflated negative binomial regression

Submission Date: 11. 18. 2020 Revised Date: (1st: 12. 16. 2020) Accepted Date: 12. 24. 2020

Ⅰ. 서 론

최근 코로나 19 범유행은 경제 전반에 부정적인 영향을 미칠 뿐만 아니라, 기업의 디지털화도 가속화 하고 있다(Seetharaman, 2020). 특히, 코로나 19로 인해 소비자는 오프라인 활동을 온라인 활동으로 대체하고 있다. 이러한 변화는 그동안 진행되어오던 디지털 적자생존 흐름을 가속화 했을 뿐만 아니라, 앞으로 디지털로의 전환이 필수적임을 보여주고 있다. 디지털화의 중요한 요소 중 하나는 소셜미디어이다. 최근 코로나 19 대응방안을 논의하기 위해 이루어진 딜로이트의 웨비나[1]에 의하면 소셜미디어는 코로나 19 이전인 2019년 1분기에 비해 2020년에는 6%에서 8%로 이용률이 증가했을 뿐만 아니라 소비자 구매 여정의 14%를 차지하는 등, 코로나 19 이후 더욱 중요해지고 있다. 또한, 소셜리스닝(Social listening)은 머신러닝 등 새로운 디지털 도구가 활용되면서 소셜미디어 환경에서 새로운 경쟁전략으로 대두되고 있다(e.g., Vermeer, Araujo, Bernritter, and van Noor, 2019).

연구자들도 소셜미디어가 기업의 다양한 성과에 긍정적인 영향을 미친다는 결과를 보고하고 있다. 소셜미디어의 활용은 매출액 등 단기적 성과(e.g., Kumar, Choi, and Greene, 2017) 뿐만 아니라 브랜드 자산 등 장기적 성과에도 긍정적인 영향을 미치는 것으로 보고되었다(e.g., Brodie, Hollebeek, Jurić, and Ilić, 2011). 이러한 소셜미디어 마케팅의 핵심적인 지표로는 고객 인게이지먼트(customer engagement) 또는 콘텐츠 인기도가 있다(De Vries, Gensler, and Leeflang 2012; Hoffman and Fodor, 2010). 선행연구는 소셜미디어 콘텐츠 성과(e.g., 고객 인게이지먼트)에 영향을 미치는 요인으로 콘텐츠의 구성적 특성을 연구하였다. 예를 들어 De Vries et al.(2012)은 생동감(vividness), 상호작용성(interactivity), 정보(information) 등의 효과를 분석하였으며, Swani and Miline(2017)은 브랜드 신호(brand Cue), 판매전략(CTA), 정보검색(information search), 소구방법(functional vs. emotional) 등의 요인을 추가하여 시장특성(B2B/B2C)에 따라 어떻게 요인들의 영향이 조절되는지 분석하였다. 또한, 고객 인게이지먼트에 영향을 미치는 요인으로 감정이 연구되었다(Berger and Milkman, 2012; Heath, Bell, and Sternberg, 2001). 예를 들어 Berger and Milkman(2012)는 뉴욕타임즈의 기사가 독자에게 더 높은 감정적 반응을 불러일으킬수록 공유될 가능성이 크다는 점을 발견하였다.

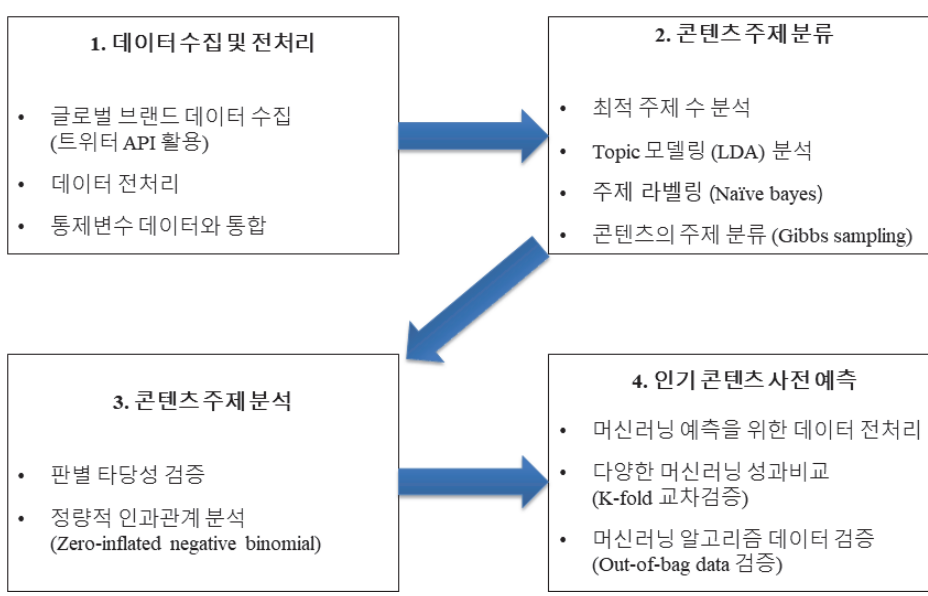
한편, 소셜미디어 메시지와 고객 인게이지먼트 간의 관계를 설명할 수 있는 잠재적 요인 중 하나로는 콘텐츠의 주제가 있다(Jalali and Papatla, 2019; Zhang, Moe, and Schweidel, 2017). 하지만, 소셜미디어 환경에서 브랜드 콘텐츠 주제가 소비자 행동에 미치는 영향은 아직 충분히 연구되지 않았다. 예외적으로 Zhang et al.(2017)은 미국 대학의 트위터 메시지와 이용자 간의 주제 적합도가 높을수록 리트윗이 많이 된다는 점을 보고하였다. 또한, Jalali and Papatla(2019)는 프로모션을 포함한 트윗 메시지가 리트윗에 긍정적인 영향을 미친다는 점과 프로모션 관련 단어가 메시지 상단에 있을수록 이러한 효과가 강화된다는 점을 발견하였다. 하지만, 두 연구 모두 소셜미디어 환경에서 브랜드가 전달할 수

1) www2.deloitte.com/si/en/pages/strategy-operations/articles/changing-consumer-digital-marketing-impact-Covid-19.html

있는 다양한 주제를 포괄하지 못했다는 한계점이 있다. 또한, Jalali and Papatla(2019)는 자동화된 기법(Malhotra, Malhotra, and See, 2012)을 활용하여 메시지의 인기도를 사전에 예측 할 수 있는 방법론(e.g., 머신러닝 모델)에 대한 향후 연구를 요청하였으나, 현재까지 충분한 연구가 이루어지지 않았다.

최근 경영학을 비롯한 다양한 분야에서 연구자들이 다양한 연구문제에 머신러닝을 적용하고 있다. 예를 들어 머신러닝 기술은 주식시장의 가격예측(Göçken, Özçalıcı, Boru, and Dosdoğru, 2016), 의학 분야의 암 탐지(Cruz and Wishart, 2006), 금융 분야의 신용 등급 예측과(Tsai and Chen, 2010) 같이 다양한 분야에서 적용되고 있다. 머신러닝은 디지털 트랜스포메이션의 대표적인 도구로 브랜드의 소셜리스닝 역량을 강화할 수 있는 대표적인 도구로 여겨지고 있다(Vermeer et al., 2019). 또한, 머신러닝 기반 방법론인 LDA(Latent Dirichlet Allocation)는 소셜미디어 콘텐츠의 주제를 식별하는데 적합한 분석방법으로 보고되고 있다(Blei, Ng, and Jordan, 2003; Tirunillai and Tellis, 2014; Jalali and Papatla, 2019; Zhang et al., 2017). Jalai and Papatla(2019)의 후속연구 요청과 같이 새로운 방법론을 활용하여 소셜미디어 주제를 식별하고 고객 인게이지먼트에 미치는 영향을 분석하는 것은 소비자 행동이 급격히 변화하는 환경에서 효과적인 소셜미디어 메시지를 개발하는데 중요한 시사점을 제공할 수 있다. 따라서, 본 연구에서는 머신러닝 방법론을 활용하여 소셜미디어 콘텐츠 주제를 분류하고 주제와 고객 인게이지먼트의 관계를 분석하고자 한다. 또한, 마케터가 콘텐츠를 발행하기 전에 인기도를 사전에 예측할 수 있는 머신러닝 모델을 제안하고자 한다.

연구의 절차는 그림 1과 같다. 첫째, 소셜미디어



〈Figure 1〉 연구방법 및 절차

환경의 고객 인게이지먼트와 머신러닝 방법론에 관한 선행연구를 검토하여 연구문제를 설정한다. 둘째, 머신러닝 기반 방법론인 토픽 모델링을 활용하여 글로벌 브랜드 트윗 메시지의 콘텐츠 주제를 분류한다. 셋째, 콘텐츠 주제가 고객 인게이지먼트에 미치는 영향을 영과잉 음이항(Zero-inflated negative binomial) 회귀분석을 통해 검증한다. 넷째, 콘텐츠 주제에 따른 트윗의 고객 인게이지먼트 성과를 다양한 머신러닝 알고리즘으로 예측하고 성과를 비교 분석한다. 다섯째, 분석결과를 토대로 이론적 실무적 시사점을 논의한다.

Ⅱ. 이론적 배경

2.1 소셜미디어 콘텐츠와 고객 인게이지먼트

소셜미디어 환경에서 고객 인게이지먼트 이론은 브랜드-고객 간의 상호작용을 설명하는 중요한 프레임워크로 여겨지고 있다(Pansari and Kumar 2016). 고객 인게이지먼트는 "특정 서비스 관계에서 대상과 상호작용함으로써 발생하는 심리적 상태"로 정의된다(Brodie et al., 2011). 또한, Muntinga et al. (2011)에 의하면 소셜미디어 환경의 고객 인게이지먼트를 소비(consuming), 기여(contributing), 생산(creating)으로 개념화 할 수 있다. 소비단계는 가장 낮은 수준으로 단순히 콘텐츠를 보는 등의 소비 행동을 의미한다. 기여단계는 소셜미디어 콘텐츠에 댓글을 다는 등의 행동을 의미하며, 생산단계는 소셜미디어 콘텐츠를 직접 작성하는 가장 상위단계의 행동을 의미한다. 소셜미디어 맥락의 선행연구는 브랜드 콘텐츠 인기지표로 소비자의 고객 인게이지먼트를 측정해왔다.

한편, 고객 인게이지먼트 문헌 이외에도 소셜미디어 콘텐츠 특성과 인기지표(i.e., 좋아요, 리트윗) 간의 관계는 소셜미디어 마케팅 문헌에서 활발히 연구되고 있다. 브랜드는 소셜미디어 환경에서 콘텐츠를 잠재 고객에게 전달하고 콘텐츠 확산을 유도함으로써 제품에 대한 인지도와 태도에 긍정적 영향을 미치기 위해 노력한다(Batra and Keller, 2016). 이러한 맥락에서 소셜미디어 콘텐츠는 브랜드에 대한 정보와 인식을 전달하는 매개체로써 소비자의 긍정적인 반응을 유도할 수 있는 중요한 마케팅 도구이다(e.g., de Vries et al. 2012; Swani and Milne, 2017).

선행연구는 대부분 소셜미디어 콘텐츠가 풍부할수록 인기지표에 긍정적인 영향을 미친다는 연구결과를 보고하였다. 예를 들어 Araujo, Neijens, and Vliegenthart(2015)는 19,343개의 글로벌 브랜드 트윗 메시지를 분석하여 정보가 풍부한 메시지가 소비자에게 긍정적인 평가를 받는다고 주장하였다. 그 밖에도 이미지나 동영상이 소셜미디어 콘텐츠에 포함되는 경우 콘텐츠의 생동감을 높여 인게이지먼트에 긍정적인 영향을 미친다는 연구결과가 보고 되었다(de Vries et al. 2012). 또한, 최근에는 Li and Xie(2019)이 소셜미디어 콘텐츠에 포함된 이미지의 실존감, 이미지 특성, 이미지와 텍스트 간의 적합도가 소셜미디어 인게이지먼트에 긍정적인 영향을 미친다는 연구결과를 제시하였다.

하지만, 소셜미디어 콘텐츠 주제에 대해 탐색한 연구는 소수이다. 선행연구는 감정 등 단편적인 변수에 초점을 맞추었다. 예를 들어, Berger and Milkman(2012)은 뉴욕타임즈 기사 중 경외심, 분노의 감정을 불러일으키는 기사가 공유될 확률이 높다는 점을 발견하였다. 유사하게 Heath et al. (2001)

도 감정적인 내용이 공유될 확률이 높다는 사실을 발견하였으며, 특히 혐오감을 일으킬 수 있는 부정적 감정도 공유될 가능성이 증가한다고 하였다. 한편, Toubia and Stephen(2013)은 이용자가 자신의 이미지를 관리하기 위해 콘텐츠를 공유하며, 자신이 원하는 이미지와 일치하는 콘텐츠를 선택적으로 공유한다는 점을 발견하였다.

본 연구와 관련된 선행연구로는 소셜미디어 콘텐츠 주제의 효과를 분석한 연구가 있다. 초기 연구로는 Malhotra et al.(2012)가 리트윗 요청(e.g., #RT_if retweet)이 포함된 트윗과 시간에 민감한 판매 홍보 트윗(e.g., 기간이 정해진 프로모션)이 높은 리트윗 수를 기록 한다는 점을 발견하였다. 또한, Araujo et al.(2015)는 사진과 동영상에 대한 링크와 해시 태그가 리트윗을 증가시킨다는 점을 보고 하였다. 이후, Zhang et al.(2016)은 메시지의 주제와 팔로워가 좋아하는 주제 사이의 적합도 수준이 리트윗에 긍정적인 영향을 미친다는 결과를 제시하였다. 가장 최근의 연구인 Jalali and Papatla (2019)은 트위터 메시지에 프로모션 관련 주제가 포함될수록 리트윗이 높아지며, 이러한 효과는 프로모션 주제와 관련된 단어가 메시지 상단에 위치할수록 강화된다는 점을 발견하였다. 하지만 선행연구는 트위터 메시지의 여러 토픽의 효과를 분석하지 못했다는 한계점이 있다. 또한, Malhotra et al.(2012)이 요청한 자동화된 방법론을 활용하여 마케터가 사전에 메시지의 인기를 예측 할 수 있는 방법을 탐색한 연구는 제한적이다. 따라서 본 연구에서는 머신러닝 방법에 관한 문헌을 리뷰 한 후, 머신러닝 방법을 활용하여, 트위터 메시지 주제 분류 및 고객 인게이지먼트 성과를 예측할 수 있는지 검증하고자 한다.

2.2 머신러닝 모델

점점 예측이 어려워지는 마케팅 환경은 머신러닝 활용의 중요성을 증가시키고 있다. Samuel(1959)은 머신러닝을 명시적으로 프로그램이 작성되지 않아도 컴퓨터가 스스로 학습할 수 있는 능력을 제공하는 학문이라고 정의하였다. 머신러닝은 특정 업무 경험을 반영해 그 성과를 향상하고 이를 평가하는 매커니즘으로 구성된다. 마케팅 문헌에서도 머신러닝을 활용한 연구들이 보고되고 있다. 선행연구는 예측(e.g., Cui and Curry, 2005; Vermeer et al., 2019), 특성 추출(Feature extraction; Tirunillai and Tellis, 2014), 기술적 해석(Trusov, Ma, and Jamal, 2016) 등 다양한 목적으로 활용되고 있다. 또한, 머신러닝의 연구방법 측면에서는 SVM (Support-vector machine; Cui and Curry, 2005), 토픽 모델링(Topic modeling; Tirunillai and Tellis, 2014), 앙상블 트리(Ensemble trees; Guo, Sriram, and Manchanda, 2018), 딥러닝(Ballestar, Grau-Carles, Sainz, 2018) 등 다양한 머신러닝 모델을 활용한 연구결과가 보고되고 있다.

첫째, 마케팅에 도입된 최초의 머신러닝 모델은 중 하나는 SVM이다. 마케팅 맥락에서 Cui and Curry (2005)는 SVM을 다항 로짓 모델과 비교하여 SVM의 예측성과가 더 뛰어나다는 결과를 제시하였다. 구체적으로 다항 로짓 모델은 연구의 시사점을 제시하는데 더 적합하지만, 대규모 데이터를 다루는 환경에서는 SVM이 더 적합하다는 점을 발견하였다. 이러한 결과는 머신러닝 모델이 예측적인 성과는 매우 뛰어나지만, 설명변수와 예측결과 간의 관계를 설명할 수 없다는 한계점을 지적하고 있다. 이후 유사하게 Huang and Luo(2016)은 Fuzzy

SVM 능동 학습 알고리즘을 사용하여 기존 방법보다 성과가 뛰어나다고 주장하였다.

둘째, 딥러닝(Deep learning) 모델은 최근 마케팅 연구에서 가장 많이 활용되는 머신러닝 모델로 텍스트 및 이미지 데이터 분석에 활용되고 있다. 예를 들어 Liu, Lee, and Srinivasan(2019)은 소비자 리뷰를 분석하여 미학(Aesthetic)과 가격 콘텐츠가 전환에 영향을 미친다는 결과를 보고하였다. 또한, Chakraborty, Kim, and Sudhir(2019)는 텍스트 데이터에서 감성특성을 추출하기 위해 Hybrid CNN-LSTM 모델을 개발하였으며, Yelp 리뷰를 대상으로 감정 분류의 적절성을 검증하였다.

셋째, 앙상블(ensemble) 방법은 여러 개별 학습자를 결합하는 학습 알고리즘으로 높은 예측 정확도를 가지는 것이 특징이다. 일반적인 방법으로 스태킹(stacking), 배깅(bagging), 부스팅(boosting)이 활용되고 있다. 스태킹은 개별 설명변수의 선형조합을 사용하여 정확도를 높이며(Breiman, 1996), 배깅은 각각의 개별 학습 트리가 부트스트랩 샘플을 활용하여 학습하며, 개별 학습 트리의 예측이 집계되어 최종 예측결과를 산출한다. 반면, 부스팅에서는 개별 학습 트리가 순서대로 훈련되고 각각의 정확도에 따라 더 강력한 학습 트리를 생성하게 된다. 또한, 적응적 부스팅(Adaptive boosting)은 이전의 잘못 분류한 내용을 조정하기 위해, 후속 학습자가 조정되는 방법이다. 종합하면, 배깅의 경우 병렬로 학습하며 부스팅은 순서대로 학습하고 학습이 끝난 후, 결과에 가중치를 부여한다. 일반적으로 부스팅이 개별 결정 트리의 성과는 높지만, 속도가 느리고 학습 데이터를 과도하게 학습하는 과적합(Overfitting)이 발생할 가능성이 크다. 인기 있는 앙상블 모델로는 랜덤 포레스트(Random-forest; Breiman, 2001)와 Gradient-boosted tree(GBM; Fridman, 2002)가 있다. 각각 배깅과 부스팅 모델을 활용한다. 랜덤 포레스트는 개별 트리가 원본 데이터의 부트스트랩 샘플을 통해 구축되며, 각 분할된 트리는 상관관계를 줄이기 위해 입력변수를 무작위로 할당하게 된다. 최종적으로 개별 트리의 예측결과를 평균화하여 최종 예측을 산출하게 된다. GBM은 여러 트리가 순서대로 훈련하며, 각 트리는 이전에 적용된 트리의 오류를 줄여서 정확도를 높인다. 희소성을 인식하는 XG부스트(eXtreme gadient boosting; XGBoost)은 Kaggle의 데이터 사이언스 대회에서 많은 우승을 한 모델이다(Chen and Guestrin, 2016).

마지막으로 LDA는 주어진 문서에 대하여 어떤 주제들이 존재하는지를 분석하는 확률적 토픽 모델링 방법 중 하나이다(Blei et al., 2003). 마케팅 문헌에서는 Tirunillai and Tellis(2014)이 LDA 방법론을 활용해 제품 리뷰에서 중요한 품질 차원을 추출하고 추출된 차원의 타당성을 검증했다. Büschken and Allenby(2016)는 일반적인 LDA 모델을 확장하여 동일한 문장 내에서 개별 단어가 하나의 주제에서만 추출되도록 제한함으로써, 일반적인 LDA의 결과보다 더 판별 타당성이 높고 일관성 있게 주제를 추출할 수 있다는 결과를 보고하였다. LDA는 텍스트뿐만 아니라 유사한 의미 구조가 존재하는 다른 마케팅 환경에도 적용되었다. 예를 들어 Jacobs, Donkers, and Fok(2016)은 온라인 소매업체의 데이터를 사용하여 LDA 모델이 소비자의 구매 예측에서 협업 필터링을 능가하며, 대규모 제품 구색으로 확장 할 수 있음을 보여주었다.

본 연구와 관련이 깊은 연구로는 소셜미디어의 메시지를 대상으로 토픽 모델링을 수행한 연구가 있다(Jalali and Papatla, 2019; Zhang et al., 2017). 두 연구 모두 확률적 토픽 모델링을 활용하

여 다양한 브랜드의 트위터 메시지의 주제를 분류하였다. 하지만 제한된 주제(e.g., 프로모션)와 리트윗(i.e., 고객 인게이지먼트) 간의 관계를 분석하였다는 한계점이 있다. 따라서 본 연구에서는 다양한 글로벌 브랜드의 트위터 메시지에 포함된 주제를 보다 포괄적으로 분류하고, 주제와 고객 인게이지먼트 간의 관계를 분석하고자 한다. 또한, 이러한 관계가 머신러닝 방법론을 통해 사전에 예측될 수 있는지 탐색하고자 한다. 따라서 아래와 같은 연구문제를 설정하였다.

연구문제 1: 글로벌 브랜드의 소셜미디어 콘텐츠 주제는 어떻게 분류되는가?

연구문제 2: 글로벌 브랜드의 소셜미디어 콘텐츠 주제에 따라 고객 인게이지먼트 성과에 차이가 있는가?

연구문제 3: 글로벌 브랜드의 소셜미디어 콘텐츠의 성과(인기도)를 머신러닝 모델을 활용하여 예측할 수 있는가?

Ⅲ. 연구방법

3.1 데이터

본 연구에서는 트위터를 대상으로 글로벌 브랜드의 콘텐츠를 수집하였다. 트위터는 대표적인 소셜미디어 플랫폼으로 소셜미디어 연구에 적합한 것으로 보고되었다(Sundstrom and Levenshus 2017; Tao and Wilson 2015). 글로벌 브랜드 데이터는 인터브랜드에 2019년 선정된 베스트 글로벌 브랜드 중 미국에 기업이 소재한 브랜드를 수집하였다. 콘텐츠 분석을 위해서는 하나의 언어를 대상으로 해야 하며, 미국 소재 브랜드(i.e., 영어 콘텐츠)를 선택한 이유는 다음과 같다. 첫째, 대부분의 글로벌 브랜드는 영어를 기반으로 대표계정을 운영하고 있다. 둘째, 인터브랜드에 포함된 브랜드 중 미국 소재 브랜드가 가장 많았다. 셋째, 영어는 전 세계 공용어로 가장 많은 수의 소비자 집단을 분석할 수 있다.

데이터는 트위터는 API를 통해 수집하였다. API를 이용하면 각 브랜드 계정당 3,300개까지 수집할 수 있다. 또한, 3,300개 중 공개되지 않은 트위터 메시지는 수집에서 제외된다. 본 연구에서는 R의 twitteR[2] 패키지를 활용하여 각 브랜드 계정에 등록된 트위터 메시지를 최신순으로 수집하였다. 최종적으로 53개 미국 소재 브랜드 중 트위터를 운영하지 않는 2개 브랜드(i.e., 애플, GE)를 제외하여, 51개 브랜드의 154,705개 데이터를 분석하였다. 수집된 브랜드와 데이터 수는 부록에 수록하였다.

3.2 측정

본 연구의 독립변수인 콘텐츠 주제는 LDA를 활용하여 콘텐츠 주제를 분류하고 메시지가 각 주제로 분류될 확률을 변수로 측정하였다(Jalali and Papatla, 2019). 종속변수인 고객 인게이지먼트는 리트윗 수로 측정하였다(e.g., Aleti, Pallant, Tuan, and van Laer, 2019; Okazaki, Díaz-Martín, Rozano, and Menéndez-Benito, 2015). 또한, 고객 인게이지먼트에 영향을 미칠 수 있는 언어적 특성을 측정하여 통제하였다. 선행연구에 따르면 주제뿐만 아

2) geoffjentry.hexdump.org/twitteR.pdf

니라, 언어적 스타일도 고객 인게이지먼트에 영향을 미칠 수 있다(Aleti et al., 2019). 언어적 특성은 LIWC 2015 소프트웨어를 활용하여 분석하였다. 선행연구를 참조하여 언어적 차원의 요약변수(Summary variable)인 분석적(Analytic), 외부초점(Clout), 감정(Tone) 변수와 LIWC 소프트웨어에서 제공하는 단어 수(WC), 문장길이(WPS), 긴 단어(SIXLTR), 사전에 포함된 단어 수(DIC)를 언어적 통제변수로 측정하였다(Aleti et al. 2019). 더불어 고객 인게이지먼트에 영향을 미칠 수 있는 브랜드 자산(Brand)은 인터브랜드 브랜드 가치 데이터를 측정하여 통제변수로 추가하였다. 마지막으로 선행연구에서 고객 인게이지먼트에 영향을 미치는 요인으로 보고된 이미지 또는 동영상 유무(media), 주말(weekend), 시간(night, morning, afternoon, evening)을 추가하였다(Li and Xie, 2019; Kanuri, Chen, Sridhar, 2018). 시간은 브랜드 콘텐츠의 작성 시간(i.e., 브랜드가 소재한 미국의 현지 작성 일시)을 기준으로 밤(0:00~5:59), 아침(6:00~11:59), 오후(12:00~17:59), 저녁(17:59~23:59)으로 구분하여 더미 변수로 측정하였다.

3.3 분석절차

분석은 다음과 같은 방법으로 진행하였다. 첫째, 글로벌 브랜드의 트위터 계정의 데이터를 API를 활용하여 수집하고 전처리 과정을 수행하였다. 데이터 전처리를 통해 불용어, URL 링크 등을 제외하였다. 그리고 브랜드 자산 및 언어적 특성변수 등 다른 통제변수 데이터와 결합하였다. 둘째, 토픽 모델링을 활용하여 콘텐츠 주제를 분류하였다. LDA 방법론을 활용하였으며, 분석 도구로는 R 패키지 textminerR을 활용하였다.[3] 우선 최적 주제 수를 분석한 결과 18개가 가장 적합한 것으로 나타났다. 이러한 분석 결과에 따라 18개의 토픽을 지정하여 토픽을 분류한 다음, 분류한 주제에 적합한 이름을 R 통계 패키지 LabelTopics[4]에서 제공하는 나이브 라벨링 알고리즘(Naive labeling algorithm)을 활용하여 산출하였다. 나이브 베이즈 모델은 베이즈 정리를 바탕으로 임의의 데이터가 특정 클래스에 속할 확률을 계산하며, 주로 텍스트 분류에 활용된다. 나이브 라벨링 알고리즘은 probable bi-grams을 기반으로 가장 적합한 토픽의 이름을 라벨링할 수 있다. 다음으로 깁스 샘플링(Gibbs sampling)을 활용하여 수집한 브랜드 메시지가 18개 주제로 분류될 확률을 측정하였다. 이러한 과정을 통해 전체 주제의 이름을 라벨링하고, 개별 메시지가 각 주제에 해당할 사후 확률을 측정할 수 있었다. 셋째, 분류된 주제가 적절한지 다른 언어적 변수를 통해 판별 타당성을 검증하였다. 다음으로 소셜미디어 콘텐츠 주제가 고객 인게이지먼트에 어떠한 영향을 미치는지 영과잉 음이항 회귀분석을 통해 검증하였다. 넷째, 머신러닝 모델을 활용하여 소셜미디어 콘텐츠의 고객 인게이지먼트를 예측하였다. 구체적으로 주요 머신러닝 알고리즘 16개의 예측성과를 비교하였다. 최종적으로 가장 우수한 머신러닝 모델인 XG부스트(eXtreme gradient boosting) 모델을 활용하여 OOB(out-of-bag) 방식으로 예측성과를 검증하였다.

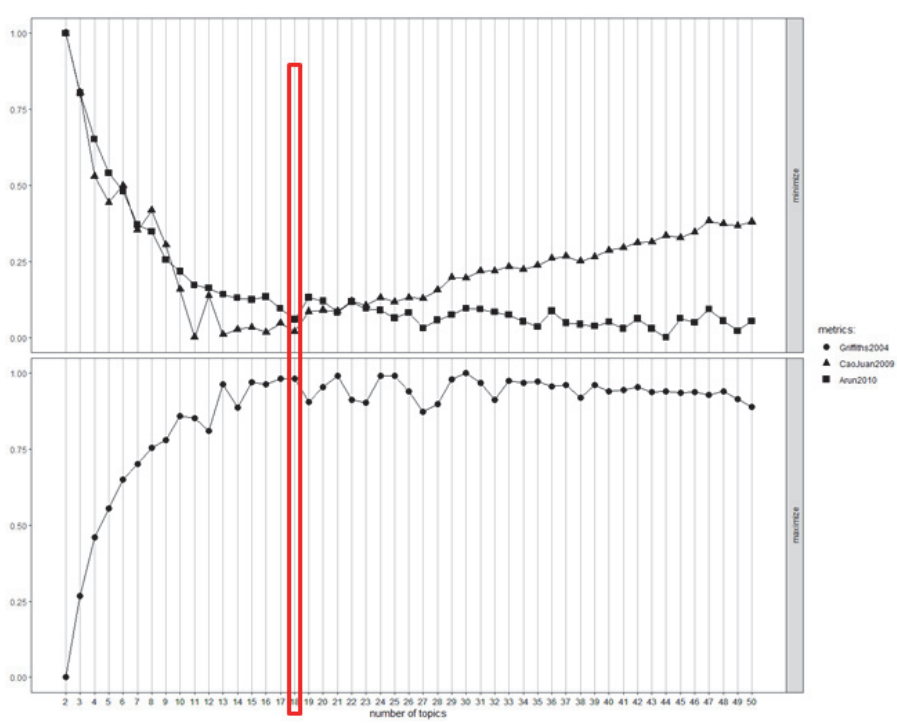
3) cran.r-project.org/web/packages/textmineR/index.html
4) www.rdocumentation.org/packages/stm/versions/1.3.6/topics/labelTopics

Ⅳ. 분석결과

4.1 연구 문제1

연구 문제1은 글로벌 브랜드의 소셜미디어 콘텐츠 주제가 어떻게 분류되는지를 탐색하기 위한 것이다. 본 연구에서는 LDA를 활용하여 메시지 콘텐츠의 주제를 확률적으로 분석하였다(Blei et al., 2003; Tirunillai and Tellis, 2014). 우선 최적 토픽의 수를 분석하기 위해 선행연구를 참조하여 3개의 적합도 지표를 분석하였다(Anandarajan, Hill, and Nolan, 2019). Griffiths와 Caojuan 지표는 분석된 값이 낮을수록 적합하며, Arun 지표는 값이 클수록 적합한 것으로 해석할 수 있다. 분석결과 아래 그림과 같이 주제를 18개로 분류하는 것이 가장 적합한 것으로 나타났다. 다음으로 18개의 주제를 지정하여 토픽 모델링을 실시하였다. 그리고 나이브 베이즈 방법을 활용하여 각 토픽의 이름을 지정하였다. 각 토픽의 이름과 고빈도 출현단어는 아래 표와 같다. 다음으로 깁스 샘플링방법을 활용하여 개별 콘텐츠가 18개 토픽으로 분류될 사후 확률을 산출하여 변수로 측정하였다.

각 토픽을 분석하기 전에 판별 타당성을 검증하였다. 선행연구에 따르면 소셜미디어 콘텐츠 주제에 따라서 언어적 특성에 차이가 있을 가능성이 크다. 따라서 본 연구에서는 언어적 스타일에 관한 선행연구의 판별 타당성 방법을 참조하여, 개별 메시지의



〈Figure 2〉 최적 토픽 수 분석

언어적 특성변수를 측정하여 주제에 따라 차이가 있는지 분석하였다(Aleti et al., 2019). 분석결과 아래 표와 같이 모든 언어적 특성변수에서 차이가 있는 것으로 나타났다. 따라서 주제의 판별 타당성이 충분한 것으로 판단하였다.

다음으로 각 18개 토픽의 주제를 주요 단어와 분류된 콘텐츠를 토대로 해석하였다. 토픽1. digital_transformation은 디지털화에 관한 주제로 구글과 같은 IT 기업에서 주로 활용하는 콘텐츠 주제로 나타났다. 특히 B2B 브랜드에서 디지털화에 대한 트렌드와 브랜드가 제공하는 서비스를 소개하는 내용이 많았다. 토픽2. conversation_glad는 브랜드가 소셜미디어 팔로워에게 다양한 토론 질문을 제시하고 의견을 듣는 내용이 많았다. 예를 들어 최근에 가장 중요한 이슈 중 하나인 코로나 19에 관한 내용이 다수 포함되어 있었다. 토픽3. support_community와 토픽14. diversity_inclusion은 브랜드의 사회적 공헌 활동에 대한 소개와 관련된 주제로 분석되었다. 주로 브랜드가 진행하고 있는 사회공헌 프로젝트에 대해 소개하는 내용이 많았다. 예를 들어 지역 문제, 직장 내 성 평등 등 사회적 문제를 해결하는 브랜드의 노력에 관한 성과와 직원 인터뷰 콘텐츠가 다수 포함되어 있었다.

토픽4. send_dm, 토픽6. glad_hear, 토픽7. phone_number, 토픽8. direct_message, 토픽9. hear_dm, 토픽11. provide_info 등은 고객의 불만

〈Table 1〉 토픽 모델링 결과 및 다빈도 출현단어

번호	토픽명	다빈도 출현 단어					N
1	digital_transformation	ai	learn	data	cloud	technology	25,036
2	conversation_glad	global	ceo	private	feel	free	7,021
3	support_community	support	community	covid	small	time	7,086
4	send_dm	contact	send	message	dm	information	6,931
5	values_standards	caption	shop	collection	discover	company	5,379
6	glad_hear	happy	glad	hear	love	great	10,954
7	phone_number	number	dm	phone	phone_number	address	5,032
8	direct_message	send	steps	surprise	message	find	4,941
9	hear_dm	dm	send	hear	email	account	9,013
10	learn_report	check	app	find	products	working	8,214
11	provide_info	team	hear	provide	call	info	5,437
12	customer_care	connecting	care	code	team	zip	3,995
13	business_page	business	learn	page	ecommerce	goglobal	3,050
14	diversity_inclusion	learn	women	people	year	proud	13,911
15	account_reach	account	watch	country	reach	time	15,214
16	chance_win	live	today	win	team	day	10,622
17	special_delivery	friends	cheers	good	special	tweet	7,067
18	click_link	day	great	make	work	chance	5,802

행동과 관련된 주제로 분석되었다. 이러한 콘텐츠 주제는 대부분 소비자의 불만에 응대하는 내용으로 서비스 실패에 대한 고객 문의에 대한 브랜드 답변이 많았다. 특히 맥도날드와 같은 서비스 브랜드에서 이러한 주제의 비율이 높은 것으로 분석되었다. 토픽5. value_standards는 브랜드가 제공하는 제품이나 서비스에 대한 우수성이나 수상내용과 관련된 주제로 제품의 품질과 관련된 홍보내용이 많은 것으로 분석되었다. 또한, 브랜드가 새롭게 출시하는 제품이나 서비스의 소개도 분류에 포함되어 있었다. 토픽10. learn_report는 글로벌 브랜드가 발행하는 여러 가지 보고서나 자료를 소개하는 내용과 관련된 것으로 분석되었다. 예를 들면 지속가능성 보고서 등이 있었다. 토픽3 support_ community와 내용적인 측면에서 유사하지만, 문서나 추가적인 페이지를 제공하는 콘텐츠가 많다는 점이 다른 것으로 분석되었다.

토픽12. customer_care는 브랜드가 고객 관리에 어떠한 노력을 기울이고 있는지를 소개하는 내용으로 나타났다. 주로 전 세계에 고객사를 두고 있는 B2B 기업에서 이러한 유형의 소셜미디어 콘텐츠가 많은 것으로 분석되었다. 토픽15. account_reach는 브랜드의 새로운 서비스나 제품의 발매 소식과 관련된 주제이다. 주로 디즈니와 같은 미디어 브랜드에서 활용하고 있었다. 토픽16. chans_win과 토픽17. special_delivery는 브랜드와 관련된 프로모션 및 이벤트에 관한 주제로 분석되었다. 이러한 토픽은 일반적인 예상과 달리 전체 콘텐츠에서 차지하는 비중이 낮은 것으로 나타났다. 토픽18 click_link는 단순히 링크를 클릭하라는 내용의 메시지가 많은 것으로 나타났다. 각 주제에 따른 대표 브랜드와 예시 메시지를 정리하면 다음 표 3과 같다.

4.2 연구 문제2

연구 문제2는 LDA를 통해 추출한 소셜미디어 콘텐츠 주제가 고객 인게이지먼트에 미치는 영향을 분석하기 위한 것으로 다수준 분석을 활용하였다. 이러한 위계 선형모형을 활용함으로써 콘텐츠 주제(1수준)의 효과를 브랜드(2수준)의 영향과 분리하여 추정할 수 있다. 즉 수준별 오차항의 독립성 가정을 만족시킴으로써 회귀계수를 추정의 신뢰성을 높일 수 있다. 또한, 본 연구의 종속변수인 고객 인게이지먼트(i.e., 리트윗 수)는 카운트 데이터로 표준편차(1,813)가 평균(68.48)을 초과하는 과다분산 형태이다. 카운트 데이터의 경우 일반적으로 포아송(Poisson) 모델이 고려 되지만 과대 산포가 있는 경우 모델의 효율성이 감소하므로 음이항(Negative binomial) 회귀모델을 권장한다. 또한, 0이 51.3%로 영이 과다한 경우 이러한 특성을 설명하기 위해 영과잉 음이항(zero-inflated negative binomial) 모델이 권장된다(Greene, 2003). 영과잉 음이항 모델은 종속변수가 1 이상인 경우를 분석하는 모형과 0인 경우를 분석하는 로짓 모형을 통합한 모형이다(Azagba and Sharaf, 2011).

본 연구에서는 이러한 데이터의 특성에 맞는 모델을 채택하기 위해 포아송 회귀분석, 영과잉 음이항 회귀분석을 수행하고 모델적합도를 비교하여 최종적으로 가설검증 방법을 채택하였다. 분석결과 영과잉 음이항 다수준 모델이 가장 적절한 것으로 나타났다.

연구 문제2를 분석한 결과는 다음과 같다. 다음 표 6의 모델(1)은 주제변수를 투입하지 않은 모델이며, 모델(2)는 주제변수를 투입한 모델이다. 모델의 적합도를 살펴보면 모델(1)에 비해 모델(2)에서 통계적으로 모델적합도가 개선된 것을 확인할 수 있다.

〈Table 2〉 언어적 특성을 통한 토픽 판별 타당성 검증

Variable	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	T13	T14	T15	T16	T17	T18	F
WC	23.38	31.86	33.27	27.97	29.52	20.40	30.67	28.79	27.61	32.37	30.95	25.14	31.06	33.05	23.92	27.65	19.81	19.46	1,186.5**
Analytic	72.64	82.62	79.57	65.61	79.70	60.47	58.87	64.10	49.55	65.19	62.93	80.49	80.28	84.79	77.92	81.98	74.86	74.82	1,119.6**
Clout	77.40	72.97	86.85	94.03	74.80	92.63	92.25	95.80	95.13	89.14	95.84	85.70	82.80	82.23	73.88	78.83	86.51	81.87	1,561.1**
Authentic	15.25	17.00	16.75	19.17	16.00	13.68	15.86	34.26	15.81	21.31	9.95	14.28	18.25	17.37	25.00	21.69	13.57	18.98	302.5**
Tone	57.32	52.53	69.18	77.68	58.59	89.82	67.62	75.89	61.65	58.71	74.74	58.23	58.95	68.82	54.24	62.89	83.35	80.98	959.3**
WPS	12.10	16.44	15.07	10.86	12.89	8.90	11.73	10.60	10.34	12.45	10.97	11.27	14.06	15.90	11.34	11.74	7.69	9 .42	1,024.2**
Sixltr	25.92	25.90	25.15	22.62	23.32	20.16	18.95	19.52	15.26	20.57	18.68	24.52	23.98	25.95	21.65	21.69	22.60	21.52	864.9**
Dic	59.24	61.19	69.73	76.83	59.10	73.86	70.99	79.41	81.27	72.26	72.54	65.19	64.53	63.11	61.70	60.02	70.85	66.22	2,453.8**

*p〈0.05; **p〈0.01

〈Table 3〉 토픽별 주요 브랜드와 트위터 예시

번호	토픽명	브랜드	예시 메시지	리트윗 수
1	digital_transformation	Accenture	• People depend on technology more than ever due to #COVID19...	9
2	conversation_glad	Goldman	• How might a national face mask mandate lower transmission rates and impact …	17
3	support_community	Gillette	• This week, we′re donating 10K face shields to healthcare organizations in need …	29
4	send_dm	Ford	• @ Could you please send us a DM with some more details regarding your …	0
5	values_standards	Tiffany	• VictoriaR vine designs prove that the best blooms are the kind that last forever …	90
6	glad_hear	Kellogg	• @ We′re happy to see you′re enjoying our promotional opportunity. Enjoy! …	0
7	phone_number	UPS	• @ Hi, we are here to assist. Please DM your tracking number, phone number …	0
8	direct_message	Google	• @ Hi there. Try recovering your username here: …	0
9	hear_dm	Uber	• @ Hey there, could you please elaborate on your concern? …	0
10	learn_report	Pampers	• We continue to be amazed by the incredible employees in our plants …	5
11	provide_info	McDonalds	• @ Sorry you were disappointed with your Sprite, Duke. Please provide some …	4
12	customer_care	Caterpillar	• Technicians are critical to keeping Cat machines up and running, just like …	4
13	business_page	DHL	• In the future, airlines and transportation companies will see new designs …	4
14	diversity_inclusion	J. and J.	• When you empower one woman, she can empower another and another …	17
15	account_reach	Disney	• Sit back, relax, and enjoy the sights and sounds of animation… now streaming…	86
16	chance_win	Harley.	• The time is here! This Prism Supply #Iron1200 could be in your garage …	30
17	special_delivery	JackDaniels	• Welcome to #TheJack. Friends, enjoy tonight′s festivities and we′ll see you …	22
18	click_link	LinkedIn	• The news you need to start your day: URL	4

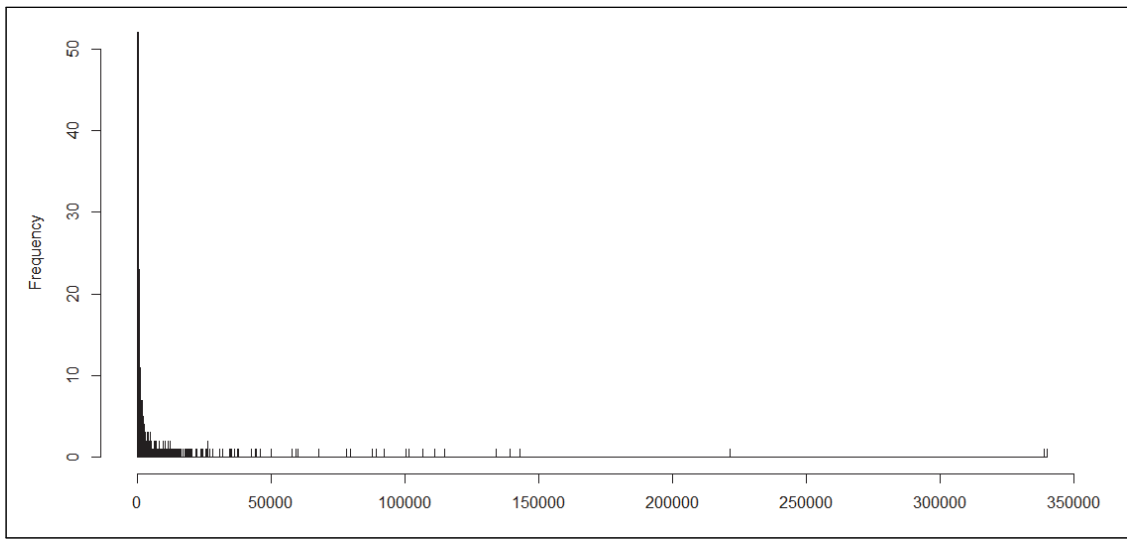
〈Table 4〉 변수의 기술통계량 및 상관관계 분석

Variables	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. Topic 1	1													
2. Topic 2	-.053**	1												
3. Topic 3	-.065**	-.051**	1											
4. Topic 4	-.098**	-.063**	-.061**	1										
5. Topic 5	-.067**	-.056**	-.047**	-.052**	1									
6. Topic 6	-.112**	-.075**	-.063**	-.060**	-.056**	1								
7. Topic 7	-.084**	-.055**	-.054**	-.035**	-.046**	-.056**	1							
8. Topic 8	-.079**	-.053**	-.048**	-.044**	-.043**	-.052**	-.040**	1						
9. Topic 9	-.108**	-.071**	-.069**	-.036**	-.058**	-.063**	-.037**	-.048**	1					
10. Topic 10	-.090**	-.071**	-.057**	-.045**	-.047**	-.045**	-.049**	-.050**	-.037**	1				
11. Topic 11	-.088**	-.059**	-.052**	-.036**	-.050**	-.055**	-.038**	-.044**	-.044**	-.043**	1			
12. Topic 12	-.063**	-.041**	-.039**	-.036**	-.034**	-.046**	-.035**	-.034**	-.039**	-.037**	-.033**	1		
13. Topic 13	-.054**	-.040**	-.036**	-.041**	-.035**	-.048**	-.038**	-.033**	-.046**	-.038**	-.038**	-.030**	1	
14. Topic 14	-.067**	-.047**	-.031**	-.093**	-.060**	-.102**	-.081**	-.076**	-.105**	-.098**	-.085**	-.058**	-.055**	1
15. Topic 15	-.109**	-.081**	-.072**	-.080**	-.042**	-.084**	-.067**	-.061**	-.088**	-.081**	-.070**	-.051**	-.054**	-.096**
16. Topic 16	-.091**	-.074**	-.058**	-.071**	-.047**	-.078**	-.064**	-.057**	-.084**	-.068**	-.066**	-.043**	-.046**	-.080**
17. Topic 17	-.090**	-.058**	-.052**	-.051**	-.046**	-.047**	-.045**	-.040**	-.057**	-.053**	-.045**	-.037**	-.037**	-.081**
18. Topic 18	-.075**	-.051**	-.046**	-.045**	-.039**	-.043**	-.041**	-.032**	-.054**	-.049**	-.042**	-.034**	-.034**	-.066**
19. Brand	.118**	-.058**	-.039**	-.045**	-.081**	.044**	.041**	.162**	-.042**	.068**	.011**	.051**	-.060**	-.043**
20. Media	.160**	.126**	.015**	-.141**	.128**	-.159**	-.120**	-.100**	-.168**	-.131**	-.114**	-.051**	.063**	.165**
21. WC	.063**	.047**	.085**	-.009**	.021**	-.152**	.022**	.000	-.004	.078**	.035**	-.035**	.019**	.120**
22. Analytic	.126**	.059**	.043**	-.065**	.043**	-.134**	-.103**	-.066**	-.207**	-.075**	-.083**	.029**	.030**	.136**
23. Clout	-.177**	-.128**	.029**	.117**	-.105**	.129**	.080**	.110**	.150**	.062**	.126**	.014**	-.013**	-.039**
24. Authentic	-.037**	-.013**	-.017**	.010**	-.019**	-.042**	-.023**	.126**	-.025**	.035**	-.072**	-.028**	.001	-.017**
25. Tone	-.103**	-.089**	.009**	.072**	-.047**	.192**	.008**	.052**	-.028**	-.061**	.046**	-.030**	-.034**	.020**
26. WPS	.154**	.126**	.093**	-.052**	.015**	-.136**	-.027**	-.052**	-.075**	-.003	-.049**	-.022**	.030**	.177**
27. Sixltr	.135**	.086**	.071**	.004	.022**	-.077**	-.064**	-.050**	-.176**	-.039**	-.076**	.031**	.025**	.130**
28. Dic	-.225**	-.091**	.032**	.145**	-.113**	.143**	.048**	.150**	.234**	.084**	.075**	-.021**	-.031**	-.101**
29. Retweets	-.004	-.006*	.005	-.007**	.006*	-.009**	-.005*	-.006*	-.009**	-.006*	-.007**	-.001	-.005	-.002
Mean	.113	.054	.050	.048	.042	.068	.037	.035	.060	.057	.042	.027	.028	.102
S.D.	.222	.163	.148	.159	.137	.181	.145	.141	.175	.160	.150	.116	.118	.205

〈Table 4〉 변수의 기술통계량 및 상관관계 분석 (계속)

Variables	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29
1. Topic 1															
2. Topic 2															
3. Topic 3															
4. Topic 4															
5. Topic 5															
6. Topic 6															
7. Topic 7															
8. Topic 8															
9. Topic 9															
10. Topic 10															
11. Topic 11															
12. Topic 12															
13. Topic 13															
14. Topic 14															
15. Topic 15	1														
16. Topic 16	-.061**	1													
17. Topic 17	-.064**	-.058**	1												
18. Topic 18	-.055**	-.050**	-.035**	1											
19. Brand	-.016**	-.065**	-.079**	.017**	1										
20. Media	.147**	.102**	.004	-.067**	-.139**	1									
21. WC	-.081**	.002	-.127**	-.123**	-.027**	.207**	1								
22. Analytic	.069**	.093**	.017**	.015**	-.084**	.296**	.147**	1							
23. Clout	-.176**	-.083**	.040**	-.019**	.085**	-.258**	.102**	-.274**	1						
24. Authentic	.100**	.042**	-.040**	.013**	-.037**	-.049**	-.006*	.092**	-.202**	1					
25. Tone	-.119**	-.031**	.108**	.091**	005	-.150**	-.023**	-.055**	.205**	-.104**	1				
26. WPS	-.033**	-.019**	-.146**	-.080**	-.047**	.211**	.577**	.199**	-.057**	.006*	-.105**	1			
27. Sixltr	-.026**	-.020**	-.002	-.028**	-.048**	.072**	-.115**	.254**	-.217**	-.024**	-.013**	.004	1		
28. Dic	-.118**	-.144**	.058**	.001	-.007**	-.410**	.080**	-.349**	.364**	.209**	.235**	-.106**	-.283**	1	
29. Retweets	.054**	.002	-.006*	-.005*	.010**	.032**	-.008**	.007**	-.026**	.008**	-.018**	.000	-.007**	-.022**	1
Mean	.085	.073	.043	.037	9.558	.270	26.942	72.490	83.860	17.910	66.119	11.951	22.538	66.760	68.48
S.D.	.199	.180	.150	.133	.879	.444	13.217	29.086	19.963	25.139	35.327	7.019	10.172	15.281	1,813

*p〈0.05; **p〈0.01



〈Figure 3〉 리트윗 변수의 분포

〈Table 5〉 모델의 적합도

모델	AIC	BIC	logLik
포아송 모델	68,084,389	68,084,707	-34,042,162
포아송 다수준 모델	47,718,778	47,719,106	-23,859,356
영과잉 음이항 모델	815,582.8	816,229.5	-407,726.4
영과잉 음이항 다수준 모델	743,989.3	744,655.9	-371,927.7

또한, 브랜드 수준의 분산을 살펴보면, 조건모델과 영과잉모델 모두 주제변수를 투입한 모델(2)가 투입하지 않은 모델(1)에 비해 임의효과 분산이 감소하였다. 따라서 본 연구에서 분류한 소셜미디어 메시지 주제는 고객 인게이지먼트의 분산을 설명하는데 유효한 변수로 판단할 수 있다.

분석결과, 주제에 따라 고객 인게이지먼트에 미치는 영향에 차이가 있는 것으로 나타났다. 우선 토픽4, 토픽6, 토픽7, 토픽8, 토픽9, 토픽11 등 고객의 불만 행동에 대한 처리와 관련된 주제는 고객 인게이지먼트에 부정적인 영향을 미치는 것으로 나타났다. 이러한 주제는 개별 고객의 불만에 대한 답변이기 때문에 인게이지먼트에 부정적인 영향을 미치는 것으로 이해할 수 있다.

반면, 브랜드의 사회적 책임 활동과 관련된 토픽3, 토픽14은 모두 고객 인게이지먼트에 긍정적인 영향을 미치는 것으로 나타났다. 선행연구에 의하면 기업의 사회적 책임 활동은 소비자를 포함한 이해관계자의 관계에 긍정적인 영향을 미치는 것으로 보고되었다. 소셜미디어 환경에서도 이러한 긍정적 효과가 나타난 것으로 이해할 수 있다. 그 밖에 Jalali and Papatla(2019)의 연구결과에서 나타난 것과 같이

〈Table 6〉 영과잉 음이항 회귀분석 결과

Conditional Model	(1)		(2)	
	Estimate	Std. Error	Estimate	Std. Error
T1 digital_transformation	Referenced		Referenced	
T2 conversation_glad			-.2224***	0.0280
T3 support_community			.6956***	0.0282
T4 send_dm			-2.2198***	0.0760
T5 values_standards			-.1562***	0.0284
T6 glad_hear			-1.0391***	0.0388
T7 phone_number			-.2696***	0.0656
T8 direct_message			-.9966***	0.0552
T9 hear_dm			-2.1997***	0.0723
T10 learn_report			-.2055***	0.0449
T11 provide_info			-.9897***	0.0626
T12 customer_care			-.0726	0.0547
T13 business_page			-1.0808***	0.0389
T14 diversity_inclusion			.128***	0.0199
T15 account_reach			-.0432	0.0234
T16 chance_win			.3149***	0.0228
T17 special_delivery			-.8697***	0.0427
T18 click_link			-.4655***	0.0419
WC	.0242***	.0006	.0176***	0.0006
Analytic	.0045***	.0003	.0035***	0.0003
Clout	-.0043***	.0003	-.0053***	0.0003
Authentic	-.0007***	.0002	-.0009***	0.0002
Tone	-.0026***	.0002	-.0022***	0.0002
WPS	-.0043***	.0009	-.003***	0.0009
Sixltr	-.0072***	.0007	-.0115***	0.0007
Dic	-.0168***	.0005	-.0155***	0.0005
log(brand)	.3938*	.2132	.4246**	0.2075
media	.0230*	.0123	.0253**	0.0122
weekend	-.0049	.0129	-.0462***	0.0127
Night	.2221***	.0196	.2328***	0.0196
Morning	-.2414***	.0310	-.1811***	0.0305
Afternoon	.1775***	.0120	.177***	0.0118
(Intercept)	.1585	2.0469	.1838	1.9932

〈Table 6〉 영과잉 음이항 회귀분석 결과 (계속)

Random variance	Brand	Variance	Std. Dev.	Variance	Std. Dev.
		1.761	1.327	1.652	1.285
Zero-inflation model		Estimate	Std. Error	Estimate	Std. Error
T1 digital_transformation		Referenced		Referenced	
T2 conversation_glad				.313**	.0280
T3 support_community				-.2434***	.0282
T4 send_dm				1.9233***	.0760
T5 values_standards				.196**	.0284
T6 glad_hear				.783***	.0388
T7 phone_number				2.8395***	.0656
T8 direct_message				1.4942***	.0552
T9 send_dm				1.5747***	.0723
T10 learn_report				1.8256***	.0449
T11 provide_info				1.7166***	.0626
T12 customer_care				1.1436***	.0547
T13 business_page				1.0468***	.0389
T14 diversity_inclusion				-1.2051***	.0199
T15 account_reach				.0462	.0234
T16 chance_win				-.2594***	.0228
T17 special_delivery				1.4433***	.0427
T18 click_link				.9916***	.0419
WC		-.0127***	.0011	-.0222***	.0006
Analytic		-.004***	.0004	-.0037***	.0003
Clout		.0155***	.0006	.0102***	.0003
Authentic		-.0031***	.0005	-.003***	.0002
Tone		.0012***	.0003	.0011**	.0002
WPS		-.0724***	.0026	-.055***	.0009
Sixltr		-.001	.0011	.002*	.0007
Dic		.0289***	.0009	.0192***	.0005
log(brand)		-.2335	.4579	-.1879	.2075
media		-3.3555***	.0572	-3.036***	.0122
weekend		-.0823***	.0248	-.061**	.0127
Night		.3749***	.0330	.2899***	.0196
Morning		.8464***	.0650	.598***	.0305
Afternoon		.1317***	.0235	.0902***	.0118
(Intercept)		-.1672	4.3943	-.1606	1.9932
Random variance	Brand	Variance	Std. Dev.	Variance	Std. Dev.
		7.769	2.787	6.54	2.557
AIC		756,831		743,989	
BIC		757,160		744,655	
logLik		-378,383		-371,927	
Chisq. Test		Δ12,910 (d.f=34)***			
Num. of Brands		51			
Num. of Tweets		154,705			

*p〈0.1; **p〈0.05; ***p〈0.01

토픽16과 같은 프로모션 주제는 고객 인게이지먼트에 긍정적인 영향을 미치는 것으로 나타났다.

또한, 영과잉 모델 분석결과에서도 조건모델과 유사하게 기업의 사회적 책임과 관련된 토픽3과 토픽14는 고객 인게이지먼트가 0일 확률에 부정적인 영향을 미치는 것으로 나타났다. 또한, 프로모션가 관련된 토픽16도 부정적인 영향을 미치는 것으로 나타났다.

통제변수를 분석한 결과를 살펴보면, 브랜드 자산은 고객 인게이지먼트에 긍정적인 영향을 미치는 것으로 나타났다. 또한, 단어 수가 많고 이미지가 포함될수록 고객 인게이지먼트에 긍정적인 영향을 미치는 것으로 나타났다. 시간적 차원에서는 주말에 포스팅된 콘텐츠일수록 고객 인게이지먼트가 감소하며, 밤이나 오후에 작성할수록 고객 인게이지먼트가 증가하는 것으로 분석되었다.

4.3 연구 문제3

연구 문제3은 최근 디지털 트랜스포메이션의 핵심적인 도구인 머신러닝을 활용하여 브랜드 콘텐츠의 고객 인게이지먼트 성과를 예측할 수 있는지 분석하기 위한 것이다. 선행연구에 따르면 다양한 머신러닝 모델 알고리즘이 존재하며 알고리즘에 따라 성과에 차이가 있을 수 있다(Hartman, Huppertz, Schamp, and Heitmann, 2019). 따라서 본격적인 분석에 앞서 본 연구에 적합한 머신러닝 알고리즘을 분석하기 위해 R의 CARET 패키지[5]를 활용하여 다양한 머신러닝 알고리즘의 예측성과를 비교하였다. 조율 모수의 후보 값을 10개로 설정하고 데이터의 교차검증(K-fold cross validation)은 5로 설정하여 수행하였다. 또한, 머신러닝 성능 비교의 지표로는 ROC(Receiver operating characteristic)를 선정하여 수행하였다(Akobeng, 2007).

머신러닝 알고리즘의 예측성과를 비교하기 위한 기준은 Vermeer et al.(2019)의 연구를 참조하여 민감도, 특이도, PPV(Positive predictive value), NPV(Negative predictive value), Prevalence 등의 지표를 활용하였다. 이러한 지표는 예측이 부정확하게 분류되었는지, 혹은 정확하게 분류되었는지에 관한 측정개념을 기반으로 한다(Esuli and Sebastiani, 2010). 해당 측정개념은 1) 실제 양성을 양성으로 판정한 경우인 TP(True Positive), 2) 실제 음성을 양성으로 판정한 경우인 FP(False Positive), 3) 음성을 음성으로 판정한 경우인 FN(False Negative), 4) 양성을 음성으로 판정한 경우인 True Negative(TN)으로 구분된다. 또한, 이러한 측정개념을 활용하여 아래와 같이 지표를 산출하여 모델을 평가하였다.

- 민감도(Sensitivity) : $\frac{TP}{TP+FN} \times 100$
- 특이도(Specificity) : $\frac{TN}{FP+TN} \times 100$
- 정확도(Accuracy) : $\frac{TP+TN}{TP+FP+FN+TN} \times 100$
- Prevalence : $\frac{TP+FN}{TP+FP+FN+TN} \times 100$
- PPV(Pose Pred Value) : $\frac{TP}{TP+FP} \times 100$
- NPV(Neg Pred Value) : $\frac{TN}{FN+TN} \times 100$

5) https://cran.r-project.org/web/packages/caret/caret.pdf

선행연구를 기반으로 총 16개의 머신러닝 알고리즘 모델을 비교하였다(Lessmann, Haupt, Coussement, and De Bock, 2019). 각 머신러닝 알고리즘에 대한 설명과 예측성과는 아래 표 7과 같다. 분석결과를 살펴보면 앙상블 모델이 상대적으로 예측성과가 우수한 것으로 나타났다. 또한, 앙상블 모델 중에서도 희소성을 인식하는 XG부스팅 모델이 가장 우수한 것으로 나타났다. 이러한 분석결과를 그래프로

〈Table 7〉 머신러닝 알고리즘의 예측성능 비교 결과

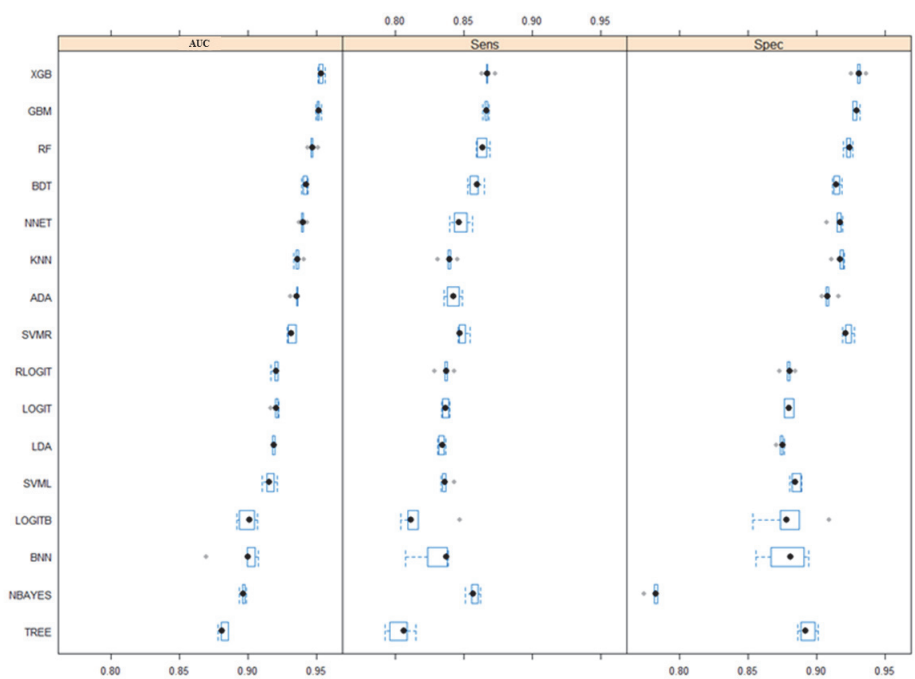
번호	머신러닝 알고리즘	설명	AUC	민감도	특이도
1	Classification tree (TREE)	• 결정규칙과 결과를 트리구조로 도식화한 결정트리를 사용하는 예측모델	.8823	.8037	.8932
2	Artificial neural network(NNET)	• 시냅스의 결합으로 네트워크를 형성한 인공 뉴런이 학습하는 모델	.9401	.8475	.9152
3	k-nearest-neighbor (KNN)	• 분류나 회귀에 사용되는 비모수 방식으로 데이터 간 거리 계산을 기반으로 하는 모델	.9366	.8390	.9168
4	Linear discriminant analysis(LDA)	• 공간상에서 클래스의 분리를 최대화하는 주축으로 차원을 축소하는 모델	.9189	.8339	.8740
5	Logistic regression (LOGIT)	• 데이터가 어떤 범주에 속할 확률을 예측하는 지도 학습 모델	.9203	.8366	.8797
6	Naive bayes (NBAYES)	• 베이즈 정리를 적용한 확률 분류기의 하나로 여러 알고리즘을 이용하여 훈련하는 모델	.8967	.8572	.7812
7	Regularized logistic regression(RLOGIT)	• 일반화 오류를 줄이기 위한 로지스틱 회귀분석의 수정 모델	.9203	.8367	.8794
8	Support vector machine with linear kernel (SVML)	• 데이터 집합을 바탕으로 새로운 데이터의 분류를 판단하는 비확률적 이진 선형 분류모델	.9161	.8368	.8848
9	Support vector machine with radial basis function kernel(SVMR)	• Gaussian kernel이라고도 불리며, SVMR과 유사하지만 내적 연산이 비선형 커널 함수로 대체됨	.9321	.8490	.9225
10	AdaBoost(ADA)	• 앙상블 기반 모델로 weak classifier를 본적으로 적용해서 data의 특징을 찾아가는 모델	.9350	.8423	.9085
11	Bagged Decision Trees (BDT)	• 기존 결정 트리의 단점인 높은 분산을 줄일 수 있는 모델	.9418	.8586	.9146
12	Bagged Neural Networks (BNN)	• 다중 predictor를 생성하는 신경망 네트워크 모델	.8965	.8289	.8776
13	Random Forest(RF)	• 앙상블 학습 방법으로 다수의 결정 트리로부터 예측치를 출력하는 모델	.9469	.8636	.9231
14	Logit Boost(LOGITB)	• 로지스틱의 손실을 최소화하는 가산적 트리 모델	.8997	.8176	.8800
15	Gradient Boosting Algorithm(GBM)	• 다수의 약한 학습기를 순차적으로 학습하면서 잘못 예측한 데이터에 가중치 부여하는 모델	.9515	.8662	.9283
16	eXtreme Gradient Boosting(XGB)	• GBM의 문제점인 속도와 과적합을 보완하기 위해 개발된 모델로 regularization이 포함됨	.9534	.8675	.9300

주. AUC(area under the cureve): 수신자 판단 특성 곡선(receiver operating characteristic curve)의 밑면적을 계산한 값

나타내면 아래 그림과 같다.

가장 성능이 우수한 XGB 모델을 대상으로 예측성과를 분석하였다. 예측성과 검증은 OOB 데이터를 활용하였다. 구체적으로 전체 데이터(154,705개)를 7/3 비율로 분리하여 108,294개(70%)의 데이터를 활용하여 머신러닝 모델을 학습하였으며, 학습에 활용되지 않은 46,411개(30%) 데이터를 대상으로 학습된 모델의 예측성과를 검증하였다. 또한, 성과변수는 평균값을 기준으로 분리하여 0과 1로 코딩하였다. 코딩된 성과변수는 1에 비해 0이 과도하게 많은 불균형 데이터 형태로 나타났다. 이러한 불균형 데이터를 학습하는 머신러닝 모델은 과반 데이터를 0으로 예측하는 경향이 있다. 따라서 2가지 데이터 처리방법을 통해 예측성과를 검증하였다. 첫째, 불균형 데이터를 그대로 활용하여 분석하였다. 둘째, SMOTE(synthetic minority oversampling technique; Chawla, Bowyer, Hall, and Kegelmeyer, 2002) 샘플링방법을 활용하여 0과 1의 비율을 균형화한 후 7/3으로 나누어 머신러닝 모델을 학습하고 검증하였다.

먼저, 원본 데이터를 예측한 결과를 보면 정확도는 96% 이상으로 나타났으나, 실제 양성을 음성으로 예측하는 경향이 심화 되어 민감도가 낮게 나타났다. 하지만 토픽 변수를 추가하지 않은 모델보다 추가한 모델이 더 정확도가 개선되는 것으로 나타났다. 특히, 민감도의 경우 기본모델은 48%로 나타났으나, 주제를 투입한 모델은 54%로 6% 정도 향상되었다. Ling, Deng, Gu, Zhou, Li, and Sun



〈Figure 4〉 머신러닝 성능 비교

〈Table 8〉 원본 데이터 예측 결과

성능지표	Baseline model	Full model
토픽 변수	NO	YES
통제 변수	YES	YES
정확도(95% 신뢰구간)	.9578(.9559, .9596)	.9618(.96, .9635)
No Information Rate	.9311	.9315
민감도	.4846	.5399
특이도	.9928	.9929
PPV(Pose Pred Value)	.8333	.8465
NPV(Neg Pred Value)	.9630	.9670
Prevalence	.0689	.0684

〈Table 9〉 샘플링방법을 활용한 균형 데이터 예측 결과

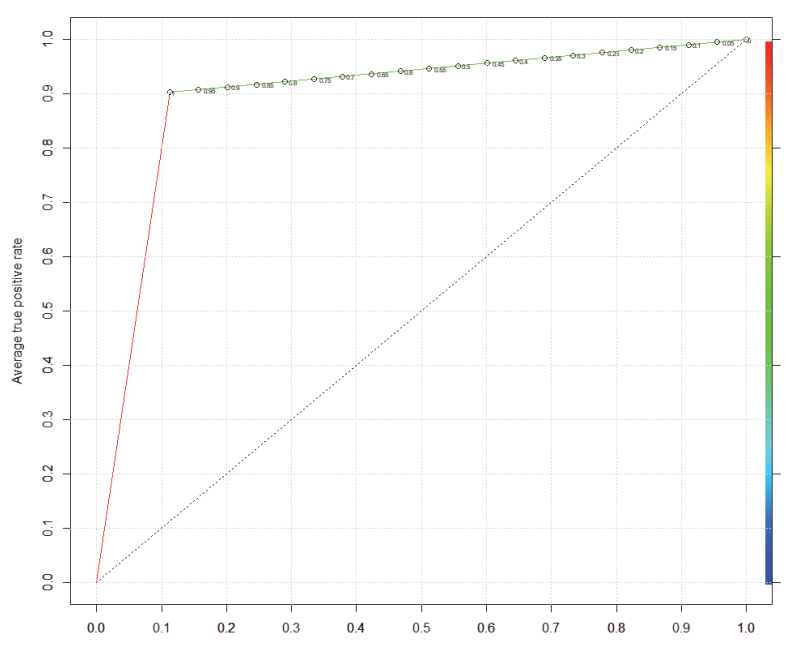
성능지표	Baseline model	Full model
토픽 변수	NO	YES
통제 변수	YES	YES
정확도(95% 신뢰구간)	.8569(.8537, .8601)	.8948(.8919, .8975)
No Information Rate	.5016	.5034
민감도	.8700	.8873
특이도	.8439	.9023
PPV(Pose Pred Value)	.8471	.9020
NPV(Neg Pred Value)	.8672	.8877
Prevalence	.4984	.5034

(2017)에 의하면 0.1%의 정확도 향상도 성과에 유의한 변화를 가져올 수 있다.

다음으로 샘플링방법을 활용하여 예측한 결과에서는 기본모델과 토픽 변수를 추가한 모델 간의 차이가 더 심화 되는 것으로 나타났다. 기본모델의 정확도는 85.69%로 나타났으며, 토픽 변수를 추가한 모델은 89.48%로 3% 이상 개선되었다. 따라서, 글로벌 브랜드의 소셜미디어 콘텐츠 성과(인기도)를 머신러닝 모델을 활용하여 약 90%의 정확도로 예측할 수 있는 것으로 분석되었다. 실제 본 연구의 테스트 데이터를 대상으로 고객 인게이지먼트 성과와 예측결과를 비교하면 아래의 표와 같다. 예측결과는 확률로 표기되며, 이러한 분석을 통해 마케터는 자신이 개발한 메시지의 고객 인게이지먼트 성과를 사전에 확인할 수 있다. 또한, 머신러닝 모델의 성능을 ROC 커브로 나타내면 다음 그림과 같다.

〈Table 10〉 테스트 데이터 예측결과 예시

토픽	브랜드	트위터 메시지	예측	실제
Digital_Transformation	Google	• Today we′re launching a redesigned @GooglePhotos to help you…	95%	93
Conversation_Glad	JP morgan	• Three questions to ask if you′re new to investing and financial plan..	7%	1
Support_Community	Microsoft	• ″We must take a stand against this. We must all fight for Black …	45%	40
Send_DM	Caterpillar	• @Mohan_KumarB Hi Mohan! The sweepstakes is only available to U.S.	-10%	0
Values_Standards	JackDaniels	• We can′t fix the cold weather, but we can help you fix a warm cocktail.	68%	83
Glad_Hear	UPS	• @lemonchronicle We love that idea, Elizabeth! Thanks so much for …	-1%	0
Learn_Report	PayPal	• Learn how to transfer money from your #PayPal Balance to your …	60%	12
Provide_Info	HarleyDavid .	• @mac_of_all_macs Hey, Mac. We are sorry to learn about the issues …	-11%	0
Customer_Care	Amazon	• Hey, India. We′re rolling out our new fleet of electric delivery rickshaws…	85%	8,813
Diversity_Inclusion	Microsoft	• LGBTQI+ people at Microsoft want to embrace the uncomfortable. A Di…	100%	108
Account_Reach	Disney	• We′re excited to share that Tokyo Disneyland will reopen on July 1 URL…	103%	1,849
Chance_Win	KFC	• LAST CHANCE to get your name on Sanders Claus′ list! Tag #secretsanders …	82%	14
Special_Delivery	Morgan Stanley	• Wishing you and your family a joyous and blessed #EidalFitr ! URL…	18%	10
Click_Link	Pepsi	• @itsSivachandru It′s great to take a day off. Click the link for your chance…	0%	0



〈Figure 4〉 머신러닝 모델의 ROC 커브

Ⅴ. 결론 및 시사점

5.1 연구의 시사점

본 연구는 소셜미디어 콘텐츠 주제와 고객 인게이지먼트 간의 관계를 분석하고, 머신러닝 모델의 활용 가능성을 탐색적으로 분석하였다. 본 연구의 이론적 시사점은 다음과 같다. 첫째, 본 연구는 글로벌 브랜드의 소셜미디어 콘텐츠 주제가 고객 인게이지먼트에 미치는 영향을 탐색적으로 분석하였다는 점에서 의미가 있다. 다수의 선행연구는 콘텐츠의 구성적 특성(e.g., 이미지 수, 동영상 포함 여부; de Vries et al., 2012)이나 불러일으키는 감정(e.g., 분노, 경외심; Berger and Milkman, 2012)이 성과에 미치는 영향에 대해 분석하였다. 하지만, 소비자의 정보처리 과정과 밀접한 관련이 있는 콘텐츠의 주제에 대해서는 충분히 연구되지 않았다. 예외적으로 Zhang et al.(2017)과 Jalali and Papatla(2019)가 LDA를 활용하여 분석하였으나, 다양한 주제의 효과를 분석하지 않았다는 한계점이 있었다. 이러한 맥락에서 본 연구는 소셜미디어의 불만 행동 관리와 관련된 메시지는 고객 인게이지먼트에 부정적인 영향을 미치며, 브랜드의 사회적 책임 활동과 관련된 주제는 긍정적인 영향을 미친다는 사실을 검증함으로써 소셜미디어 마케팅 및 고객 인게이지먼트 문헌에 기여하였다. 둘째, 본 연구는 머신러닝 모델을 활용하여 소셜미디어 콘텐츠의 고객 인게이지먼트 성과를 예측할 수 있음을 제시하였다. 최근 마케팅 맥락에서 머신러닝을 비롯한 다양한 기술의 적용 가능성을 검증하는 것은 중요한 연구문제로 제시되고 있다. 특히 소셜미디어 마케팅 맥락에서 Malhotra et al.(2012)와 Jalali and Papatla (2019)은 소셜미디어 콘텐츠의 성과를 사전에 예측할 수 있는 방법론에 관한 연구를 요청한 바 있다. 본 연구는 이러한 연구요청에 응답하고 다양한 머신러닝 모델을 비교하여 XG부스트 모델이 소셜미디어 콘텐츠 성과를 예측하는데 가장 적합하다는 점을 발견하였다. 또한, 약 90%의 높은 정확도로 예측할 수 있다는 점을 제시함으로써 마케팅 방법론에 기여하였다.

본 연구의 실무적 시사점은 다음과 같다. 첫째, 소셜미디어 채널은 코로나 19 이후 더욱 중요해지고 있으며, 소셜 리스닝은 머신러닝 등 새로운 디지털 도구가 활용되면서 새로운 경쟁전략으로 대두되고 있다. 본 연구결과는 이러한 맥락에서 브랜드 마케터가 소셜미디어 메시지 개발에 활용할 수 있는 가이드 라인을 제공한다. 마케터는 브랜드와 관련된 다양한 사회적 책임 활동과 프로모션에 대한 소식을 전달함으로써 고객 인게이지먼트를 증가시킬 수 있을 것이다. 또한, 단어 수가 많고 이미지가 많이 포함될수록 고객 인게이지먼트에 긍정적인 영향을 미칠 수 있다. 더불어 브랜드 메시지 일정관리의 측면에서는 주말을 피하고, 밤이나 오후에 작성하는 것이 유리하다. 둘째, 본 연구결과는 브랜드 마케터가 머신러닝을 활용하여 브랜드 콘텐츠를 개발하는데 시사점을 제공한다. 본 연구에서는 16가지 머신러닝 모델의 예측성과를 비교 분석함으로써 XG부스트 모델이 콘텐츠 성과의 예측에 가장 우수하다는 결과를 제시하였다. 브랜드 마케터는 소셜미디어 메시지 개발에 도움을 줄 수 있는 머신러닝 모델을 개발하는데 본 연구결과를 초기 가이드라인으로 활용할 수 있을 것이다.

5.2 연구의 한계점 및 향후 연구방향

본 연구의 한계점 및 향후 연구 방향은 다음과 같다. 첫째, 본 연구는 트위터의 데이터를 수집하여 활용하였다. 하지만, 소셜미디어 종류에 따라 연구결과에 차이가 있을 수 있다. 따라서 향후 연구에서는 다양한 소셜미디어를 대상으로 연구가 이루어질 필요가 있다. 둘째, 본 연구에서는 16개의 주요 머신러닝 알고리즘을 비교하였으며, CARET 패키지를 활용하여 조율 모수의 후보 값을 10개로 설정하였다. 하지만, 본 연구에서 다루지 않은 알고리즘과 조율 모수에 따라 가장 적합한 머신러닝 모델은 달라질 수 있다. 따라서 향후 연구에서는 본 연구에서 다루지 않은 머신러닝 모델과 함께 더 넓은 범위의 모수 후보를 탐색하는 것이 필요하다. 셋째, 본 연구는 미국의 브랜드만을 대상으로 하고 있다. 하지만 문화에 따라 콘텐츠 주제가 고객 인게이지먼트에 미치는 영향에는 차이가 있을 수 있다. 따라서 향후 연구에서는 다양한 국가를 대상으로 연구될 필요가 있다. 넷째, 본 연구에서는 리트윗을 단순히 고객 인게이지먼트 성과로 측정하였으나, 향후 연구에서는 긍정적 리트윗과 부정적 리트윗을 구분하여 분석할 필요가 있다. 다수의 선행연구에서도 고객 인게이지먼트는 리트윗 수로 측정하였으나(e.g., Aleti, Pallant et al., 2019; Okazaki et al., 2015), 이러한 리트윗 중에는 부정적인 리트윗이 포함되어 있을 수 있다. 특히 본 연구에서 제시한 토픽4. send_dm, 토픽6. glad_hear, 토픽7. phone_number, 토픽8. direct_message, 토픽9. hear_dm, 토픽11. provide_info 등은 고객의 불만 행동과 관련된 주제로 부정적인 고객 인게이지먼트가 발생할 가능성이 크기 때문에 향후 연구에서 다룰 필요가 있다.

참고문헌

Aleti, T., J. I. Pallant, A. Tuan, & T. van Laer (2019), "Tweeting with the stars: Automated text analysis of the effect of celebrity social media communications on consumer word of mouth," *Journal of Interactive Marketing*, 48(1), pp.17-32.

Anandarajan, M., C. Hill, & T. Nolan(2019), Probabilistic topic models, In *Practical Text Analytics* (pp. 117-130). Springer, Cham.

Araujo, T., P. Neijens, & R. Vliegenthart(2015), "What Motivates Consumers To Re-Tweet Brand Content?: The impact of information, emotion, and traceability on pass-along behavior," *Journal of Advertising Research*, 55(3), pp.284-295.

Azagba, S., & M. F. Sharaf(2011), "Psychosocial working conditions and the utilization of health care services," *BMC Public Health*, 11(1), pp.1-7.

Ballestar, M. T., P. Grau-Carles, & J. Sainz(2018), "Customer segmentation in e-commerce: Applications to the cashback business model," *Journal of Business Research*, 88, pp.407-414.

Batra, R., & K. L. Keller(2016), "Integrating marketing communications: New findings, new lessons, and new ideas," *Journal of Marketing*, 80(6), pp.122-145.

Berger, J., & K. L. Milkman(2012), "What makes online content viral?," *Journal of Marketing Research*, 49(2), pp.192-205.

Blei, D. M., A. Y. Ng, & M. I. Jordan(2003), "Latent dirichlet allocation," *Journal of Machine Learning Research*, 3(Jan), pp.993-1022.

Breiman, L.(1996), "Bagging predictors," *Machine Learning*, 24(2), pp.123-140.

Breiman, L.(2001), "Random forests," *Machine Learning*, 45(1), pp.5-32.

Brodie, R. J, L. D. Hollebeek, B. Jurić, & A. Ilić (2011), "Customer engagement: Conceptual domain, fundamental propositions, and implications for research," *Journal of Service Research*, 14(3), pp.252-271.

Büschken, J., & G. M. Allenby(2016), "Sentence-based text analysis for customer reviews," *Marketing Science*, 35(6), pp.953-975.

Chakraborty, I., M. Kim, & K. Sudhir(2019), "Attribute Sentiment Scoring with Online Text Reviews: Accounting for Language Structure and Attribute Self-Selection", *SSRN Electronic Journal*, available at: http://doi.org/10.2139/ssrn.3395012.

Chawla, N. V., K. W. Bowyer, L. O. Hall, & W. P. Kegelmeyer(2002), "SMOTE: synthetic minority over-sampling technique," *Journal of Artificial Intelligence Research*, 16, pp.321-357.

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785-794).

Cruz, J. A., & D. S. Wishart(2006), "Applications of machine learning in cancer prediction and prognosis," *Cancer Informatics*, 2, 11769351 0600200030.

Cui, D., & D. Curry(2005), "Prediction in marketing using the support vector machine," *Marketing Science*, 24(4), pp.595-615.

De Vries, L., S. Gensler, & P. S. Leeflang(2012), "Popularity of brand posts on brand fan pages: An investigation of the effects of social media marketing," *Journal of Interactive Marketing*, 26(2), pp.83-91.

Friedman, J. H.(2002), "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, 38(4), pp.367-378.

Göçken, M., M. Özçalıcı, A. Boru, & A. T. Dosdoğru (2016), "Integrating metaheuristics and artificial neural networks for improved stock price prediction," *Expert Systems with Applications*, 44, pp.320-331.

Greene, W. H.(2003), *Econometric analysis*, Pearson Education India.

Guo, T., S. Sriram, & P. Manchanda(2018), "The effect of information disclosure on industry payments to physicians," *Available at SSRN 3064769*.

Hartmann, J., J. Huppertz, C. Schamp, & M. Heitmann(2019), "Comparing automated text classification methods," *International Journal of Research in Marketing*, 36(1), pp.20-38.

Heath, C., C. Bell, & E. Sternberg(2001), "Emotional selection in memes: the case of urban legends," *Journal of Personality and Social Psychology*, 81(6), pp.1028-1041.

Hoffman, D. L., & M. Fodor(2010), "Can you measure the ROI of your social media marketing?," *MIT Sloan Management Review*, 52(1), pp. 41-49.

Huang, D., & L. Luo(2016), "Consumer preference elicitation of complex products using fuzzy support vector machine active learning," *Marketing Science*, 35(3), pp.445-464.

Jacobs, B. J., B. Donkers, & D. Fok(2016), "Model-based purchase predictions for large assortments," *Marketing Science*, 35(3), pp.389-404.

Jalali, N. Y., & P. Papatla(2019), "Composing

tweets to increase retweets," *International Journal of Research in Marketing*, 36(4), pp.647-668.

Kanuri, V. K., Y. Chen, & S. Sridhar(2018), "Scheduling content on social media: Theory, evidence, and application," *Journal of Marketing*, 82(6), pp.89-108.

Kumar, V., J. B. Choi, & M. Greene(2017), "Synergistic effects of social media and traditional marketing on brand sales: capturing the time-varying effects," *Journal of the Academy of Marketing Science*, 45(2), pp.268-288.

Lessmann, S., J. Haupt, K. Coussement, & K. W. De Bock(2019), "Targeting customers for profit: An ensemble learning framework to support marketing decision-making," *Information Sciences,* In press.

Li, Y., & Y. Xie(2020), "Is a picture worth a thousand words? An empirical study of image content and social media engagement," *Journal of Marketing Research*, 57(1), pp.1-19.

Ling, X., W. Deng, C. Gu, H. Zhou, C. Li, & F. Sun(2017, April), "Model ensemble for click prediction in bing search ads," In *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 689-698.

Liu, X., D. Lee, & K. Srinivasan(2019), "Large-scale cross-category analysis of consumer review content on sales conversion leveraging deep learning," *Journal of Marketing Research*, 56(6), pp.918-943.

Malhotra, A., C. K. Malhotra, & A. See(2012), "How to get your messages retweeted," *MIT Sloan Management Review*, 53(2), pp.61-66.

Muntinga, D. G., M. Moorman, & E. G. Smit(2011), "Introducing COBRAs: Exploring motivations for brand-related social media use," *International Journal of Advertising*, 30(1), pp. 13-46.

Okazaki, S., A. M. Díaz-Martín, M. Rozano, & H. D. Menéndez-Benito(2015), "Using Twitter to engage with customers: a data mining approach," *Internet Research,* 25(3), 416-434.

Pansari, A., & V. Kumar(2017), "Customer engagement: the construct, antecedents, and consequences," *Journal of the Academy of Marketing Science*, 45(3), pp.294-311.

Samuel, A. L.(1959), "Some studies in machine learning using the game of checkers," *IBM Journal of Research and Development*, 3 (3), pp.210-229.

Seetharaman, P.(2020), "Business models shifts: Impact of Covid-19," *International Journal of Information Management*, 54, 102173.

Swani, K., & G. R. Milne(2017), "Evaluating Facebook brand content popularity for service versus goods offerings," *Journal of Business Research*, 79, pp.123-133.

Tirunillai, S., & G. J. Tellis(2014), "Mining marketing meaning from online chatter: Strategic brand analysis of big data using latent dirichlet allocation," *Journal of Marketing Research*, 51(4), pp.463-479.

Toubia, O., & A. T. Stephen(2013), "Intrinsic vs. image-related utility in social media: Why do people contribute content to twitter?," *Marketing Science*, 32(3), pp.368-392.

Trusov, M., L. Ma, & Z. Jamal(2016), "Crumbs of the cookie: User profiling in customer-base analysis and behavioral targeting," *Marketing Science*, 35(3), pp.405-426.

Tsai, C. F., & M. L. Chen(2010), "Credit rating by hybrid machine learning techniques," *Applied*

Soft Computing, 10(2), pp.374-380.

Vermeer, S. A., T. Araujo, S. F. Bernritter, & G. van Noort(2019), "Seeing the wood for the trees: How machine learning can help firms in identifying relevant electronic word-of-mouth in social media," *International Journal of Research in Marketing*, 36(3), pp.492-508.

Zhang, Y., W. W. Moe, & D. A. Schweidel(2017), "Modeling the role of message content and influencers in social media rebroadcasting," *International Journal of Research in Marketing*, 34(1), pp.100-119.

• The author Jungwon Lee is a Ph.D. Candidate of Corporate Management at Korea University and an Instructor at Sungsnin Women' University and Dankook University. He received his B.B.A in International Business from Chungbuk National University and M.S in e-business from Korea University. His research interests include digital marketing, social media, machine learning, and he has published papers in Korean Management Review, Korean Marketing Review, Journal of IT Service, and Information Systems Research.

• The author Cheol Park is a Professor of Global Business at Korea University Sejong. He received his B.A. in Economics, M.B.A. and Ph.D. in Business Administration from Seoul National University. He had worked for Samsung as assistant manager of global marketing team before joining academic area. He has been a visiting scholar at Vanderbilt University, University of Hawaii, Mongolia International University, and University of Jinan in China. His research interests include digital marketing and online consumer behaviors. He has published papers in influential journals such as International Journal of Information Management, Journal of Interactive Marketing, International Marketing Review, and Journal of Business Research.

〈부 록〉

〈Table 1〉 브랜드별 데이터

번호	기업명	브랜드 (M, $)	N	팔로워	개설일
1	Apple	234,241	NA	NA	NA
2	Google	167,713	3,194	22,102,494	2009-02
3	Amazon	125,263	3,039	3,341,970	2009-02
4	Microsoft	108,847	3,058	8,967,080	2009-09
5	Coca-Cola	63,365	3,058	3,315,373	2009-03
6	McDonald's	45,362	3,191	3,618,260	2009-09
7	Disney	44,352	3,159	6,774,564	2009-08
8	IBM	40,381	2,725	591,079	2009-01
9	Intel	40,197	3,037	4,820,218	2007-03
10	Facebook	39,857	3,177	13,414,131	2007-03
11	Cisco	35,559	3,174	705,791	2008-08
12	Nike	32,376	2,948	8,203,715	2011-11
13	Oracle	26,288	3,133	773,802	2007-03
14	GE	25,566	NA	NA	NA
15	American E.	21,629	3,174	866,777	2009-05
16	Pepsi	20,488	2,822	2,993,972	2008-12
17	J.P. Morgan	19,044	3,142	527,600	2013-02
18	UPS	18,072	3,192	236,997	2010-06
19	Accenture	16,205	3,180	501,681	2007-11
20	Budweiser	16,018	1,475	17,497	2014-02
21	Pampers	15,773	2,799	154,375	2009-07
22	Ford	14,325	3,135	1,248,384	2008-07
23	Gillette	13,753	3,178	131,577	2009-04
24	Adobe	12,937	3,135	665,332	2009-08
25	Citi	12,697	3,184	915,275	2009-10
26	eBay	12,010	3,172	725,011	2009-01
27	Starbucks	11,798	3,058	11,066,657	2006-11
28	Goldman. S.	11,352	3,023	803,182	2011-02
29	HP	10,891	3,122	1,092,595	2008-11
30	Visa	10,756	2,838	379,367	2012-04
31	Kellogg's	10,419	3,189	100,627	2012-03
32	Mastercard	9,430	2,782	479,619	2009-09
33	Dell	9,086	3,068	735,602	2009-07
34	3M	9,035	3,062	1,419,548	2011-09
35	Netflix	8,963	2,968	8,361,291	2008-10
36	Colgate	8,824	2,690	75,483	2009-02
37	Morgan. S.	8,185	3,196	568,816	2011-12
38	Salesforce	8,004	3,044	511,350	2009-04
39	HP	7,909	3,157	70,629	2010-10
40	PayPal	7,604	3,179	639,574	2009-04
41	FedEx	6,998	3,185	298,731	2010-04
42	Caterpillar	6,791	3,062	136,772	2008-06
43	Corona	6,369	3,035	43,027	2011-04
44	Jack Daniel's	6,347	3,170	200,130	2010-09
45	DHL	5,987	3,195	32,634	2014-05
46	John Deere	5,883	2,969	190,556	2009-01
47	J &J	5,720	3,193	198,220	2009-02
48	Uber	5,714	2,712	1,042,019	2009-01
49	Discovery	5,525	3,140	8,707,686	2008-12
50	KFC	5,509	3,123	1,429,795	2008-07
51	Tiffany &Co.	5,335	3,165	1,634,884	2009-11
52	LinkedIn	4,836	3,081	1,551,345	2008-02
53	Harley-David	4793	2,818	445,259	2008-11
총합			154,705		

주. 53개 브랜드 중 데이터를 수집할 수 없는 Apple과 GE를 제외하여 51개 브랜드 데이터를 분석하였다.