JDCS
디지털콘텐츠학회논문지
Journal of Digital Contents Society
Vol. 22, No. 9, pp. 1477-1486, Sep. 2021

Check for updates

토픽 모델링을 활용한 게임 리뷰 데이터의 감성 분류

김 태 국[1] · 김 길 환[2*]
[1]상명대학교 경영공학과 석사과정
[2*]상명대학교 경영공학과 부교수

Sentiment Classification of Game Review Data Using Topic Modeling

Tae-Kook Kim[1] · Kilhwan Kim[2*]
[1]Master's Course, Management Engineering, Sangmyung University, Republic of Korea
[2*]Associate Professor, Management Engineering, Sangmyung University, Republic of Korea

[요 약]

온라인 시장이 성장함에 따라 사용자의 경험은 리뷰로 표출되고 있고, 비사용자는 사용자의 리뷰를 보고 구매를 하는 경향이 많다. PC게임 시장의 경우 사용자가 직접 구매를 해 이용할 수 있는 시장이기 때문에 더욱 리뷰에 의존하게 된다. 이에 본 연구는 텍스트 마이닝 기법을 이용하여 게임 리뷰를 긍정과 부정으로 분류하는 방법을 제안한다. 이를 위해 최근 3년간의 상위 게임의 리뷰를 수집하고, 토픽 모형을 이용하여 게임 장르 별로 리뷰의 주요 토픽을 파악한다. 그리고 토픽 모형의 결과를 이용하여 다양한 분류기법으로 긍정 및 부정 리뷰를 분류하는 모형을 생성한다. 분류모형은 게임 장르에 따라 67~73%의 분류 정확도를 보였다. 이러한 연구 결과를 게임 산업에서 활용한다면 게임 수요자의 평가를 신속히 반영하는 데 도움이 되리라 기대된다.

[Abstract]

As the online market grows, users' experiences are expressed as reviews, and prospective users tend to make their purchases after seeing other users' reviews. Since PC games are directly purchased online by users, prospective users tend to rely more on reviews. In response, this work proposes a way to classify game reviews as positive and negative using text mining techniques and identify the main factors in game selection. We collect reviews of top games over the past three years and employ the topic model to identify the main topics of reviews by games genre. We then use the results of the topic model to generate a model that classifies positive and negative reviews using various classification techniques. The classification model result in 67~73% classification accuracy depending on the game genre. If the results are used in the game industry, it is expected to help develop and market games that quickly reflect customer requirements by analyzing game consumers' evaluations.

색인어 : 게임 리뷰, PC 게임, 감성 분류, 텍스트마이닝, 토픽 모델링

Keyword : Game Review, PC games, Sentiment Classification, Text Mining, Topic Modeling

http://dx.doi.org/10.9728/dcs.2021.22.9.1477

Received 04 August 2021; **Revised** 23 August 2021
Accepted 23 August 2021

***Corresponding Author; Kilhwan Kim**

Tel: +82-41-550-5444
E-mail: khkim@smu.ac.kr

Ⅰ. 서 론

오늘날 게임 산업은 대표적인 고부가가치 콘텐츠 산업이다. 콘솔, PC, 모바일 등의 다양한 방식의 게임들이 존재하지만 1997년 외환위기 이후에 PC방 창업이 급증하면서 PC 게임을 중심으로 게임 산업이 발전하였다. 게임 산업 전반의 매출 상향 속에서 2018년 이후에 다소 두드러지는 흐름은 플랫폼 비중의 다변화다[1]. 2010년대 들어 큰 폭의 성장세를 이루며 게임 산업 전반의 매출을 견인해 왔던 모바일게임은 여전히 높은 매출 비중을 차지하고는 있지만, 전년 대비 비중은 정체나 감소한 결과를 드러냈다. 물론 아직 국내 콘솔 시장은 여전히 전체 점유율에서 낮은 수준이지만 한동안 매출의 중심을 이루었던 모바일 플랫폼의 성장 둔화가 예견되는 시점에서 플랫폼 다변화는 어느 게임사건 고려하지 않을 수 없는 변수가 되었다[2]. 2018년 게임 산업의 게임 제작 및 배급업이 86.7%의 비중을 차지하는 가운데 모바일게임이 46.6%를 차지하고 그 뒤를 이어 PC게임이 35.2%를 차지하고 있다. 이처럼 모바일 게임의 성장세가 높긴 하지만 PC게임의 매출액 또한 꾸준히 성장하는 추세이다[1]-[2].

한편 소비재 시장에서 온라인 리뷰에 대한 접근성이 증가함에 따라 다양하고 방대한 리뷰들이 작성되고 있고, 온라인 리뷰는 소비자에게 유용한 정보를 제공함과 동시에 기업이 소비자와 소통하고 질 높은 상품을 만들 수 있도록 소통의 장을 제공하였다. 연구조사기업 Channel Advisor에서 2010년과 2011년에 발표한 소비자 행위 보고서에 따르면 상품 리뷰가 온라인 소비자의 구매 결정에 미치는 영향은 점점 더 커지고 있는데, 2011년 보고서에서는 설문조사에 참여한 이용자의 74%가 상품 리뷰를 읽어본다고 응답했다[3]. 그중 43%의 이용자가 상품 리뷰 정보가 구매 결정에 영향을 미쳤다고 응답했다. 소비자들은 제품을 구매할 때 정보를 탐색하고 대안을 비교해 최종적으로 하나의 제품을 선택하는 과정을 거친다[3]. 이미 여러 선행 연구들에서 밝혀진 바와 같이 [1]-[3], 온라인 리뷰어의 긍정적 또는 부정적 상품평이 소비자들의 정보탐색 과정은 물론이고, 제품의 구매 결정, 해당 온라인 기업의 이미지 등에 커다란 영향을 미치고 있으며, 인터넷에서 제공하는 사용 후기나 전문가들의 추천 등이 일반적인 소비자들의 제품 선택과정에서 결정적인 선택기준으로서 작용하고 있다. 온라인상에서 얻을 수 있는 게임 평가 정보는 리뷰, 제품 정보 등 다양화되어있다.

마찬가지로, 게임 산업에서 소비자들의 리뷰는 게임 제작에 많은 영향을 끼치고 있다. 그러므로 최근에 게임 리뷰 데이터를 분석하는 연구가 활발히 진행되었다 [4]-[6]. 하지만 게임 분야에서는 기존 출시된 게임에 대한 게임 평론가, 전문가와 소비자의 평이 다른 게임도 많고, 소비자들이 게임을 구매할 때 게임의 어떤 요소들이 중점적으로 영향을 받는지 알기 어려우므로, 사용자의 리뷰에 나타난 감성을 분류하고 그러한 사용자 감성에 자리 잡은 심층적 요인이 무엇인지 파악할 필요가 있다. 특히, 게임에 대한 리뷰는 작성자의 평점이 같이 제공되어 리뷰가 긍정적 리뷰인지 부정적 리뷰인지를 쉽게 파악할 수 있는 예도 있지만(https://www.metacritic.com 등), 평점이 부여되지 않는 리뷰만 제공되는 예도 있다 (www.gamemeca.com 등). 또한 게임 리뷰는 트위터, 페이스북 등 다양한 SNS에 상에도 게재가 된다. 따라서 다양한 원천에서 획득되는 사용자의 게임 리뷰를 신속하게 긍정과 부정 리뷰로 나누고, 긍정과 부정 리뷰에 나타난 주요 토픽이 무엇인지를 파악하는 것은 매우 중요하다. 그러나 기존 연구들은 리뷰의 평점에 국한된 단순한 감성 분석에 치우치거나 [4], 리뷰에 잠재된 토픽이나 관심 요소를 도출하였지만 이를 리뷰의 감성을 예측하는 연구로 발전하지 못했다[5]-[6].

이에 본 연구는 최근 3년간 게임소비자들이 높은 평가를 한 상위 200개의 게임의 리뷰데이터를 기반으로 토픽 모델링을 적용한 후, 소비자들이 최근에 평가를 한 게임들의 주요 토픽들을 분석하고, 긍정, 부정 토픽에 해당하는 게임 요소들이 무엇인지 분석하여 고객이 게임의 구매 시 가장 영향을 끼치는 요소들을 확인한다. 더 나아가 토픽을 예측 변수로 이용하여 새로운 리뷰에 잠재된 토픽을 분석하고 자동으로 리뷰의 감성을 분류하는 모형을 개발한다. 아울러 제시된 토픽 모형 기반의 감성 분석 모형과 기존의 감성 분석 모형의 성능을 비교·분석한다. 이를 통해 제시된 감성 분석 방법이 기존 연구와 비교하면 두 가지 장점이 있다는 것을 보이려고 한다. 첫째는 정성적 장점으로, 이 연구에서 제시하는 방법은 사용자 리뷰의 긍·부정 분류가 토픽에 근거함으로써 사용자의 만족과 불만족에 근원이 되는 심층적 요인(토픽)이 무엇인지 파악할 수 있다는 것이다. 둘째는 정량적 장점으로, 토픽 모형을 사용한 감성 분류가 기존의 방식에 비교해 더 짧은 데이터 처리 시간에도 불구하고 분류의 정확도가 기존 방식과 같거나 더 나은 결과를 보인다는 것이다(표 7과 표 8 참조). 따라서 본 연구에서 제시된 방법이 활용되면 향후 PC게임 분야 전문가가 PC게임의 주요 만족, 불만족 요소가 무엇인지를 실시간으로 심층적으로 파악할 수 있게 함으로써, 소비자들의 니즈를 충족시킬 게임을 개발하는 데 소비자 의견을 분석, 활용하는 주기를 단축하는 데 도움을 줄 수 있을 것이다.

Ⅱ. 관련 연구

게임 분야의 온라인 리뷰를 통해 사용자의 니즈가 무엇인지, 게임에 대해 얼마나 만족하는지에 대한 연구가 활발히 진행됐다. B.Straat et al[4]는 두 개의 게임 프랜차이즈의 모든 사용자 리뷰를 수집하여, 감정 분석을 하여 데이터 세트가 충분히 큰 경우 사용자 리뷰의 높은 등급이 해당 감성과 상관관계가 있음을 보여주고, 기존 게임의 성공 또는 실패 요인에 관해 기술하였다. 게임 플랫폼인 스팀(STEAM)의 사용자 리뷰데이터를 이용한 연구도 많이 진행되고 있는데, M.Y.Wui et al[5]는 VR(Virtual Reality) 산업의 성장세가 주춤한 원인이 사용자를 끌어들일 수 있는 콘텐츠가 부족하다고 보고 연

구를 진행하였다. 리뷰데이터에 텍스트 마이닝과 네트워크 분석을 적용해 VR 게임 사용자의 관심 요소에 대해 연구하였다. T.S.Kim[6]은 모바일 게임의 온라인 리뷰 분석을 통해 사용자 관점의 중요 게임 요소와 토픽을 파악하고, 액션 장르 3가지의 모바일 게임의 온라인리뷰를 사례로 연구를 진행해 사용자가 관심을 두는 게임 요소와 토픽을 추출하고 토픽별 만족도를 분석했다. 그러나 서론에서도 밝혔듯이, 기존 연구들은 리뷰의 평점에 국한된 단순한 감성 분석에 치우치거나[4], 리뷰에 잠재된 토픽이나 관심 요소를 도출하였지만 이를 리뷰의 감성을 예측하는 연구로 발전하지 못했다 [5]-[6]. 본 연구는 게임 리뷰 데이터에 대하여 토픽 모델링을 기반으로 감성 분류 모형을 개발하였다는 데에서 기존 연구와 차별성이 있다.

텍스트 마이닝의 기법의 하나인 토픽 모델링은 여러 문서의 텍스트에서 의미 있는 토픽을 추출하는 것을 의미한다. 통계적 추론을 이용한 방법으로 문장들 속 잠재된 주제들을 찾기 위해 고안된 방법이다. 특히 데이터의 특성에 따른 주제 비교에 매우 효과적인 방법이다.

표 1. 온라인 리뷰 및 게시물을 분석한 기존 연구
Table 1. Existing research analyzing online review and posts

Author	Data	Research Purpose
M.Y.Wui et al[5]	Steam Game Review	Identifying factors of interest in VR games
T.S.Kim[6]	Mobile game Review	Identifying public opinion in the mobile game market
J.E.Park[7]	Amazon eco Review	Smart speaker UX improvement
S.H.Park[8]	Levothyroxine Review	Derivation of customized medication guidance plan
D.W.Kim et al[9]	Jobplanet Review	Comparison of Job Satisfaction Factors
S.H.Park[10]	Amazon Review	Presenting a rating prediction model for review

이런 장점 때문에 최근 많은 연구가 온라인 리뷰 분석을 위해 토픽 모델링 기법을 활용 중이다[6]-[10]. 토픽 모델링은 문서 내에서 맥락과 관련 있는 단어들을 이용해 유사한 의미가 있는 단어를 군집화하는 방식을 통해 주제를 추론, 확률 불포화 할 수 있다[11]-[13]. 리뷰를 이용해 토픽을 찾아내는 토픽 모델링을 이용한 기존 연구를 표 1에 정리하였다.

그런데 토픽 모델을 사용하여 리뷰를 분석한 대부분의 연구도 토필 모형을 사용하여 잠재된 토픽을 찾는 데 주안점을 두었고 토픽 모델의 결과를 사용하여 감성 분류모형을 개발하지는 않았다[5]-[9]. 반면 S.H.Park[10]의 연구는 토픽 모형을 사용하여 리뷰의 평점을 예측하였다는 데에서 본 연구와 방법론의 유사성이 있으나, 단어-토픽 행렬을 인공신경망 모형으로 학습해 평점이라는 점수를 예측하는 회귀 모형을 구축한 데 반해 본 연구에서는 긍·부정 리뷰를 분류하는 분류모형을 구축하였다는 데에 차이점이 있다. 아울러 본 연구는 토픽 모델링의 단어-토픽 확률값을 이용해 부정, 긍정을 나타내는 토픽과 토픽에 분포된 단어들을 파악해 장르 별 게임에 대해 개선점을 도출할 수 있다는 점이 S.H.Park[10]와의 차이점이라 할 수 있다. 아울러 본 연구에서는 토픽 모형을 사용한 감성 분류모형만을 제시하는 것에 그치지 않고, 토픽 모형을 사용하지 않는 분류모형과의 성능 차이를 비교하였으며, 비지도 학습에서의 최적의 토픽의 수와 분류 등의 지도학습에서 사용하기 위한 최적의 토픽 수가 다를 수 있다는 점을 보인 점도 S.H.Park[10]의 연구와 차별점이 있다.

토픽 모델링을 대표하고 기초가 되는 LDA(Latent Dirichlet Allocation)는 D.Blei[11]이 제안한 방법으로 디리클레 분포에 기반을 둔 확률적 토픽 모델링 알고리즘이다. LDA 모형은 관련된 단어가 토픽별로 속할 확률을 계산, 계산된 단어 분포를 바탕으로 문서를 분석함으로 문서가 어떤 토픽을 가지고 다루고 있는지 예측할 수 있다. LDA는 텍스트를 단어의 순서는 무시한 단어 주머니(bag of words)로 간주한다.

그림 1은 LDA 모형의 대략적인 구조를 보여준다. β는 토픽의 단어 분포 확률이며, M은 문서의 개수, Θ는 토픽으로 구성되는 문서, z는 단어들로 구성되는 토픽, ω는 실제 관측된 단어, α는 디리클레 분포의 모수, 사전확률을 뜻한다. 여기서 LDA는 각각의 문서에 대해 다음과 같은 생성과정을 가정한다.

(1) 총 단어 수 $N \sim Passion(\xi)$을 선택
(2) 문서 Θ 선택 $\Theta \sim Dir(\alpha)$
(3) N 단어 ω_n 각각은 (3-A)와 (3-B)의 과정에 의해 생성
(3-A) 토픽 $Z_e n \sim Mltinomial(\Theta)$
(3-B) 토픽 $Z_e n \sim P(\omega_n | Z_e n, \beta)$에서 단어 ω_n 을 선택

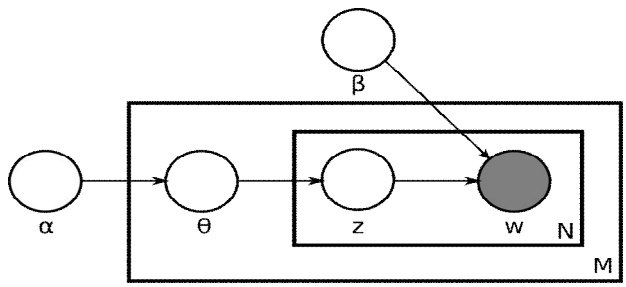

그림 1. LDA 모델 [11]
Fig. 1. Graphical representation of the LDA model [11]

단어 ω_n은 우리가 관측할 수 있는 값이고 α는 K개의 양의 정수로 이루어진 벡터의 값이고, 나머지는 관찰 불가능한 변수이다. 잠재변수 Θ의 분포로 디리클레 분포를 사용하는 이유는 사후 확률의 분포를 추정하는 데 있어 쉽기 때문이다. LDA 기반 토픽 모델링의 특징은 문서의 토픽 혹은 정의된 토픽과 키워드를 대상으로 학습 과정 없이 문서 속에서 숨겨진 주제의 구조를 찾는 것이다. 자주 나타나는 단어의 그룹을 하나의 토픽으로 간주한다. 토픽 모델링은 문서 집합에서 단어의 동시 출현을 바탕으로 같은 의미 부류에 속하는 용어를 토픽으로 묶어준다. LDA가 산출한 토픽이 얼마나 잘 산출되었는지를 평가하는 지표로 Perplexity가 있다.

Perplexity는 사전적으로 혼란도(혼란한 정도)라고 한다. 이 수치는 특정 확률 모델이 실제로 관측되는 값을 얼마나 잘 예측하는지를 평가할 때 사용한다. 토픽 모형도 문서 집합 내 단어 출현 횟수를 바탕으로 하는 문서 내 토픽 출현 확률과 토픽 내 단어 출현 확률을 가지고 있는 확률 모형이므로, 확률 모형 평가 척도를 이용한다. Perplexity의 정의는 다음 (1) 과 같다.

$$Perplexity = 2^{\frac{-\sum LL(w)}{N}} \quad (1)$$

단, $LL(w)$는 로그 우도로써, 토픽 모형 내에서 특정 단어가 해당 토픽으로 부여될 확률값에 로그를 취한 값이고, 이 값이 작을수록 해당 토픽 모형은 실제 결과를 잘 반영한다는 것으로 학습이 잘 되었다고 평가할 수 있다. 이 값은 LDA 등 적절한 K를 정하기 어려울 때 유용하게 쓰인다.

그런데 흔히 비지도 학습으로 수행된 최적 결과가 연구의 최종목적인 분류와 회귀 모형에서 최적이 아닐 수 있다. 따라서 본 연구에서는 연구 방법에서 제시된 방법을 사용하여 감성 분석에 최적인 토픽의 수를 정한다.

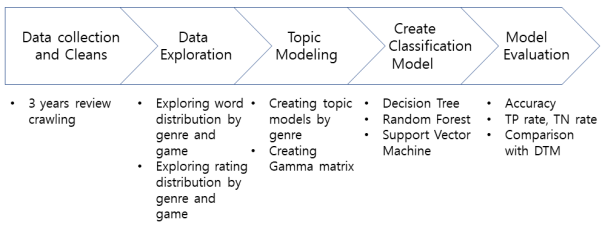

그림 2. 연구 과정
Fig. 2. Research process

Ⅲ. 연구 방법

본 연구는 그림 2와 같이 온라인 리뷰 데이터수집과 정제 과정을 거쳐, 토픽 모델링 분석 및 분류 모델 생성과 모델 평가 순서로 진행되었다.

최근 3년간 상위 200개의 게임에 대한 온라인 리뷰를 수집했다. 게임의 장르와 관계없이 PC게임으로 지정하여 추출하여 리뷰를 수집했다. 최초 수집 시점인 2017년부터 2019년까지 출시한 게임에 대한 모든 리뷰내용과 ID, 해당 리뷰의 평점을 수집했다. 장르별 수집 결과는 33,075개의 리뷰이다.

정제 과정은 컴퓨터가 데이터를 처리하기 쉽게 변환하는 작업이다. 온라인 리뷰는 대부분이 정제되지 않은 데이터이기 때문에 데이터 그대로 분석에 이용하게 되면, 컴퓨터가 불필요한 결과를 도출하게 되므로 특수문자 등은 제거되어야 한다. 수집된 데이터는 전치사 같은 불 용어를 제거하고 'game','play' 같은 전체 리뷰에서 높은 비율을 가지는 공통 단어는 문서에서 상당한 비율을 차지하지만, 문서 파악에 있어 의미가 없어 문서별로 50% 이상 출현한 단어는 제거했다.

데이터 탐색은 데이터를 분석하기 위해서 시각화하고 측정값을 여러 가지 기법을 사용해 탐색하는 것이다. 분석에 앞서서 변수의 유형에 따라서 변수의 속성을 파악한다. 먼저 전체 데이터와 장르별로 단어와 평점 분포를 살펴본다. 그리고 평점별 그룹을 나누고 상, 중, 하 그룹에 대한 분포를 살펴본다.

토픽 모델링은 수집한 리뷰 텍스트 내에서 토픽 추출을 목적으로 하며 본 연구는 텍스트마이닝 기법의 하나인 LDA를 사용한다. 각 문서에 토픽이 존재하는 확률을 추론하고, 문서의 토픽 분포와 토픽의 단어 분포가 존재하며 두 분포 모두 디리클레 분포를 따른다고 가정한다. LDA 분석에서는 토픽의 개수(K)를 지정해야 한다. 최적의 토픽 수(K)를 구하는 방법으로 Perplexity 지표를 참고한다. 본 연구에서는 먼저 최종 분류모형과 무관하게 교차 검증을 이용해 Perplexity를 계산해 최적의 토픽 수 K를 구한다. 이후 토픽 모델링의 결과를 이용하여 문서-토픽 행렬을 생성한다. 생성된 토픽 모델을 이용하여 사용자 리뷰에 잠재된 주요 토픽의 특성을 살펴본다.

대표적인 분류기법인 의사결정 나무, 랜덤 포레스트와 서포트 벡터 머신(SVM; Support Vector Machine)을 이용해 감성 리뷰를 분류하는 분류모형을 생성한다. 의사결정 나무는 데이터가 가진 속성으로 분할기준을 판별하고 분할기준 속성에 따라 트리 형태로 모델링하는 분류 예측 모형이다 [14]-[15]. 수치형 변수와 변수 모두 사용이 가능하다. 하지만 연속형 변수를 비연속적인 값으로 취급하기 때문에 분리 지점에서는 예측 오류가 클 가능성이 있다. 랜덤 포레스트는 더 많은 무작위성을 주어 1개의 나무가 아닌 여러 개의 나무를 이용한 기계학습 기법이며 예측력이 매우 높다 [14]-[15]. SVM은 데이터를 분리하는 초평면(Hydroplane) 중에서 데이터들과 거리가 가장 먼 초평면을 선택해 분리하는 지도학습 기반의 이진 선형 분류모형이다 [14]-[15]. 본 연구는 각 분류기를 활용해 토픽 모형의 결과인 문서-토픽 행렬을 이용하여 감성 리뷰를 분류하는 감성 분류모형을 개발한다. 그런데 비지도 학습인 LDA에서 최적인 토픽 수가 분류모형의 성능에 최적이 아닐 수 있다. 따라서 교차 검증을 통해 분류 성능을 최적화하는 토픽 수 K를 결정하는 과정을 거친다. 그리고 각 감성 분류모형에서 감성 분류에 가장 중요한 2개의 토픽을 살펴봄으로써 사용자의 감성에 잠재된 주요 토픽이 무엇인지를 살펴본다.

토픽 모형을 이용한 분류모형의 성능을 파악하기 위해 그림 3과 같은 혼동 행렬을 이용한 평가지표를 사용한다. 혼동 행렬은 분류모형에서 예측한 범주와 데이터의 실제 분류 범주를 교차 표 형태로 정리한 행렬이다. 본 연구에서는 정확도와 TP rate, TN rate를 이용한다. 정확도는 실제 범주를 정확하게 예측한 비율로 전체 예측 중 참 긍정(TP)과 참 부정(TN)이 차지하는 비율이다. TP rate는 실제 긍정인 범주 중 긍정을 바르게 예측한 비율이다. TN rate는 실제 부정인 범주 중 부정을 바르게 예측한 비율이다. 본 연구에서는 이 세 가지의 평가지표를 이용해 개발된 감성 분류모형의 성능을 비교해본다.

아울러 토픽 모형을 이용한 감성 분류모형과 토픽 모형을 이용하지 않고 문서-토픽 행렬을 이용한 감성 분류모형의 성

능 차이를 파악하기 위해, 앞에서 논의한 세 가지 성능 지표와 처리 시간을 측정하여 비교 분석한다.

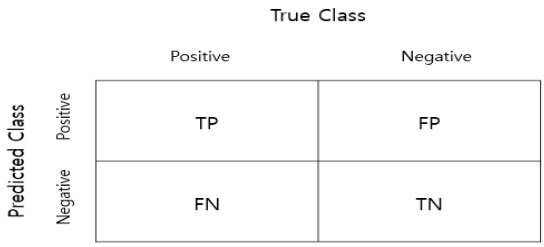

그림 3. 혼동 행렬
Fig. 3. Confusion matrix

Ⅳ. 데이터 탐색 및 전처리

토픽 모델링을 진행하기 위해 메타크리틱(https://www.metacritic.com)에서 3년간 상위 200개의 게임에 대한 리뷰데이터를 추출하였고 데이터 수집 결과는 표 2와 같이 22개 장르의 총 33,075개의 리뷰가 수집되었다. 수집된 데이터 중 영어가 아닌 리뷰는 제거하였다. 이후 데이터 토큰화를 실시하고 토큰화 과정 중 의미가 없는 불용어는 제거하였다. 이후 32,579개의 문서별로 상위 50% 이상 공통으로 출현한 단어는 유의미한 결과를 낼 수 없다고 판단해 제거하였다. 하지만 'game'과 'games'는 다른 단어이지만 같은 의미의 단어라 볼 수 있으므로 50%가 되지 않아도 제거하였다. 데이터 정제 결과는 아래 표 3과 같다.

표 2. 데이터 수집 결과
Table 2. Data collection results

Genre	Number of Review	Average word count	Rating Average
Action	10,867	50.15	5.07
Action Adventure	9,697	47.01	5.83
Sports	242	64.73	3.28
Strategy	1,970	61.22	6.43
Role Playing	4,774	64.69	6.75
17 other Genres	5,525	66.92	6.55
Total	33,075	59.12	5.65

표 3. 데이터 정제 결과
Table 3. Data refinement results

Refinement	Refined number of Review
Removing non-English Review	33,075 review -> 32,579 review
Tokenization	32,579 review -> 4,786,909 words
Removing stopwords	4,786,909 words -> 1,779,772 words
Removing common words	1,779,772 words -> 1,593,105 words

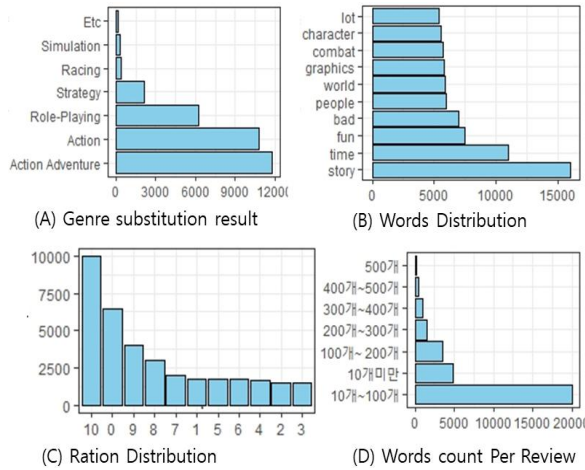

그림 4. 전체 데이터 통계
Fig. 4. Data statistics

Arcade나 Horror 장르의 경우 인디 게임인 경우가 많아 리뷰의 개수는 적다. 본 연구의 데이터로 사용하기 위해 리뷰의 개수가 적은 장르의 경우 다른 장르로 대체하였다. 예를 들어 A 게임의 경우 Action 장르도 가지고 있으면서 Arcade 장르도 가질 수 있다. 그래서 소수의 리뷰를 가진 장르의 경우 명시되어 있는 다른 장르가 있으면 다른 장르로 대체하였다.

장르를 대체한 후 결과는 그림 4의 (A) 와 같다. 22개의 장르에서 7개로 축소되었다. Etc 의 경우 대체가 불가능한 Sports 장르이거나 Puzzle 장르 등이 포함되어 있다. 데이터 탐색을 위해 전체장르의 단어 분포와 평점 분포, 리뷰가 많은 3개 장르의 단어 분포와 평점 분포를 살펴본다.

표 4. 장르별 데이터 탐색
Table 4. Data exploration by genre

Genre	Rating Average	Standard Deviation	Top words
Action	5	16.41	Story, time, fun
Action Adventure	6.07	16.82	Story, time, character
Role Playing	6.75	18.07	Story, character, combat

표 5. 그룹별 리뷰의 개수
Table 5. Number of Review per group

Rating	Group	Number of Review
0~1	Negative	8,443
2~7	Neutral	8,000
8~10	Positive	16,136

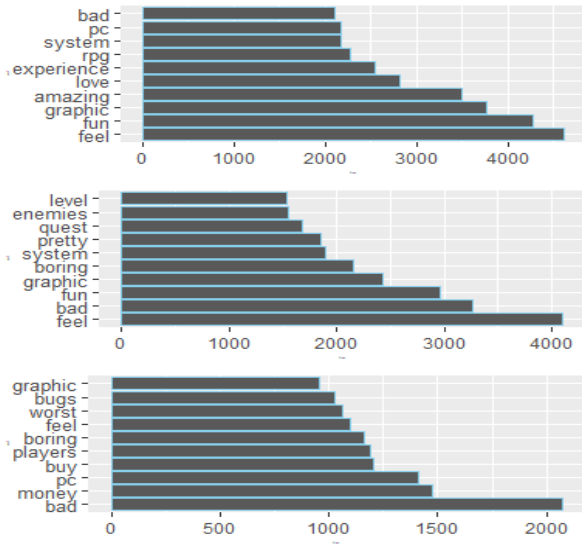

그림 5. 상, 중, 하 그룹의 상위 10개 단어 분포
Fig. 5. Distribution of top 10 words by group

전체 리뷰의 단어 분포는 그림 4의 (B) 와 같으며 보았을 때 단어 story가 15,690개로 가장 많이 출현하였고 뒤를 이어 time, fun이 나타났다. 전체적으로 보았을 때 불용어 제거가 끝난 데이터이기 때문에 게임에 관련된 평가 단어가 대부분을 차지했다.

평점 분포는 그림 4의 (C) 와 같다. 0점, 10점이 대부분을 차지하고 있으며 중위수는 7, 3사분위수가 10으로 높은 평점이 리뷰 중 많은 데이터를 포함한다는 것을 알 수 있었다.

단어-문서행렬로 살펴본 리뷰 별 단어 개수는 그림 4의 (D)와 같다. 10개 미만의 단어를 포함한 리뷰가 5,000건 정도이고 10개 이상 100개 미만의 단어를 포함한 리뷰가 대부분인 것을 알 수 있다. 이에 10개 미만의 단어도 분석에 영향을 미칠 수 있으므로 3개 미만의 단어를 가진 리뷰도 포함한다.

3가지 장르에 대해 평점 분포를 살펴보았을 때 0점과 10점에 대해 치우침이 심해 극단성을 가지고 있는 분포가 대부분이다. 이에 변수별로 상관관계를 알기 위해 0~1점과 2~7점, 8~10점으로 그룹으로 상, 중, 하 그룹을 나누어 고 평점과 저 평점의 단어 수와 단어 분포를 살펴본다. 먼저 평점 상, 중, 하 그룹의 리뷰 개수를 살펴보면 표 5와 같다.

표 4와 표 5는 분석을 시행할 데이터의 기초적인 정보다. 세 장르의 평균 평점은 5점~6점이고, 가장 많은 단어는 'story','time'이 공통으로 많이 출현하고 있다. 중, 하 그룹의 리뷰 개수가 상위 그룹의 리뷰의 개수의 1/2수준이다. 총 32,579개의 리뷰 중 75%에 해당하는 리뷰가 상, 하위 그룹으로 나뉘어 극단적으로 분포하고 있다.

상 그룹(평점 8~10점)의 단어 분포는 그림 5의 위와 같다. 전체 단어 분포와 비슷한 단어 분포를 나타내고 있으며 'story'의 출현이 가장 많고 'fun', 'amazing'과 같이 긍정의 단어가 많이 출현하고 있다. 하 그룹(평점 0~1점)의 단어 분포를 살펴보면 'bad', 'boring', 'worst' 등 부정의 단어가 많이 출현하고 있다.

Ⅴ. 토픽 모델링 및 감성 분류 결과

상위 3개의 장르의 토픽 모델링을 실시한다. 먼저 장르별로 7:3의 비율로 훈련 집합과 테스트 집합으로 나누고 70%의 훈련 집합에서 토픽 모델링의 토픽 개수인 최적의 K를 찾기 위해 5-fold 교차 검증을 하여 토픽의 개수에 따른 각 훈련 집합을 생성해 검증 집합에 대해 토픽 모델링 성능 지표인 Perplexity를 계산해 최적의 K를 찾고 최적의 토픽 모형의 개수를 결정한다. 이후 토픽 모델링의 결과를 이용해 문서-토픽 행렬을 추출한 후 문서-토픽 행렬의 값 이용해 긍정, 부정 분류를 모형을 생성한다. 그림 6은 Role Playing 장르의 토픽 모델링의 Perplexity 결과이다.

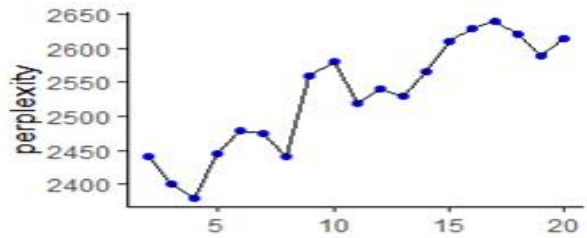

그림 6. Role Playing 장르의 Perplexity 결과
Fig. 6. Perplexity results for role-playing games

그림 7. Role Playing 장르의 LDA 결과
Fig. 7. LDA results for role-playing games

Role Playing 장르의 Perplexity 결과로 4개의 토픽으로 모델링을 실시하는 것이 분석하는 데 더 좋은 결과를 가지고 올 것으로 보아 Role Playing 장르의 토픽 개수는 4개 토픽으로 선정하였다.

그림 7에서 볼 수 있듯이, Role-Playing 장르의 토픽 모델링 결과는 1번 rpg와 관련된 긍적 토픽, 2번 토픽의 경우 경험에 대한 토픽, 3번 토픽은 게임 시스템에 대한 토픽, 4번 토픽의 경우 확장팩, 퀘스트등 게임 진행에 대한 토픽을 나타내고 있다. 토픽 모델링의 결괏값인 문서-토픽 행렬의 값과 기존의 리

뷰 점수를 이용해 문서-토픽 행렬의 기존 리뷰 점수를 태깅한 후 8점~10점인 상 그룹은 “긍정”, 2~7점인 중 그룹은 “중립”, 0~1점인 하 그룹은 “부정”으로 변환한다. 이후 태깅 값을 예측 변수로 각 문서-토픽 행렬의 값으로 분할 예측을 시행한다.

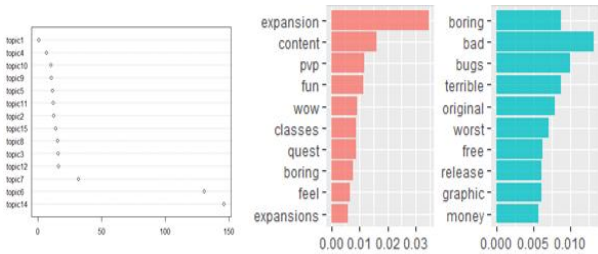

그림 8. Role Playing 장르의 의사결정 나무 결과
Fig. 8. Decision tree results for role-playing games

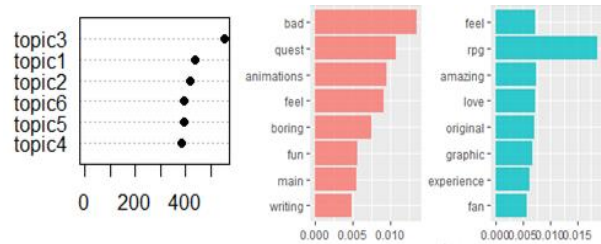

그림 9. Role Playing 장르의 랜덤포레스트 결과
Fig. 9. Random forest results for role-playing games

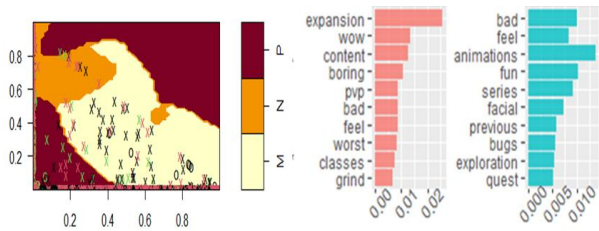

그림 10. Role Playing 장르의 SVM 결과
Fig. 10. SVM results for role-playing games

표 6. Perplexity vs 지도학습
Table 6. Perplexity vs Supervised Learning

Classification method	Genre	Perplexity		Supervised Learning	
		Topic	Accuracy	Topic	Accuracy
Decision Tree	Action Adventure	4	62%	8	68.12%
	Action		60%	6	67.31%
	Role Playing		61%	15	67.87%
Random Forest	Action Adventure		65%	8	70.04%
	Action		62%	4	71.33%
	Role Playing		64%	6	71.02%
SVM	Action Adventure		65%	3	71.88%
	Action		63%	4	72.66%
	Role Playing		64%	11	72.08%

그림 8은 분류모형의 기본 모형인 의사결정 나무를 실시한 결과이다. 완전 나무의 경우 분할이 진행될수록 적은 데이터로 분할이 이루어지며 과적합이 일어나는 경우가 많다. 따라서 가지치기를 사용한다. 가지치기는 완전 나무의 CP 값을 이용해 최적의 CP 값을 정한다.

Role-Playing 장르의 높은 정확도를 가진 토픽 모형의 토픽 개수는 15개이며 이를 이용해 만든 완전 나무를 생성하고 이에 대한 최적 CP를 선택해 분류기를 학습시킨 결과 변수 중요도는 그림 8과 같다.

Role Playing 장르의 최종 의사결정 나무의 가장 높은 변수 중요도를 가진 토픽은 6번 토픽과 14번 토픽이었으며, 왼쪽 6번 토픽의 경우 “expansion”,”content”와 같은 긍정적 단어가 주를 이루었으며 오른쪽 14번 토픽의 경우 “boring”, “bad” 등 부정적 단어가 분포되어 있다.

랜덤 포레스트는 분류나 회귀 등에 사용되는 앙상블 모형의 일종으로, 훈련 시 구성된 다수의 결정 나무로부터 분류 또는 회귀 분석을 출력함으로써 동작하는 알고리즘이다. 랜덤 포레스트의 경우 분석 과정 중 복잡성, 정확도가 높아 우수한 기법으로 평가되고 있으며, 다양한 분야에서 널리 적용 중이다.

Role Playing 장르의 경우 분류모형을 실시할 최적 토픽 개수는 6개이며, 토픽 모형으로 랜덤 포레스트 실시 결과는 그림 9와 같고, 1번과 3번 토픽이 가장 높은 변수 중요도가 나타났다, “pc”, “graphic” 등의 게임 요소 단어들과 “feel”, “bad”, “fun” 등의 감성 단어들이 분포되어 있다.

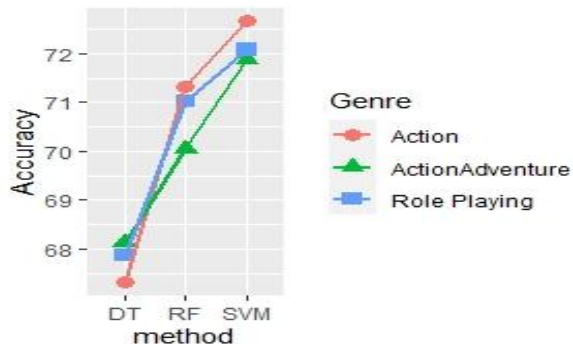

그림 11. 장르별 분류기에 따른 정확도
Fig. 11. Accuracies of classifiers by genre

SVM은 기계학습 분야 중 하나로 지도학습 모형이며, 주로 분류모형을 위해 사용한다. 간단히 말해서 이진 분류기라고 할 수 있으며, SVM의 목표는 초평면을 기준으로 한쪽 면으로 같은 데이터가 놓일 수 있게 평평한 경계를 만드는 것이다. SVM은 분포의 가정이 없어 어떠한 형태의 데이터에도 적용할 수 있고, 예측 정확도가 높은 모형으로 평가받고 있다. 하지만, 커널 선택이 어렵고, 고차원으로 갈수록 해석이 어려워진다는 단점이 있다.

Role Playing 장르의 SVM 결과는 그림 10과 같고, 4번 토픽과 5번 토픽이 “긍정”,“중립”,“부정”을 결정하는 데 있어 주요 토픽으로 선택되었다.

Role Playing 장르의 주요 토픽인 4, 5번 토픽의 단어 분포는 1번 토픽의 경우 “expansion”이 주요 단어였으며, 2번 토픽의 경우 “animation”에 대한 감정 단어가 주를 이루었다. 결과적으로 Role Playing 장르의 SVM 결과는 게임에 콘텐츠 등 즐길 거리와 볼거리 등 요소가 부정적 감정을 나타내게 하는 요소라고 할 수 있다.

표 6은 Perplexity를 참고한 토픽 모형과 지도학습을 이용한 토픽 모형으로 분류모형을 생성한 결과다. Perplexity의 경우 3개 장르의 최적 토픽 수는 4개지만 지도학습의 경우 각 장르와 분류기마다 최적 토픽의 수가 달랐고, 정확도 측면에서도 지도학습을 이용한 방법이 높은 정확도를 나타냈다.

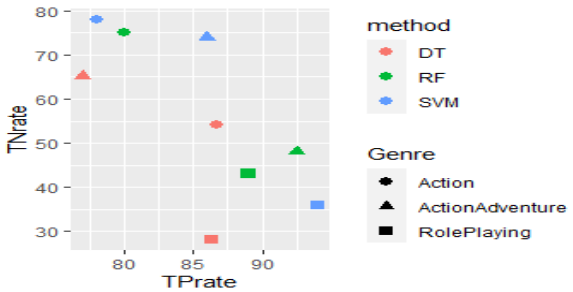

그림 12. 장르별 TP rate, TN rate
Fig. 12. TP rate, TN rate by genre

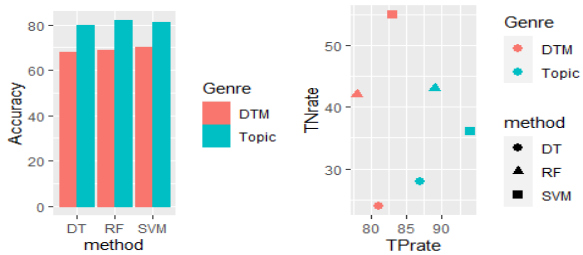

그림 13. 토픽 모형과 문서-단어 행렬 기반 분류모형 비교
Fig. 13. Comparison between the classifiers based on topic models and those based on DTMs

각 장르에 토픽 모델링을 실시하고, 분류모형으로 의사결정 나무, 랜덤 포레스트, SVM을 적용한 결과 예측 정확도는 그림 11과 같다. 의사결정 나무의 경우 가장 정확도가 낮았으며, 랜덤 포레스트와 SVM의 경우 의사결정 나무를 이용했을 때 보다 더 나은 정확도를 보였고, 가장 높은 정확도를 가진 분류모형은 SVM이었다.

각 장르의 TP rate와 TN rate는 아래 그림 12와 같다. 실제 긍정을 긍정이라고 예측한 비율이 부정을 부정이라고 예측한 비율보다 훨씬 높게 나타났으며, Role Playing 장르의 경우 TN rate가 다른 장르와 비교해 현저히 낮았다. 이는 Role playing 장르의 리뷰 중 실제 부정을 나타내는 데이터가 다른 장르보다 개수가 더 작으므로 나타난 현상이다.

토픽 모형을 사용한 감성 분류모형과 토픽 모형을 사용하지 않고 기본적인 텍스트 분석 형식인 문서-단어 행렬(DTM; Document Term Matrix)을 이용한 분류모형의 성능 차이는 그림 13과 같다. DTM으로 분류를 하였을 때 보다 토픽 모델링을 실시하였을 때 5%~10%가량 성능이 더 좋아지는 것을 확인 할 수 있다. DTM의 경우 데이터가 크면 클수록 연산 속도가 현저히 떨어진다는 단점이 있고, 데이터의 커질수록 분석 자체가 불가능하다. 반면 토픽 모델링의 경우 데이터를 간소화시켜주는 장점과 처리 속도도 상승하는 효과를 얻을 수 있고, 분류작업에 있어 더 좋은 성능을 보여준다.

TP rate와 TN rate를 보았을 때 DTM의 경우가 부정을 예측하는 데 있어 조금 더 높은 정확도를 가지고 있지만, 긍정 예측에서는 토픽 모델링이 더 좋은 성능을 보인다. Role Playing 장르의 경우 부정적 리뷰가 긍정적 리뷰보다 더 적은 데이터를 가지고 있으면서 토픽 모델링으로 더 데이터 축소를 시키기 때문에 부정을 예측하기에는 DTM을 데이터로 이용해서 성능이 더 우수해 보이지만, 데이터가 더 많아지면 토픽 모델링이 긍정, 부정 예측에 있어 더 좋은 성능을 보일 것이다.

표 7. 문서-단어 행렬을 이용한 감성 분류기의 성능
Table 7. Performance results of the sentiment classifiers using DTMs

Genre	Decision Tree		Random Forest		SVM	
	Accur-acy(%)	Processing-Time (sec)	Accur-acy(%)	Processing-Time (sec)	Accur-acy(%)	Processing-Time (sec)
Action Adventure	72.7	347.41	73.4	511.71	76.3	421.37
Action	72.8	384.50	74.2	508.25	75.7	433.76
Role Playing	73.4	259.12	75.1	510.67	76.3	429.13
Average	72.9	330.34	74.2	510.21	76.1	428.09

표 8. 토픽 모형을 이용한 감성 분류기의 성능
Table 8. Performance results of the sentiment classifiers using topic models

Genre	Decision Tree		Random Forest		SVM	
	Accur-acy(%)	Processing-Time (sec)	Accur-acy(%)	Processing-Time (sec)	Accur-acy(%)	Processing-Time (sec)
Action Adventure	74.2	5.28	76.3	25.13	77.2	5.19
Action	75.4	5.02	77.5	24.87	77.7	4.98
Role Playing	74.6	4.89	78.3	22.45	79.6	4.31
Average	74.7	5.06	77.3	24.15	77.5	4.83

표 7은 DTM과 토픽 모형을 비교하기 위해 장르별 1,000개의 리뷰를 무작위 비 복원 추출을 하고 같은 데이터에서 DTM을 만들어서 TF-IDF 확률값을 이용해 행렬을 만들고

각 분류기에 학습시켜 예측까지의 처리 시간과 결과이고, 표 8은 토픽 모형을 이용해 단어-토픽 행렬을 만들어 분류기에 학습시키고 예측까지의 처리 시간과 결과다. 정확도 측면에서 차이가 크게 나지 않지만, 처리 시간 측면에서는 차이가 크게 났다. 그리고, DTM을 이용한 분석의 경우 정 분류율이 높다고 하더라도 DTM의 예측 변수가 8,000개가 넘기 때문에 데이터가 많을수록 해석에 있어 효율성이 떨어졌으며 토픽 모형의 경우는 토픽에 분포된 단어를 통해 결과해석에 용이했다.

Ⅵ. 결 론

본 연구는 게임을 이용한 사용자들이 작성한 리뷰의 평점과 리뷰를 가지고 긍정, 부정 예측 모형을 제시한 연구이다. 메타크리틱 에서 Action Adventure, Action, Role Playing 장르의 게임들을 대상으로 하고, 장르별 200가지의 게임의 리뷰를 크롤링해 데이터 집합을 생성하였다. 데이터는 정제 과정을 통해서 불용어, 숫자 등 불필요한 단어들을 제거하였고, 이후 LDA 모형을 이용해 잠재 토픽을 추출하고 결과로 생성된 문서-토픽 행렬을 이용해 각 분류기에 학습시켰다. 훈련데이터와 테스트 데이터를 7:3의 비율로 랜덤하게 나누었고, 학습에는 훈련데이터를 이용해 교차 검증을 이용하여 사용되었다. 이후 테스트 데이터를 이용해 분류기를 검증하였다.

본 연구의 결론은 크게 세 가지로 볼 수 있다.

첫째 토픽 모형을 사용한 분류기가 사용하지 않은 분류기보다 더 높은 성능을 보여주었다. 기존 DTM 행렬을 이용하여 분류기를 학습시키고 검증한 결과보다, LDA를 이용해 잠재 토픽 추출 및 문서-토픽 행렬을 이용해 분류기를 학습시켰을 때 약 5% 이상의 정확도가 상승한 것을 확인하였다. DTM 행렬을 이용한 방법은 데이터 자체의 크기도 방대하고 처리 속도가 현저히 떨어지는 것도 확인했다. LDA를 활용 시 데이터의 크기를 줄일 수 있고, 이에 따른 데이터 처리 속도도 향상되므로 LDA 이용해 새로운 리뷰의 긍정, 부정을 예측한다면 정확도 높은 결과를 얻을 수 있을 것이다.

둘째, 비지도 학습에 최적인 토픽 수와 지도학습에 최적인 토픽의 수는 차이가 있었다. Perplexity가 최소인 비지도 학습상에서 최적의 토픽 수를 지도학습에 이용할 때 최적의 성능이 나타나지 않았다. 이는 토픽 모형을 리뷰의 분류모형에 이용할 때는 Perplexity 등의 비지도 학습의 성능 지표보다는 교차 검증으로 최적의 토픽 수를 결정하는 것이 더 적절함을 시사한다.

셋째, 게임 리뷰에 대해 긍정, 부정 예측을 진행하면서 장르별 영향을 주는 토픽이 다른 것을 알 수 있다. ‘Action Adventure’ 장르의 경우 ‘pc’, ‘launcher’ 등 게임 실행에 관련된 단어들이 많이 분포되어 있고, 분류기에 가장 많은 영향을 미친 토픽이다. ‘Action’ 장르의 경우 ‘graphic’ 등 게임의 시각적 요소에 대한 단어 분포가 주된 토픽이었다. ‘Role Playing’ 장르의 경우 ‘expansion’, ‘content’와 같은 게임의 즐길 거리에 대한 단어들이 분류기 적합에 가장 많은 영향을 끼쳤다. 이는 장르별 소비자들의 요구사항이 다르다는 것을 알 수 있다.

본 연구의 공헌은 크게 두 가지로 볼 수 있다.

첫째, 토픽 모델링을 이용한 리뷰 분류와 머신 러닝을 융합한 분류모형을 제시하였다. 토픽 모형을 이용하는 것이 분류모형에 좋다는 것을 알 수 있었다. 따라서 본 연구의 결과는 긍정, 부정 예측과 토픽 모델링의 새로운 활용법에 대한 방법을 제시하였다는데 의의가 있다.

둘째, 주요 장르에 대해 비교분석을 시행하였다. 기존 시행된 게임의 리뷰에 대한 감성 분석 연구는 모바일의 몇몇 게임 중심이거나, ‘Action’ 장르 중심 등 대표적인 게임에 관한 연구가 진행되었다. 본 연구는 장르 세분화를 통해 장르별 고객들이 긍정과 부정적 감성에 중요하게 생각하는 요소들을 토픽 모델을 통해 비교분석을 했다는데 의의가 있다.

마지막으로 본 연구의 향후 연구 방향을 제시한다.

첫째, 본 연구에서는 Action Adventure, Action, Role Playing인 3가지 장르를 가지고 있는 게임에 대해서만 연구를 진행하였고, 특수 장르를 가지고 있는 게임에 대해서는 데이터의 양이 3가지 장르보다 매우 적기 때문에 분석 시행을 하지 못했다. 장르도 크게 Action 장르로 분류되어 있지만, FPS, TPS 등 여러 세분화된 장르가 존재한다. 향후 연구에서는 게임 장르를 더욱 세분화해 리뷰 분석 모형을 더 정교화 할 수 있을 것이다. 이와 같은 분류 예측을 할 때, 세분화된 장르별 게임에 대해 더 중요한 요소들이 존재할 것이며, 중요한 토픽을 찾아볼 수 있으리라 기대한다.

둘째, 본 연구에 데이터로 메타크리틱의 리뷰를 이용하였지만, 리뷰는 다양한 형태를 띠고 있다. 짧은 리뷰도 있고, 길이가 긴 리뷰도 있다. 문서 내에는 여러 단어가 존재하고, 각 단어는 여러 문서에 포함된다. 단어로부터 토픽을 추출하기 때문에 토픽들은 서로 상관관계를 가질 수 있다. LDA는 토픽 간 상관관계를 모형화할 수 없고 이 경우 상관 토픽 모형(CTM; Correlated Topic Model)을 이용하는 것이 더 좋은 방법일 수 있다. 또한, 단문의 경우 BTM(Biterm Topic Model)을 이용하는 것이 더 향상된 모형을 생성할 수 있다. 이처럼 리뷰의 형태에 따라 각기 다른 토픽 모형을 사용하는 것이 분석하면서 더 좋은 성능을 보일 수 있다. 향후 연구에서 리뷰의 형태를 분석해 맞는 토픽 모형을 이용해 볼 수 있을 것이다.

참고문헌

[1] Korea Creative Content Agency, 2019 Contents Industry Statistical Survey, Korea Creative Content Agency, 2019.

[2] Korea Creative Content Agency, Global Game Industry Trend (January+February 2020 issue), Korea Creative Content Agency, 2020.

[3] CNNIC, "China Network Shopping Market Research Report," Beijing: China Internet Network Information Center, pp.20-21, 2016.

[4] B.Straat and, H. Verhagen, "Exploring Video Game Design and Player Retention-a Longitudinal Case study," in *Proceedings of the 22nd international Academic Mindtrek Conference*, pp.39-48, 2018.
[5] M-Young Wui, Ji Young Na, and Young Il Park, "A study on the Elements of Interest for VR Game Users Using Text Mining and Text Network Analysis – Focused on STEAM User Review Data," *Journal of Korea Game Society*, Vol. 18, No.6, pp. 69-82, 2018.
[6] Tae Sun Kim, "Elements and Topics of Mobile Games Extracted from Mining User Online Reviews and Online Ratings," *Journal of Product Research*, Vol. 37, No. 5, pp. 67-76, 2019.
[7] Jeong Eun Park, Online review analysis to improve smart speaker user experience : using the topic modeling analysis method, M.S. UX. dissertation, Graduate School of Information, Yonsei University, 2019.
[8] So Hyun Park, Exploratory study on patients experience evaluation using topic modeling: focus on patient review analysis of Levothyroxine, M.S. Pharmacy. dissertation, College of Pharmacy Seoul National University, 2018.
[9] Dong Wook Kim, Ju Young Kang, and Jay Ick Lim, "Comparative Analysis of Job Satisfaction Factors, Using LDA Topic Modeling by Industries : The Case Study of Job Planet Reviews," *Jounal of Information Technology Services*, Vol. 15, No.3, pp. 157-171, 2016.
[10] Sang Hyun Park, Classification and application of online review data based on topic modeling and neural networks, MBA dissertation, Graduation School Kyung Hee University, 2017.
[11] D. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," *Advances in neural information processing systems*, Vol.1, No.14 pp. 601-608, 2002.
[12] J. Chang, S. Gerrish, C. Wang, J. Boyd-Graber, and D. Blei, "Reading tea leaves: How humans interpret topic models," *Neural Information Processing Systems*, pp. 288-296, 2009.
[13] B. Liu, "Sentiment Analysis and Opinion Mining," *Synthesis Lectures on Human Language Technologies*, Vol.5, No.1, pp. 1-167, 2012.
[14] I. H. Witten, E. Frank, and M. A. Hall, *Data mining: practical machine learning tools and techniques,* 4th ed. Cambridge, MA: Elsevier, 2017.
[15] M. Kuhn and K. Johnson, *Applied predictive modeling,* New York, NY: Springer, 2013.

김태국(Tae-Kook Kim)

2020년 : 상명대학교 (공학학사)
2020년 : 상명대학교 대학원 (공학석사-경영공학과)

20020년~현 재: 상명대학교 경영공학과 석사과정
※관심분야 : 텍스트마이닝(Text Mining), 머신러닝(Machine Learning), 딥러닝(Deep Learning) 등

김길환(Kilhwan Kim)

1996년 : KAIST 산업경영학과 (공학석사)
2009년 : KAIST 산업공학과 (공학박사)

1997년~2003년: LG CNS 선임 컨설턴트
2003년~2005년: 기업정보화지원센터 선임 컨설턴트
2009년~2012년: 한국전자통신연구원 선임 연구원
2012년~현 재: 상명대학교 경영공학과 부교수
※관심분야 : 데이터마이닝, 텍스트마이닝, 확률모형 등