리뷰의 의미적 토픽 분류를 적용한 감성 분석 모델

(Sentiment Analysis Model with Semantic Topic Classification of Reviews)

임명진*, 김판구**, 신주현***
(Myung Jin Lim, Pankoo Kim, Ju Hyun Shin)

요 약

지상파에 한정되어 방영되었던 과거와는 달리 현재는 케이블 채널과 인터넷 웹에서도 수많은 드라마가 방영되고 있다. 드라마를 보고난 후 시청자들은 리뷰를 통해 적극적으로 자신의 의견을 표현하고 이러한 리뷰의 분석에 관련된 연구들이 활발하게 진행되고 있다. 드라마의 특성상 장르가 뚜렷하지 않고 시청자의 다양한 연령층으로 인해 다른 시청자들의 리뷰와 평가는 어떤 드라마를 볼 것인지 결정하는데 도움이 된다. 하지만 많은 리뷰를 시청자가 일일이 확인하고 분석하는 것은 어렵기 때문에 자동으로 분석하기위한 데이터 분석 기법이 필요하다. 이에 본 논문에서는 드라마 선택에 중요한 영향을 미치는 리뷰의 토픽을 분류하고 단어의 의미 유사도에 따라 의미적 토픽으로 재분류한다. 그리고 리뷰를 의미적 토픽에 따른 문장으로 분류한 다음 감성 단어를 통해 감성을 분석하는 모델을 제안한다.

■ 중심어 : 리뷰 ; 토픽 ; 의미적 토픽 ; 토픽 분류 ; 감성 분석

Abstract

Unlike the past, which was limited to terrestrial broadcasts, many dramas are currently being broadcast on cable channels and the Internet web. After watching the drama, viewers actively express their opinions through reviews and studies related to the analysis of these reviews are actively being conducted. Due to the nature of the drama, the genre is not clear, and due to the various age groups of viewers, reviews and ratings from other viewers help to decide which drama to watch. However, since it is difficult for viewers to check and analyze many reviews individually, a data analysis technique is required to automatically analyze them. Accordingly, this paper classifies the topics of reviews that have an important influence on drama selection and reclassifies them into semantic topics according to the similarity of words. In addition, we propose a model that classifies reviews into sentences according to semantic topics and sentiment analysis through sentiment words.

■ keywords : review ; topic ; semantic topic ; topic classification ; sentiment analysis

Ⅰ. 서 론

드라마는 지상파에만 한정되었던 과거와는 달리 현재는 케이블 채널뿐만 아니라 인터넷 웹을 통하여 수많은 드라마가 방영되고 있다. 시청자들은 드라마를 시청한 후 리뷰를 통해 자신의 생각이나 느낌을 자유롭게 표현하고 공유하며 리뷰의 분석에 관련된 연구들이 활발하게 진행되고 있다[1]. SNS와 온라인 게시판, 커뮤니티를 통해 시청자들은 드라마에 대한 평가에 적극적으로 참여하고 있으며[2], 이러한 리뷰 데이터는 시청자가 어떤 드라마를 볼 것인지 선택하는데 영향을 준다. 하지만 드라마의 특성상 장르가 뚜렷하지 않고 시청자의 다양한 연령층으로 인해 리뷰 데이터로 시청자들의 취향을 분석한다는 것은 어렵다. 또한 온라인상의 수많은 리뷰 데이터들을 시청자가 모두 읽고 분석하려면 많은 시간이 소요되기 때문에 자동으로 분석하기 위한 데이터 분석 기법이 필요하다[3]. 드라마 선택에 중요한 영향을 미치는 리뷰를 토픽에 따라 분류하면 시청자의 다양한 취향을 파악할 수 있고, 토픽에 대한 감성 단어를 통해 감성을 분석하여 긍·부정을 판별하면 리뷰 데이터의 자동 분석이 가능하다. 따라서 본 논문에서는 리뷰 데이터의 토픽을 정의하고 의미적 토픽으로 재분류 한 후 리뷰를 토픽에 따른 문장으로 분류한 다음 감성 단어를 통해 감성을 분석하는 모델을 제안한다.

* 정회원, 조선대학교 컴퓨터공학과 대학원생
** 정회원, 조선대학교 컴퓨터공학과 교수
*** 정회원, 조선대학교 신산업융합학부 교수
이 성과는 정부(과학기술정보통신부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구(No. 2019R1F1A1057325) 이며 조선대학교 학술연구비의 지원을 받아 연구되었음(2020년도)

접수일자 : 2020년 05월 20일
수정일자 : 2020년 06월 01일
게재확정일 : 2020년 06월 01일
교신저자 : 신주현, e-mail : jhshinkr@chosun.ac.kr

본 논문의 구성은 다음과 같다. 2장에서는 리뷰의 토픽 분류와 LDA 토픽 모델링, Word2Vec와 감성 분석에 관련된 연구에 관하여 기술하고, 3장에서는 본 논문에서 제안하는 리뷰의 의미적 토픽 분류를 통한 감성 분석 모델을 기술한다. 4장에서는 제안한 모델을 적용한 실험 및 평가 결과를 기술하며, 5장에서는 결론 및 향후 연구에 대해 기술하고 마무리한다.

Ⅱ. 관련연구

1. 리뷰의 속성 분류

리뷰 전체에 대해 긍정이나 부정으로 분석하게 되면 여러 문제가 생길 수 있다. 한 사람이 제품 리뷰에 대해 여러 가지 장점을 나열했지만 마지막 문장으로 '그러나 이 제품을 또 사고 싶지는 않을 것 같다'라고 남겼다면 이 리뷰는 결론은 부정적인 리뷰지만 상품에 관해 긍정적인 문장이 더 많기 때문에 긍정 리뷰로 분석될 것이다. 또한 단순히 상품에 대한 긍·부정을 아는 것에서 나아가 사람들은 무엇이 좋고 무엇이 싫은지 구체적으로 알고 싶어 한다. 이러한 문제를 해결하기 위해서는 리뷰 전체의 긍·부정 여부가 아닌 사람들이 관심 있는 속성 단위의 텍스트를 분석하는 방법이 필요하다[4]. 표 1은 영화의 흥행과 리뷰 특성에 관한 연구에서 제시한 세 가지로 구분된 영화의 속성을 나타낸다.

표 1. 영화의 속성

구분	내용
핵심 속성	내용, 감독, 연기, 배우, 장르
주변 속성	촬영, 의상, 특수효과, 배경음악, 문화적 코드
커뮤니케이션 속성	포스터, 예고편, 웹사이트, 홍보 광고, 구전, 시사회

표 1에 따르면 영화가 가지고 있는 핵심 속성과 주변 속성, 커뮤니케이션 속성을 알 수 있다. 이러한 속성들은 영화의 흥행과 리뷰 특성과의 연관성이 있고 영화 선택의 결정적인 요인이 된다[5]. 따라서 리뷰에 포함된 여러 속성에 관련된 감성 분석을 하면 평가 대상에 대한 사람들의 의견이나 감정을 세부적으로 잘 파악할 수 있다. 속성 단위로 감정을 분석하기 위해서는 속성명을 추출한 다음 속성에 따른 감성어를 찾은 뒤 감성어에 대한 감성 분석이 필요하다[6].

2. LDA 토픽 모델링

LDA(Latent Dirichlet Allocation)는 2003년 David Blei, Andrew Ng, Michaeall.Jordan이 제안한 모형으로 자연어 처리에서 문서들에 대해 각 문서에 어떤 주제들이 포함되어 있는지에 대한 확률 모형이다[7]. 그림 1은 LDA 모델을 나타낸다.

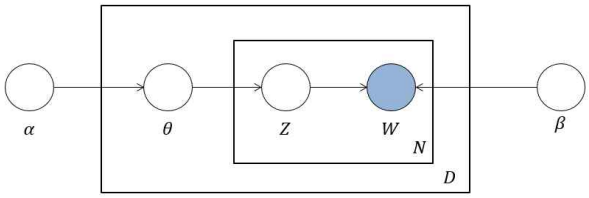

그림 1. LDA Model

그림 1에서 N은 단어의 개수이고 D는 문서의 개수를 나타낸다. 문서 집합에서 관측된 W을 이용하여 숨겨진 θ와 β를 추론한다. 각 문서들이 갖는 주제인 θ를 확률적으로 분석하며 각 주제에 해당하는 단어들의 확률 분포인 Z도 나타낼 수 있다. 본 논문에서는 문서 내 주제 분포를 분석하고 주제 내 단어 분포를 분류하기 위하여 LDA를 사용하였다.

단어가 어떤 주제에 포함하게 될 확률을 찾아내는 방법으로 LDA 클러스터링을 기반으로 TV 프로그램의 줄거리를 이용하여 콘텐츠를 추천하는 연구에서는 시청자들의 시청 기록을 LDA를 이용하여 사용자가 각 TV프로그램을 시청할 확률을 모델로 생성한 후 사용자의 성향에 따라 집합을 생성하고, 추천 대상이 되는 사용자와 비교하는 방법을 사용하였다[8]. 하지만 사용자의 시청기록을 수집하는데 어려움이 있었고, 시청성향이 넓은 경우 예측하지 못했다.

3. Word2Vec

Word2Vec은 Word Embedding 모델의 하나로 빠른 학습 속도를 가지고 있다[9]. 그림 2는 Word2Vec의 두 가지 학습 모델 구조이다.

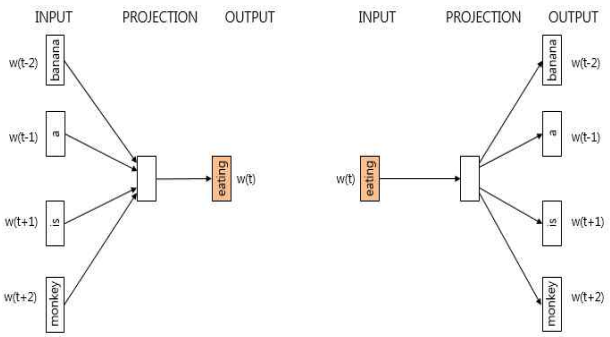

그림 2. CBOW 모델과 Skip-gram 모델

그림 2에 따르면 Word2Vec은 크게 두 가지 학습모델이 있다. 먼저 CBOW 모델은 주변에 있는 단어들로 대상 단어를 예측하고, Skip-gram 모델은 대상 단어로 주변 단어들을 예측

한다[10]. 따라서 Word2vec은 단어의 의미와 문장에서의 맥락을 함께 고려하여 단어를 벡터로 나타내기 때문에 의미가 유사한 단어들끼리 가까운 벡터 공간에 위치하게 된다. 이것은 같은 단어라도 의미와 맥락에 따라서 다른 벡터 공간에 위치한다는 것을 의미한다[11]. 본 논문에서는 단어들의 벡터 사이의 거리를 코사인 유사도로 계산하여 단어들 간의 의미적 유사도를 구하기 위해 특정 토픽에 따라 전후 문맥 단어들을 예측하는 skip-gram 방법을 사용하였다.

4. 감성 분석

텍스트 마이닝의 한 분야로서 감성 분석은 감성 분류 또는 오피니언 마이닝으로 불리며 문서에서 사람들의 의견과 태도, 성향 등을 분석하여 긍정과 부정에 대한 감성을 분류하고 추측하는 방법이다[12]. 감성 분석은 문서의 최소 단위인 단어의 감성 극성에 기반을 두어 이루어진다[13]. 따라서 단어의 감성 극성을 정확하게 적용한 감성 사전을 사용하는 것이 중요하다. 표 2는 감성 분석 3단계를 나타낸다.

표 2. 감성 분석 3단계

단계	내용
데이터수집	리뷰게시판, 블로그, SNS에서 리뷰 데이터를 수집
주관성탐지	사용자의 주관이 드러난 부분만을 추출
극성탐지	추출한 감성 데이터를 긍정·부정의 감성으로 분류

표 2에 따르면 감성 분석은 웹이나 게시판에서 다양한 리뷰 데이터를 수집하여 감성 분석에 사용될 텍스트 요소만을 분류하고 주어진 데이터를 긍정과 부정으로 판단하는 단계로 구성된다.

상품평을 기반으로 의류, 영화 등의 사용자의 감성을 분석할 수 있는 감성 사전을 자동으로 구축하는 시스템을 구현한 연구가 있었으며[14], 감성 분석을 통해 추천할 영화를 랭킹으로 정렬하여 추천해주는 연구가 있었다[15]. 이러한 연구들에서 감성 분석 방법은 전체 리뷰를 대상으로 극성을 분류하는 방법을 사용하기 때문에 사람들의 취향을 판단하기에는 어렵다. 따라서 리뷰를 사용자의 주관이 드러난 토픽에 따라 분류하여 감성을 판별하는 방법이 필요하다.

III. 의미적 토픽 분류를 적용한 감성 분석

1. 시스템 구성도

본 논문에서는 시청자의 드라마 선택에 중요한 영향을 미치는 리뷰를 의미적 토픽 분류를 적용하여 감성 분석하는 모델을 제안한다. 분석 단계는 크게 4단계로 데이터 수집 및 전처리 단계와 토픽 분류 단계, 의미적 토픽 재분류 단계와 토픽 감성 단어를 통한 감성 분석 단계로 구성된다. 그림 3은 본 논문에서 제안하는 감성 분석 모델의 시스템 구성도이다.

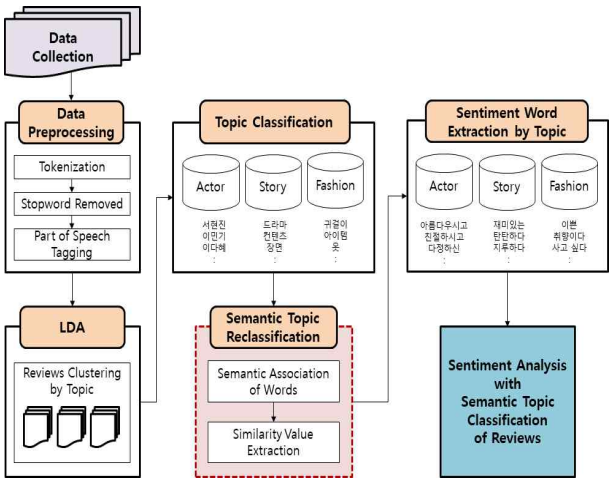

그림 3. 시스템 구성도

드라마 리뷰 데이터를 웹 크롤링으로 수집하여 .txt 파일로 저장하고, 전처리 과정으로 KoNLP 라이브러리를 사용하여 한국어 처리한다. 다음 단계로 문장을 어절 단위로 토큰화 한 후 불용어를 제거하고 단어들의 품사를 판별하여 보통 명사들만 추출한다. LDA 기반 토픽 모델링 단계로 문서 내 주제의 분포와 주제 내 단어의 분포를 추출하여 토픽을 분류한다.

Word2Vec 학습을 통해 단어를 벡터화한 후 단어 간의 의미적 유사도를 구한다. 토픽과 단어 간 의미적 유사도가 더 높은 단어는 의미적 토픽으로 재분류한다. 의미적 토픽에 따라 분류한 문장에서 감성 단어를 통해 긍·부정을 판별하여 토픽별 감성을 분석한다.

기존의 감성 분석 모델은 전체 리뷰의 감성 단어를 긍·부정으로 분류하였다. 하지만 드라마 리뷰의 특성상 개인의 성향 즉 토픽에 따라 느끼는 감정이 다르기 때문에 본 논문에서는 이러한 점을 보안하기 위해 의미적 토픽을 적용하여 감성 분석하였다. 의미적 토픽 분류를 적용한 감성 분석 모델은 전체 리뷰를 분석하는 방법보다 시청자의 성향에 맞는 세밀한 결과를 도출할 수 있다.

2. 데이터 수집 및 전처리

실험을 위한 드라마 리뷰 데이터는 2018년 10월 JTBC에서 방영되었고 현재 Netflix 인기 드라마로 방영되고 있는 드라마 '뷰티인사이드'의 시청 소감 게시판[16]과 NAVER에서 제공하는 실시간 TALK[17]에서 17,467개의 리뷰 데이터를 웹

크롤링으로 수집하여 .txt 파일로 저장하였다. 그림 4는 수집한 뷰티인사이드 드라마의 리뷰 데이터를 보여준다.

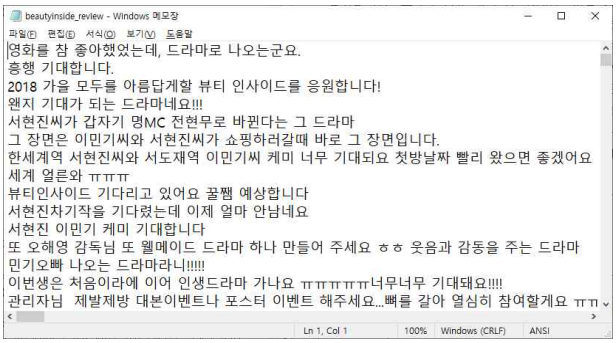

그림 4. 리뷰 데이터

리뷰 데이터는 KoNLP 라이브러리를 사용하여 한국어 자연어 처리를 한다. 전처리 작업으로 어절 단위로 문장을 토큰화(Tokenizing)한 후 숫자, 영어, 특수문자, 완전한 글자가 아닌 단어를 불용어로 포함하여 제거한다. 형태소 분석을 통해 단어들의 품사를 판별하기 위한 Pos Tagging 작업을 수행하여 보통 명사를 추출한다. 표 3은 리뷰 데이터의 전처리 과정을 비교하여 보여준다.

표 3. 리뷰 데이터의 전처리 과정

데이터	리뷰
원본	서현진 배우님 나와서 다 너무 재미있어요...
토큰화	서현진, 배우, 님, 나와서, 다, 너무, 재미있어요, ...
불용어 제거	서현진, 배우, 나와서, 너무, 재미있어요
명사 추출	서현진, 배우

3. LDA 기반 토픽 분류

LDA 토픽 모델링에서 토픽 수는 사용자가 지정하는 하이퍼파라미터로서 최적의 토픽 수는 실험을 통해 구해야 한다. 본 논문에서는 적합한 토픽개수 선정을 위하여 토픽들 사이의 거리를 계산하여 서로간의 관련성을 최소화하는 방법을 적용하였다[18]. 토픽 수를 2~10개로 변경하여 실험한 결과 3개로 지정했을 때 가장 분류가 잘 되었다. 3개의 토픽에 따라 3개의 단어 집합이 생성되었고 같은 토픽 내에 있는 단어들은 서로 동일한 주제 범위를 갖게 된다. 그림 5는 토픽 거리 맵과 토픽별 단어 집합을 보여준다. 원의 수는 토픽의 수를 의미하고, 각 원은 하나의 토픽을 나타내며 각 원의 넓이는 코퍼스 내에서 전체 키워드들에 대한 비율을 나타낸다[19].

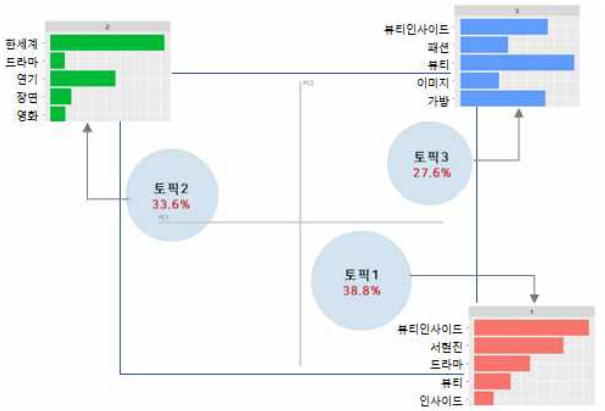

그림 5. LDA 토픽 거리 맵과 토픽별 단어 집합

그림 5에서 토픽 1은 38.8%의 단어들이 집합을 이루었고 '뷰티인사이드', '서현진', '드라마' 등의 단어들이 분류되어 있었다. 토픽별로 분류된 단어들을 수집하여 토픽별 단어 집합을 만든다[20]. 표 4는 토픽별 상위 10개의 단어를 추출한 결과이다.

표 4. 토픽별 상위 10개 단어

토픽ID	토픽 이름	상위 10개 단어
토픽1	배우	뷰티인사이드, 서현진, 드라마, 뷰티, 인사이드, 이미지, 이다희, 안재현, 배우, 장면
토픽2	스토리	한세계, 드라마, 연기, 장면, 영화, 캐릭터, 정보, 내용, 케미, 뷰티인사이드
토픽3	패션	뷰티인사이드, 패션, 뷰티, 이미지, 가방, 옷, 원피스, 셔츠, 귀걸이, 드라마

토픽별 단어 집합을 분석한 결과 토픽1은 배우, 토픽2는 스토리, 토픽3은 패션에 관련된 단어들이 분류되어 있었다. 이는 드라마 리뷰의 내용이 주로 배우, 스토리, 패션을 중심으로 작성되어 있음을 보여준다. 따라서 본 논문에서는 '배우', '스토리', '패션'을 리뷰의 토픽으로 정의하였다.

4. 의미적 토픽 재분류

앞에서 분류된 LDA 기반 토픽 분류는 확률에 의한 분류 방법으로 단어의 의미가 고려되지 않았으며 토픽과 관련이 없거나 중복되는 단어가 있다. 따라서 본 논문에서는 이러한 문제를 해결하기 위해 단어의 의미를 고려한 의미적 토픽으로 재분류하는 방법을 제안한다. 데이터를 Word2Vec을 이용하여 200차원, Skip-gram 방식으로 학습하고 단어들의 벡터 값을 Vector Space Model로 구축했다. 그 결과 의미적으로 유사한 단어들이 근접한 벡터 공간에 위치하는 것을 볼 수 있었고 단어의 문맥적 의미가 보존되는 것을 알 수 있었다. 단어 간 유사도를 구하기 위해 단어의 벡터 값을 cosine similarity로 계산하고

단어 벡터들 간의 거리를 측정하였다[21]. cosine similarity는 두 벡터의 각도를 cosine 값으로 계산하여 벡터 간의 유사한 정도를 구한다. 0~1의 값으로 1에 가까울수록 두 단어가 유사하다는 것을 의미한다[22]. 식 (1)은 두 벡터 A, B의 cosine similarity를 구하는 계산식을 나타낸다.

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \cdot B_i}{\sqrt{\sum_{i=1}^{n}(A_i)^2} \cdot \sqrt{\sum_{i=1}^{n}(B)^2}} \quad (1)$$

표 4에서 분류된 단어들 중 '뷰티인사이드', '뷰티', '인사이드', '드라마'와 같이 드라마 제목이나 토픽을 분류할 수 없는 포괄적인 의미를 갖고 있는 단어들은 제외하고 둘 이상의 토픽에 공통으로 포함되는 중복 단어에 대해 토픽과 단어 간의 cosine similarity를 구한다. 표 5는 토픽과 중복 단어에 대한 유사도를 계산한 결과이다.

표 5. 토픽과 단어 간 유사도

단어	배우	스토리	패션
이미지	0.9998723268	-	0.9821032285
장면	0.9833170175	0.9999205598	-
⋮	⋮	⋮	⋮

표 5와 같이 '이미지'는 '배우'와 '패션' 토픽에 중복되는 단어로 '배우'와 유사도가 더 높은 것을 볼 수 있다. 따라서 '이미지'는 '배우' 토픽으로 분류한다. '장면'은 유사도가 더 높은 '스토리' 토픽으로 분류한다. 이러한 방법으로 토픽별 중복되는 단어를 유사도가 더 높은 의미적 토픽으로 재분류한 결과는 표 6과 같다.

표 6. 의미적 토픽 재분류

토픽ID	토픽	재분류된 단어
토픽1	배우	서현진, 이미지, 이다희, 안재현, 배우, 류화영, 연기, 이민기, 주인공, 김성령
토픽2	스토리	연기, 장면, 영화, 캐릭터, 정보, 내용, 케미, 스토리, 사랑, 천주교
토픽3	패션	패션, 가방, 옷, 원피스, 셔츠, 귀걸이, 제이에스티나, 신발, 선글라스, 청바지

표 6과 같이 토픽을 분류할 수 없는 포괄적인 단어는 제외하고 중복되는 단어를 재분류한 의미적 토픽을 적용하여 감성 분석을 진행한다.

5. 감성 단어를 통한 감성 분석

텍스트로부터 극성 값을 판별하는 과정으로 감성 분석은 감성 사전을 사용하고 문서의 최소 단위인 어휘의 극성으로 분석한다. 본 논문에서는 앞에서 정의한 리뷰의 의미적 토픽인 '배우', '스토리', '패션'에 따라 Word2Vec을 이용하여 각 토픽에 관련된 문장별로 분류한다. 토픽별 리뷰 문장의 예시는 표 7과 같다.

표 7. 토픽별 리뷰

토픽	리뷰
배우	서현진 배우님 친절하시고 아름다우시고 이민기 배우님도 너무 다정함! 첫방 정말 기대하고 있어요!
패션	뷰티인사이드 시상식 드레스인데 웨딩드레스 빰치게 이쁘네요. 병원씬에서 노랑 가디건과 반지는 어디꺼?
스토리	영화 뷰티인사이드를 너무 재미있게 봐서 드라마도 기대가 크다. 영화와 다른 캐릭터 설정이 훨씬 더 재미있는 스토리를 만들어낼 것 같다.

표 7과 같이 토픽별로 분류된 리뷰에서 감성을 분석하기 위해 감성 단어를 추출한다. 리뷰에서 사람들이 주로 감성을 드러내는 품사인 '형용사'와 '동사'를 추출하여 감성 단어로 설정하였다. 표 8은 토픽별 추출된 감성 단어들을 나타낸다.

표 8. 토픽별 감성 단어

토픽	감성 단어
배우	아름다운, 친절하다, 다정한, 나쁘다, 예쁘다
스토리	재미있는, 탄탄하다, 지루한, 좋은, 억지스럽다
패션	예쁜, 취향이다, 아쉬운, 완벽하다, 아쉽다

표 8과 같이 추출된 토픽에 따른 감성 단어를 공개된 KNU 긍·부정 사전[23]을 사용하여 긍·부정을 판별할 수 있다. 리뷰에서 긍·부정 단어가 몇 번 나왔는지 분류하여 차를 통해 극성 값을 판별한다. 속성이 포함된 문장을 'x', 긍·부정을 판단하는 'class'는 'c'라 한다. 그러므로 문장 'x'가 긍정이면 +1, 부정이면 -1의 값을 가진다[24]. 수식 (2)는 긍·부정을 구하는 계산식을 나타낸다.

$$\begin{aligned} &p(x_1,x_2,x_3,...x_n|c) = p(x_1|c) + p(x_2|c) + p(x_3|c)...p(x_n|c) \\ &p(x_1,x_2,x_3,...x_n|c) > 0 \text{ 긍정} \\ &p(x_1,x_2,x_3,...x_n|c) < 0 \text{ 부정} \end{aligned} \quad (2)$$

수식 (2)의 긍·부정 계산식과 R의 easySenti 패키지를 이용하여 리뷰를 의미적 토픽 분류에 따라 감성을 판별하였다. 표 9는 토픽 분류에 따른 리뷰별 감성 분석 결과이다. 리뷰에서 토픽 분류에 따른 감성 단어가 몇 차례 나왔는지 그리고 긍정인지 부정인지 판별하여 합이 양수이면 긍정(POS), 음수면 부정(NEG)으로 구분하여 리뷰의 의미적 토픽 분류를 적용한 감성을 분석한다.

표 9. 리뷰별 감성 분석 결과

리뷰ID	배우	패션	스토리	합	POS/NEG
리뷰1	긍정(+1)	-	긍정(+1)	+2	POS
리뷰2	부정(-1)	-	부정(-1)	-2	NEG
리뷰3	긍정(+1)	패션(+1)	긍정(+1)	+3	POS
⋮	⋮	⋮	⋮	⋮	⋮

IV. 실험 및 평가

1. 실험 결과 및 성능 평가

본 절에서는 제안하는 의미적 토픽 분류의 성능 평가를 위해 토픽의 일관성(Coherence)을 측정하였다. 토픽의 일관성은 두 가지로 평가할 수 있다. 먼저, TC-PMI는 토픽 내 단어들의 응집성을 의미하며 PMI를 이용하여 토픽 내 단어들의 공기(co-occurence)를 측정한다. TC-W2V는 Word2Vec를 이용하여 토픽 내의 단어 간 연관성(relatedness)을 측정한다[25]. 식 (3)은 TC-PMI를 구하는 식을 나타낸다.

$$TC-PMI=\frac{1}{N\cdot K}\sum_{k=1}^{K}\sum_{j=2}^{N}\sum_{i=1}^{j-1}\log\frac{p(w_{ki},w_{kj})}{p(w_{ki})\cdot p(w_{kj})} \quad (3)$$

식 (3)에서 N은 토픽 내 상위 k개의 단어를 의미하며, K는 전체 토픽의 수를 나타낸다. $p(w_i,w_j)$는 전체 문서에서 두 단어 w_i,w_j가 같이 나올 확률을 나타내고 $p(w_i)$는 단어 w_i가 나올 확률을 의미한다[25]. 식 (4)는 TC-W2V를 구하는 식을 나타낸다. $cor(w_i,w_j)$는 Word2Vec으로부터 계산된 두 단어 w_i와 w_j사이의 유사도 값을 의미한다.

$$TC-W2V=\frac{1}{N\cdot K}\sum_{k=1}^{K}\sum_{j=2}^{N}\sum_{i=1}^{j-1}cor(w_{kj},w_{ki}) \quad (4)$$

성능 평가를 위해 LDA 기반 토픽 분류와 의미적 토픽 분류를 비교 평가 하였다. PMI는 드라마 리뷰를 통해 계산하였고, Word2Vec는 200차원으로 학습하였다. 또한 토픽 내에서 상위 30개의 단어를 해당 토픽의 대표 단어라 여기고, N을 30으로 설정하였다. 학습될 토픽의 개수 K는 3으로 설정하였다. 표 10은 LDA 기반 토픽 분류와 의미적 토픽 분류의 토픽의 일관성 결과인 TC-PMI와 TC-W2V를 나타낸다. LDA 기반 토픽 분류 보다 본 논문에서 제안하는 의미적 토픽 분류 모델이 토픽의 응집성과 연관성에서 전반적으로 뛰어난 성능을 보이고 있음을 확인할 수 있다.

표 10. 토픽 일관성 결과

	LDA 기반 토픽 분류		의미적 토픽 분류	
	TC-PMI	TC-W2V	TC-PMI	TC-W2V
배우	1.1473	0.0247	1.4247	0.0321
스토리	1.6451	0.0354	1.9585	0.0454
패션	1.6712	0.0535	1.9531	0.0773

그림 6은 표 10의 결과를 그래프로 나타낸 토픽의 일관성 그래프이다. LDA 기반 토픽 분류와 비교하여 의미적 토픽 분류 모델이 TC-PMI와 TC-W2V의 성능이 더 뛰어난 것을 확인할 수 있다. 이는 제안한 방법으로 구성된 토픽 단어 집합이 같이 나올 확률이 높은 단어들로 구성되어 있고 의미상으로도 서로 유사하다는 것을 의미한다.

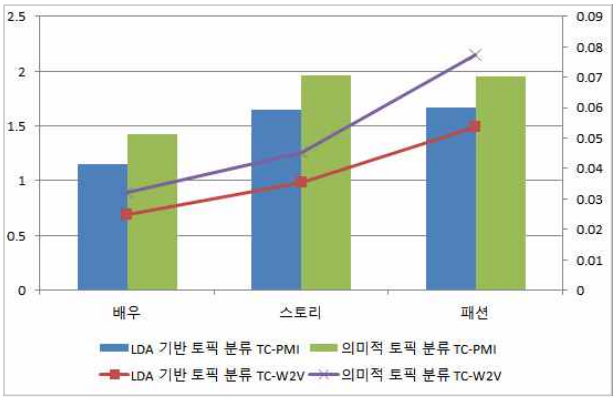

그림 6. 토픽 일관성 그래프

2. 비교 평가

본 절에서는 제안한 의미적 토픽 분류를 적용한 감성 분석 모델에 대한 비교 평가를 위해 리뷰에 전체에 대한 감성 분석과 LDA 기반 토픽 분류, 그리고 의미적 토픽 분류에 따른 감성 분석 결과를 비교하는 실험을 진행한다. 표 11은 세 가지 모델의 감성 분석 결과를 비교하여 리뷰별 감정 단어에서 긍정 단어와 부정 단어가 차지하는 비율을 보여준다. 전체 리뷰를 감성 분석하면 62.10%가 '긍정'으로 판별되지만 LDA 기반 토픽 분류와 의미적 토픽 분류에 따른 감성 분석 결과는 '스토리'와 '패션'은 긍정이지만 '배우'에서 부정으로 판별되는 것을 볼 수 있다. 이러한 경우 전체 리뷰의 감성 분석은 드라마를 시청할 때 '배우'를 중요시하게 보는 시청자에게는 잘못된 정보일 수 있다.

표 11. 감성 분석 비교

구분			Count	Ratio(%)	POS/NEG
전체 리뷰		POS	10847	62.10	POS
		NEG	6620	37.90	
LDA 기반 토픽 분류	배우	POS	2473	40.85	NEG
		NEG	3581	59.15	
	스토리	POS	3542	69.59	POS
		NEG	1546	30.41	
	패션	POS	3849	62.12	POS
		NEG	2347	30.41	
의미적 토픽 분류	배우	POS	2251	37.18	NEG
		NEG	3803	62.82	
	스토리	POS	3774	74.15	POS
		NEG	1316	25.85	
	패션	POS	4084	65.91	POS
		NEG	2112	34.09	

표 12는 세 가지 모델의 감성 분석 결과인 긍·부정 비율의 차를 절댓값으로 나타낸 결과이다. 전체 리뷰를 감성 분석한 결과 긍·부정 비율의 차는 24.20이고, LDA 기반 토픽 분류에서 차의 평균은 29.73, 의미적 토픽 분류는 35.25임을 볼 수 있다. 이처럼 의미적 토픽 분류의 긍·부정 비율의 차가 가장 크다는 것은 다른 모델에 비해 감성 분석이 더욱 명확하게 잘 분류된다는 것을 의미한다. 따라서 리뷰를 의미적 토픽 분류에 따라 감성을 판별하게 되면 시청자의 성향에 맞는 세밀한 결과를 도출할 수 있어서 사용자 맞춤형 분석과 추천이 가능하다.

표 12. 감성 분석 차

	전체 리뷰	LDA 기반 토픽 분류			의미적 토픽 분류		
		배우	스토리	패션	배우	스토리	패션
|POS-NEG|	24.20	18.30	39.17	31.71	25.64	48.29	31.83
평균	24.20	29.73			35.25		

V. 결론 및 향후 연구

본 논문에서는 드라마를 시청한 후 시청자들이 남긴 리뷰 데이터에서 토픽을 분류하고, 단어의 의미적 유사도를 적용하여 의미적 토픽으로 재분류 한 후, 리뷰를 토픽에 따른 문장으로 분류하고 감성 단어를 통해 감성을 판별하는 모델을 제안하였다. 총 17,467개의 데이터를 웹 크롤링하여 드라마 리뷰를 수집하고, 전처리 과정을 통하여 두 글자 이상의 보통 명사만 추출하고 불용어를 제거하였다. 전처리한 데이터를 LDA 토픽 모델링으로 실험한 결과 토픽 수를 3개로 했을 때 가장 분류가 잘 되었다. 각 토픽별 단어들을 분석한 결과 배우, 스토리, 패션에 관련된 단어들이 분류되어 있었고, '배우', '스토리', '패션'을 리뷰의 토픽으로 정의하였다. 토픽을 분류할 수 없는 포괄적인 단어들은 제거하고 토픽별 중복되는 단어는 토픽과 Word2Vec을 활용하여 의미적 유사도를 구하고 유사도가 높은 토픽으로 재분류한다. 재분류된 토픽에 따라 리뷰 문장을 분류하고 사람들이 감성을 드러내는 품사인 '형용사'와 '동사'를 감성 단어로 추출하여 감성을 판별하는 모델을 제안하였다. 기존 LDA 기반 토픽 분류와 제안하는 의미적 토픽 분류를 성능 평가 한 결과 제안하는 방법이 토픽의 응집성과 연관성에서 전반적으로 뛰어난 성능을 보였다. 그리고 전체 리뷰와 LDA 기반 토픽 분류, 의미적 토픽 분류의 감성 분석을 비교 평가한 결과 전체 리뷰는 긍정이더라도 토픽 분류에 따라 '배우'는 부정임을 확인할 수 있었다. 또한 감성 분석의 결과인 긍·부정 비율의 차를 비교하여 의미적 토픽 분류가 LDA 기반 토픽 분류보다 감성 분석이 잘 되는 것을 확인할 수 있었다. 이러한 감성 분석 모델을 활용하여 시청자의 성향을 더욱 정확하게 파악하고 세밀한 결과를 도출하여 맞춤형 서비스가 가능하고 시청자의 취향에 따른 드라마 추천 시스템에 활용이 가능하다.

REFERENCES

[1] Shuai Li, Fei Hao, Hee-Cheol Kim, "Online Social Media Review Mining for Living Items with Probabilistic Approach: A Case Study," *스마트미디어저널*, 제2권, 제2호, 20-27쪽, 2013년 6월

[2] 박승수, "온라인 리뷰와 머신러닝을 활용한 드라마 시청률 예측 모델 연구", *연세대학교 정보대학원 석사학위 논문*, 2017. 2

[3] 홍택은, 김정인, 신주현, "인스타그램 이미지와 텍스트 분석을 통한 사용자 감정 분류," *스마트미디어저널*, 제5권, 제1호, 61-68쪽, 2016년 3월

[4] 신수정, "글에서 감정을 읽다 감성 분석의 이해," *IDG Tech Report*, 1-11쪽, 2014

[5] 김혜원, "영화흥행과 온라인 구전 특성의 관련성에 관한 연구," *한국엔터테인먼트산업학회논문지*, 제4권, 제2호, 1-14쪽, 2010년 6월

[6] 장재영, "온라인 쇼핑몰의 상품평 자동분류를 위한 감성분석 알고리즘," 한국전자거래학회지, 제14권, 제4호, 19-33쪽, 2009년 11월

[7] D.M. Blei, A.Y. Ng, and M.I. Jordan, "Latent Dirichlet Allocation," *Journal of Machine Learning Research*, vol. 3, pp. 993-1022, 2003.

[8] 박창용, 이재동, 박진희, 이지형, "TV 프로그램

줄거리를 이용한 LDA 클러스터링 기반의 콘텐츠 추천 기법," *한국HCI학회 학술대회*, 618-621쪽, 2013년 1월

[9] T. Mikolov, K. Chen, G. Corrado and J. Dean, "Efficient Estimation of Word Representations in Vector Space," arXiv preprint, arXiv:1301.3781, 2013.

[10] T. Mikolov, I. Sutskever, K. Chen, G. Corrado and J. Dean, "Distributed Representations of Words and Phrases and their Compositionality," *Proc. of International conference on neural information processing systems,* pp. 3111-3119, 2013.

[11] 김선미, 나인섭, 신주현, "단어 연관성 가중치를 적용한 연관 문서 추천 방법," *멀티미디어학회논문지*, 제22권, 제2호, 250-259쪽, 2019년 2월

[12] 남민지, "SNS 해시태그를 이용한 사용자 감정 분류 방법에 관한 연구", *조선대학교 산업기술융합대학원 석사학위 논문,* 2015. 2

[13] 이상훈, 최정, 김종우, "영역별 맞춤형 감성사전 구축을 통한 영화리뷰 감성분석," *지능정보연구*, 제22권, 제2호, 97-113쪽, 2016년 6월

[14] 송종석, 이수원, "상품평 극성 분류를 위한 특징별 서술어 긍정/부정 사전 자동 구축," *정보과학회논문지:소프트웨어 및 응용*, 제38권, 제3호, 157-168쪽, 2011년 3월

[15] 오성호, 강신재, "사용자 영화평의 감정어휘 분석을 통한 영화검색시스템," *한국산학기술학회 논문지*, 제14권, 제3호, 1422-1427쪽, 2013년 3월

[16] 시청소감 : 뷰티인사이드 : 프로그램 : JTBC(2018), http://tv.jtbc.joins.com/board/pr10010943/pm10049817 (accessed Mar., 15, 2020).

[17] TV 뷰티 인사이드 :: 네이버 TV 연예(2018), https://entertain.naver.com/tvBrand/6482355 (accessed Mar,. 10, 2020).

[18] 니우한잉. "*LDA를 이용한 온라인 리뷰의 다중 토픽별 감성분석*", 부산대학교 대학원 석사학위 논문, 2018. 2

[19] Topic Modeling Using R(2019), https://entertain.naver.com/tvBrand/6482355 (accessed Mar., 4, 2020).

[20] 임명진, 신주현, "리뷰 속성 분류를 통한 감성 판별 방법," *한국스마트미디어학회 학술대회*, 제8권, 제1호, 22-24쪽, 2019년 4월

[21] 홍택은, 신주현, "이미지와 텍스트 정보의 카테고리 분류에 의한 SNS 팔로잉 추천 방법," *스마트미디어저널*, 제5권, 제3호, 54-61쪽, 2016년 9월

[22] cosine similarity(2020), https://en.wikipedia.org/wiki/Cosine_similarity (accessed Mar., 17, 2020).

[23] KNU 한국어 감성사전(2020), http://dilab.kunsan.ac.kr/knusl.html (accessed Mar., 20, 2020).

[24] 한두진. "드라마 리뷰 속성별 감성분류 방법", *조선대학교 산업기술융합대학원 석사학위 논문*, 2019. 2

[25] 곽창욱, 김선중, 박성배, 김권양, "무한 사전 온라인 LDA 토픽 모델에서 의미적 연관성을 사용한 토픽 확장," *정보과학회 컴퓨팅의 실제 논문지*, 제22권, 제9호, 461-466쪽, 2016년 9월

저 자 소 개

임명진(정회원)

2000년 군산대학교 컴퓨터과학과 학사 졸업.
2018년 조선대학교 소프트웨어융합공학과 석사 졸업.
2018년~현재 조선대학교 컴퓨터공학과 박사 과정.

<주관심분야 : 빅데이터 처리, 데이터마이닝, 자연어처리, 머신러닝 등>

김판구(정회원)

1994년 서울대학교 컴퓨터공학과 박사 졸업.
2007년~현재 조선대학교 컴퓨터공학과 교수

<주관심분야 : 지능형 정보처리, 시맨틱 웹, 온톨로지, 자연어처리, 데이터마이닝 등>

신주현(정회원)

2007년 조선대학교 전자계산학과 박사 졸업.
2018년~현재 조선대학교 신산업융합학부 부교수.

<주관심분야 : 데이터베이스, 데이터마이닝, 자연어처리, 인공지능 등>