Information Systems Review
Vol. 24, No. 1 February 2022 http://dx.doi.org/10.14329/isr.2022.24.1.001

소셜미디어 감성분석을 위한 베이지안 속성 선택과 분류에 대한 연구

Investigating the Performance of Bayesian-based Feature Selection and Classification Approach to Social Media Sentiment Analysis

강 창 민 (Chang Min Kang) 성균관대학교 경영대학 석사과정
어 균 선 (Kyun Sun Eo) 성균관대학교 경영대학 박사과정
이 건 창 (Kun Chang Lee) 성균관대학교 글로벌 경영학과/삼성융합의과학원(SAIHST) 융합의과학과 교수, 교신저자

요 약

온라인 사용자들이 소셜 미디어상에 올린 온라인 리뷰 속 숨겨진 감정을 분석하는 감성분석은 소셜미디어의 확산에 힘입어 많은 관심을 받고 있다. 본 연구는 기존 연구들과 차별화된 방법으로 감성분석을 시도하기 위하여 베이지안 네트워크에 기반한 감성 분석 모델을 제안한다. 모델에는 MBFS (Markov Blanket-based Feature Selection)가 속성 선택 기법으로 사용된다. MBFS의 성과를 실증적으로 증명하기 위하여 소셜미디어인 Yelp의 리뷰 데이터를 활용하였다. 벤치마킹 속성 선택 기법으로는 상관관계기반 속성 선택, 정보획득 속성 선택, 획득비율 속성 선택을 사용하였다. 한편, 해당 속성선택 방법을 토대로 4개의 머신러닝 알고리즘을 이용하여 분류성과를 비교하였다. 나아가 MBFS로 선택된 속성들 간 인과관계를 확인하고자 베이지안 네트워크를 통해 What-if 분석을 실시하였다. 본 연구에서 택한 머신러닝 분류기는 베이지안 네트워크 기반의 TAN (Tree Augmented Naïve Bayes), NB (Naïve Bayes), S-Spouses(Sons & Spouses), A-markov (Augmented Markov Blanket)이다. 성과분석 결과 본 연구에서 제안한 MBFS 방법이 정확도, 정밀도, F1점수 측면에서 벤치마킹 방법보다 더 우수한 성과를 나타내었다.

키워드: 소셜 미디어 감성분석, 속성선택, 마코브 블랭킷, 머신러닝, 베이지안 네트워크, 조건-결과 분석

Ⅰ. 서 론

정보의 시대, 제4차 산업혁명의 시대로 접어들며 온라인 서비스에 대한 사람들의 의존도는 높아졌다(Chan *et al.*, 2016). 의존도의 증가는 자연스레 온라인 플랫폼 내 관계의 망 형성으로 이어졌다. 그리고 이는 소셜미디어라고 칭해진다. 소셜미디어 속 형성된 관계의 망에서는 단순한 대화뿐 아니라, 정보와 의견의 공유가 이루어진다(Asur and

† 이 논문은 2019년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임(NRF-2019S1A5A2A01046529).

Huberman. 2010). 소셜미디어를 사용하는 사람들이 늘어남에 따라 소셜미디어를 마케팅의 수단으로 활용하는 기업들이 생겼다. 적은 비용으로 보다 효과적인 광고를 할 수 있다는 장점은 소셜미디어 마케팅을 기업의 필수 요소로 발전시켰다(Cho *et al.*, 2012). 소셜미디어 마케팅은 소셜미디어를 근간으로 하기 때문에 1차 소비자들의 반응 및 감정을 기업뿐 아니라 2차, 3차 소비자들도 즉각적으로 확인할 수 있다. 이러한 반응의 투명성은 기업에게 리뷰의 중요성을 각인시켜주는 역할을 하였으며, 온라인 리뷰를 활용한 감성분석 연구에 이목을 집중시켰다(Yoo *et al.*, 2018).

COVID-19의 확산은 감성분석의 중요성을 다시 한 번 고찰하게 한다. 세계보건기구에 의하면 신종 코로나 바이러스 확진자 수는 4억 명이 넘으며, 사망자는 580만 명을 상회한다. COVID-19는 전 세계적인 문제로, 이에 대한 뉴스가 소셜미디어 웹 사이트 전체에 퍼지고 있다. 소셜미디어는 발병 관련 사건들을 전달하고 사용자들은 소셜미디어 속에서 다양한 의견 및 감정을 공유하고 경험한다(Chakraborty *et al.*, 2020). 바이러스에 대한 정보, 자국민들이 필요로 하는 정보를 파악하고 전달하는 것은 국민들을 바이러스에 대한 무지로부터 벗어나게 해줌과 동시에 국민들에게 경각심을 안겨준다. 그리고 이는 예방적 차원에서 필요하다. 따라서 소셜미디어 속 빅 데이터를 분석하는 것은 국가적으로 중요하다.

소셜미디어 속 빅 데이터 분석의 중요성이 재조명된 것은 기업적 차원에서도 마찬가지이다. COVID-19의 빠른 확산은 오프라인 대인관계를 단절시켰으며, 온라인 관계의 망을 더욱 크고 견고하게 만들었다. 즉, 소셜미디어에 대한 사람들의 의존도가 폭발적으로 증가한 것이다. 모든 것이 온라인으로 이루어진다는 것은 더 이상 과장이 아니다. 그리고 이는 기업의 소셜미디어 마케팅 증가로 이어진다(Wang *et al.*, 2020).

감성분석의 필요성이 증가하는 시대적 흐름에 발맞춰 보다 자세하고 효과적인 분석 기법이 필요하다. 감성분석 프로세스는 다음과 같다. 우선, 감정을 추출할 소스를 선택한다. 즉, 분석하고자 하는 소셜미디어 플랫폼을 정한다. 두번째 단계는 데이터 수집 프로세스이다. 트위터를 통해 감성분석을 하는 경우 해시태그를 활용하여 데이터를 수집하는 것이 그 예이다. 다음 단계는 전처리이며, 이 단계에서 행하는 대표적인 처리는 속성 선택이다. 마지막 단계는 머신러닝 분류기를 통해 감성분석을 실시하는 것이다(Alamoodi *et al.*, 2020).

리뷰 작성은 소비자의 구매 행위에 따른 사후적 행위이다. 리뷰는 구매에 선행하여 발생하지 않기 때문에, 사후발생적 관점에서 리뷰를 연구할 필요성이 있다. 그러나 리뷰 감성 분석에 있어 이러한 사후확률론적 접근은 아직 이루어지지 않고 있다. 따라서 본 연구는 대표적 사후확률론인 베이지안 네트워크 중심의 모델을 효과적인 감성 분석 모델로 제안하며, Yelp 리뷰 감성 분석을 통해 이를 입증한다. 우선, 리뷰 전문의 다양한 단어들을 베이지안 속성 선택 방법인 마코브 블랭킷 속성 선택(Markov Blanket-based Feature Selection, MBFS)으로 선정한다. 그리고 베이지안 네트워크 기반의 분류기를 사용하여 감성분석 성능을 비교한다. 마지막으로 MBFS로 선택된 속성들의 인과관계를 What-if 분석으로 확인한다. 본 연구의 연구질문은 다음과 같다.

연구질문 1: 감성분석에서 베이지안 네트워크 기반의 속성선택 모델은 어떠한 성과를 보이는가?

연구질문 2: 감성분석에 베이지안 네트워크를 적용하여 What-if 분석을 실시할 때 어떠한 인과관계를 보이는가?

Ⅱ. 관련연구

2.1 감성분석

이모션 마이닝(Emotion Mining)은 문장 혹은 문서에서 나타나는 집필자의 감정을 추출 및 분석하

여 이해하고 나아가 예측하는 것이다. 이모션 마이닝은 감정을 기쁨, 놀라움, 분노, 슬픔 등 다양한 클래스로 분류한다(Yassine and Hajj, 2010). 감성분석(Sentiment Analysis, Opinion Mining)은 집필자의 감정을 분석한다는 목적차원에서 이모션 마이닝과 동일하다. 그러나 감정을 긍정과 부정으로 분류한다는 차별점이 존재한다(Prabowo and Thelwall, 2009). 정보기술의 혁신과 온라인 서비스의 활성화는 이러한 감성분석을 온라인 리뷰 및 SNS(Social Network Service)에 적용시키려는 노력으로 이어졌다(Feldman, 2013; Min, 2020). 온라인 데이터를 활용한 감성분석은 오늘날에도 큰 관심을 받고 있으며 지속적으로 성장 중이다(Yadav and Vishwakarma, 2020). <표 1>은 온라인 데이터 기반 최근(2015~2020)의 감성분석 연구들을 나타낸 것이다.

Gokalp *et al.*(2020)은 감성분석 연구를 위해 Iterated Greedy를 토대로 한 새로운 래퍼(Wrapper) 기법을 제안하였다. 세 종류의 베이지안 분류기를 활용하여 각종 리뷰 데이터를 분석한 결과 새로운 래퍼기법을 적용한 방식이 감성분석에서 높은 경쟁력을 보였다. Costello and Lee(2020)는 전기자동

〈표 1〉 감성분석 선행연구

Authors	Dataset	FS Method	Classifier	Purpose / Contribution
Gokalp *et al.* (2020).	9 Public Sentiment Dataset and 4 Amazon Reviews	CFS, Chi-square, GR, IG, ReliefF, Symmetric uncertainty	Bayesian Logistic Regression, Naïve Bayes, Multinomial Naïve Bayes	Proposed a novel iterated greedy based feature selection algorithm for sentiment analysis.
Costello and Lee (2020)	Electric Vehicles Social Media Data (Youtube)	CFS, Chi-square, IG, ReliefF	Decision Tree, Naïve Bayes, Support Vector Machines, Logistic Regression, Bagging, Random Forrest, Random SubSpace, Adaptive Boosting	Provided a real-world example of social media sentiment analytics. It can be adopted in other areas of research and business.
Eo and Lee (2019).	5 Amazon Reviews (apparel, book, dvd, electronic, kitchen)	CFS, IG, ReliefF	Logistic Regression, Decision Tree, Naïve Bayes, Neural network, Support Vector Machines, Bagging, Random Forrest, Random SubSpace, Stacking	Proposed a machine learning model to predict positive and negative opinions of text using opinion mining.
Parlar *et al.* (2018)	Movie Reviews, Product Reviews (Turkish, English)	Chi-square, IG, Document Frequency Difference, Optimal Orthogonal Centroid, Query Expansion Ranking	Multinomial Naïve Bayes, Support Vector Machines, Maximum Entropy Modelling, Decision Tree	Proposed a new feature selection method - query expansion ranking. It's based on query expansion term weighting methods which is from the information retrieval field.
Sihwi *et al.* (2018)	Twitter (12 popular movies)	IG	Naïve Bayes	Proposed IG + NB model for Movie tweets. Information Gain was chosen in order to increase the run time efficiency. Naive Bayes Classifier was chosen due to greater accuracy.

〈표 1〉 감성분석 선행연구(계속)

Authors	Dataset	FS Method	Classifier	Purpose / Contribution
Yousefpour *et al.* (2017)	Movie reviews, Amazon reviews (Book, Electronic, Kitchen, Music)	Chi-square, IG, OIFV, FIFS	Naïve Bayes, Support Vector Machines, Maximum Entropy, Linear Discriminant Function	Proposed two methods for integrating feature selection in sentiment analysis. - Ordinal-based integration of different feature vectors (OIFV) - Frequency-based integration of different feature subsets (FIFS)
Liu *et al.* (2017)	Movie reviews(1-8) Short informal texts (9-12)	Document Frequency, Chi-square, IG, GR	Decision Tree, Naïve Bayes, Support Vector Machine, Radial Basis Function Neural Network, K-Nearest Neighbor	Proposed that gain ratio performs best among four feature selection algorithms. support vector machine performs best among five learning algorithms(Accuracy)
Lee and Hong (2015)	Amazon reviews (Movie, Book)	Document Frequency, Chi-square, IG, Term Frequency - Inverse Document Frequency	Support Vector Machine	SVM model based on Chi-square Feature Selection shows the most superior performance.
This Study	Yelp reviews	Benchmark: CFS, IG, GR, RQ: MBFS	Tree Augmented Naïve Bayes Naïve Bayes Sons & Spouses, Augmented Markov Blanket	Proposing that Bayesian Network based model will show the most effective results for sentiment analysis.

차 소셜미디어(Youtube)의 데이터로 감성분석을 실시하였다. 분류기를 단일 분류기와 앙상블 분류기로 나누어 분석을 실시하였으며, 속성 선택 기법으로 상관관계기반 속성 선택, 카이제곱 속성 선택, 정보획득 속성 선택, ReliefF를 사용하였다. 단일 분류기인 서포트 벡터 머신과 로지스틱 회귀분석이 유의미한 결과를 나타냈다. Eo and Lee (2019) 또한 앙상블 분류기와 단일 분류기를 구분하여 감성분석을 실시했다. 단일 분류기로는 로지스틱 회귀분석, 의사결정나무, 나이브 베이즈, 신경망, 서포트 벡터 머신을 사용하였으며, 배깅, 랜덤 포레스트, 랜덤 서브스페이스, 스태킹을 앙상블 분류기로 활용하였다. 이 중 정보 획득 속성 선택 바탕의 스태킹 분류기가 좋은 결과를 보였다. Parlar *et al.*(2018)은 감성분석에 효율적인 속성 선택 기법으로 QER을 새롭게 제안하였다. QER은 쿼리 확장 용어 가중치를 기반으로 한 속성 선택 기법으로 다항분포 나이브 베이즈 분류기와 함께 사용되었을 때 효과적인 결과를 보여주었다. 마찬가지로 Yousefpour *et al.*(2017)은 두 가지의 통합적 속성 선택 기법, OFIV와 FIFS를 제안하고 이를 카이제곱 속성 선택, 정보획득 속성 선택의 결과와 비교분석하였다. 이 밖에도 Sihwi *et al.*(2018)은 영화 관련 트위터를 감성분석 할 때 정보획득 속성 선택과 나이브 베이즈를 사용하는 것이 높은 정확도를 보인다고 주장했다. 이는 82.19%의 높은 정확도로 뒷받침 되었다. 또한 Liu *et al.*(2017)은 4가지의 속성 선택 기법을 5가지의 분류기로 성과 비교하였다. 속성 선택 기법에선 획득 비율 속성 선택, 분류기에선 서포트 벡터 머신(SVM)이 높은

정확도를 보였다. 마지막으로 Lee and Hong(2015)은 4가지의 속성 선택 기법을 비교분석하여 카이 제곱 속성 선택 기반의 서포트 벡터 머신 모델이 가장 효율적임을 나타냈다.

본 연구는 최근 감성분석 연구들의 동향을 파악하고 도입하고자 한다. 따라서 선행연구들에서 주로 사용된 상관관계, 정보획득, 획득비율 속성 선택 기법을 벤치마킹하여 MBFS 기법과 비교한다. 그리고 이는 TAN(Tree Augmented Naïve Bayes), 나이브 베이즈, Sons & Spouses, 그리고 Augmented Markov Blanket 총 4가지의 분류기 성과분석을 통해 이루어진다.

2.2 속성 선택

2.2.1 상관관계기반 속성 선택 (Correlation based Feature Selection)

속성 선택은 학습에 있어 중요한 속성들을 구분한다. 그리고 분석 및 예측을 하는데 가장 유용한 데이터에 학습의 알고리즘을 집중시킨다. 따라서 모델의 중요 속성 집합을 식별하는 속성 선택은 머신러닝의 핵심이다(Kira and Rendell, 1992). CFS는 속성들을 식별하는 대표적인 속성 선택 기법으로 상관관계를 그 근간으로 한다. 이는 속성이 타겟 변수와 상관관계가 높을수록 기여도 또한 높을 것으로 생각하기 때문이다. 구체적으로 CFS는 "좋은 속성 집합은 목표변수와 높은 상관관계를 가지지만 속성 간에는 서로 상관관계가 없는 집합이다"를 기본 가정으로 한다. 이러한 가정을 바탕으로 상관관계(Correlation) 기반 휴리스틱 평가 함수에 의해 속성 선택이 이루어진다. 그러나 이 같은 방법은 작은 예측 값을 보인 속성이 큰 예측 값을 보여주는 속성에 가려져 중요 속성으로 선택되지 않는다는 맹점이 있다. CFS의 속성 평가 함수는 다음과 같다(Hall, 1999).

$$M_S = \frac{k\overline{r_{cf}}}{\sqrt{k + k(k-1)\overline{r_{ff}}}}$$

M_S = Heuristic "merit" of a feature subset S containing k features

$\overline{r_{cf}}$ = Average feature-class correlation($f \in S$)

$\overline{r_{cf}}$ = Average feature-feature intercorrelation

2.2.2 정보획득(Information Gain)과 획득비율(Gain Ratio)

정보획득(Information Gain, IG)은 텍스트의 불확실성(엔트로피)을 활용하는 알고리즘이다. IG 알고리즘에 의하면 텍스트의 불확실성이 커질수록 클래스(감성 값)의 불확실성도 커진다. 따라서 IG는 속성이 갖는 불확실성이 낮을수록 중요도가 높아진다는 원리를 바탕으로 속성을 선택한다. 구체적인 계산 과정은 다음과 같다.

우선 Y에 대한 불확실성은 다음과 같이 계산된다.

$$H(Y) = -\sum_{y \in Y} p(y)\log_2(p(y)).$$

X를 관찰 한 후의 Y에 대한 불확실성을 계산하는 공식은 다음과 같다.

$$H(Y|X) = -\sum_{x \in X} p(x)\sum_{y \in Y} p(y|x)\log_2(p(y|x)).$$

마지막으로 위의 두 공식을 활용하여 IG값을 도출하는 공식은 다음과 같다.

$$IG(Class, Attribute) = H(Class) - H(Class|Attribute)$$

그러나 IG 알고리즘은 일변량 방식으로 속성을 선택하므로 중복 속성을 처리할 수 없다는 한계를 갖는다(Tang, 2014).

획득비율(Gain Ratio, GR)은 IG 알고리즘을 속성의 불확실성으로 나눈 것으로 IG를 정규화한 기법이다. GR은 IG의 편향을 감소시킨다(Dağ *et al.*, 2012; Karegowda *et al.*, 2010; Quinlan, 1986).

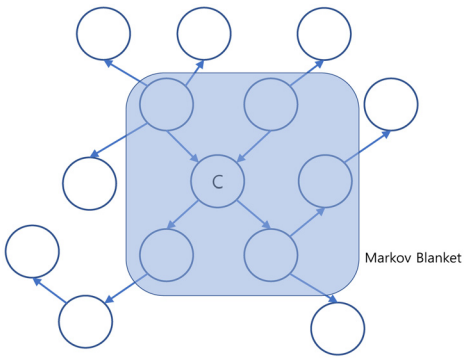
$$GR(Class, Attribute) = (H(Class) - H(Class|Attribute)) / H(Attribute)$$

2.2.3 래퍼(Wrapper)

상관관계와 같은 필터방식의 속성선택 기법은 변수 간의 연관성을 측정한다. 반면, 래퍼(Wrapper)는 변수 집합의 유용성을 측정하는 것을 근간으로 한다. 즉, 래퍼는 정확도, 정밀도, F1점수와 같은 분류 성능 척도를 기준으로 가능한 모든 속성의 조합을 평가하여 최적의 속성 집합을 찾는다. 따라서 래퍼는 필터방식보다 모델의 성능 향상 측면에서 좋은 결과를 보인다. 그러나 이러한 특징으로 인해 표본의 수가 충분하지 않을때 과적합의 위험이 있다(Kohavi and John, 1997).

2.2.4 마코브 블랭킷 속성 선택(Markov Blanket-based Feature Selection)

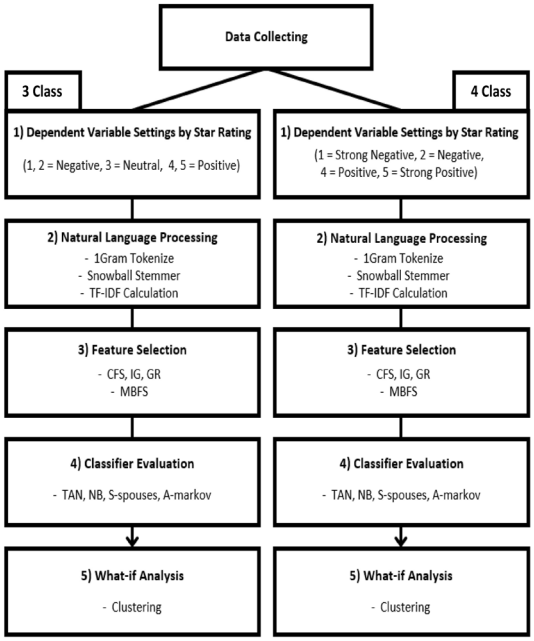
베이지안 네트워크에서 노드는 변수를 의미하며, 노드를 연결하는 선은 변수 간 조건부 의존성을 나타낸다. 타겟 노드와의 관계를 통해 노드 집합을 설정하는 마코브 블랭킷(MB)은 속성 선택의 기법으로도 사용 된다(Koller and Sahami, 1996). MB를 통해 설정된 집합은 타겟 노드(노드C)의 부모 노드, 자식 노드, 그리고 배우자 노드(자식의 다른 부모 노드)로 구성된다. 자식 노드와 부모 노드가 MB로 설정되는 이유는 두 개의 노드가 타겟 노드와 직접적 연관이 있기 때문이다. 반면, 배우자 노드는 두 가지와 다르게 타겟 노드와의 관계에서 V-Structure를 보인다. 그럼에도 MB으로 설정되는 이유는 자식노드의 정보가 결정되어도 타겟 노드와 의존관계를 보여주기 때문이다. MB로 설정된 노드들은 타겟 노드를 나머지 노드, 즉 독립적인 노드들의 네트워크로부터 보호해준다. 이는 MB가 타겟 노드의 결과를 예측하는데 필요한 유일한 정보임을 의미한다. MB는 데이터 집합의 변수가 많을 때 효율적인 속성 선택 기법으로 사용될 수 있다(Wang *et al.*, 2020). MB의 구조는 <그림 1>과 같다.



〈그림 1〉 마코브 블랭킷

Ⅲ. 연구방법

본 연구의 분석 절차는 다음 <그림 2>와 같다.



〈그림 2〉 분석 절차

3.1 종속변수 정의

본 연구는 Yelp.com에서 제공하는 리뷰 데이터를 대상으로 감성분석을 시행하였다. Yelp는 소비

자들이 식당과 다른 사업체에 대한 리뷰를 남길 수 있는 웹사이트이다(Luca, 2016). Yelp에서 사용자들은 모든 식당을 점수 1에서 5까지 평가하고 리뷰를 작성할 수 있다. Yelp 데이터의 예시는 다음 그림과 같다. 예시 속 고객은 샌프란시스코 근처에 위치한 식당을 이용하고 식당에 대한 별점을 등록한 후, 리뷰 글을 작성하였다. 본 연구는 Yelp로부터 약 8,000건의 데이터를 수집하였으며, 중립적, 긍정적 그리고 부정적 의견을 모두 파악하기 위해 종속변수를 3클래스와 4클래스로 나누었다. 3클래스는 중립적 리뷰를 포함하여 구성했고, 4클래스는 긍정과 부정을 세분화하였다. 우선 3클래스는 별점 4와 5에 해당하는 리뷰는 긍정적 리뷰, 3의 리뷰는 중립적 리뷰, 그리고 1과 2의 리뷰는 부정적 리뷰로 설정하였다(별점 1,2=0 / 별점 3=1 / 별점 4,5=2). 반면, 4클래스는 별점 5에 해당하는 리뷰를 강한 긍정 리뷰, 4에 해당하는 리뷰를 약한 긍정 리뷰로 분류하였으며, 별점 1의 리뷰를 강한 부정 리뷰, 2의 리뷰를 약한 부정 리뷰로 구분하였다(별점 1=0 / 별점 2=1 / 별점 4=2 / 별점 5=3). 이러한 세분화 과정을 통해 3클래스 7,497건, 4클래스 6,590건의 데이터를 추출, 분석하였다.

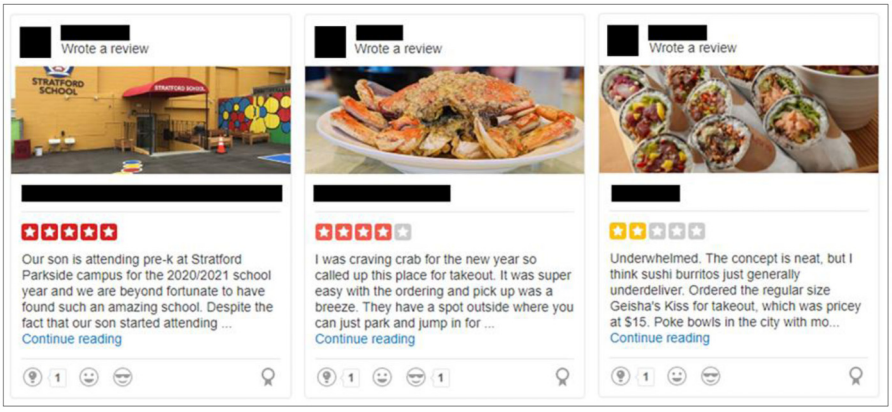
본 연구에 사용된 데이터의 기술통계량은 <표 2>와 같다.

3.2 텍스트 전처리

텍스트 데이터를 처리하는 것은 다음과 같이 세 단계로 구성된다. 첫째, 토큰화, 둘째, 어간 추출(Stemmer), 셋째, TF-IDF 산출이다. 토큰화 단계는 문장을 단어 단위로 분리하는 단계이다. 단어의 집합으로 구성된 문장들은 토큰화 단계를 거쳐 단어로 분리된다. 그리고 분리된 단어들은 빈번하게 등장하는 단어 또는 거의 등장하지 않은 단어들로 구분된다. 어간 추출은 어형이 변형된 단어에서 접사를 제거하고 단어의 어간을 분리해내는 것이다. 예를 들어, “dogs”의 어간으로는 “dog”가 추출된다. 마지막으로 TF-IDF는 단어의 빈도 및

〈표 2〉 기술통계량

기술통계	N	최소값	최대값	평균	표준편차	분산
원본 데이터(별점1~5)	7952	1	5	3.72	1.5456	2.389
4클래스 데이터	6590	0	3	2.06	1.1655	1.358
3클래스 데이터	7497	0	2	1.41	0.8544	0.730



〈그림 3〉 Yelp 리뷰

역 문서 빈도를 사용해 문서 내 각 단어마다 중요한 정도를 나타내는 방법이다. 이를 수식으로 표현하면 다음과 같다(Erra *et al.*, 2015).

$f(t,d)$는 d문서 속 t단어의 빈도를 나타내며 $|\{d|t\in d\}|$는 t단어가 포함된 문서 d의 수를 의미한다.

$$tf(t,d) = \frac{f(t,d)}{|d|}$$

$$idf(t,D) = \log\frac{|D|}{|\{d|t\in d\}|}$$

$$tfidf(t,d,D) = tf(t,d) \times idf(t,D)$$

3.3 속성 선택 적용

자연어 처리 단계를 거친 데이터는 3클래스, 4클래스 각 1,339개, 1,261개의 속성으로 구성되어 있다. 이 중 불필요한 변수를 선별하기 위해 다음의 속성 선택 기법을 적용하였다. 적용된 속성 선택 기법은 CFS, IG, GR, MBFS이며, 이 중 IG와 GR은 영향을 미치는 속성만을 선택하기 위해 임계값을 0보다 높은 값으로 설정하였다. 선택된 속성들의 수는 <표 3>과 같다. 3클래스의 경우 총 1,339개의 속성 중 CFS를 통해 61개, IG와 GR을 통해 576개, 그리고 MBFS를 통해 77개의 속성들이 선택되었다. 4클래스의 경우 총 1,261개의 속성 중 CFS: 62개, IG: 569개, GR: 569개, MBFS: 82개의 속성들이 선택되었다. IG와 GR의 경우 단어별 영향을 미치는 정도에 차이는 있었으나 속성의 구성에는 차이가 없었다. 이는 IG의 편향을 감소시키는 것이 속성의 구성을 바꾸지는 못하였음을 의미한다.

〈표 3〉 선택된 속성의 수

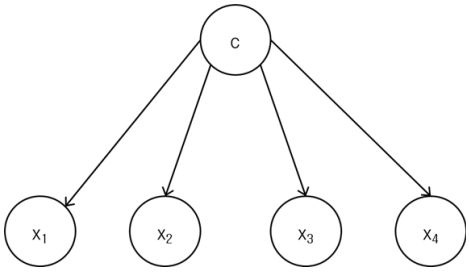
Number of Selected Features	3 Class	4 Class
Before	1339	1261
CFS	61	62
IG	576	569
GR	576	569
MBFS	77	82

3.4 분류기 평가

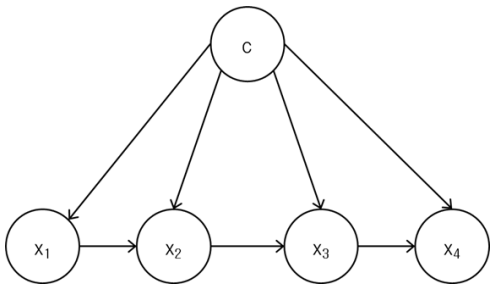
본 연구는 분류기를 사용하여 속성 선택 기법의 유효성을 검증하였다. 사용된 분류기는 TAN, NB, S-Spouses, A-markov이다. 속성들 간의 방향성, 인과관계를 확인하는 것이 목적이기 때문에 표준통계 기법을 제외하고 베이지안 네트워크 기반의 기법들만 사용하였다.

베이지안 네트워크는 확률분석과 그래프 이론에 기반한 불확실성 인과관계 표현 및 추론 모델이다. 모델은 노드와 호(Arc)로 구성되어 있으며 호는 노드 간 확률적 상관관계를 나타낸다. 또한, 베이지안 네트워크는 그래프의 노드로 표시되는 모든 속성에 대한 전반적 분포를 나타낸다. 그리고 이 분포는 조건부 확률표(Conditianl probaility table, CPT)에 의해 구성된다(Li and Abdul Rahman, 2018). TAN은 트리와 유사한 나이브 베이즈 형태로, 클래스 노드가 모든 속성노드와 직접 연결되어 있으며 속성노드는 상, 하위노드로 구성되어 있다. 속성노드 간에는 호의 제한이 없다(Jiang *et al.*, 2012). S-Spouses는 클래스 노드가 잠재적으로 다른 배우자 노드 집합을 가진 구조이다. <그림 6>을 살펴보면 자식 노드에는 클래스 노드를 포함한 여러 부모 노드가 연결되어 있다. 이 모델은 자식 노드들이 우선순위를 고정하지 않고 대상의 한계 의존성에 따라 검색되는 A-NB(Augmented Naïve Bayes)의 확장된 방법이다(Costello *et al.*, 2020; Prabhakaran *et al.*, 2016). A-markov는 모델을 마코브 블랭킷 구조로 초기화한 다음, 각 속성 노드 사이에 유지되는 확률적 관계를 비지도 탐색으로 찾는 기법이다. 비지도 탐색 방법은 추가적인 시간비용이 발생하지만 초기버전보다 더 나은 예측결과를 보여준다(Conrady and Jouffe, 2015).

또한 보다 견고한 분석을 위하여 10겹 교차검증 방법을 사용한다. 10겹 교차검증은 전체 데이터 셋을 10분할한 다음 9분할은 학습용으로 사용하고 나머지 1분할은 검증용으로 활용한다(Arlot and Celisse, 2010). 이 때 검증용 분할은 순차적으로 진행되며, 모든 데이터가 검증된다.

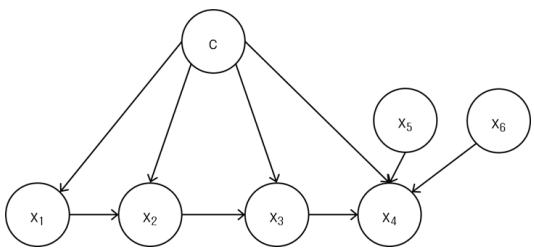


〈그림 4〉 Naïve Bayes



〈그림 5〉 Tree Augmented Naïve Bayes

다음으로 분류기에 관한 성능을 평가하기 위해 정확도(Accuracy), 정밀도(Precision), F1점수(F-Measure)를 지표로 적용한다. 평가지표들은 <표 4>의 혼동 행렬을 바탕으로 산출된다. TP(True Positive)는 긍정리뷰를 분류기가 긍정으로 분류한 건수이고, FP(False Positive)는 부정 리뷰를 긍정으로 분류한 건수이다. TN(True Negative)은 부정 리뷰를 분류기가 부정으로 분류한 건수이다. 마지막으로, FN(False Negative)은 긍정 리뷰를 분류기가 부정으로 분류한 건수이다. 정확도, 정밀도, F1점수의 수식은 다음과 같다.



〈그림 6〉 Sons & Spouses

〈표 4〉 혼동 행렬

		Predicted Class	
		Positive	Negative
Actual Class	Positive	TP	FN
	Negative	FP	TN

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{F-Measure} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Ⅳ. 연구결과

4.1 분류 결과

본 연구에서는 베이지안 기반 감성분석의 효과성을 알아보기 위해 기존의 머신러닝 방법에서 사용되는 CFS, IG, GR과 비교하였다. 비교는 분류성과 분석으로 진행되었으며 사용된 알고리즘은 TAN, NB, S-Spouses, A-markov이다. <표 5>는 3클래스 분석을 나타내고, <표 6>은 4클래스 분석을 보여준다. 속성선택 전의 값보다 높은 값을 보인 결과들은 음영 처리를 하였다. 그리고 알고리즘별 가장 높은 성과가 도출된 값은 강조 처리를 하였다. 또한 성능의 우수성을 보다 명확하게 비교하기 위해 2클래스 분류를 추가로 실시하였으며, 2클래스 분석의 결과는 <부록 A>와 같다.

3클래스 분석에서는 MBFS를 적용한 결과가 적용하기 전보다 모든 평가지표에서 높은 결과를 나타냈다. CFS를 사용한 결과 또한 적용하기 전의 결과와 비교하여 높게 나타났다. 정확도는 MBFS를 적용한 A-markov의 결과 77.84로 가장 높게 산출되었고 정밀도와 F1점수는 CFS를 사용한 A-markov의 결과가 0.83과 0.80으로 가장 높았다.

〈표 5〉 3클래스 평가지표 결과

Accuracy	TAN	NB	S_Spouses	A_markov
Before	70.83	69.51	73.50	77.23
CFS	**77.59**	77.24	76.96	77.36
IG, GR	71.78	70.21	73.40	77.26
MBFS	**77.44**	**77.28**	**77.42**	**77.84**
Precision	TAN	NB	S_Spouses	A_markov
Before	0.70	0.68	0.73	0.80
CFS	0.80	0.80	0.80	0.83
IG, GR	0.70	0.68	0.73	0.80
MBFS	**0.80**	**0.80**	**0.80**	**0.81**
F-measure	TAN	NB	S_Spouses	A_markov
Before	0.70	0.69	0.73	0.78
CFS	0.79	0.78	0.78	0.80
IG, GR	0.71	0.69	0.73	0.78
MBFS	**0.79**	**0.79**	**0.79**	**0.79**

〈표 6〉 4클래스 평가지표 결과

Accuracy	TAN	NB	S_Spouses	A_markov
Before	61.70	61.37	63.57	65.49
CFS	64.87	64.93	64.37	64.25
IG, GR	64.51	63.67	63.38	65.84
MBFS	**65.37**	**65.37**	**65.42**	65.42
Precision	TAN	NB	S_Spouses	A_markov
Before	0.61	0.61	0.66	0.68
CFS	**0.69**	**0.69**	0.68	**0.70**
IG, GR	0.64	0.63	0.65	0.69
MBFS	0.68	0.68	**0.68**	0.69
F-measure	TAN	NB	S_Spouses	A_markov
Before	0.61	0.61	0.65	0.67
CFS	0.67	0.67	0.66	0.67
IG, GR	0.64	0.64	0.64	0.67
MBFS	**0.67**	**0.67**	**0.67**	**0.67**

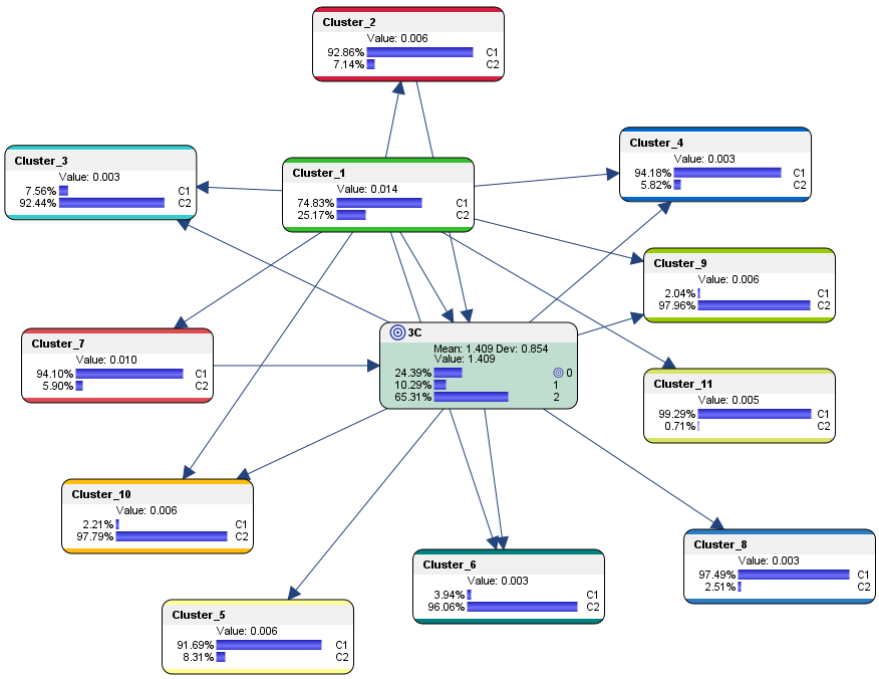
4클래스에서 MBFS 모델의 결과 중 A-markov를 제외한 TAN, NB, S-Spouses가 속성선택 전보다 높은 정확도를 보였다. 그리고 IG와 GR을 사용한 A-markov 방법이 가장 높은 결과를 나타냈다. 정밀도 결과에서는 MBFS와 CFS가 속성선택 전과 비교하여 상승하였다. 마지막으로, F1점수에서 MBFS 모델은 A-markov를 제외한 모든 알고리즘에서 성과의 향상을 보이며 정확도의 결과와 비슷한 양상을 나타냈다.

4.2 What-if 분석 결과

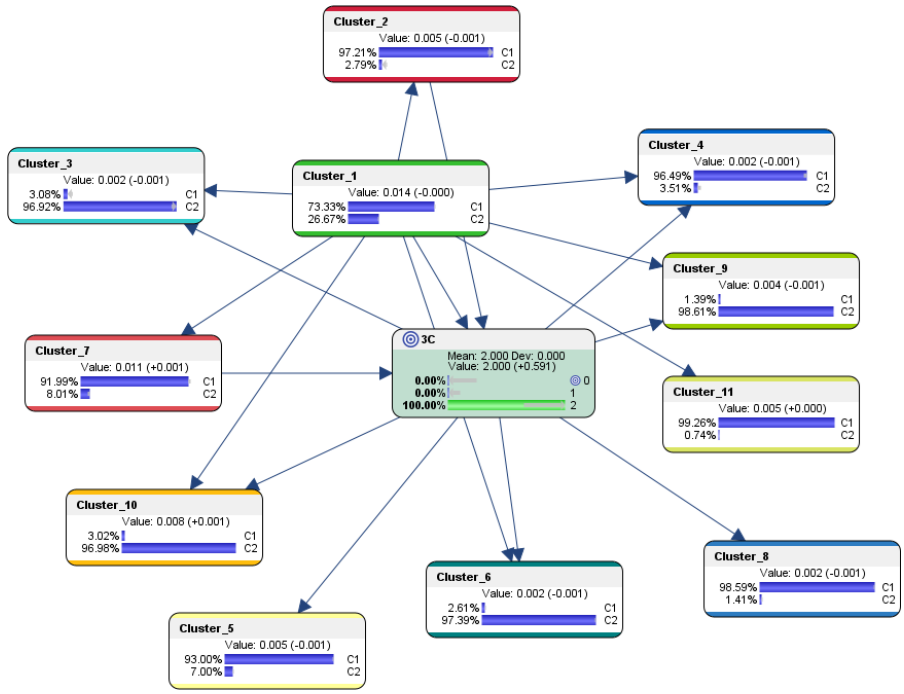
타겟 변수와 속성 간 인과관계를 자세히 알아보기 위해 What-if 분석을 실시하였다. What-if 분석을 시행하기에 앞서 MBFS로 선택된 속성들을 클러스터링 하였으며, 클러스터링은 분류분석에서 보편적으로 사용되는 계층적 결합 알고리즘(Hierarchical agglomerative clustering algorithm)을 통해 이루어졌다(Murtagh and Contreras, 2012). 계층적 결합 클러스터링은 독립적인 클러스터에서 점차 가까운 클러스터를 탐색하여 군집을 이루어 나가는 상향식 알고리즘으로, 이를 통해 유사한 성향을 갖고 있는 단어들을 하나로 군집화하여 군집 간의 관계를 파악할 수 있다. 3클래스의 경우 58개의 단어가 11개의 클러스터로 분류되었으며, 4클래스의 경우 56개의 단어가 8개의 클러스터로 지정되었다. 클러스터 별 구체적인 단어 구성은 <부록 B>, <부록 C>와 같다.

<그림 7>은 3클래스 모델의 초기관계를 보여주고, <그림 8>은 목표변수의 긍정(별점 4, 5)을 100%로 조절하였을 경우 What-if 분석 모델을 나타낸다. What-if 분석 결과 긍정리뷰를 100%로 조절하였을 때, 양(+)의 변동을 보이는 클러스터로는 클러스터 7, 10, 11이 있었다. 즉, 클러스터 7, 10, 11은 긍정리뷰와 정의 관계를 보이는 것으로 확인되었다. 클러스터 7을 구성하는 단어들로는 amaze, delicious, favorite, perfect, yummy 등이 있다. 반면 나머지 8개의 클러스터들은 긍정리뷰를 100%로 조절하였을 때, 음(-)의 변동을 보였다.

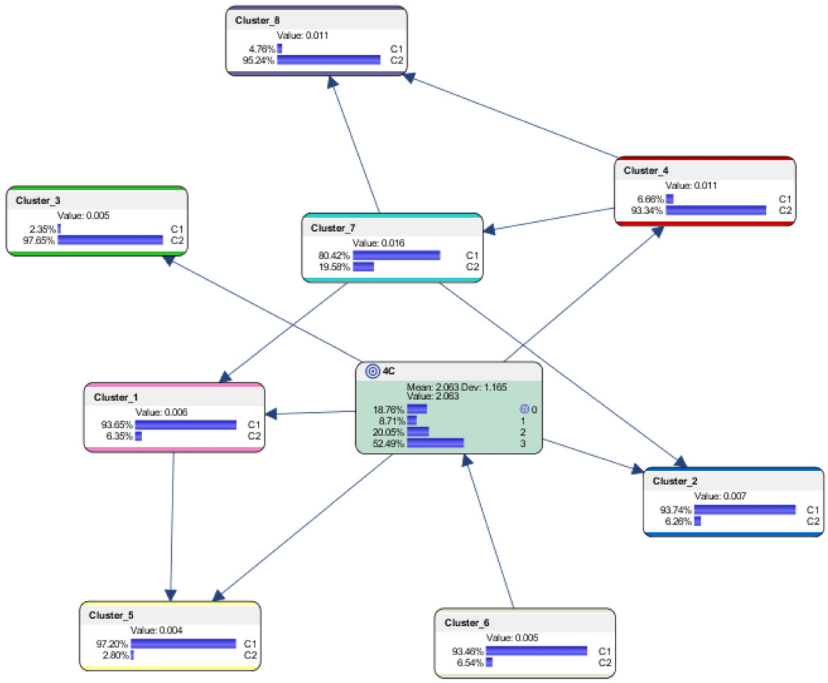
<그림 9>와 <그림 10>은 4클래스 모델의 What-if 분석 전과 후를 도식화한 것이다. <그림 10>은 강한 긍정리뷰를 100%로 조절한 경우로, 8개의 클러스터 중 클러스터 4와 클러스터 8만이 양(+)의 변동을 보였다. 나머지 6개의 클러스터들은 모두 음(-)의 변동을 나타냈다. 그 중 클러스터 6은 -0.003으로 가장 큰 변동 값을 보였다. 클러스터 6의 구성 단어들은 bad, horrible, manage, rude, terrible, worst이다.



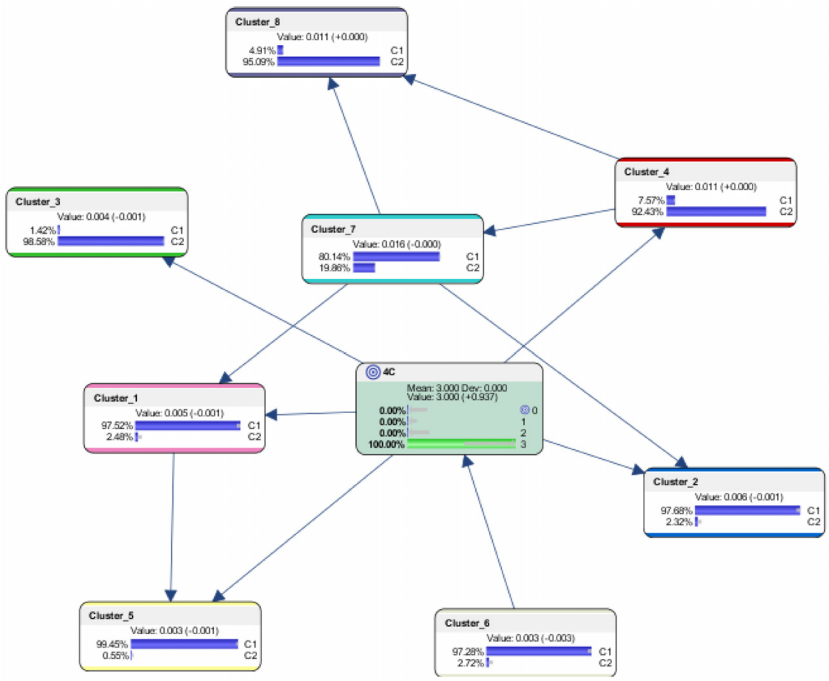
〈그림 7〉 3클래스 초기 모델



〈그림 8〉 3클래스 What-if 분석 모델 - 긍정리뷰 100%



〈그림 9〉 4클래스 초기 모델



〈그림 10〉 4클래스 What-if 분석 모델 - 강한 긍정리뷰 100%

Ⅴ. 토의 및 결론

본 연구의 결과 요약은 다음과 같다. 첫째, 성과분석에서 모든 속성을 사용하는 것보다 MBFS를 통해 속성 선택을 하는 것이 감성분석에 효과적이다. 3클래스 분석 시 정확도와 정밀도 그리고 F1점수에서 향상을 보였으며, 4클래스 또한 A-markov를 제외하고 모두 향상된 결과가 도출되었기 때문이다. 둘째, 3클래스와 4클래스 모든 경우에서 MBFS 모델은 벤치마킹 속성 선택 기법을 사용한 모델들과 유사하거나 높은 결과를 보였다. 이러한 결과는 베이지안 네트워크를 근간으로 하는 MBFS 모델의 우수성을 증명한다. 셋째 베이지안 네트워크를 적용하여 What-if 분석을 실시한 결과, 정의 관계와 부의 관계를 보이는 구체적인 단어 클러스터들을 확인할 수 있었다. 즉, 베이지안 네트워크를 활용하여 감성분석을 실시하였을 때 높은 분류 성과를 보여줄 뿐 아니라 변수들 간 확률적 인과관계를 확인할 수 있었다. 이는 베이지안 네트워크를 적용한 모델이 감성분석의 효과성 측면에서 적합하다는 것을 의미한다.

본 연구의 학술적 의의는 다음과 같다. 첫째, 베이지안 네크워크는 사회과학 분야에서 과거부터 변수 간의 인과관계를 파악하기 위한 도구로써 주로 사용되었다. 본 연구에서는 이를 감성분석 연구에 적용하여 기존의 머신러닝 방법으로 사용되는 속성 선택 기법들보다 향상된 성과를 나타낼 수 있음을 확인하였다.

둘째, 감성분석에서는 텍스트를 정량화 하는 단계를 필수적으로 거쳐야 한다. 데이터 처리 과정 동안 수많은 속성벡터가 생성되므로 이를 바탕으로 속성 선택을 사용하여야 한다. 속성 선택 기법을 사용함으로써 컴퓨터의 연산량을 줄여줄 뿐만 아니라, 불필요한 비용을 줄일 수 있다. 본 연구에서는 초기 1,300여 개에 해당되는 속성을 CFS로 60여 개, IG와 GR로 600여 개, 마지막으로 MBFS을 사용하여 80여개로 줄일 수 있었다.

그러나 이러한 학술적 의의에도 불구하고 본 연구는 다음과 같은 한계가 존재한다. 첫째, 본 연구는 Yelp로부터 수집한 리뷰 데이터를 바탕으로 진행되었다. Yelp는 이용후기에 관한 데이터만을 제공하기 때문에 페이스북, 트위터 등 다른 소셜미디어에 비해 다양성을 반영하지 못한다. 또한 영어데이터를 바탕으로 분석을 시행하였으므로 국내 실정과 다를 수 있다. 따라서 후속연구에서는 한국어 감성분석 등의 한글 컨텐츠에 대한 심도 깊은 연구가 필요하다. 둘째, 본 연구에서는 베이지안 네트워크를 적용한 감성분석 모델의 효과성 검증을 속성선택 기법 간의 비교를 통해서만 하였다. 인과관계를 확인하는 것에 중점을 두고 베이지안 네트워크 기반의 분류기법만을 사용하였기 때문이다. 따라서 향후 연구에서는 베이지안 네트워크 모델과 표준통계 방식을 따르는 알고리즘 분류 모델 간 비교분석이 필요하다.

참 고 문 헌

[1] Alamoodi, A. H., B. B. Zaidan, A. A. Zaidan, O. S. Albahri, K. I. Mohammed, R. Q. Malik, E. M. Almahdi, M. A. Chyad, Z. Tareq, A. S. Albahri, H. Hameed, and M. Alaa, “Sentiment analysis and its applications in fighting COVID-19 and infectious diseases: A systematic review”, *Expert Systems with Applications*, Vol.167, 2020, p. 114155.

[2] Arlot, S. and A. Celisse, “A survey of cross-validation procedures for model selection”, *Statistics Surveys*, Vol.4, 2010, pp. 40-79.

[3] Asur, S. and B. A. Huberman, “Predicting the future with social media”, *2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology,* Vol. 1, 2010, pp. 492-499, https://doi.org/10.1109/WI-IAT.2010.63.

[4] Chakraborty, K., S. Bhatia, S. Bhattacharyya, J. Platos, R. Bag, and A. E. Hassanien, “Sentiment analysis of COVID-19 tweets by Deep Learning

Classifiers-A study to show how popularity is affecting accuracy in social media, *Applied Soft Computing*, Vol.97, 2020, p. 106754. Available at https://doi.org/10.1016/j.asoc.2020.106754.

[5] Chan, S. F., B. R. Barnes, and K. Fukukawa, "Consumer control, dependency and satisfaction with online service", *Asia Pacific Journal of Marketing and Logistics*, Vol.28, No.4, 2016, pp. 594-615.

[6] Cho, T. J., H. J. Yun, and C. C. Lee, "Twitter and retweet context: User characteristics and message attributes of Twitter for PR and marketing", *Information Systems Review*, Vol.14, No.1, 2012, pp. 21-35.

[7] Conrady, S. and L. Jouffe, *Bayesian Networks and BayesiaLab: A Practical Introduction for Researchers*, Bayesia USA, 2015.

[8] Costello, F. J., C. Kim, C. M. Kang, and K. C. Lee, "Identifying high-risk factors of depression in middle-aged persons with a novel sons and spouses bayesian network model", *Healthcare*, Vol.8, No.4, 2020, p. 562.

[9] Costello, F. J. and K. C. Lee, "Exploring the sentiment analysis of electric vehicles social media data by using feature selection methods", *Journal of Digital Convergence*, Vol.18, No.2, 2020, pp. 249-259, Available at https://doi.org/10.14400/JDC.2020.18.2.249.

[10] Dağ, H., K. E. Sayin, I. Yenidoğan, S. Albayrak, and C. Acar, "Comparison of feature selection algorithms for medical data", *2012 International Symposium on Innovations in Intelligent Systems and Applications*, 2012, pp. 1-5.

[11] Eo, K. S. and K. C. Lee, "Exploring an optimal feature selection method for effective opinion mining tasks," *Journal of the Korea Society of Computer and Information*, Vol.24, No.2, 2019, pp. 171-177, Available at https://doi.org/10.9708/JKSCI.2019.24.02.171.

[12] Erra, U., S. Senatore, F. Minnella, and G. Caggianese, "Approximate TF-IDF based on topic extraction from massive message stream using the GPU", *Information Sciences*, Vol.292, 2015, pp. 143-161, Available at https://doi.org/10.1016/j.ins.2014.08.062.

[13] Feldman, R., "Techniques and applications for sentiment analysis", *Communications of the ACM*, Vol.56, No.4, 2013, pp. 82-89.

[14] Hall, M., *Correlation based feature selection for machine learning* (Doctoral dissertation), University of Waikato, Dept. of Computer Science, 1999.

[15] Jiang, L., Z. Cai, D. Wang, and H. Zhang, "Improving tree augmented naive bayes for class probability estimation", *Knowledge-Based Systems*, Vol.26, 2012, pp. 239-245.

[16] Karegowda, A. G., A. S. Manjunath, and M. A. Jayaram, "Comparative study of attribute selection using gain ratio and correlation based feature selection", *International Journal of Information Technology and Knowledge Management*, Vol.2, No.2, 2010, pp. 271-277.

[17] Kira, K. and L. A. Rendell, "A practical approach to feature selection", *Machine Learning Proceedings*, Morgan Kaufmann, 1992, pp. 249-256.

[18] Kohavi, R. and G. H. John, "Wrappers for feature subset selection", *Artificial Intelligence*, Vol.97, No.1-2, 1997, pp. 273-324.

[19] Koller, D. and M. Sahami, *Toward optimal feature selection*, Stanford InfoLab, 1996.

[20] Lee, T. and T. Hong, "Terms based sentiment classification for online review using support vector machine", *Information Systems Review*, Vol.17, No.1, 2015, pp. 49-64, Available at https://doi.org/10.14329/isr.2015.17.1.049.

[21] Li, L. X. and S. S. Abdul Rahman, "Students' learning style detection using tree augmented na-

ive Bayes", *Royal Society Open Science*, Vol.5, No.7, 2018, p. 172108.

[22] Liu, Y., J. W. Bi, and Z. P. Fan, "Multi-class sentiment classification: The experimental comparisons of feature selection and machine learning algorithms", *Expert Systems with Applications*, Vol.80, 2017, pp. 323-339, Available at https://doi.org/10.1016/j.eswa.2017.03.042.

[23] Luca, M., "Reviews, reputation, and revenue: The case of Yelp.Com, *Harvard Business School NOM Unit Working Paper* 12-016, 2016.

[24] Min, J. Y., "The Amplifying Aspects of SNS Comments: An exploratory study through the sentiment comparison between news site comments and SNS comments", *Information Systems Review*, Vol.22, No.4, 2020, pp. 163-184.

[25] Murtagh, F. and P. Contreras, "Algorithms for hierarchical clustering: An overview", *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, Vol.2, No.1, 2012, pp. 86-97.

[26] Parlar, T., S. A. Özel, and F. Song, "QER: A new feature selection method for sentiment analysis", *Human-centric Computing and Information Sciences*, Vol.8, No.1, 2018, p. 10, Available at https://doi.org/10.1186/s13673-018-0135-8.

[27] Prabhakaran, R., R. Krishnaprasad, M. Nanda, and J. Jayanthi, "System safety analysis for critical system applications using Bayesian networks", *Procedia Computer Science*, Vol.93, 2016, pp. 782-790.

[28] Prabowo, R. and M. Thelwall, "Sentiment analysis: A combined approach", *Journal of Informetrics*, Vol.3, No.2, 2009, pp. 143-157.

[29] Quinlan, J. R., "Induction of decision trees", *Machine Learning*, Vol.1, No.1, 1986, pp. 81-106.

[30] Sihwi, S. W., I. P. Jati, and R. Anggrainingsih, "Twitter sentiment analysis of movie reviews using information gain and Naïve Bayes classifier, *2018 International Seminar on Application for Technology of Information and Communication*, 2018, pp. 190-195, Available at https://doi.org/10.1109/ISEMANTIC.2018.8549757.

[31] Tang, J., S. Alelyani, and H. Liu, "Feature selection for classification: A review", *Data classification: Algorithms and Applications*, Vol.37, 2014, pp. 1-29.

[32] Wang, H., Z. Ling, K. Yu, and X. Wu, "Towards efficient and effective discovery of Markov blankets for feature selection", *Information Sciences*, Vol.509, 2020, pp. 227-242.

[33] Wang, Y., A. Hong, X. Li, and J. Gao, "Marketing innovations during a global crisis: A study of China firms' response to COVID-19", *Journal of Business Research*, Vol.116, 2020, pp. 214-220.

[34] Yadav, A. and D. K. Vishwakarma, "Sentiment analysis using deep learning architectures: A review", *Artificial Intelligence Review*, Vol.53, No.6, 2020, pp. 4335-4385.

[35] Yassine, M. and H. Hajj, "A framework for emotion mining from text in online social networks", *In 2010 IEEE International Conference on Data Mining Workshops* 2010, pp. 1136-1142.

[36] Yoo, S., J. Song, and O. Jeong, "Social media contents based sentiment analysis and prediction system", *Expert Systems with Applications*, Vol. 105, 2018, pp. 102-111, Available at https://doi.org/10.1016/j.eswa.2018.03.055.

[37] Yousefpour, A., R. Ibrahim, and H. N. A. Hamed, "Ordinal-based and frequency-based integration of feature selection methods for sentiment analysis", *Expert Systems with Applications*, Vol.75, 2017, pp. 80-93, Available at https://doi.org/10.1016/j.eswa.2017.01.009.

〈부 록〉

〈부록 A〉 2클래스 평가지표 결과

Accuracy	TAN	NB	S_Spouses	A_markov
Before	79.62	78.25	82.39	86.76
CFS	85.87	85.42	85.61	86.19
IG, GR	81.82	81.02	81.94	86.69
MBFS	**87.07**	**86.95**	**87.03**	**86.82**
Precision	TAN	NB	S_Spouses	A_markov
Before	0.79	0.78	0.82	0.87
CFS	0.86	0.85	0.86	**0.87**
IG, GR	0.81	0.81	0.82	**0.87**
MBFS	**0.87**	**0.87**	**0.87**	**0.87**
F-measure	TAN	NB	S_Spouses	A_markov
Before	0.79	0.78	0.82	0.87
CFS	0.86	0.85	0.86	0.86
IG, GR	0.82	0.81	0.82	**0.87**
MBFS	**0.87**	**0.87**	**0.87**	**0.87**

2클래스 분류성과 분석은 6,590개의 데이터를 통해 이루어졌다. 별점 4와 5에 해당하는 리뷰를 긍정적 리뷰, 1과 2의 리뷰는 부정적 리뷰로 설정하여 시행하였다. 가장 높은 성과가 도출된 값은 강조 처리를 하였다.

〈부록 B〉 3클래스 클러스터링

3C	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6	Cluster7	Cluster8	Cluster9	Cluster10	Cluster11
1	always	ask	bill	bad	care	already	amaze	charge	manage	happy	pretty
2	awesome	given	complete	basic	customer	another	delicious	arrive	terrible	quick	tasty
3	best	happen	cost	clear	knowl-edge	due	favorite	guess			
4	great	min	horrible	didn't	phone	however	perfect	minute			
5	love	never	money	left	poor	lack	yummy				
6	pro-fession	nothing	paid	okay							
7	recom-mend	sign	pay	seem							
8	select	told	rude								
9	thank		worst								

〈부록 C〉 4클래스 클러스터링

4C	Cluster1	Cluster2	Cluster3	Cluster4	Cluster5	Cluster6	Cluster7	Cluster8
1	ask	care	call	delicious	another	bad	amaze	bit
2	hair	disappoint	happen	favorite	bill	horrible	awesome	enjoy
3	maybe	don’t	instead	love	charge	manage	best	friend
4	ok	either	left	lunch	fine	rude	great	good
5	pay	money	min	perfect	last	terrible	profession	
6	said	never	minute	spot	pay	worst	recommend	
7	still	seem	see	yummy	sign		thank	
8	wasn’t	understand	star					
9			told					

Information Systems Review
Volume 24 Number 1
February 2022

Investigating the Performance of Bayesian-based Feature Selection and Classification Approach to Social Media Sentiment Analysis

Chang Min Kang* · Kyun Sun Eo** · Kun Chang Lee***

Abstract

Social media-based communication has become crucial part of our personal and official lives. Therefore, it is no surprise that social media sentiment analysis has emerged an important way of detecting potential customers' sentiment trends for all kinds of companies. However, social media sentiment analysis suffers from huge number of sentiment features obtained in the process of conducting the sentiment analysis. In this sense, this study proposes a novel method by using Bayesian Network. In this model MBFS (Markov Blanket-based Feature Selection) is used to reduce the number of sentiment features. To show the validity of our proposed model, we utilized online review data from Yelp, a famous social media about restaurant, bars, beauty salons evaluation and recommendation. We used a number of benchmarking feature selection methods like correlation-based feature selection, information gain, and gain ratio. A number of machine learning classifiers were also used for our validation tasks, like TAN, NBN, Sons & Spouses BN (Bayesian Network), Augmented Markov Blanket. Furthermore, we conducted Bayesian Network-based what-if analysis to see how the knowledge map between target node and related explanatory nodes could yield meaningful glimpse into what is going on in sentiments underlying the target dataset.

Keywords: Social Media Sentiment Analysis, Feature Selection, Markov Blanket, Machine Learning, Bayesian Network, What-If analysis

* Master Student, SKK Business School, Sungkyunkwan University
** Ph.D Student, SKK Business School, Sungkyunkwan University
*** Corresponding Author, Professor, Global Business Administration/Department of Health Sciences & Technology / SAIHST(Samsung Advanced Institute for Health Sciences & Technology), Sungkyunkwan University

◐ 저 자 소 개 ◑

강 창 민 (77aktrp3@gmail.com)

성균관대학교 경영학과(학사)를 졸업하고, 현재 성균관대학교 경영대학에서 경영정보전공으로 석사과정 재학 중이다. 주요 관심분야는 딥러닝, 빅 데이터 분석, 감성분석, 뉴로 사이언스 등이다.

어 균 선 (eokyunsun@gmail.com)

강릉원주대학교 산업정보경영공학과(공학사)를 졸업하고, 성균관대학교 경영대학에서 경영정보전공으로 석사학위를 취득하였다. 현재 성균관대학교 경영학과 박사과정에 재학 중이며, 주요 관심분야는 딥러닝, 앙상블 학습, 감성분석 등이다.

이 건 창 (kunchanglee@gmail.com)

KAIST 경영과학과에서 석사 및 박사학위를 취득하였고, 현재 성균관대학교 경영대학 글로벌 경영학과/삼성융합의과학원(SAIHST) 융합의과학과 교수로 재직하고 있다. Journal of MIS, Decision Support Systems, Computers in Human Behavior, IEEE Transactions on Engineering Management, Frontiers in Psychology 등 다수의 국외 저널에 논문을 게재하였으며, 주요 연구분야는 인공지능, 빅 데이터분석, 헬스인포매틱스, 감성분석 등이다

논문접수일 : 2021년 02월 16일 **게재확정일 : 2021년 11월 25일**
1차 수정일 : 2021년 10월 30일