Journal of Digital Convergence
Vol. 17. No. 2, pp. 163-170, 2019

ISSN 1738-1916
https://doi.org/10.14400/JDC.2019.17.2.163

속성선택방법과 워드임베딩 및 BOW (Bag-of-Words)를 결합한 오피니언 마이닝 성과에 관한 연구

어균선[1], 이건창[2*]
[1]성균관대학교 경영대학 박사과정
[2]성균관대학교 글로벌경영학과/삼성융합의과학원 융합의과학과 교수

Investigating Opinion Mining Performance by Combining Feature Selection Methods with Word Embedding and BOW (Bag-of-Words)

Kyun Sun Eo[1], Kun Chang Lee[2*]
[1]Doctoral Student, SKK Business School, Sungkyunkwan University
[2]Professor, Global Business Administration/Dept of Health Sciences & & Technology, SHAIHST Sungkyunkwan University

요 약 과거 10년은 웹의 발달로 인한 데이터가 폭발적으로 생성되었다. 데이터마이닝에서는 대용량의 데이터에서 무의미한 데이터를 구분하고 가치 있는 데이터를 추출하는 단계가 중요한 부분을 차지한다. 본 연구는 감성분석을 위한 재표현 방법과 속성선택 방법을 적용한 오피니언 마이닝 모델을 제안한다. 본 연구에서 사용한 재표현 방법은 백 오즈 워즈(Bag-of-words)와 Word embedding to vector(Word2vec)이다. 속성선택(Feature selection) 방법은 상관관계 기반 속성선택(Correlation based feature selection), 정보획득 속성선택(Information gain)을 사용했다. 본 연구에서 사용한 분류기는 로지스틱 회귀분석(Logistic regression), 인공신경망(Neural network), 나이브 베이지안 네트워크(naive Bayesian network), 랜덤포레스트(Random forest), 랜덤서브스페이스(Random subspace), 스태킹(Stacking)이다. 실증분석 결과, electronics, kitchen 데이터 셋에서는 백 오즈 워즈의 정보획득 속성선택의 로지스틱 회귀분석과 스태킹이 높은 성능을 나타냄을 확인했다. laptop, restaurant 데이터 셋은 Word2vec의 정보획득 속성선택을 적용한 랜덤포레스트가 가장 높은 성능을 나타내는 조합이라는 것을 확인했다. 다음과 같은 결과는 오피니언 마이닝 모델 구축에 있어서 모델의 성능을 향상시킬 수 있음을 나타낸다.

주제어 : 워드 임베딩, 오피니언 마이닝, 감성분석, 속성선택, 머신러닝

Abstract Over the past decade, the development of the Web explosively increased the data. Feature selection step is an important step in extracting valuable data from a large amount of data. This study proposes a novel opinion mining model based on combining feature selection (FS) methods with Word embedding to vector (Word2vec) and BOW (Bag-of-words). FS methods adopted for this study are CFS (Correlation based FS) and IG (Information Gain). To select an optimal FS method, a number of classifiers ranging from LR (logistic regression), NN (neural network), NBN (naive Bayesian network) to RF (random forest), RS (random subspace), ST (stacking). Empirical results with electronics and kitchen datasets showed that LR and ST classifiers combined with IG applied to BOW features yield best performance in opinion mining. Results with laptop and restaurant datasets revealed that the RF classifier using IG applied to Word2vec features represents best performance in opinion mining.

Key Words : Word embedding, Opinion mining, Sentiment analysis, Feature selection, Machine learning

*본 논문은 2017년도 정부(미래창조과학부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(No. 2017R1A2B4010956).
(This study was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIP; Ministry of Science, ICT & Future Planning) (No. 2017R1A2B4010956)).
*Corresponding Author : Kun Chang Lee(kunchanglee@gmail.com)
Received November 13, 2018 Revised January 14, 2019
Accepted February 20, 2019 Published February 28, 2019

1. 서론

오피니언 마이닝(Opinion mining)은 텍스트가 지니는 긍정적 또는 부정적인 의견을 분석하는 감성분석(Sentiment analysis) 분야이다[1-3]. 인터넷 뉴스, 블로그, 소셜미디어의 발전과 더불어 사용자가 작성한 콘텐츠는 폭발적으로 증가하게 되었다. 그 중 텍스트는 콘텐츠에서 상당 부분 차지하고 있어서 오피니언 마이닝은 중요한 연구분야로 대두되었다[4].

감성 분석은 상품 및 서비스에 대한 리뷰를 긍정적 또는 부정적인지 자동으로 분류하는 것으로 머신러닝을 이용해 높은 성과를 나타내고 있다. 기업의 입장에서 상품 또는 서비스에 대한 반응을 파악하는 것은 매우 중요한 요소이다. 상품에 대해 긍정적인 반응이면 장점을 극대화 할 수 있고 부정적인 반응이면 개선안을 마련할 수 있다. 전통적 감성분석은 백 오브 워즈 방법을 이용해 N-gram의 조합을 만들어 감성분석을 수행한다. 하지만 이 방법은 문장의 의미론적(Semantic) 특성을 반영하지 못한다는 단점이 있다. 감성분석에는 백 오브 워즈 방법의 단점을 보완할 수 있는 Word2vec 방법이 있다. Word2vec는 벡터공간에 단어를 표현할 수 있기 때문에 단어 간 유사성 또는 관계를 파악 가능해진다[5]. 본 연구에서는 Wor2vec방법을 이용해 의미론적 특성을 반영할 수 있는 오피니언 마이닝 방법을 제안한다. 그리고 속성선택(Feature selection)방법을 적용해 종속변수에 영향을 주는 속성만 선택하여 효율적인 감성분석을 시도한다[6,7].

본 연구에서는 단일 또는 앙상블 분류기 모형을 벤치마킹 하여 재표현, 속성선택, 분류 및 검증을 통해 속성선택방법과 재표현 방법 간의 최적의 조합을 제안 한다. 단일분류기는 로지스틱 회귀분석, 인공신경망, 나이브 베이지안 네트워크이고, 앙상블 분류기는 랜덤포레스트, 랜덤서브스페이스, 스태킹이다.

본 연구에서 제시하는 연구질문(Research question, RQ)는 다음과 같다.

RQ1 : 오피니언 마이닝 모델을 구축할 경우, 재표현 방법 중 어느 방법의 성능이 높은가?

RQ2 : 오피니언 마이닝 모델을 구축할 경우, FS방법을 적용한 경우 분류기의 성능은 향상되는가?

RQ3 : 오피니언마이닝 모델을 구축할 경우, 어느 재표현 방법과 FS방법의 분류기 조합 성능이 가장 높은가?

본 연구의 구성은 다음과 같다. 2장에서는 머신러닝을 이용한 오피니언 마이닝에 대한 선행연구와 Word2vec에 대해서 설명한다. 3장에서는 머신러닝 분류기에 대해 설명한다. 4장에서는 연구절차와 사용한 데이터에 대해 설명한다. 5장에서는 연구결과에 대해 논의 한다. 마지막으로 6장에서는 결론 및 본 연구의 한계점, 향후 연구에 토의한다.

2. 관련연구

2.1 오피니언 마이닝(Opinion mining)

오피니언 마이닝은 텍스트에 내포된 작성자의 의견이 긍정, 부정, 혹은 중립적인지를 분석하는 것이다[3].

소셜미디어의 발달은 오피니언 마이닝과 같은 감성분석 분야를 급속도로 발전시키는 계기가 되었다. 소셜미

Table 1. Opinion mining studies

author	Data	study method	BOW	W2V	FS
Ghiassi et al. 2013	Twitter	DAN2, SVM	O	X	O
Da Silva et al. 2014	Twitter	Comparing of Bow and FH	O	X	X
Wang et al. 2014	Amazon review	ensemble method with Bagging, Boosting, Random Subspace	O	X	X
Yoo et al. 2018	twitter	prediction system for users' sentimental trajectories for events analyzed in real time	O	X	X
Garcia-Pablos et al. 2018	SemEval2016	W2VLDA(Word2vec with Latent dirichlet allocation)	X	O	X
This study	Amazon review, SemEval2014	Comparison performance of representation method with single and ensemble classifiers.	O	O	O

디어중 하나인 트위터는 지인이나 친구에게 한 문장 또는 짧은 문장으로 감정이나 상황을 전달하는 도구로 자리 잡았다[8,9]. 트위터는 전 세계 5억명 이상이 사용하는 마이크로블로그 서비스로서 오피니언 마이닝에서 중요하게 다루는 데이터 소스이다. 오피니언 마이닝에 관한 선행 연구는 다음 Table 1과 같다.

Ghiassi et al. (2013)는 트위터를 기반으로 의견을 분류하기 위해서 서포트 벡터머신과 동적 인공신경망모델을 사용하였다. 단어의 빈도에 따른 속성 선택 방법을 사용함으로써 불필요한 속성을 제거하였다[10].

Da Silva et al. (2014)는 백 오브 워즈 방법과 피쳐해싱(Feature hashing)방법을 비교하였다[11]. 전반적으로 피쳐해싱은 좋은 재표현 방법으로 나타났지만 정확도 측면에서는 백 오브 워즈가 피쳐해싱보다 좋은 성능을 내는 것을 확인했다.

Wang et al. (2014)는 앙상블 분류기를 오피니언 마이닝에 적용하였다. 배깅 (bagging), 부스팅(Boosting), 랜덤서브스페이스를 사용한 앙상블 분류기는 단일분류기보다 오피니언 마이닝 모델의 성능을 향상시켰다[12].

Yoo et al. (2018)는 실시간으로 축적되는 소셜 미디어 콘텐츠를 효율적으로 관리할 수 있는 감성 예측 시스템을 제안하였다[13]. 감성예측 시스템은 사용자의 감성 경로를 분석하고 예측하며, 특정 이벤트를 실시간으로 발견하는 시스템이다.

Garcia-Pablos et al. (2018)는 Wor2vec과 토픽모델링 방법인 LDA(Latent dirichlet allocation)를 결합한 방법인 W2VLDA를 제시했다. W2VLDA는 사전 학습이 필요 없으며, 특정 도메인이나 언어적 자원의 필요 없이 감정 분류를 수행하는 시스템이다[14].

본 연구에서는 단일 및 앙상블 분류기 모형을 사용 하여 재표현, 속성선택 방법의 성능 검증을 통해 속성선택 방법과 재표현 방법 간의 최적의 조합을 제안 한다.

2.2 워드임베딩(Word embedding)

Word2vec(Word embedding to vector)은 Tomas Mikolov와 동료들에 의해 연구된 딥러닝 방법을 적용하여 만든 단어 임베딩 모델이다[5]. 머신러닝 분류기 모델을 적용하기 위해서는 단어를 벡터로 나타내는 과정이 필요하다. 단어를 벡터로 표현하는 방법 중에는 텍스트 내에 있는 단어를 포착해 해당 단어가 존재하면 1로 표현하고 존재하지 않으면 0으로 표현하는 백 오브 워즈 (bag-of-words, 즉 BOW) 방법이 있다. 나아가 단어의 빈도수를 측정하는 TF (Term frequency)와, 문서 내 단어 중요도를 나타내는 TF-IDF (Term frequency - Inverse document frequency) 를 측정해 단어벡터를 구성한다. 하지만 이 백 오브 워즈 방법은 단어와 단어 간의 관계를 파악할 수 없고, 문서 전체의 단어를 이용해 속성을 생성하기 때문에 벡터의 크기가 증가한다. 이와 같은 과정은 벡터의 희박성(sparsity) 문제로 견고한 머신러닝 모델을 만드는 것은 어렵다. 백 오브 워즈 방법과 달리 Word2vec을 이용해 단어를 학습할 경우, 문맥상 비슷한 의미를 가진 단어들은 서로 가까운 공간벡터를 가진다. Word2vec은 워드임베딩을 표현 하는 2가지 방법을 제공한다. 모델은 Fig. 1 과 같다. 첫 번째는 Continuous bag-of-words (CBOW) 모델이고, 두 번째는 skip-gram모델이다. CBOW모델은 전체 텍스트에서 단어의 주변 단어들을 이용하여 단어를 예측하는 모델 구조이고, skip-gram 모델은 주어진 단어를 통해 주변 단어를 예측한다.

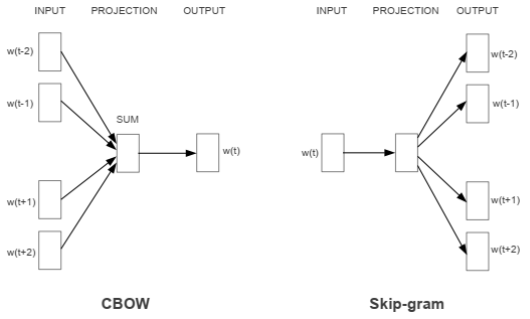


Fig. 1. Word2vec

3. 분류기(Classifiers)

3.1 단일분류기

본 연구의 목적을 위해 사용한 분류기 모델은 다음과 같다. 단일분류기는 로지스틱 회귀분석, 인공신경망, 나이브 베이지안 네트워크이고, 앙상블 분류기는 랜덤포레스트, 랜덤서브스페이스, 스태킹이다.

로지스틱 회귀분석은 선형 및 비선형 분류의 용도로 사용하는 회귀분석 분류기이다. 해당하는 목적변수에 속한 학습용 인스턴스의 출력을 1로 설정하고 소속되지 않은 인스턴스의 출력은 0으로 설정해 목적변수에 대한

회귀를 수행한다. 이 과정의 결과로 선형 함수가 도출된다. 그 후에는, 학습되지 않은 목적변수에 대해 테스트를 할 때, 선형 함수의 결과를 계산하여 가장 큰 값을 채택한다[15].

인공신경망은 인간 뇌의 뉴런의 구조를 모방한 알고리즘이다. 인공신경망은 다층 퍼셉트론(multilayered perceptron) 으로 입력층 (input layer), 은닉층 (hidden layer), 출력층 (output layer)으로 이루어져 있다. 입력층에서는 각 변수에 대응하는 노드들로 구성되어 있다. 은닉층은 입력층으로부터 전달되는 변수 값들을 비선형방정식으로 처리하여 출력층에 전달한다. 출력층은 클래스에 대응하는 마디를 가진다[16].

나이브 베이지안 네트워크의 목표는 목표변수 정보가 포함되어 있는 학습용 인스턴스를 이용하여 검증용 인스턴스의 목표변수를 정확하게 예측하는 것이다. 나이브 베이지안 네트워크는 두 가지 단순화 가정에 의존한다. 베이지안 네트워크는 나이브의 형태로 변화한다. 특히 예측용 변수는 해당 목표변수 별로 조건적으로 독립적이라고 가정한다. 파악된 변수 또는 잠재된 변수는 분류 과정에 영향을 미친다. 나이브 베이지안 네트워크는 목표변수로부터 예측 가능한 변수 까지 모든 아크가 연결되어 있다[17].

3.2 앙상블분류기

랜덤 포레스트는 회귀 및 분류 작업을 수행하는 앙상블 방법이다. 랜덤포레스트는 학습하는 동안 다수의 의사결정나무를 구성하며, 분류 또는 개별 의사결정나무의의 평균 예측인 클래스를 출력한다[18].

텍스트 분석의 다수 중첩 값으로 인하여 랜덤서브스페이스는 다른 앙상블 학습방법과 비교할 때, 감성 분류에 더 적합한 학습방법으로 알려져 있다. 랜덤의 공간에서 학습 데이터는 배깅 앙상블 학습방법과 같은 알고리즘을 사용하면서 수정한다[19]. 수정은 인스턴스가 아닌 속성 공간에서 진행된다. 랜덤서브스페이스 방법은 주어진 속성중에서 랜덤으로 몇 개의 속성을 취하여 이를 속성 서브스페이스로 하여 학습하는 방법이다. 이러한 방법으로 단일분류기별로 성능을 테스트하면 최적의 속성 서브스페이스와 그에 따른 적정 단일분류기를 확인할 수 있는 장점이 있다.

동일한 학습방법의 모델을 조합한 배깅과는 다르게, 스태킹은 다양한 종류의 학습 알고리즘 방법을 적용하여 다른 알고리즘의 분류를 결합하는 앙상블 학습방법이다[20]. 스태킹은 메타 학습 분류기를 이용해서 어느 분류기가 신뢰도 있는지 성능을 추정한 후에 가장 높은 성능을 나타내는 분류기를 선택해서 조합한다. 분류기의 조합을 통해 분류기 간 의 장점과 단점을 상호보완 할 수 있다.

4. 연구방법 및 결과

4.1 연구절차

본 연구는 백 오브 워즈 방법과 Word2vec방법을 통해 오피니언 마이닝을 위한 속성선택 방법과 머신러닝 분류기의 조합의 성능을 비교한다. 본 연구는 다음 Fig. 2과 같은 단계로 구성된다.

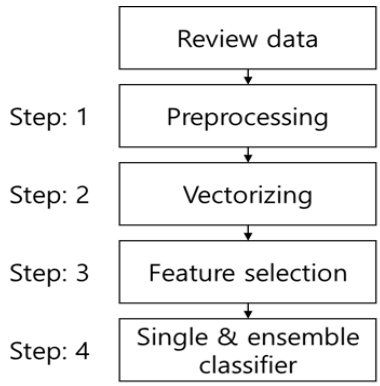


Fig. 2. Procedures

Step 1: 데이터 전처리

데이터의 전처리 단계로 리뷰 데이터에 있는 문장을 단어로 분리하고 분석에 불필요한 텍스트를 제거한다.

Step 2: 재표현 방법을 통한 벡터화

전처리된 텍스트는 재표현 방법을 통해 벡터화 한다. 본 연구에서는 백 오브 워즈와 Word2vec 두가지 벡터화 방법을 사용했다.

Step 3: 속성 선택(FS)

벡터화된 데이터 셋은 두가지 속성선택 방법을 사용해 속성의 수를 줄인다. 본 연구는 상관관계 기반 속성선택과 정보획득 속성선택 방법을 적용한다. 목표 변수에 관련이 없는 속성은 제거하고 관련이 있는 속성만 선택한다.

Step 4: 머신러닝 분류기 학습 & 검증

백 오브 워즈와 Word2vec으로 된 데이터 셋을 통해 분류기모형의 성능을 측정한다.

4.2 데이터

본 연구에 사용한 데이터는 아마존 상품의 리뷰 데이터이다. Blizer et al. (2007)는 아마존 리뷰를 수집해 각각의 도메인 주제에 따라 긍정의미, 부정의미로 레이블링 했다[21]. 본 논문에서는 electronic, kitchen 2가지 데이터 셋을 사용했고 다음으로 SemEval2014 데이터를 사용했다. SemEval2014 데이터는 Laptop, Restaurant 데이터 셋으로 이루어져 있으며 Laptop은 총 1853건, Restaurant는 2969건이다[22]. 두 도메인 모두 긍정적인 리뷰, 부정적인 리뷰로 구성되어 있다. 본 연구에서는 아마존 데이터에서 2개의 도메인을 사용, SemEval2014에서 2개의 도메인을 사용 총 4개의 도메인에 대한 분석을 실시했다.

4.3 연구결과

본 연구에서 사용한 모델의 성능을 검증하기 위해서 10-fold cross validation을 통해 단일분류기 및 앙상블 분류기 모형을 검증했다. 백 오브 워즈 방법과 Word2vec 방법의 속성선택전후의 결과를 비교했다. 본 연구의 결과는 다음 Table 2, Table 3 와 같다.

이하 백 오브 워즈는 BOW, Word2vec은 WE로 표기한다. LR은 로지스틱 회귀분석, NN은 인공신경망, NBN은 나이브 베이지안 네트워크이고, RF는 랜덤포레스트, RS는 랜덤서브스페이스, ST는 스태킹이다. 그리고 CFS는 상관관계 기반 속성선택 방법, IG는 정보획득 속성선택방법이다.

4.3.1 RQ1에 대한 결과

RQ1 : 감성분석을 위한 재표현 방법중 어느 방법의 성능이 높은가?

electronic 데이터의 경우, LR + before에서는 BOW는 79.55, WE는 76.96으로 BOW가 높은 성능을 나타냈다. LR + CFS에서는 BOW는 79.55, WE는 76.96으로 BOW가 높은 성능을 나타냈다. NN + before에서는 BOW는 58.68, WE는 66.48로 WE의 성능이 더 높았다. NN + before 조합을 제외한 분류기와 FS방법의 모든 조합에서 BOW 방법의 성능이 더 높았다. kitchen 데이터의 경우, 분류기와 FS방법의 모든 조합에서 BOW방법이 WE 방법 보다 높음을 확인했다. laptop 데이터의 경우, LR + CFS에서는 BOW는 76.85, WE는 75.77로 BOW 방법이 높았지만, 다른 분류기와 FS방법의 조합에서 WE가 BOW보다 높은 성능을 나타냈다. 전반적으로 restaurant 데이터의 경우에서는 분류기와 FS방법의 모든 조합에서 BOW보다 WE가 높은 성능을 나타냈다.

결과적으로 아마존 상품리뷰 데이터인 electronic과 kitchen에서 WE 방법 보다 BOW방법의 성능이 높음을 확인했다. 그리고 SemEval 데이터인 laptop과 restaurant에서는 BOW보다 WE방법의 성능이 높음을 확인했다.

4.3.2 RQ2에 대한 결과

RQ2 : 감성분석을 위한 FS방법을 적용한 경우 분류기의 성능은 향상되는가?

BOW에서는 CFS를 적용한 NN의 경우 electronics는 58.68에서 77.43으로, kitchen은 68.15에서 75.60로, laptop은 65.08에서 74.16으로, restaurant는 72.82에서 74.76으로 상승했다. BOW에서 IG를 적용한 경우는 LR, NN, RS, ST가 상승했다. IG를 적용한 LR의 경우 electronics는 79.55에서 81.15로, kitchen은 77.35에서 78.65로, laptop은 73.82에서 76.60으로 restaurant는 74.71에서 75.31로 상승했다. IG를 적용한 NN의 경우 electronics는 58.68에서 75.26으로, kitchen은 68.15에서 74.90으로, laptop은 65.08에서 71.54로 restaurant는 72.82에서 74.63으로 상승했다. RS의 경우 electronics는 78.98에서 79.49로, kitchen은 74.30에서 76.45로, laptop은 73.49에서 76.01로 restaurant는 73.96에서 74.38로 상승했다. ST의 경우에서 electronics는 79.08에서 81.35로, kitchen은 77.80에서 78.10으로, laptop은 71.46에서 76.52로 restaurant는 74.00에서 76.15로 상승했다.

WE에서는 CFS를 적용한 NBN의 경우 electronics는 61.62에서 63.33으로, kitchen은 59.50에서 62.05로, laptop은 70.05에서 73.67로, restaurant는 61.23에서 64.90으로 상승했다.

WE에서 IG적용한 NBN의 경우 electronics를 제외하고, kitchen은 59.50에서 59.90으로, laptop은 70.05에서 70.70으로, restaurant는 61.23에서 61.27로 상승했다.

WE에서 CFS를 적용한 RF의 경우 kitchen을 제외하고, electronics는 64.15에서 65.91로, laptop은 81.43에서 81.81로, restaurant는 82.59에서 82.96으로 상승했다. IG를 적용한 RF의 경우 kitchen을 제외하고, electronics는

Table 2. BOW results

BOW	electronics					
before	LR	NN	NBN	RF	RS	ST
ACC	79.55	58.68	75.42	80.79	78.98	79.08
AUC	0.87	0.85	0.81	0.88	0.87	0.87
CFS						
ACC	80.48	77.43	77.32	76.08	78.51	79.08
AUC	0.88	0.86	0.85	0.84	0.87	0.88
IG						
ACC	**81.15**	75.26	76.76	78.15	79.49	**81.35**
AUC	0.89	0.85	0.86	0.85	0.87	0.89
BOW	kitchen					
before	LR	NN	NBN	RF	RS	ST
ACC	77.35	68.15	72.80	77.70	74.30	77.80
AUC	0.86	0.84	0.79	0.86	0.83	0.86
CFS						
ACC	77.35	75.60	76.65	74.20	74.75	77.25
AUC	0.85	0.83	0.84	0.81	0.83	0.85
IG						
ACC	**78.65**	74.90	75.20	76.95	76.45	**78.10**
AUC	0.87	0.85	0.84	0.85	0.85	0.87
BOW	laptop					
before	LR	NN	NBN	RF	RS	ST
ACC	73.82	65.08	66.33	75.00	73.49	71.46
AUC	0.80	0.82	0.78	0.83	0.82	0.80
CFS						
ACC	**76.85**	74.16	66.07	74.58	75.25	**76.68**
AUC	0.84	0.84	0.83	0.82	0.84	0.84
IG						
ACC	76.60	71.54	66.41	74.42	76.01	76.52
AUC	0.84	0.83	0.83	0.82	0.84	0.84
BOW	restaurant					
before	LR	NN	NBN	RF	RS	ST
ACC	74.71	72.82	56.09	74.00	73.96	74.00
AUC	0.75	0.75	0.72	0.75	0.72	0.75
CFS						
ACC	74.97	74.76	54.57	73.87	73.58	75.60
AUC	0.73	0.73	0.72	0.70	0.72	0.73
IG						
ACC	**75.31**	74.63	54.36	74.00	74.38	**76.15**
AUC	0.74	0.74	0.74	0.71	0.73	0.74

Table 3. WE results

WE	electronics					
before	LR	NN	NBN	RF	RS	ST
ACC	76.96	66.48	61.62	64.15	62.65	76.35
AUC	0.84	0.72	0.66	0.71	0.70	0.84
CFS						
ACC	67.77	65.81	63.33	65.91	63.22	69.01
AUC	0.75	0.71	0.70	0.72	0.69	0.75
IG						
ACC	**76.76**	67.46	61.00	64.83	63.07	76.24
AUC	0.84	0.73	0.66	0.71	0.70	0.84
WE	kitchen					
before	LR	NN	NBN	RF	RS	ST
ACC	75.75	66.70	59.50	62.70	63.50	75.60
AUC	0.84	0.73	0.64	0.68	0.67	0.84
CFS						
ACC	70.15	66.70	62.05	61.30	61.70	68.65
AUC	0.76	0.73	0.67	0.68	0.67	0.76
IG						
ACC	**76.50**	68.40	59.90	62.45	63.25	**76.65**
AUC	0.84	0.76	0.65	0.68	0.68	0.84
WE	laptop					
before	LR	NN	NBN	RF	RS	ST
ACC	77.07	80.36	70.05	81.43	81.70	77.01
AUC	0.84	0.87	0.78	0.90	0.89	0.84
CFS						
ACC	75.77	78.20	73.67	**81.81**	80.41	75.82
AUC	0.84	0.84	0.82	0.90	0.89	0.84
IG						
ACC	78.58	80.04	70.70	**81.92**	81.17	77.60
AUC	0.85	0.86	0.78	0.90	0.88	0.85
WE	restaurant					
before	LR	NN	NBN	RF	RS	ST
ACC	79.42	82.69	61.23	82.59	80.77	79.19
AUC	0.83	0.85	0.68	0.87	0.85	0.83
CFS						
ACC	76.25	79.35	64.90	**82.96**	80.30	75.11
AUC	0.77	0.79	0.72	0.87	0.84	0.77
IG						
ACC	79.35	82.59	61.27	**83.03**	80.43	79.25
AUC	0.84	0.84	0.68	0.87	0.85	0.84

64.15에서 64.83으로, laptop은 81.43에서 81.92로, restaurant는 82.59에서 83.03으로 상승했다. IG를 적용한 ST의 경우 electronics를 제외하고, kitchen은 75.60에서 76.65로, laptop은 77.01에서 77.60으로, restaurant는 79.19에서 79.25로 상승했다.

4.3.3 RQ3에 대한 결과

RQ3 : 오피니언마이닝의 경우, 어느 재표현 방법과 FS방법의 분류기 조합 성능이 가장 높은가?

BOW에서 데이터 별 가장 높은 성능을 내는 조합은 다음과 같다.

electronics는 BOW에서 IG의 ST에서 81.35로 가장 높고 다음으로는 IG의 LR이 81.15로 두 번째로 높았다. kitchen에서는 BOW에서 IG의 LR에서 78.65로 가장 높았고 두 번째로는 IG의 ST이 78.10로 높았다.

laptop에서는 WE에서의 IG방법의 RF가 81.92로 가장 높았고, restaurant는 WE에서의 IG방법의 RF가 83.03으로 가장 높았다. 결과적으로, electronics, kitchen에서는

BOW의 IG방법 LR과 ST가 높은 성능을 나타냄을 확인했다. laptop과 restaurant에서는 WE의 IG방법의 RF가 가장 높은 성능을 나타내는 조합이라는 것을 확인했다.

5. 토의 및 결론

본 연구에서는 오피니언 마이닝 예측 모델 구축을 위해 백 오브 워즈 방법과 Word2vec 방법의 성능을 비교 분석했다. 그리고 속성선택 방법을 적용해 단일 또는 앙상블 분류기의 성능을 상승시키고자 했다. 단일분류기는 로지스틱 회귀분석, 인공신경망, 나이브 베이지안 네트워크이고, 앙상블 분류기는 랜덤포레스트, 랜덤서브스페이스, 스태킹이다.

백 오브 워즈에서는 상관관계 기반 속성선택을 적용한 인공신경망의 경우 모든 데이터에서 상관관계 기반 속성선택을 적용하기 전보다 성능이 상승함을 확인했다. 정보획득 속성선택을 적용한 경우는 로지스틱 회귀분석, 인공신경망, 랜덤서브스페이스, 스태킹이 모든 데이터에서 성능이 상승했다. 백 오브 워즈에서 데이터 별 가장 높은 성능을 내는 조합은 다음과 같다. electronics 데이터 셋은 정보획득 속성선택 방법을 적용한 스태킹에서 81.35로 가장 높고 다음으로는 정보획득 속성선택을 적용한 로지스틱 회귀분석이 81.15로 두 번째로 높았다. kitchen 데이터 셋에서는 정보획득 속성선택을 적용한 로지스틱 회귀분석에서 78.65로 가장 높았고 두 번째로는 정보획득 속성선택을 적용한 스태킹이 78.10로 높았다. laptop 데이터 셋은 상관관계 기반 속성선택을 적용한 로지스틱 회귀분석에서 76.85로 가장 높았고 두 번째로는 상관관계 기반 속성선택을 적용한 스태킹이 76.68로 높았다. restaurant 데이터 셋은 정보획득 속성선택을 적용한 스태킹에서 76.15로 가장 높았고 두 번째로는 정보획득 속성선택을 적용한 스태킹이 75.31로 높았다. laptop 데이터 셋을 제외한 electronics, laptop, restaurant 데이터 셋에서는 정보획득 속성선택을 적용한 로지스틱 회귀분석과 스태킹이 높은 성능을 나타냄을 확인했다. 반면 laptop 데이터 셋에서는 상관관계 기반 속성선택을 적용한 로지스틱 회귀분석과 스태킹이 높은 성능을 나타냄을 확인했다.

Word2vec에서는 상관관계 기반 속성선택을 적용한 나이브 베이지안 네트워크의 경우 모든 데이터에서 성능이 향상됨을 확인했다. Word2vec에서 정보획득 속성선택을 적용한 나이브 베이지안 네트워크의 경우 electronics 데이터 셋을 제외하고, 남은 3개의 데이터에서 성능이 상승함을 확인했다. Word2vec에서 상관관계 기반 속성선택을 적용한 랜덤포레스트의 경우 kitchen 데이터 셋 을 제외하고, 남은 3개의 데이터 셋에서 상승함을 확인했다. 정보획득 속성선택 방법을 적용한 랜덤포레스트의 경우 kitchen 데이터 셋을 제외하고, 남은 데이터에서 성능이 상승함을 확인했다. 정보획득 속성선택을 적용한 스태킹의 경우 electronics 데이터 셋을 제외하고, 남은 3개의 데이터에서 성능이 상승함을 확인했다.

electronics, kitchen 데이터 셋에서는 백 오브 워즈 방법의 정보획득 속성선택방법을 적용한 로지스틱 회귀분석과 스태킹이 높은 성능을 나타내는 것을 확인했다. laptop과 restaurant 데이터 셋에서는 Word2vec의 정보획득 속성선택방법을 적용한 랜덤포레스트가 가장 높은 성능을 나타내는 조합이라는 것을 확인했다. 본 연구의 한계는 다음과 같다. 연구에 사용된 데이터는 electronics, kitchen 데이터 셋과 laptop, restaurants 데이터 셋이기 때문에, 다른 도메인에 적용하는데 문제가 발생할 수 있다. 향후 연구과제는 본 연구에서 수행한 데이터와 다른 영역의 도메인에 대해서도 감성분류를 위한 방법을 연구할 필요가 있다.

REFERENCES

[1] M. Kang, J. Ahn & K. Lee. (2018). Opinion mining using ensemble text hidden Markov models for text classification. *Expert Systems with Applications, 94,* 218-227.

[2] J. R. Piñeiro-Chousa, M. Á. López-Cabarcos & A. M. Pérez-Pico. (2016). Examining the influence of stock market variables on microblogging sentiment. *Journal of Business Research, 69(6),* 2087-2092.

[3] A. Yadollahi, A. G. Shahraki & O. R. Zaiane. (2017). Current state of text sentiment analysis from opinion to emotion mining. *ACM Computing Surveys (CSUR), 50(2),* 25.

[4] M. Y. Chen & T. H. Chen. (2017). Modeling public mood and emotion: Blog and news sentiment and socio-economic phenomena. *Future Generation Computer Systems.*

[5] T. Mikolov, K. Chen, G. Corrado & J. Dean. (2013). Efficient estimation of word representations in vector

space. arXiv preprint arXiv:1301.3781.

[6] L. P. Ni, Z. W. Ni & Y. Z. Gao. (2011). Stock trend prediction based on fractal feature selection and support vector machine. *Expert Systems with Applications, 38(5),* 5569-5576.

[7] Y. Liu, J. W. Bi & Z. P. Fan. (2017). Multi-class sentiment classification: The experimental comparisons of feature selection and machine learning algorithms. *Expert Systems with Applications, 80,* 323-339.

[8] F. Corea. (2016). Can Twitter Proxy the Investors' Sentiment? The Case for the Technology Sector. *Big Data Research, 4,* 70-74.

[9] Y. Ruan, A. Durresi & L. Alfantoukh. (2018). Using Twitter trust network for stock market analysis. *Knowledge-Based Systems, 145,* 207-218.

[10] M. Ghiassi, J. Skinner & D. Zimbra. (2013). Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network. *Expert Systems with applications, 40(16),* 6266-6282.

[11] N. F. Da Silva, E. R. Hruschka & E. R. Hruschka Jr. (2014). Tweet sentiment analysis with classifier ensembles. *Decision Support Systems, 66,* 170-179.

[12] G. Wang, J. Sun, J. Ma, K. Xu & J. Gu. (2014). Sentiment classification: The contribution of ensemble learning. *Decision support systems, 57,* 77-93.

[13] S. Yoo, J. Song & O. Jeong. (2018). Social media contents based sentiment analysis and prediction system. *Expert Systems with Applications, 105,* 102-111.

[14] A. García-Pablos, M. Cuadros & G. Rigau. (2018). W2vlda: almost unsupervised system for aspect based sentiment analysis. *Expert Systems with Applications, 91,* 127-137.

[15] S. Menard. (2002). *Applied logistic regression analysis,* 106, Sage.

[16] R. J. Schalkoff. *Artificial neural networks*, 1, New York: McGraw-Hill.

[17] N. Friedman, D. Geiger & M. Goldszmidt. (1997). Bayesian network classifiers. *Machine learning, 29(2-3),* 131-163.

[18] L. Breiman. (2001). Random forests. *Machine learning, 45(1),* 5-32.

[19] T. K. Ho. (1998). The Random Subspace Method for Constructing Decision Forests, *IEEE Trans. Pattern Analysis and Machine Intelligence, 20(8),* 832-844.

[20] D. H. Wolpert. (1992). Stacked generalization. *Neural networks, 5(2),* 241-259.

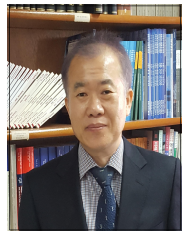
[21] J. Blitzer, M. Dredze & F. Pereira. (2007). Biographies, bollywood, boom-boxes and blenders: Domain adaptation for sentiment classification. *In Proceedings of the 45th annual meeting of the association of computational linguistics*, (pp. 440-447).

[22] S. Poria, E. Cambria & A. Gelbukh. (2016). Aspect extraction for opinion mining with a deep convolutional neural network. *Knowledge-Based Systems, 108,* 42-49.

어 균 선(Eo, Kyun Sun) [정회원]

- 2016년 2월 : 강릉원주대학교 산업경영공학과 (공학사)
- 2018년 2월 : 성균관대학교 경영학과 (경영학 석사)
- 2018년 2월 ~ 현재 : 성균관대학교 경영학과 박사과정
- 관심분야 : 데이터 마이닝, 감성분석, 인공지능
- E-Mail : eokyunsun@gmail.com

이 건 창(Lee, Kun Chang) [정회원]



- 1984년 2월 : 카이스트 경영과학과 (공학석사-의사결정지원)
- 1988년 8월 : 카이스트 경영과학과 (공학박사-인공지능)
- 성균관대학교 경영대학 및 삼성융합의과학원 (SAIHST) 융합의과학과 교수
- 관심분야 : 창의성과학, 인공지능, 헬스 인포매틱스, 감성분석 등
- E-Mail : kunchanglee@gmail.com