Journal of The Korea Society of Computer and Information
Vol. 24 No. 2, pp. 171-177, February 2019
www.ksci.re.kr
https://doi.org/10.9708/jksci.2019.24.02.171

Exploring an Optimal Feature Selection Method for Effective Opinion Mining Tasks

Kyun Sun Eo*, Kun Chang Lee**

Abstract

This paper aims to find the most effective feature selection method for the sake of opinion mining tasks. Basically, opinion mining tasks belong to sentiment analysis, which is to categorize opinions of the online texts into positive and negative from a text mining point of view. By using the five product groups dataset such as apparel, books, DVDs, electronics, and kitchen, TF-IDF and Bag-of-Words(BOW) fare calculated to form the product review feature sets. Next, we applied the feature selection methods to see which method reveals most robust results. The results show that the stacking classifier based on those features out of applying Information Gain feature selection method yields best result.

▸Keyword: Opinion mining, Sentiment analysis, Feature selection

I. Introduction

온라인 쇼핑은 인터넷의 발달로 인해 대중화 되었다. 이는 상품의 구매를 오프라인에서 온라인으로 급격히 변화시켰다. 스마트폰 사용이 보편화되면서 소비자는 시간과 공간의 제약을 받지 않고 온라인 소비활동을 한다. 소비자들은 온라인으로 상품을 구입하고 이용후기를 남기며, 서로 제품에 대한 정보를 공유한다[1].

온라인 리뷰는 소비자에게 합리적인 의사결정을 하도록 필요한 정보를 제공한다. 온라인 리뷰는 텍스트, 이미지, 비디오 등과 같은 멀티미디어 형태로 생성된다[2]. 온라인 리뷰를 분석하는 것은 기업이 소비자의 반응을 파악할 수 있는 기회를 찾을 수 있도록 하고, 기업이 소비자에 대한 대응방안을 수립할 수 있도록 한다. 온라인 리뷰분석을 통하여 기업과 소비자는 상품 및 서비스에 대한 정보를 면밀히 파악할 수 있다[3].

상품 및 서비스의 리뷰를 분석함으로서 리뷰의 의견이 긍정적인지 또는 부정적인지와 같은 반응을 파악하는 것을 오피니언 마이닝(Opinion mining)이라 한다[4]. 오피니언 마이닝은 온라인 리뷰 분석을 통하여 해당 제품에 대한 사용자의 의견, 즉 해당 제품에 대해 긍정적인가 부정적인가를 분석할 수 있다. 소비자의 의견을 나타내는 감성단어를 오피니언 마이닝을 이용해 추출하고 선별하는 것은 기업의 입장에서 소비자의 반응을 효율적으로 파악할 수 있도록 한다. 선행연구에 따르면, 텍스트 내의 중요한 단어를 선택하기 위해 머신러닝을 이용한 속성선택(Feature Selection, FS)을 시도했다[5]. FS는 불필요한 단어를 걸러내고 중요한 단어를 선택에 분류성능을 높일 수 있다.

본 연구는 감성분석을 위한 효과적인 오피니언 마이닝 모델을 제안하기 위해서 오피니언 마이닝 모델 구축에 필요한 위한 학습 분류기와 속성선택 방법FS의 조합을 제안한다. 따라서 본 연구에서는 감성분석을 위한 오피니언 마이닝 모델 구축시, 오피니언 마이닝에 적합한 속성 선택은 어떤 것인지 제안하고자 한다.

본 연구에서 사용한 FS는 다음과 같다. Correlation based feature selection(CFS), Information gain(IG), ReliefF을 사용한다. 분류에 사용한 분류기는 다음과 같다. Logistic regression(LR), Decision tree(DT), Neural network(NN),

• First Author: Kyun Sun Eo, Corresponding Author: Kun Chang Lee
*Kyun Sun Eo (eokyunsun@gmail.com), SKKU Business School, Sungkyunkwan University.
**Kun Chang Lee (kunchanglee@gmail.com), Professor at SKKU Business School/SAIHST (Samsung Advanced Institute of Health Sciences & Technology), Sungkyunkwan University.
• Received: 2018. 11. 13, Revised: 2019. 01. 03, Accepted: 2019. 01. 07.
• This work 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).

Table 1. Study of Opinion mining

author	Data	LR	DT	NN	SVM	NBN	RF	BA	ST	RS	Feature selection
Ghiassi et al., 2013	Twitter	x	x	O	O	x	x	x	x	x	filtering by terms of frequency
Da Silva et al., 2014	Twitter	O	x	x	O	O	O	x	x	x	
Wang et al., 2014	Amazon Product review	x	x	x	x	x	x	O	x	O	
Liu et al., 2017	Cornell Movie review	x	O	x	O	O	x	x	x	x	IG, Gain ratio (GR)
This study	Amazon product review	O	O	O	O	O	O	O	O	O	IG, Cfs, ReliefF

Support vector machine(SVM), Naive bayesian network(NBN), Random forest(RF), Bagging, Stacking, Random subspace(RS)

본 논문의 구성은 다음과 같다. 2장에서 오피니언 마이닝 및 FS 관련 연구에 대해 소개한다. 3장에서는 실험 방법 및 결과에 대해 설명한 후, 마지막으로 4장에서 결론 및 토의, 향후 연구에 대해 설명한다.

II. Previous Studies

1. Opinion Mining

오피니언 마이닝은 문장속의 긍정 및 부정으로 된 의견에 대해 분석하는 것이다[4].

Ghiassi et al. (2013) 은 트위터 데이터를 수집하여 오피니언 마이닝을 시도했다[6]. 트위터는 전 세계적으로 5억명 이상이 사용하는 소셜미디어이다. 사람들은 트위터를 통해 지인이나 친구에게 자신의 감정이나 상태를 전달한다. 서포터벡터머신과 Dynamic architecture for artificial neural network (DAN2) 모델을 제안하여 분류를 진행하였으며, 단어의 빈도에 따라 속성을 제거하는 방법을 사용하였다. da silva et al. (2014)는 BOW방법과 Feature hashing(FH) 등 재표현 방법 간의 비교를 했다[7]. 감성 사전인 lexicon을 적용해 감성분석을 진행했다. Wang et al. (2014)은 오피니언 마이닝 모델의 성능을 높이기 위해 앙상블 분류기를 사용하였다[8]. 데이터는 아마존의 다양한 상품별 리뷰 데이터를 사용하였고, 분류기는 Bagging, 부스팅(Boosting), RS등의 앙상블 방법을 사용하였다. Liu et al. (2017)은 IG와 gain ratio(GR)와 같은 FS방법 간의 분류기 성과비교를 하였다[9]. 오피니언 마이닝 관련연구는 다음 Table 1. 과 같다.

2. Feature Selection (FS)

FS는 수십만 개의 속성을 포함하는 데이터 세트의 출현으로 생긴 연구 분야이다[5]. FS는 데이터 생성 및 관리의 프로세스를 보다 잘 모델링하고 변수를 확보하는 비용을 줄이기 위해 수행한다. 머신러닝 알고리즘의 관점에서 속성 선택은 차원수를 줄이면서 알고리즘의 성능을 유지하거나 성능을 향상시키는데 사용할 수 있다.

속성 선택 방법에 대한 분류는 다음과 같다.

(1)필터(Filter) 학습: 데이터의 일반적인 속성에 의존하고 유도 알고리즘이 독립적인 전처리 단계로 학습한다. 이 모델은 효율적은 계산과정과 좋은 산출물을 낼 수 있다고 평가된다.

(2)래퍼(Wrapper) 학습: 알고리즘을 블랙박스로 간주하여 사용하며, 예측 성능을 사용하여 변수 하위 집합의 상대적 유용성을 평가한다. 즉, 속성 선택 알고리즘을 호출하여 각 피처의 서브 세트를 평가하는 서브 루틴으로서 학습 방법을 사용한다. 그러나 분류기와 상호작용은 필터보다 더 나은 성능 결과를 얻는 경향을 보여준다.

(3)임베딩(Embedding): 교육과정에서 FS를 수행하며 대게 지정된 학습 시스템에만 적용된다. 따라서 최적의 하위 집합이 분류 기준 구성에 포함되어 표시된다. 속성의 부분집합과 가설의 결합된 공간에서 검색하는 이 방법은 래퍼보다 빠르게 계산되며, 종속성을 찾을 수 있다.

본 연구는 독립적인 전처리 과정의 일환인 필터 방법의 속성 선택 방법을 사용하여 감성분석을 위한 최적의 속성 선택 방법을 찾고자 한다. 필터 방법은 감소된 속성을 가진 데이터에 최종적으로 사용될 예측 인자로부터 직접적인 피드백 없이 데이터로부터 직접 계산된 성능 평가 메트릭을 기반으로 한다. 본 연구는 필터 속성 선택 방법을 기반으로 IG, CFS, ReliefF 방법을 기반으로 기술한다.

2.1 Information gain (IG)

Information Gain Filter는 평가 속성의 가장 일반적인 단일 변수 방법 중 하나이다. 이 방법은 정보획득에 따라 기능을 평가하고 한 번에 하나의 기능만 고려한다. 엔트로피 측정은 다음과 같이 고려된다[5].

$$H(Y) = -\sum p(y) log_2(p(y))$$

[수식 1]

여기서 p(y)는 확률변수 Y에 대한 한계확률 밀도 함수이다. 학습 데이터 세트 S에서 관측된 Y값이 두 번째 속성 X의 값에 따라 분할될 경우, X에 의해 유도된 파티션에 대해 Y의 엔트로피가 나누어지기 이전의 Y의 엔트로피보다 작다면, X와 Y의 특징 사이에 관계가 있다. 그러면 X를 관찰한 후에 Y는 다음과 같다.

$$H(Y|X) = \sum p(x) \sum p(y|x) log_2(p(y|x))$$

[수식 2]

여기서 p(y|x)는 주어진 x의 조건부 확률이다. 엔트로피가 훈련 집합 S에서 "불순물"의 기준으로 주어지면, Y의 엔트로피가 감소하는 양을 나타내는 X에 의해 제공된 Y에 관한 추가 정보를 반영하는 측정을 정의할 수 있다. 이 측정은 X와 Y 사이의 종속성을 나타내는 지표로서 IG라고 한다.

$$IG = H(Y) - H(Y|X) = H(X) - H(X|Y)$$

[수식 3]

이 방법은 모든 속성의 규칙적인 분류를 제공하고, 얻어진 순서에 따라 속성의 수를 선택하기 위한 임계값이 필요하다.

2.2 Correlation based feature selection (CFS)

CFS는 상관관계 기반 휴리스틱 평가 함수에 따라 속성 서브세트를 랭킹하는 간단한 다 변수 필터 알고리즘이다. 이는 포함된 속성들 간에 부분 집합이 서로 상관관계가 있는지 혹은 없는지에 대한 오류를 평가하는 방법이다. 휴리스틱 평가는 목표변수에 대한 분류의 성능을 저해시키는 관련성이 낮은 변수를 다룬다. 관련성이 낮은 속성은 클래스와 상관관계가 낮기 때문에 배제한다. 속성선택은 독립적으로 수행하므로 변수간 강한 상호작용에 대한 구분은 할 수 없다[10].

2.3 ReliefF

ReliefF는 다중 클래스 문제를 처리할 수 있는 Relief 알고리즘의 확장이며, 불완전하고 노이즈가 많은 데이터를 처리할 수 있는 보다 견고한 기능을 가지고 있다. ReliefF는 데이터에서 인스턴스 R_i를 무작위로 선택한 다음 동일한 클래스의 k개의 가장 가까운 이웃(H_j)을 찾고, 다른 클래스에서 가장 가까운 가장 근접한 인스턴스는 $M_j(C)$를 찾아 작동한다. R_i, H_j 및 $M_j(C)$에 대한 값에 따라 모든 속성 A에 대한 품질 평가 W[A]가 업데이트 된다.

인스턴스 R_i와 H_j는 속성 A의 다른 값을 갖는다면, 이 속성은 동일한 클래스의 인스턴스를 분리하므로, 품질평가 W[A]는 감소할 것이다. 반대로 인스턴스 R_i와 M_j가 클래스에 대한 속성 A의 값이 다른 경우, 속성 A는 다른 클래스 값으로 두 개의 인스턴스를 분리한다. 이는 품질평가 W[A]를 증가시킨다. ReliefF는 멀티 클래스 문제를 고려하기 때문에 모든 인스턴스에 기여도를 평균화한다[11].

3. Machine learning classifiers

3.1 Logistic regression(LR)

LR은 선형 또는 비선형 형태의 분류를 목적으로 사용되는 회귀분석 방법이다. 이 방법은 각 클래스에 속한 트레이닝 인스턴스의 출력을 1로 설정하고 비 소속 인스턴스의 출력을 0으로 설정하여 클래스에 대해 회귀를 수행한다. 이러한 결과로 선형 방정식이 도출된다. 그런 다음, 미지의 클래스 검증을 할 때, 각 선형 방정식의 결과를 계산하고 가장 큰 값을 선택한다[12].

3.2 Naive bayes network(NBN)

NBN의 목표는 클래스 정보가 포함되어 있는 트레이닝 인스턴스를 학습시켜 테스트 인스턴스의 클래스를 정확하게 예측하는 것이다. NBN은 두가지 중요한 단순화 가정에 의존하기 때문에 베이지안 네트워크가 Naive의 형태로 특화한다. 특히 예측 속성은 주어진 클래스 별로 조건적으로 독립적이라고 가정하고, 파악된 속성 또는 잠재된 속성은 예측 프로세스에 영향을 미친다. 따라서 그림으로 묘사된 NBN은 다음 그림과 같이 타겟 속성으로부터 예측 가능한 속성 까지 모든 호가 연결된 형태를 가지고 있다[13].

3.3 Neural network(NN)

NN는 인간 뇌의 뉴런을 모방한 알고리즘이다. 가장 잘 사용되는 알고리즘은 다층 신경망(multilayer perceptron)으로 입력층(input layer), 은닉층(hidden layer), 출력층(output layer)로 구성된다. 입력층에서는 각 변수에 대응하는 마디(node)들로 구성되어 있다. 다음 은닉층은 입력층으로부터 전달되는 변수 값들을 비선형함수로 처리하여 출력층에 전달한다. 출력층은 목표변수에 대응하는 노드를 갖는다[14].

3.4 Random Forest(RF)

RF는 회귀 및 기타 작업을 수행하기 위한 앙상블 학습 방법이다. 이 분류기는 훈련 기간 동안 다수의 의사결정 나무가 구성되며, 분류 모드 또는 개별 트리의 평균 예측인 클래스 출력에 의해 작동된다[15].

3.5 Random subspace(RS)

텍스트 분류의 많은 중첩 값으로 인해 RS는 다양한 앙상블 분류기와 비교할 때, 정서 분류에 더 적합한 분류기로 알려져 있다. 무작위 공간에서 교육 데이터 세트는 Bagging과 같은 알고리즘을 사용하여 수정된다. 그러나 수정은 인스턴스 공간이 아닌 속성 공간에서 진행된다. RS방법은 기본 분류기를 구성하고 집계하기 위해 임의의 부분 공간을 사용하는 경우가 유용하다. 데이터 집합에 다수의 중복 또는 관련이 없는 속성이 있는 경우 원래의 기능 공간과 비교하여 임의의 부분 공간에서 보다 효과적인 기본 분류기를 선택할 수 있다. 이 기본 분류기의 결합된 결정은 전체 속성 세트의 학습 데이터 세트에서 작성된 단일 분류기 보다 뛰어나다[16].

3.6 Decision tree(DT)

DT는 분류 또는 회귀작업에 적용할 수 있는 효율적인 비모수적 방법이다. 그것들은 종속 변수를 예측하기 위해 입력 공간이 지역 영역으로 분리되는 감독학습을 위한 계층적 데이터 구조이다. 결정 트리는 유한한 비공유 노드 집합과 모서리 집합으로 구성된 그래프이다[17].

3.7 Support vector machine(SVM)

SVM은 분류 작업을 해결하기 위한 최적의 분리 초평면 상태를 제공한다. 수리적 분석은 노은 차원의 분리된 초평면에서 선형문제로 입력 공간과 관련된 비선형 문제를 나타내기 때문에 우수하다. SVM은 분류 문제에 대한 예측 정확도가 우수한 측면에서 NN과 유사하지만 상대적으로 여러 가지 이점을 갖는다. 첫 번째, NN과 달리 SVM은 구조적 위험을 최소화하여 과도한 문제를 피할 수 있다. 둘째, NN은 많은 양의 가중치가 필요하기 때문에 학습을 위해 많은 양의 데이터를 필요로 하는 반면, SVM은 학습을 위한 소수의 데이터만을 사용하고, 소량의 학습 데이터는 우수한 예측성능을 보인다[18].

3.8 Bagging

Bagging은 Bootstrap aggregating의 약자로 균일한 확률분포에 따라 반복적으로 샘플링을 한 후, 의사 결정트리 모형을 조합하여 투표(Voting)에 따라서 분류 예측을 한다. 배깅은 주로 회귀분석에서 사용되며 안정성과 정확도를 향상시키고 분산을 줄여준다[19].

3.9 Stacking(ST)

동일한 타입의 모델을 조합하는 Bagging과는 달리, ST는 다양한 학습 알고리즘을 학습하여 다른 여러 알고리즘의 예측을 결합하는 작업이다. 특히, 메타 학습기를 이용해서 어떤 분류기를 신뢰할 수 있는지를 성능을 추정 후 최고의 성능을 내는 분류기를 찾아내서 조합한다. 이러한 조합을 통해 서로의 장점과 약점을 상호보완가능하다[20].

III. The Proposed Scheme

본 연구에서는 웨카(Weka)을 통해 분석을 진행하였다. 웨카는 뉴질랜드 와이카토 대학에서 만든 데이터마이닝 오픈소스 도구이다.

1. Data

실험에 사용한 데이터는 아마존 상품 리뷰 자료이다. Blizer et al. (2007) 이 연구한 자료로 각각의 도메인에 따라 긍정, 부정으로 레이블이 달린 2,000개 리뷰이다[21]. 총 apparel, book, dvd, electronic, kitchen 총 5가지 도메인에 대해 분석을 실시했다.

Table 2. Results of Accuracy

apparel									
	LR	DT	NN	SVM	NBN	RF	BA	ST	RS
Before	79.24	74.39	77.89	79.59	75.04	79.84	78.09	79.29	76.69
CFS	79.14	76.64	79.09	77.89	75.33	75.74	78.64	79.29	77.09
IG	81.64	77.44	81.54	80.94	77.64	77.89	78.54	**81.84**	78.64
ReliefF	78.19	73.84	77.44	77.84	73.79	77.99	76.74	77.99	77.24
book									
	LR	DT	NN	SVM	NBN	RF	BA	ST	RS
Before	73.75	67.20	72.00	72.25	70.70	74.30	71.15	73.05	72.30
CFS	74.10	70.80	73.75	72.60	71.25	71.45	72.90	74.20	71.40
IG	76.55	70.40	76.15	76.25	73.90	73.15	73.85	**76.75**	72.85
ReliefF	72.30	67.00	72.20	71.95	70.50	72.15	70.00	71.80	70.85
dvd									
	LR	DT	NN	SVM	NBN	RF	BA	ST	RS
Before	75.59	68.17	75.19	76.44	71.08	77.19	73.83	75.64	74.49
CFS	74.79	70.28	74.79	74.94	72.43	70.53	71.13	74.79	73.14
IG	78.65	69.58	77.80	78.80	75.09	74.39	75.14	**78.85**	73.78
ReliefF	75.69	67.87	74.74	76.19	70.83	75.64	73.28	75.54	71.78
electronic									
	LR	DT	NN	SVM	NBN	RF	BA	ST	RS
Before	79.55	73.04	77.69	79.70	75.42	80.79	77.43	79.08	78.61
CFS	80.48	76.65	79.86	77.69	77.32	76.08	78.00	80.37	78.46
IG	81.15	77.17	80.94	80.58	76.76	78.15	78.05	**81.46**	79.49
ReliefF	79.13	73.97	78.10	79.70	75.93	78.62	77.28	78.31	75.36
kitchen									
	LR	DT	NN	SVM	NBN	RF	BA	ST	RS
Before	77.35	70.55	76.25	77.00	72.80	77.70	75.30	78.00	73.89
CFS	77.35	75.15	77.60	76.30	76.65	74.20	75.85	77.25	75.90
IG	78.65	73.55	78.50	78.40	75.20	76.95	76.75	**78.65**	76.45
ReliefF	76.70	73.30	76.65	75.95	75.25	75.05	75.00	76.35	75.85

2. Evaluation Criteria

2.1 AUC (Area under the ROC)

The receiver operation characteristic curve (ROC)는 세로축에는 민감도(Sensitivity)와 가로축에는 1-특이도(Specificity)로 그려지는 곡선을 뜻한다[22]. 민감도는 결과가 positive인 경우에서 몇 %가 정말 positive라고 예측했는지를 말하고, 반대로 특이도는 결과가 negative인 경우에서 몇 %가 정말 negative라고 예측하는지를 말한다. Area under the curve(AUC)는 ROC curve의 밑면적을 계산한 값으로 AUC값이 1.00이면 가장 완벽하다고 할 수 있다.

2.2 10 fold-cross validation

본 연구는 교차검증 방법을 이용하여 머신러닝 분류기 모델을 검증했다[23]. 10 fold-cross validation방법은 원 데이터를 10 폴더로 나눈 다음, 9개 폴더는 모델을 학습하는데 사용하고 남은 1개 폴더는 테스트하여 총 10번을 반복하는 검증방법이다.

3. Results

3.1 Results for RQ

FS방법을 적용한 결과 속성의 수는 다음 Table 3. 와 같다. apparel, book, dvd, electronic, kitchen 순서로 CFS는 25개, 24개, 27,개,23개, 22개로 대폭 줄었고, ReliefF의 경우 107개, 189개, 215개, 109개, 70개로 소폭 감소했다. IG방법은 51개, 43개, 48개, 45개, 45개로 속성의 수가 CFS보다는 적지만 ReliefF방법보다는 많이 감소했다.

Table 3. The number of features

	apparel	book	dvd	electronic	kitchen
Before	204	326	348	240	181
CFS	25	24	27	23	22
IG	51	43	48	45	45
ReliefF	107	189	215	109	70

본 연구는 감성분류를 위해서 10 fold-cross validation 방법을 이용해 각 단일 분류기와 앙상블 분류기의 결과 값을 측정했다. 실험결과에 이용한 지표는 accuracy와 AUC를 이용했다. accuracy는 IG방법을 사용 했을 대, ST가 가장 높았다. 이 결과는 모든 도메인에서 동일했다. 각 도메인 별로 가장 높은 값은 apparel 81.84%, book 76.75%, dvd 78.85%, electronic 81.46%, 마지막으로 kitchen은 78.65%이다.

상품들마다 FS전과 FS후를 비교할 때, CFS와 IG가 FS전과 비교해서 accuracy가 상승했다. 전반적으로 IG 방법이 다른 FS 방법보다 분류기의 성능이 향상했다. 특히 electronic 의 DT가 73.04%에서 77.17%로 가장 상승폭이 높았다. 상품실험 결과 accuracy는 다음 Table 3. 와 같다. AUC의 결과 는 다음 Table 4. 와 같다. AUC 역시 IG방법에서 LR, NN, ST이 각 도메인 별로 높다. apparel은 0.89, book 0.84, dvd 0.86, electronic 0.89, kitchen 0.87로 가장 높다. accuracy와 같이 AUC또한 IG방법에서 다른 FS 방법에 비해 높은 결과값을 보였다.

3.2 T-test

FS전과 FS후의 10 fold-cross validation 결과 집단을 T-test 분석 (신뢰수준 0.05)을 이용하여 통계검증을 수행했

Table 4. Results of AUC

apparel									
	LR	DT	NN	SVM	NBN	RF	BA	ST	RS
Before	0.88	0.78	0.87	0.80	0.82	0.88	0.85	0.88	0.85
CFS	0.87	0.82	0.86	0.78	0.84	0.83	0.85	0.87	0.85
IG	**0.89**	0.83	**0.89**	0.81	0.86	0.85	0.86	**0.89**	0.86
ReliefF	0.86	0.77	0.86	0.78	0.81	0.86	0.84	0.86	0.85
book									
	LR	DT	NN	SVM	NBN	RF	BA	ST	RS
Before	0.81	0.70	0.81	0.72	0.75	0.83	0.78	0.81	0.80
CFS	0.82	0.76	0.82	0.73	0.80	0.79	0.79	0.82	0.79
IG	**0.84**	0.75	**0.84**	0.76	0.82	0.81	0.81	**0.84**	0.80
ReliefF	0.80	0.70	0.80	0.72	0.75	0.80	0.77	0.80	0.78
dvd									
	LR	DT	NN	SVM	NBN	RF	BA	ST	RS
Before	0.84	0.71	0.83	0.76	0.76	0.85	0.81	0.84	0.82
CFS	0.84	0.73	0.84	0.75	0.82	0.80	0.80	0.84	0.81
IG	**0.86**	0.74	**0.86**	0.79	0.83	0.83	0.82	**0.86**	0.83
ReliefF	0.84	0.69	0.83	0.76	0.76	0.84	0.79	0.84	0.80
electronic									
	LR	DT	NN	SVM	NBN	RF	BA	ST	RS
Before	0.87	0.78	0.86	0.80	0.81	0.88	0.86	0.87	0.87
CFS	0.88	0.82	0.88	0.77	0.85	0.84	0.86	0.88	0.87
IG	**0.89**	0.83	**0.89**	0.80	0.86	0.85	0.87	**0.89**	0.87
ReliefF	0.87	0.79	0.86	0.79	0.82	0.86	0.85	0.87	0.84
kitchen									
	LR	DT	NN	SVM	NBN	RF	BA	ST	RS
Before	0.86	0.74	0.85	0.77	0.79	0.86	0.83	0.86	0.83
CFS	0.85	0.81	0.85	0.76	0.84	0.81	0.83	0.85	0.83
IG	**0.87**	0.78	**0.87**	0.78	0.84	0.85	0.84	**0.87**	0.85
ReliefF	0.85	0.77	0.85	0.76	0.81	0.84	0.83	0.85	0.83

Table 5. Results of T-test

apparel	LR	DT	NN	SVM	NBN	RF	BA	ST	RS
Cfs	0.940	0.083	0.315	0.200	0.853	0.012*	0.718	0.999	0.823
IG	0.040*	0.025*	0.001*	0.269	0.028*	0.099	0.709	0.029*	0.272
book	LR	DT	NN	SVM	NBN	RF	BA	ST	RS
Cfs	0.687	0.012*	0.924	0.269	0.747	0.004*	0.688	0.911	0.069
IG	0.096	0.059	0.039*	0.040*	0.049*	0.445	0.084	0.029*	0.666
dvd	LR	DT	NN	SVM	NBN	RF	BA	ST	RS
Cfs	0.620	0.024*	0.797	0.254	0.244	0.000*	0.011*	0.558	0.234
IG	0.065	0.326	0.104	0.070	0.007*	0.069	0.192	0.023*	0.542
electronic	LR	DT	NN	SVM	NBN	RF	BA	ST	RS
Cfs	0.360	0.033*	0.062	0.116	0.143	0.001*	0.674	0.270	0.921
IG	0.103	0.009*	0.004*	0.458	0.291	0.026*	0.655	0.038*	0.602
kitchen	LR	DT	NN	SVM	NBN	RF	BA	ST	RS
Cfs	0.880	0.002*	0.339	0.623	0.035*	0.018*	0.534	0.612	0.129
IG	0.411	0.036*	0.124	0.264	0.189	0.638	0.210	0.647	0.059

다. 결과는 다음 Table 5. 과 같다. IG방법 적용 후와 FS전의 분류기 성능을 비교해 봤을 때 apparel의 경우 IG방법이 모든 분류기의 결과가 통계적으로 유의하다. 이는 IG방법을 적용했을 때 모든 분류기가 상승했음을 나타낸다. book의 경우 IG방법에서 SVM과 ST가 0.033과 0.029로 통계적으로 유의한 결과로 확인했다. dvd는 IG방법에서 NBN, RS가 0.007, 0.023로 통계적으로 유의하다. electronic에서 IG방법의 적용 결과로 DT, NN, RS가 0.009, 0.004, 0.038로 통계적으로 유의하다. 마지막으로 kitchen에서는 DT가 0.036으로 통계적으로 유의하다.

IV. Conclusions

본 연구는 감성분석을 위한 오피니언 마이닝 모델을 제안한다. 이를 위해 머신러닝 관점에서 속성 선택방법을 적용해 효율적인 감성분석이 가능함을 밝힌다. 오피니언 마이닝을 이용해 텍스트의 긍정의견, 부정의견을 예측하기 위한 머신러닝 모델을 제시한다.

연구의 목적을 위해 온라인 상에 공개된 apparel, book, dvd, electronic, kitchen 총 5가지 상품군의 온라인 리뷰를 수집하고, 텍스트의 TF-IDF를 계산하고 BOW 형태로 구성하여 상품 리뷰 속성 셋을 구성하였다. 속성 선택을 사용하여 텍스트의 오피니언을 잘 설명할 수 있는 속성을 고르고, 선택된 속성을 분류기 학습을 통하여 분류 성과를 측정하였다.

오피니언 마이닝의 결과 오피니언 마이닝의 분류 정확도는 IG 방법으로 속성을 선택하고, 앙상블 분류기의 한 종류인 ST을 사용하는 것이 5가지 상품에 대해 가장 높은 성능을 보였다.

IG 속성 선택은 오피니언 분류 성능은 ST 외의 다른 분류기에서도 더 높은 성능을 달성하였다. 이것은 오피니언 분류 관점에서 IG이 가장 적합한 속성을 잘 선택하는 것을 의미한다.

본 연구의 성과는 다음과 같다. 첫째, 본 연구는 속성 선택 및 단일 및 앙상블 분류기를 총합적으로 적용하는 연구이다. 둘째, 감성분석을 위한 오피니언 마이닝에는 속성 선택 방법 중 IG과 ST의 조합이 가장 뛰어난 것으로 나타났다.

기업들은 자사의 서비스나 상품을 사용하는 고객이 가지고 있는 상품 및 서비스에 대한 호불호와 감정을 파악하기 위해 많은 시간과 자원을 소비한다. 기업은 속성선택을 이용하여 문장 속에 있는 단어가 의견에 영향을 미치는 단어인지 미치지 않는 단어인지 파악하므로써 의사결정을 효율적으로 할 수 있고, 향후 오피니언 마이닝을 구축할 경우 FS를 통해 선택된 속성을 바탕으로 오피니언 마이닝 모델을 구축할 수 있다. 따라서, 본 연구의 실무적인 의의는 다음과 같다.

첫째, 텍스트에 나타난 상품 및 서비스에 대한 오피니언을 효과적으로 분류해 고객에 대한 대응이 가능하다. 긍정 감성을 나타내는 고객에는 추가적인 상품/서비스에 대한 구매를 권유할 수 있고, 부정 감성을 나타내는 고객에는 이탈을 방지하기 위한 대응을 할 수 있다.

둘째, 오피니언과 이모션을 분류하는 분류기를 사용하여 보다 즉각적인 대응을 할 수 있다. 최근에 음성의 텍스트 변환 기술의 발달로 인해, 콜센터 상담이나 고객의 대면 상황에서 발현되는 음성을 실시간적으로 텍스트 변환이 가능하고, 변환된 텍스트에 오피니언 및 이모션 모형을 적용하여 실시간으로 오피니언과 이모션을 추론할 수 있다. 이를 통해 콜센터 상담사의 상담 스크립트의 제공이나 대면 시 권유 상품을 결정할 수 있다.

본 연구의 한계는 다음과 같다. 본 연구에 사용한 데이터는 미국의 아마존에서 수집한 영문 리뷰 텍스트로 국내의 한글 텍스트에 바로 적용하기 위한 추가적인 연구가 필요하다. 또한, 본 연구에서 다룬 상품은 apparel, book, dvd, electronic, kitchen의 5가지 상품군 이지만 일반화된 결과를 얻기 위해 더욱 다양한 상품군에 대한 연구가 필요하다. 추가적으로 본 연구에서 사용한 FS방법과 분류기를 추가하여 일관성 있는 결과를 도출하여야 한다.

또한 상품을 가격이 높은 그룹과 낮은 그룹을 나눠 그룹 간 비교 및 경험재와 탐색재를 비교할 수 있다. 마지막으로 텍스트의 전처리 단계에서 오피니언과 이모션을 추론하기 위한 키워

드를 선택하기 위한 정지 단어 및 n-Gram의 조합 등에 대한 연구가 필요하다.

REFERENCES

[1] A. Yadollahi, A. G. Shahraki, & O. R. Zaiane, "Current state of text sentiment analysis from opinion to emotion mining". Association for computing machinery computing surveys, Vol. 50, No. 2, Article 25, 2017.

[2] M. V. Mantyla, D. Graziotin, & M. Kuutila, "The evolution of sentiment analysis? A review of research topics, venues, and top cited papers", Computer Science Review, Vol. 27, pp. 16-32, 2018.

[3] C. Catal, & M. Nangir, "A sentiment classification model based on multiple classifiers". Applied Soft Computing, Vol. 50, pp. 135-141, 2017.

[4] M. Kang, J. Ahn, & K. Lee, "Opinion mining using ensemble text hidden Markov models for text classification". Expert Systems with Applications, Vol. 94, pp. 218-227, 2018.

[5] Z. Li, W. Xu, L. Zhang, & R. Y. Lau, "An ontology-based Web mining method for unemployment rate prediction", Decision Support Systems, Vol. 66, pp. 114-122, 2014.

[6] M. Ghiassi, J. Skinner, & D. Zimbra, "Twitter brand sentiment analysis: A hybrid system using n-gram analysis and dynamic artificial neural network.", Expert systems with applications, Vol. 40, No. 16, pp. 6266-6282, 2013.

[7] N. F. Da Silva, E. R. Hruschka, & E. R. Hruschka, "Tweet sentiment analysis with classifier ensembles." Decision support systems, Vol. 66, pp. 170-179, 2014.

[8] G. Wang, J. Sun, J. Ma, K. Xu, & J. Gu, "Sentiment classification: The contribution of ensemble learning.", Decision support systems, Vol. 57, pp. 77-93, 2014.

[9] Y. Liu, J. W. Bi, & Z. P. Fan, "Multi-class sentiment classification: The experimental comparisons of feature selection and machine learning algorithms.", Expert systems with applications, Vol. 80, pp. 323-339, 2017.

[10] M. A. Hall, "Correlation-based feature selection for machine learning", 1999.

[11] M. Robnik-Sikonja, & I. Kononenko, "Theoretical and empirical analysis of ReliefF and RReliefF". Machine learning, Vol. 53, No. (1-2), pp. 23-69, 2003.

[12] S. Menard, "Applied logistic regression analysis, Vol. 106, Sage", 2002.

[13] W. L. Buntine, "Operations for learning with graphical models". Journal of Atificial Intelligence Research, Vol. 2, pp. 159-225, 1994.

[14] M. Ballings, D. Van den Poel, N. Hespeels, & R. Gryp, "Evaluating multiple classifiers for stock price direction prediction". Expert Systems with Applications, Vo. 42, No. 20, pp. 7046-7056, 2015.

[15] L. Breiman, "Random forests. Machine learning, Vol. 45, No. 1, pp. 5-32, 2001.

[16] T.K. Ho, "The Random Subspace Method for Constructing Decision Forests," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 20, no. 8, pp. 832-844, 1998.

[17] S. K. Murthy, "Automatic construction of decision trees from data: A multi-disciplinary survey". Data mining and knowledge discovery, Vol. 2, No. 4, pp. 345-389, 1998.

[18] V. Vapnik, "The nature of statistical learning theory. Springer science & business media", 2013.

[19] L. Breiman, "Bagging predictors". Machine learning, Vol. 24, No. 2, pp. 123-140, 1996.

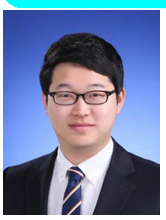
[20] D. H. Wolpert, "Stacked generalization". Neural networks, Vol. 5, No. 2, pp. 241-259, 1992.

[21] J. Blitzer, M. Dredze, & F. Pereira, "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, 2007.

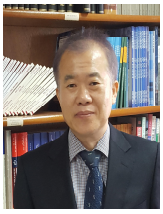
[22] A. P. Bradley, "The use of the area under the ROC curve in the evaluation of machine learning algorithms", Pattern recognition, Vol. 30, No. 7, pp. 1145-1159, 1997.

[23] S. Arlot, & A. Celisse, "A survey of cross-validation procedures for model selection", Statistics surveys, Vol. 4, pp. 40-79, 2010.

Authors



Kyun Sun Eo is a Ph.D. student in SKK Business School at Sungkyunkwan University. He is interested in data mining, machine learning, sentiment analysis, and artificial intelligence.



Kun Chang Lee is a full professor of MIS in SKK Business School at Sungkyunkwan University. He is now in charge of Creativity Science Research Institute (CSRI) and Health Mining Research Center (HMRC) as well, Sungkyunkwan University. His recent research interested in data mining, health informatics, creativity science, Human-Robot Interaction (HRI), and artificial intelligence techniques in decision making analysis.