Journal of Digital Convergence
Vol. 17. No. 3, pp. 107-117, 2019

ISSN 1738-1916
https://doi.org/10.14400/JDC.2019.17.3.107

효과적 이모션마이닝을 위한 속성선택 방법에 관한 연구

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

Exploring Feature Selection Methods for Effective Emotion Mining

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, SAIHST Sungkyunkwan University

요 약 블로그, 소셜 미디어 등의 발달로 인해 점점 더 많은 사람들이 본인의 의견이나 감정을 표현하기 위해 온라인상에서 텍스트 문장을 작성한다. 그리고 이같은 온라인 텍스트 문장속에 숨겨져 있는 긍정 또는 부정등의 감성을 찾아내는 연구분야를 감성분석 이라고 한다. 그중에서도 이모션 마이닝은 사람들의 구체적인 이모션을 찾아내는데 초점을 맞춘 연구분야이다. 본 연구에서는 속성선택 방법과 단일 및 앙상블 분류기를 조합하여 효과적인 이모션 마이닝 예측모델을 제시하고자 한다. 이를 위해 두가지 대표적인 오픈 데이터인 Tweet와 SemEval2007 데이터를 이용하여 TF-IDF를 계산하고 백 오브 워즈(BOW: bag-of-words) 형태로 속성 셋을 구성하였다. 그리고 효과적인 이모션 마이닝이 될 수 있는 최적의 속성을 선택하기 위하여 상관관계 기반 속성선택(CFS), 정보획득 속성선택 (IG), 그리고 ReliefF 등 세가지 속성선택 방법을 적용하였다. 선택된 속성을 이용하여 아홉가지 분류기 모델로 이모션 마이닝의 정확도를 비교하였다. 실험 결과, Tweet 데이터는 의사결정나무(DT)가 CFS, IG, ReliefF에 의한 속성을 이용할 경우 정확도가 상승했고, 랜덤서브스페이스(RS)는 CFS, IG에 선택된 속성을 사용할 경우 정확도가 상승했다. SemEval2007 데이터는 ReliefF에 의해 선택된 속성으로 로지스틱 회귀분석(LR)을 적용하였을 때 정확도가 상승했고, 나이브 베이지안 네트워크(NBN)은 CFS, IG에 의한 속성을 사용할 경우 정확도가 상승하였다.

주제어 : 텍스트 마이닝, 속성선택, 감성분석, 이모션 마이닝, 분류기

Abstract In the era of SNS, many people relies on it to express their emotions about various kinds of products and services. Therefore, for the companies eagerly seeking to investigate how their products and services are perceived in the market, emotion mining tasks using dataset from SNSs become important much more than ever. Basically, emotion mining is a branch of sentiment analysis which is based on BOW (bag-of-words) and TF-IDF. However, there are few studies on the emotion mining which adopt feature selection (FS) methods to look for optimal set of features ensuring better results. In this sense, this study aims to propose FS methods to conduct emotion mining tasks more effectively with better outcomes. This study uses Twitter and SemEval2007 dataset for the sake of emotion mining experiments. We applied three FS methods such as CFS (Correlation based FS), IG (Information Gain), and ReliefF. Emotion mining results were obtained from applying the selected features to nine classifiers. When applying DT (decision tree) to Tweet dataset, accuracy increases with CFS, IG, and ReliefF methods. When applying LR (logistic regression) to SemEval2007 dataset, accuracy increases with ReliefF method.

Key Words : Text mining, Feature selection, Sentiment analysis, Emotion mining, Classifiers

*Corresponding Author : Kun Chang Lee(kunchanglee@gmail.com)
Received November 28, 2018 Revised February 25, 2019
Accepted March 20, 2019 Published March 28, 2019

1. 서론

다양한 소셜 미디어 서비스가 등장하면서 일반 소비자들도 특정 회사의 제품이나 서비스에 대해서 다양한 의견을 표현할 수 있게 되었다. 기업들도 이같은 사회적 현상에 민감하게 대응하고 있다. 소셜 미디어상에서 표출되는 소비자들의 다양한 이모션과 의견은 곧 해당 기업에 대한 여론을 형성하고 그 여론은 기업의 재무성과에 많은 영향을 주기 때문이다. 이같이 온라인상에서 일반 소비자들이 표출하는 다양한 리뷰데이터에서 소비자의 숨어있는 이모션과 의견등을 분석하여 그 내용이 긍정적인지 부정적인지를 알아내고자 하는 분야를 감성분석 (sentiment analysis)라고 한다[1].

오피니언 마이닝(Opinion mining)도 감성분석중 한 분야이다. 이는 문장으로부터 제품 및 서비스에 대한 긍정적인 의견 또는 부정적인 의견을 추출하는데 사용된다[2]. 또한 온라인 리뷰 문장속에서 소비자들이 갖고 있는 즐거움, 놀라움, 분노, 혐오, 두려움, 슬픔과 같은 다양한 감정을 분석하는 것으로 특화한 연구분야를 이모션 마이닝 (Emotion mining)이라고 한다[3].

본 연구에서는 문장 속에 숨겨져 있는 작성자의 감정을 효과적으로 예측하기 위한 이모션 마이닝을 하고자 한다. 이를 위하여 감정예측을 효과적으로 수행하는데 기여하는 속성선택 (FS: Feature Selection) 방법을 실증적으로 찾고자 한다. FS는 불필요한 단어를 제거하고 중요한 의미를 가지는 단어를 선택함으로서 분류성능을 높일 수 있다[4]. FS 방법으로 적정한 속성을 선택한후 최종 이모션 마이닝 결과는 분류기를 사용하여 구한다. 기존의 감성분석 실험에서는 주로 나이브 베이지안 네트워크 (NBN: Naive Bayesian Network)과 서포트벡터머신 (SVM: Support Vector Machine)과 같은 분류기가 많이 사용된 바 있다[5].

본 연구에서는 CFS, IG, ReliefF 등 세가지 FS방법과 아홉가지의 분류기를 이용하여 최적의 이모션마이닝 모델을 제시하고자 한다. 따라서, 본 연구에서 제안하는 연구질문(Research Question: RQ)은 다음과 같다.

RQ1: 효과적인 이모션 마이닝 모델을 구축하고자 할 경우 적합한 FS 방법은 무엇인가?

RQ2: RQ1에서 확인된 FS방법에서 도출되는 속성을 이용할 경우, 최적의 이모션 마이닝 모델을 제시하는 분류기는 무엇인가?

본 논문은 다음과 같이 구성된다. 2장에서 이모션 마이닝 관련 연구, FS와 머신러닝 분류기에 대해 소개한다. 3장에서는 실험방법 및 모형 평가에 대해 설명한다. 4장에서는 연구결과, 마지막으로 5장에서는 결론, 한계점 및 향후 연구에 대해 토의한다.

2. 관련 연구

2.1 이모션마이닝(Emotion mining)

이모션 마이닝은 문장에서 나타나는 감정 즉, 기쁨, 놀라움, 분노, 슬픔 등을 분석하는 것을 말한다[3].

Danisman & Alpkocak (2008)은 문서를 벡터로 표현하고, 각 차원은 유니그램 단어로 구성하는 벡터 스페이스 모델(Vector Space Model, VSM)을 사용했다[5]. VSM에서는 특정 감정의 영역은 다른 영역과 겹치지 않는다는 것을 가정한다. 이와 같은 방법을 사용하여 다양한 감정을 분류할 수 있다. VSM은 다중 감정분석 문제에서 SVM과 NBN 분류기보다 성능이 높았다. Strapparava & Mihalcea (2008)는 SemEval 2007 데이터 셋을 이용하여 비지도 학습방법을 통한 뉴스 헤드라인의 감정 예측을 시도하였다. 어휘 의미론적 분석을 통해 감정을 예측하였다[6]. Gupta et al. (2013)는 부스트 계열의 일종인 Boostexter 알고리즘을 사용하였다[7]. Boostexter의 각 기본 분류기는 각 인스턴스에 대한 예측과 함께 신뢰 값을 할당한다. 테스트 인스턴스의 경우, 최종 분류기는 클래스당 모든 분류기에 대한 신뢰도를 출력한다. 이와 같은 방법을 이용하여 고객 서비스 부서에 보내는 고객의 이메일 데이터 셋을 분석하고, 어떠한 언어적 속성이 사용되는지 알 수 있다. 고객 데이터 세트를 분석한 속성에는 다양한 부정적 상황 및 감정, 부정 의견 또는 고객 관리 영역과 관련된 표현이 포함되어 있다.

Hasan et al. (2014)은 134,000개의 트위터 데이터 셋에서 해쉬 태그를 통하여 감정을 추출하였다[8]. 해쉬 태그를 통하여 감정을 추출한 그룹과 심리학자가 감정을 지정한 그룹을 비교분석하였다. 이와 같은 방법을 통해 심리학자가 지정한 감정은 더 일관성이 있었으며, 확인하였으며, 해쉬 태그와 더 많은 일치성을 보였다. 해당 연구는 EmoTex라는 분류기를 소개하였다. EmoTex는 k-최근접 이웃 알고리즘(k-NN)과 SVM 알고리즘을 적용하였다. Quan & Ren (2016)은 HMM (Hidden Markov

Table 1. Emotion mining studies

Study	Dataset	Emotions	Method	FS	SC	EC
Danisman & Alpkocak 2008	SemEval 2007	anger, disgust, fear, joy, sadness	Vector Space Model	x	o	x
Strapparava & Mihalcea 2008	SemEval 2007	anger, disgust, fear, joy, sadness, surprise	knowledge based, Naive Bayes	x	o	x
Gupta et al. 2013	customer emails	factual, emotional	Boosting	x	o	o
Hasan 2014	tweet	active, inactive, happy, unhappy, support	Vector Machine, k-NN	x	o	x
Quan &Ren 2016	Blog article	expectation, joy, love, surprise, anxiety, sorrow, anger, hate	HMM	x	o	x
Yang & Lin 2018	Movie	very negative, negative, neutral, positive, very positive	SVM	o	o	x
This study	Tweet, SemEval 2007	anger, fear, joy, love, sadness, surprise, thankfulness	Comparison of FS method(CFS, IG, ReliefF)	o	o	o

Model)을 사용하여 문장에서 복합 감정을 효율적으로 식별하는 방법을 탐색했다. 문장 속의 감정은 시간에 따라 다르게 변화되어 문서 범위에 걸친 일련의 스펙트럼 벡터로 인코딩 한다. HMM은 문서에 내포된 감정의 변화를 파악하기 위해 가중치를 적용한다. HMM은 전통적인 BOW방법보다 높은 성능을 나타냈다[9]. Yang, & Lin (2018)은 다중 감정 문제를 해결하려고 했다. 여러 장르의 영화에 따라 리뷰를 분석할 수 있는 새로운 학습 기반 의견 수렴 프레임 워크를 제안하였다. 장르를 고려한 다중 감정 분석은 기존의 감정 분석보다 높은 성능을 나타냈다[2].

Table 1에 이모션 마이닝 관련 주요 선행연구가 요약되어 있다. 특히 맨 마지막 행에 본 연구가 선행연구와 어떠한 점에서 차별화되는지를 표시하였다. 즉, 이모션 마이닝 관련 선행연구에서는 본 연구에서와 같이 CFS, IG, ReliefF등 세가지 FS 방법을 적용하여 최적 속성을 선택하고 이를 단일분류기(SC: Single Classifier)와 앙상블 분류기 (EC: Ensemble Classifier)에 동시에 적용하여 성과를 비교하는 연구는 없다.

2.2 속성선택(Feature selection, FS)

데이터의 폭발적인 증가로 인해 수십만 속성을 가진 데이터를 관리하기 위해서는 FS, 즉 속성선택이 필요하다[4]. FS는 분류기법을 실행하기 전에 불필요한 속성을 제거하고 실제 목표 속성에 영향을 주는 속성들을 선택하는 것이다. FS는 해당데이터의 차원 수를 감소시키면서 분류기의 성능을 유지하거나 대부분의 경우 성능을 향상시키는데 사용할 수 있다. 본 연구에서는 FS의 비교를 통해 이모션 마이닝에 적합한 FS가 무엇인지를 제시하고자 한다.

2.2.1 상관관계기반 속성선택(Correlation based FS, CFS)

CFS는 상관관계 기반의 휴리스틱 평가 함수에 따라 속성을 선택한다. 즉, 타겟변수인 클래스와 해당 속성들 간의 다양한 부분 셋간의 상관관계를 구하고 그 우선순위에 따라 적정 속성을 선택한다. 이는 해당 상관관계가 클수록 그만큼 속성이 타겟변수인 클래스에 대한 기여가 클 것으로 간주하기 때문이다. 반면 관련도가 적은 속성은 클래스와의 상관관계 또한 낮을 것으로 전제하여 배제된다[10]. 하지만 CFS는 속성간의 상호작용에 대해서는 이렇다할 통제 메커니즘이 없다는 한계가 있다.

2.2.2 정보획득(Information gain, IG)

IG는 감성분석에서 많이 사용되는 속성선택 알고리즘 중 하나이다[4]. IG 알고리즘에서는 텍스트의 불확실성이 클수록 텍스트 문장이 갖는 감성값 (긍정 또는 부정), 즉 클래스의 불확실성도 커지게 된다. 따라서, IG 알고리즘은 속성이 갖는 감성정도에 대한 불확실성 (즉, 엔트로피)가 낮아질수록 해당 속성의 중요도가 증가하는 원칙을 적용하여 해당 속성의 IG값을 측정한다. 텍스트 속성 t가 갖는 IG(t)값은 다음과 같이 계산된다.

$$IG(t) = -\sum_{t=1}^{m} P(C_i)\log P(C_i) + P(t)\sum_{t=1}^{m} P(C_i|t)\log P(C_i|t) + P(\bar{t})\sum_{t=1}^{m} P(C_i|\bar{t})\log P(C_i|\bar{t})$$

여기서 $P(C_i)$는 텍스트 속성 c_i가 발생할 확률이다. $P(t)$는 텍스트 속성 t가 발생할 확률이다. $P(\bar{t})$는 텍스트 속성 t가 발생하지 않을 확률을 나타낸다. m은 감성값 분류의 전체 총 수를 나타낸다.

2.2.3 ReliefF

ReliefF는 결측치가 있는 학습자료와 멀티클라스 문제를 모두 다룰 수가 있어서 FS방법에서 널리 사용된다[11]. ReliefF는 오리지널 Relief FS방법을 개선한 여섯가지 방법을 A에서 F까지 알파벳 순서로 나열할 때 마지막에 해당되는 FS방법이다. 이는 여타 FS방법과 비교할때에 상대적으로 계산속도가 빠르고 속성간 인터랙션이 생길 때에도 이를 효과적으로 처리하므로써 의미있는 속성이 선택되도록 한다. ReliefF는 속성점수를 -1에서 +1까지의 계산하는데 +1에 가까울 수록 해당 속성은 우수한 것으로 평가한다.

ReliefF 방법이 의미있는 속성을 선택하는 과정은 다음과 같다. 우선 주어진 학습자료에서 임의로 m개의 인스턴스 셋을 랜덤하게 선택한다. 그리고 해당 인스턴스 셋내에서 한개의 타겟 인스턴스를 선택하고 해당 타겟 인스턴스의 최근접이웃 인스턴스들중에서 타겟 인스턴스와 같은 클라스를 갖는 인스턴스를 최근접 히트(nearest hit), 상이한 클라스를 갖는 인스턴스를 최근접 미스 (nearest miss)라고 한다. 타겟 인스턴스의 속성이 최근접 미스 인스턴스의 속성과 다르면 속성점수를 올리고, 반면 타겟 인스턴스의 속성이 최근접 히트 인스턴스의 속성과 다르면 속성점수를 내린다. 이같은 과정을 통하여 속성점수가 높은 속성을 선택한다.

2.3 머신러닝 분류기(Classifiers)

2.3.1 로지스틱 회귀분석(LR)

로지스틱 회귀분석인 LR(Logistic regression)은 회귀분석 방법 중 하나이다. 일반 회귀분석과는 달리 LR은 종속변수가 범주형 데이터이기 때문에 분류문제에 적합하다. LR은 종속변수가 2개의 클래스를 갖는 이른바 2-클래스 문제에 주로 많이 사용되어 왔다. 그러나, 3개 이상의 클래스를 갖는 문제에도 적용될 수 있는데 이때는 다항로지스틱 회귀(Multinominal logistic regression)라고도 한다[12].

2.3.2 의사결정트리(DT)

의사결정트리인 DT(Decision Tree)는 머신러닝 분류기중에서 실무에서 널리 사용되는 대표적 분류기이다. 특히 대용량의 학습자료로부터 적정한 형태의 DT를 도출하여 속성간 관계를 이해하기 쉬운 트리형태로 보고자 할때 널리 사용된다[13]. DT의 최대장점은 실무자들이 현장에서 쉽게 이해하고 적용할 수 있다는 점이다.

DT는 주어진 학습자료로부터 도출되는 DT의 최종 타겟변수가 이산형 분류클래스로 이뤄진 경우는 분류트리, 반면 최종 타겟변수가 실수형 값으로 이뤄진 경우는 리그레션 트리라고 한다. 일반적으로는 DT라고 할때에는 대개 분류트리를 의미하지만 경우에 따라 리그레션 트리도 많이 사용되기 때문에 분류트리와 리그레션 트리를 동시에 의미하는 트리로서 CART (Classifiction And Regression Tree)라는 일반적 형태의 DT를 사용한다. 이외에도 실무에서 널리 사용되는 DT는 대개 C4.5로 잘 알려져 있는 알고리즘을 이용한다. 한편, 주어진 학습자료로부터 최적의 DT를 도출하기 위하여 사용되는 측정치로는 지니 임퓨리티 (Gini Impurity), 정보획득 (Information Gain), 분산감소 (Variance Reduction) 등이 있다.

2.3.3 인공신경망(NN)

Neural Network은 인간 두뇌가 학습하는 과정을 모방한 머신러닝 알고리즘이다. 흔히 사용되는 알고리즘은 다층 퍼셉트론 (multi-layered perceptron)으로서 입력층(input layer), 은닉층(hidden layer), 출력층(output layer)으로 구성되어 있다.

학습과정은 다음과 같이 진행된다[14]. 입력층은 주어진 학습자료를 구성하는 변수를 나타내는 노드로 구성되어 있으며 해당 입력층으로부터 학습자료가 입력된다. 은닉층은 입력층에서 들어오는 학습자료를 연결가중치로 처리하고 이를 다시 시그모이드 함수와 같은 전이함수로 출력값을 만들어 이를 출력층으로 전달한다. 출력층에서는 주어진 입력자료에 대한 최종적인 출력값을 계산하고 이를 실제값과 비교한 후 그 차이인 오차가 정해진 임계치보다 작은지 여부를 계산한다. 오차가 임계치보다 클 경우 출력층으로부터 은닉층, 그리고 입력층까지 단계적으로 내려가면서 연결가중치를 순차적으로 수정하므로써 주어진 학습자료를 학습한다.

2.3.4 나이브 베이지안 네트워크(NBN)

나이브 베이지안 네트워크인 NBN(Naive Bayesian Network)은 베이즈 정리를 적용한 분류기로서 독립변수 간의 사이는 독립적이라는 것을 가정한다. NBN은 멀티 클래스 문제를 분류하는데 효과적으로 적용되는 분류기이다. NBN 그래프는 속성, 호, 조건부 확률테이블로 구성되어 있고, 부모노드와 자식노드 간의 조건부 확률을 이용한다[15]. 분류에 사용될 인스턴스들은 N개의 독립변수를 타나내는$X=(x_1,...,x_n)$로 표현된다. 베이즈 정리를 이용하여 X 가 클래스 k (즉, C_k)로 분류될 조건부 확률은 다음과 같다.

$$p(C_k|X)=\frac{p(C_k)\,p(X|C_k)}{p(X)}$$

이때 NBN은 클래스 k, 즉 C_k에 대해서 다음 수식을 통해 최대 확률을 갖는 클래스 k를 찾아낸다.

$$\hat{y}=\underset{k\in 1,...,K}{\arg max}\; p(C_k)\prod_{i=1}^{n}p(x_i|C_k)$$

2.3.5 서포트 벡터 머신 (SVM)

서포트 벡터 머신, 즉 SVM(Support Vector Machine)은 분류 예측 문제를 해결하기 위한 최적의 하이퍼 평면(Hyperplane)을 구성하는 분류기이다[16]. 2개 또는 그이상의 범주 중에서 하나의 범주에 속한 데이터의 집합이 있을 경우, SVM은 주어진 데이터를 바탕으로 새로운 데이터가 어느 범주에 속할지 판단하는 비확률적 이진 선형 분류모델을 구성한다. SVM은 타겟변수의 멀리 클래스를 구분하기 위한 최적의 분리 하이퍼 평면을 구하기 위해 해당 하이퍼 평면간의 거리를 최대화 한다. 예를 들어 $f(x)=w\bullet x-b$라는 수식에 의해 점과 $f(x)$의 거리를 $\frac{1}{\|w\|}$로 나타낼 수 있다면, SVM은 $\|w\|$를 최소화하여 하이퍼 평면의 간격이 최대가 되도록 하는 최적의 하이퍼 평면을 구한다[17].

2.3.6 배깅(BA)

배깅, 즉 BA (Bagging)는 Bootstrap aggregating의 약자이다. BA는 오버피팅을 피하기 위해 분산을 줄이도록 설계된 앙상블 분류기이다[18]. 원래 BA는 DT를 적용하여 앙상블을 하는 것이 일반적이지만 다른 여타 분류기를 적용해도 무방하다. BA의 적용 프로세스는 주어진 학습자료로부터 리플레이스먼트를 허락하면서 균일분포 (uniform distribution) 를 따르도록 하는 복수개의 부분집합을 만드는 것으로 부터 시작된다. 이러한 부분집합을 부트스트랩 샘플 (bootstrap sample)이라고 한다. 이같은 복수개의 부트스트래프 샘플에 분류기를 적용하고 해당 분류결과를 평균하거나 (이는 리그레션 문제의 경우) 또는 보팅 (voting) (분류문제의 경우)하여 최종결과를 도출한다.

2.3.7 스태킹(ST)

스태킹 (ST: Stacking)은 Stacked Generalization을 줄여서 사용하는 이름이다. ST는 베이스 분류기 (만약 머신러닝 문제가 분류가 아닌 리그레션 이라면 베이스 리그레서. 리그레서는 Regressor를 의미함)와 메타 분류기 (주어진 문제가 분류가 아닌 리그레션 이라면 메타 리그레서) 두단계를 거쳐서 앙상블을 하는 방법이다 [19]. 즉, 주어진 학습자료를 이용하여 복수개의 베이스 분류기를 적용하고 분류결과를 도출한다 그런 다음 메타 분류기는 베이스 분류기의 분류결과를 입력자료로 하여 학습한 후 최종결과를 도출한다. 이와 같이 스태킹은 서로 다른 분류기를 사용하여 앙상블을 하기 때문에 이산형 앙상블이라고 말할 수 있다. 메타 분류기로 자주 사용되는 분류기는 LR이지만 주어진 문제의 성격에 따라 다양한 분류기를 메타 분류기로 사용할 수 있다.

2.3.8 랜덤포레스트(RF)

대표적인 앙상블 분류기인 랜덤 포레스트, 즉 RF (Random Forest)는 주어진 학습자료로부터 다수의 의사결정트리를 추출하여 분류성과를 향상시키도록 설계된 앙상블 분류기이다. RF는 학습과정 동안 다수의 의사결정트리를 생성한 다음, 또 다른 앙상블 분류기인 배깅 (Bagging)을 이용해 이들 다수의 의사결정트리의 결과를 앙상블 하여 분류 예측 모형을 구축한다. RF는 데이터의 특성, 속성 간의 상호작용, 그리고 비선형 모델 등의 문제를 효과적으로 다룰 수가 있기 때문에 이상치에도 분류성능이 크게 떨어지지 않는 견고성이 있는 앙상블 분류기이다[15].

2.3.9 랜덤서브스페이스(RS)

랜덤 서브스페이스인 RS (Random Subspace)는 다른 앙상블 분류기와 비교할때 감성분석 문제에 상대적으로 더 적합한 분류기로 알려져 있다. 랜덤 서브스페이스는

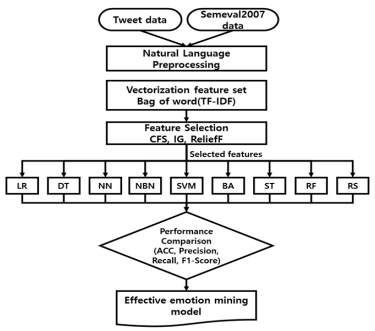


Fig. 1. Procedure of this study

이름 그대로 학습자료간 부분집합이 아닌 속성들간 부분집합을 통하여 배깅 (bagging)을 하고 분류성과를 올리는 앙상블 분류기이다[21]. 따라서, RS는 속성 배깅 이라고도 한다. 즉, 학습자료가 아닌 속성들간의 부분집합을 만들고 이들 속성 부분집합별로 분류기 결과값을 평균하여 최종 결과를 도출한다. 이같이 RS는 속성들간의 부분집합을 만들어서 앙상블을 하는 반면에 앞에서 설명한 BA는 자료들간의 부분집합을 토대로 앙상블을 한다.

RS가 갖는 대표적 장점은 분류성과가 좋은 특정 속성 부분집합이나 또는 상대적으로 분류성과가 안좋은 속성 부분집합에 대해서 민감하게 반응을 하지 않는다는 점이다. RS의 이같은 특징때문에 속성의 갯수가 학습자료의 수보다 더 큰 특이한 구조의 학습자료로부터 적정한 분류기를 찾고자 할 때 유용하다. 대표적으로 바이오 헬스 분야에서 널리 사용되는 기능적 자기공명영상 (fMRI) 자료나 유전자 자료등으로부터 적합한 분류기를 도출할때에 RS는 유용하게 사용된다. 또한 RS는 재무분야에서 시장상황에 맞는 적정한 포트폴리오 선택을 하는 문제에도 성공적으로 적용된다.

3. 연구 방법

본 연구는 다음 Fig. 1과 같은 절차를 가진다.

(a) 자연어 처리 단계

자연어 처리단계는 문장을 단어로 분리하고 단어의 형태소를 구분한다. 문장속의 단어는 명사, 동사, 형용사, 부사 등으로 구성되어 있으며 각단어마다 단어의 형태소를 태깅한다. 문장속에 있는 단어중에서 분석에 영향을 미칠 수 있는 불필요한 단어를 제거한다.

(b) 속성벡터 변환단계

자연어처리를 거친 데이터는 재표현 방법을 통해 속성 셋으로 변환한다. 본 연구에서는 백 오브 워즈를 이용하여 속성셋을 구성하고 TF-IDF를 이용해 가중치를 계산하였다.

(c) 속성 선택(FS) 단계

본 연구는 CFS방법과 IG방법 그리고 ReliefF방법을 사용한다. 다음 FS방법을 이용하여 걸러진 속성을 사용한다. FS방법을 적용해 선택된 속성의 수는 다음 Table 3과 같다.

(d) 머신러닝 분류기 학습 & 검증

본연구에서 사용한 분류기는 다음과 같다. SC는 LR, DT, NN, NBN, SVM이고, EC는 BA, ST, RF, RS이다. 본 단계에서는 FS단계에서 선택된 속성을 사용해 분류기 모형의 성능을 측정한다.

3.1 데이터

본 연구목표를 달성하기 위하여 사용한 자료는 이모션을 타겟변수로 사용하는 Tweet 자료와 SemiEval2007 자료라고 하는 공개된 오픈자료를 이용하였다. Tweet 자료는 250만개의 tweet을 크롤링한 오픈 자료로서, 타겟변수는 분노, 두려움, 기쁨, 사랑, 슬픔, 놀람, 감사 등 일곱 개의 이모션 클래스로 구성되어 있다[22]. 한편, SemEval2007 자료는 뉴스 헤드라인, 구글 뉴스 또는 CNN과 같은 웹사이트 뉴스로부터 추출한 문장으로 구성된 오픈 자료이다 [23]. SemEval2007 자료의 타겟변수는 이모션의 강도 점수에 따라 강한 긍정, 긍정, 부정, 강한부정 총 4개의 이모션 클래스로 구성되어 있다.

3.2 성과 평가 (Performance evaluation)

3.2.1 혼동행렬 (Confusion Matrix)

본 연구에서 사용한 평가지표는 정확도(Accuracy), 정밀도(Precision), 재현율(Recall), F1점수(F1-Score)이다. 평가지표는 다음 Table 2와 같은 혼동행렬을 이용한다.

TP(True Positive)는 Positive를 Positive로 분류한 경우이다. TN(True Negative)는 Negative를 Negative로 분류한 경우이다. FN은 Positive를 Negative로 분류한 경우이고, 마지막으로 FP는 Negative를 Positive로 분류한 경우이다. 정확도는 전체 중에서 Positive와 Negative

를 올바르게 분류한 비율이다. 정밀도는 Positive라고 분류한 경우에서 실제 Positive인 비율이다. 재현율은 Positive 중에서 Positive로 분류한 비율이다.

Table 2. Confusion matrix

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

3.2.2 교차검증(10-fold cross validation)

10-폴드 교차검증 방법은 데이터를 10개의 폴더로 분할 후 9개 영역은 분류기 모델을 학습에 사용되고 나머지 1개 폴더는 학습된 분류기 모델을 검증하여 총 10번을 반복하는 방법이다. 본 연구는 10 겹 교차검증방법을 이용하여 단일 및 앙상블 분류기 성능을 검증했다[24].

Table 3. The number of selected features

	Tweet	SemEval
Before	18	339
CFS	2	23
IG	2	28
ReliefF	15	181

4. 연구 결과

4.1 RQ1에 대한 결과

FS방법을 이용해 속성을 선택한 결과는 다음 Table 3와 같다. Tweet 데이터는 18개에서 CFS 2개, IG 2개 , ReliefF 15개로 줄일 수 있었다. SemEval2007 데이터는 339개에서 CFS 23개, IG 28개, ReliefF 181개로 줄었다.

4.2 RQ2에 대한 결과

4.2.1 Tweet 데이터 결과

본 연구에서는 교차검증 방법을 통해 결과를 도출했다. Tweet 데이터를 분석한 결과는 다음 Table 4 와 같다. CFS 방법과 IG 방법을 사용한 속성을 이용한 분류기 중 NBN과 ST를 제외한 모든 분류기에서 정확도가 상승함을 보였다. 가장 높은 결과로는 IG방법의 SVM과 RS에서 28.35%를 보였으며, CFS와 IG 방법에서 DT가 21.19%에서 27.91%로 가장 높은 상승폭을 보였다. 정밀도의 경우, CFS방법과 IG방법을 적용했던 속성을 사용한 DT, RF에서 상승했다. DT는 0.24에서 0.27로, RF는 0.23에서 0.27로 상승했다. 재현율의 경우 ReliefF방법을 적용한 속성을 사용했던 DT가 0.26에서 0.27로, NN이 0.24에서 0.25로, RF가 0.24에서 0.27로, 마지막으로 RS가

Table 4. Tweet results

Tweet									
Accuracy (%)	LR	DT	NBN	NN	SVM	ST	RF	RS	BA
before	27.55	21.19	27.79	26.13	27.45	27.83	26.12	26.61	25.94
CFS	**27.66**	**27.91**	27.28	**27.68**	**28.35**	27.64	**28.19**	**28.35**	**28.04**
IG	**27.66**	**27.91**	27.28	**27.68**	**28.35**	27.64	**28.19**	**28.35**	**28.04**
ReliefF	27.36	**23.14**	27.91	**26.90**	**27.89**	27.95	**27.22**	27.83	25.20
Precision	**LR**	**DT**	**NBN**	**NN**	**SVM**	**ST**	**RF**	**RS**	**BA**
before	0.25	0.24	0.24	0.23	0.21	0.10	0.23	0.25	0.24
CFS	**0.28**	**0.28**	**0.28**	**0.28**	**0.28**	**0.28**	**0.28**	**0.28**	**0.28**
IG	**0.28**	**0.28**	**0.27**	**0.28**	**0.28**	**0.28**	**0.28**	**0.28**	**0.28**
ReliefF	0.25	**0.25**	0.24	**0.25**	0.11	0.08	0.26	**0.27**	0.23
Recall	**LR**	**DT**	**NBN**	**NN**	**SVM**	**ST**	**RF**	**RS**	**BA**
before	0.12	0.26	0.13	0.24	0.17	0.04	0.24	0.22	0.26
CFS	**0.28**	**0.28**	**0.28**	**0.28**	**0.28**	**0.28**	**0.28**	**0.28**	**0.28**
IG	**0.28**	**0.28**	**0.27**	**0.28**	**0.28**	**0.28**	**0.28**	**0.28**	**0.28**
ReliefF	0.10	**0.27**	0.12	**0.25**	0.05	0.02	**0.27**	0.22	0.24
F1_Score	**LR**	**DT**	**NBN**	**NN**	**SVM**	**ST**	**RF**	**RS**	**BA**
before	0.16	0.25	0.17	0.24	0.19	0.05	0.24	0.23	0.25
CFS	**0.28**	**0.28**	**0.28**	**0.28**	**0.28**	**0.28**	**0.28**	**0.28**	**0.28**
IG	**0.28**	**0.28**	**0.27**	**0.28**	**0.28**	**0.28**	**0.28**	**0.28**	**0.28**
ReliefF	0.14	**0.26**	0.16	**0.25**	0.07	0.04	**0.27**	**0.25**	0.24

0.25에서 0.27로 상승했다. F1점수의 경우 ReliefF방법을 적용한 속성을 사용했던 DT, NN, RF, RS가 각각 상승했다. AUC의 경우에는 CFS방법과 IG방법을 적용한 속성을 사용한 DT, SVM, NBN, RF가 상승했다.

4.2.2 SemEval2007 데이터 결과

SemEval 2007 데이터 분석 결과는 다음 Table 5 와 같다. 정확도의 경우, ReliefF방법을 적용한 속성을 사용한 경우 LR에서 45.57이 가장 높다. SemEval2007 데이터 분석결과, 정확도에서 ReliefF방법을 적용한 속성을 사용한 경우, DT와 NBN을 제외한 모든 분류기들의 상승했다. CFS, IG 방법을 적용한 속성을 사용한 경우의 LR, NN, NBN, ST가 상승했다. 정밀도의 경우 모든 FS방법들을 적용한 속성을 사용한 결과가 전반적으로 상승하지 않았지만 IG방법과 NBN을 적용한 속성을 사용한 경우 0.36에서 0.82로 가장 높이 상승했다. 재현율의 경우 CFS와 IG방법을 적용한 속성을 사용했을 때 대부분 분류기에서 0.95 이상의 수치를 보였다.

4.3 통계검증

본 연구에서는 교차검증의 결과를 FS적용 전과 적용 후의 T-test 분석 (신뢰수준 0.05)을 실시했다. 결과는 다음 Table 6 와 같다. Tweet 데이터에서 DT는 FS방법을 적용한 속성을 사용한 경우 통계적으로 유의하며, 이는 FS전과 FS후의 결과가 차이난다는 것을 나타낸다. ReliefF방법을 적용한 속성을 사용한 결과는 DT를 제외하고 모두 통계적으로 유의하지 않음을 확인했다. CFS와 IG방법을 적용한 속성을 사용한 경우 DT, NN, RF,

Table 5. SemEval 2007 results

SemEval 2007									
Accuracy (%)	LR	DT	NBN	NN	SVM	ST	RF	RS	BA
before	40.54	33.90	32.50	41.95	43.35	40.65	44.98	44.17	41.97
BF	**43.37**	32.39	**41.35**	**43.16**	41.25	**43.07**	42.26	43.36	**43.07**
IG	**43.37**	32.39	**40.55**	**43.37**	41.85	**43.37**	42.46	42.46	**43.37**
ReliefF	**45.57**	**35.01**	32.09	**42.95**	**44.77**	**44.67**	**45.38**	**45.47**	**44.47**

Precision	LR	DT	NBN	NN	SVM	ST	RF	RS	BA
before	0.44	0.32	0.36	0.41	0.40	0.39	0.42	0.39	0.39
Cfs	0.36	0.32	**0.63**	0.36	0.35	0.36	0.36	0.36	0.36
IG	0.36	0.32	**0.82**	0.36	0.35	0.36	0.36	0.36	0.36
ReliefF	0.41	**0.33**	**0.40**	0.40	0.40	**0.41**	0.42	**0.40**	**0.40**

Recall	LR	DT	NBN	NN	SVM	ST	RF	RS	BA
before	0.47	0.93	0.15	0.57	0.58	0.62	0.56	0.69	0.65
Cfs	**0.95**	**1.00**	**0.51**	**0.96**	**0.97**	**0.95**	**0.95**	**0.96**	**0.95**
IG	**0.95**	**1.00**	0.06	**0.96**	**0.97**	**0.95**	**0.95**	**0.96**	**0.95**
ReliefF	**0.66**	**0.94**	0.15	**0.84**	**0.72**	**0.68**	**0.66**	**0.80**	**0.72**

F1_Score	LR	DT	NBN	NN	SVM	ST	RF	RS	BA
before	0.45	0.48	0.21	0.46	0.47	0.48	0.48	0.50	0.48
Cfs	**0.52**	0.48	**0.33**	**0.52**	**0.52**	**0.52**	**0.52**	**0.52**	**0.52**
IG	**0.52**	0.48	0.11	**0.52**	**0.52**	**0.52**	**0.52**	**0.52**	**0.52**
ReliefF	**0.51**	**0.49**	0.21	**0.54**	**0.51**	**0.51**	**0.51**	**0.53**	**0.51**

Table 6. T-test Result of Accuracy

Tweet	LR	DT	NBN	NN	SVM	ST	RF	RS	BA
Cfs	0.885	**0.000***	0.562	0.066	0.158	0.739	0.021	**0.001***	0.030
IG	0.885	**0.000***	0.562	0.066	0.158	0.739	0.021	**0.003***	0.030
ReliefF	0.829	**0.006***	0.905	0.395	0.489	0.799	0.257	0.064	0.480

SemEval	LR	DT	NBN	NN	SVM	ST	RF	RS	BA
Cfs	0.210	0.131	**0.001***	0.440	0.340	0.276	0.157	0.689	0.639
IG	0.224	0.131	**0.002***	0.369	0.502	0.231	0.206	0.423	0.560
ReliefF	**0.042***	0.401	0.827	0.406	0.509	0.091	0.801	0.555	0.291

BA, 그리고 RS가 통계적으로 유의하다. SemEval2007 데이터의 경우, CFS와 IG방법을 적용한 속성을 사용한 경우에는 NBN이 통계적으로 유의함을 확인했고 ReliefF를 적용한 속성을 사용한 경우 LR이 통계적으로 유의함을 확인했다.

4.4 Tweet 데이터와 SemEval2007 데이터 결과 비교

본 연구는 FS 방법과 9개의 분류기를 조합하여 효과적인 이모션 마이닝 모델을 제시하는 연구이다. Tweet 데이터와 SemEval2007 데이터에 대한 결과는 다음과 같다. 첫째, Tweet 데이터에서 DT와 RS는 CFS, IG방법을 적용한 속성을 사용한 결과와 사용하기전 결과가 통계적으로 유의하게 차이가 났다. 이같은 결과는 FS방법을 적용한 속성 2개만으로 DT와 RS의 성과를 상승시킬 수 있음을 나타낸다. SemEval2007 데이터의 경우, CFS와 IG 방법을 적용한 속성을 사용했을 때의 결과가 해당 FS를 적용하기 전 결과와 비교해 볼 때 NBN에서 통계적으로 유의하게 나타났다. 이때 CFS에 의한 속성수는 23개이고 IG에 의한 속성은 28개였다. 한편, ReliefF방법을 적용할 경우 선택된 속성수는 181개이고 이때 LR분류기에서 FS이전과 이후의 결과가 통계적으로 유의하게 차이가 났다. 둘째, Tweet 데이터의 경우, CFS와 IG를 적용한 속성을 사용한 RS의 정확도가 28.35로써 가장 높았다. CFS와 IG를 적용한 속성을 사용한 SVM의 정확도도 28.35로써 가장 높았지만 FS적용전 결과와 통계적으로 유의하지 않았다. SemEval2007 데이터의 경우에는 ReliefF방법에 의하여 선택된 속성을 적용한 LR의 정확도가 45.57로서 가장 높았다. Tweet 데이터는 anger, joy, love, sadness, surprise, thankfulness, fear의 7가지 감정이 타겟 클래스이고, SemEval2007 데이터의 타겟변수는 강한 긍정, 긍정, 부정, 강한부정 등 4가지의 감성 클래스를 갖고 있다. 따라서, 이 두 자료는 분류문제상 난이도가 높으며 이에 따라 어떠한 FS방법을 적용하고 어떠한 분류기를 사용하느냐에 따라 결과가 차이가 나는 것으로 판단된다.

5. 결론

지금까지 본 연구에서는 두 개의 오픈 자료인 Tweet 데이터와 SemEval2007 데이터를 이용하여 이모션 마이닝 작업 결과를 설명하였다. 이때 CFS, IG, ReliefF등 세 개의 FS방법을 적용하였고 분류기는 LR, DT, NN, NBN, SVM, BA, ST, RF, RS를 사용하였다.

이모션 마이닝의 결과 Tweet데이터는 DT가 CFS, IG, ReliefF를 적용한 속성을 사용한 경우 정확도가 상승하였고, RS는 CFS, IG를 적용한 속성을 사용할 때 정확도가 증대되었다. SemEval2007 데이터에서는 LR이 ReliefF를 적용한 속성을 사용한 경우 정확도가 상승했고, NBN의 경우는 CFS, IG를 적용한 속성을 사용할 때 정확도가 통계적으로 유의하게 향상되었다. Tweet 데이터에서는 CFS와 IG에 의한 속성을 사용한 RS의 정확도가 28.35로 가장 높았다. 반면, SemEval2007 데이터는 ReliefF에 의한 속성을 사용한 LR에서 45.57의 가장 높은 정확도가 도출되었다.

본 연구의 학문적 시사점은 다음과 같다.

첫째, FS방법을 적용하여 속성의 수를 효과적으로 줄일 수 있었고, 아울러 분류기를 사용한 데이터처리 시간을 크게 줄일 수 있었다. Tweet 데이터의 경우 속성수는 원래 총 18개였다. 이때 CFS, IG를 적용하면 속성수가 2개로 줄고, ReliefF를 이용하면 속성수가 15개로 줄어든다. SemEval2007 데이터에서의 원래 속성수는 총 339개인데, CFS를 이용하면 23개로 줄고, IG를 적용하면 28개로 줄어든다. 반면, ReliefF를 사용하면 181개로 줄어든다.

둘째, FS를 적용하면 FS전과 비교하여 분류기의 성능을 통계적으로 유의하게 향상될 수 있다. Tweet 데이터의 경우, DT는 CFS, IG, ReliefF를 적용한 속성을 사용할 때 정확도가 통계적으로 유의하게 증대된다. 반면, RS는 CFS, IG를 적용한 속성을 사용할 경우, 정확도가 유의하게 향상된다. SemEval2007 데이터에서는 LR이 ReliefF를 적용한 속성을 사용함으로써 정확도가 유의하게 증대되었고, NBN은 CFS, IG를 적용한 속성을 사용할 때 정확도가 향상되었다.

본 연구에서의 실무적인 의의는 다음과 같다.

첫째, 기업은 이모션 마이닝을 활용하여 고객이 상품 및 서비스에 대해 느끼는 상세한 감정과 의견을 효율적으로 분류할 수 있다. 기업은 이모션 마이닝을 이용해 소비자의 반응에 대한 대응전략을 수립할 수 있다. 즐거움, 감사, 사랑 등과 같은 긍정감정을 표현하는 소비자에게는 추가적인 구매를 추천할 수 있고, 분노, 슬픔과 같은 부정감정을 표출하는 고객에게는 원인을 파악하여 문제

를 해결할 열쇠를 찾을 수 있다.

둘째, FS를 통해 선택한 단어를 활용하여 해당 도메인에 대한 이모션 분류 모델을 구축할 수 있다. FS를 통해 선택된 단어는 소비자가 주로 감정표현에 사용하는 단어이므로 해당단어를 데이터화함으로서 고객의 감정을 파악하고 관리할 수 있다.

본 연구의 한계점은 다음과 같다. 이모션마이닝을 위해서 사용한 데이터는 Tweet 데이터와 SemEval2007 데이터이다. 이 두 데이터는 영문 리뷰이다. 그러므로 향후 국내의 한글 감성분석을 위한 한글 이모션 마이닝 모델에 적용하기 위해서는 한글 감성분석에 대한 추가적인 연구가 필요하다

REFERENCES

[1] J. A. Balazs & J. D. Velásquez. (2016). Opinion mining and information fusion: a survey. *Information Fusion, 27,* 95-110.

[2] H. L. Yang & Q. F. Lin. (2018). Opinion mining for multiple types of emotion-embedded products/services through evolutionary strategy. *Expert Systems with Applications, 99,* 44-55.

[3] M. V. Mäntylä, D. Graziotin & M. Kuutila. (2018). The evolution of sentiment analysis—A review of research topics, venues, and top cited papers. *Computer Science Review, 27,* 16-32.

[4] 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.

[5] T. Danisman & A. Alpkocak. (2008, April). Feeler: Emotion classification of text using vector space model. *In AISB 2008 Convention Communication, Interaction and Social Intelligence* (Vol. 1, p. 53).

[6] C., Strapparava & R. Mihalcea. (2007). Semeval-2007 task 14: Affective text. *In Proceedings of the 4th international workshop on semantic evaluations* (pp. 70-74). Association for Computational Linguistics.

[7] N. Gupta, M. Gilbert & G. D. Fabbrizio. (2013). Emotion detection in email customer care. *Computational Intelligence, 29(3),* 489-505.

[8] M. Hasan, E. Agu & E. Rundensteiner. (2014). Using hashtags as labels for supervised learning of emotions in twitter messages. *In ACM SIGKDD Workshop on Health Informatics*, New York, USA.

[9] C. Quan & F. Ren. (2016). Weighted high-order hidden Markov models for compound emotions recognition in text. *Information Sciences, 329,* 581-596.

[10] M. A. Hall. (1999). *Correlation-based feature selection for machine learning.*

[11] M. Robnik-Šikonja & I. Kononenko. (2003). Theoretical and empirical analysis of ReliefF and RReliefF. *Machine learning, 53(1-2),* 23-69.

[12] D. R. Cox. (1958). The regression analysis of binary sequences. *Journal of the Royal Statistical Society. Series B (Methodological),* 215-242.

[13] S. K. Murthy. (1998). Automatic construction of decision trees from data: A multi-disciplinary survey. *Data mining and knowledge discovery, 2(4),* 345-389.

[14] E. C. Bae & K. C. Lee. (2016). Predicting Stock Liquidity by Using Ensemble Data Mining Methods", *Journal of The Korea Society of computer and Information, 21(6),* 9-19,

[15] S. Park, K. M. Yang & S. B. Cho. (2013). A Hierarchical CPV Solar Generation Tracking System based on Modular Bayesian Network. *Journal of KIISE: Software and Applications,* 41.

[16] V. Vapnik. (2013). The nature of statistical learning theory. Springer science & business media.

[17] M. H. Song, J. Lee, S. P. Cho & K. J. Lee. (2005). SVM Classifier for the Detection of Ventricular Fibrillation, *The Institute of Electronics Engineers of Korea - System and Control, 42(5),* 27-34.

[18] M. Ballings, D. Van den Poel, N. Hespeels & R. Gryp. (2015). Evaluating multiple classifiers for stock price direction prediction. *Expert Systems with Applications, 42(20),* 7046-7056.

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

[20] J. H. Lee & J. G. Baek. (2018). RTC(Real-Time Contrast) Control Chart using Random Forest based Multi-Class Classifier, *Journal of the Korean Institute of Industrial Engineers, 44(4),* 306-315.

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

[22] W. Wang, L. Chen, K. Thirunarayan & A. P. Sheth. (2012). Harnessing twitter big data for automatic emotion identification. *In Privacy, Security, Risk and Trust (PASSAT), 2012 International Conference on and 2012 International Confernece on Social Computing (SocialCom)*, IEEE, 587-592.

[23] 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.

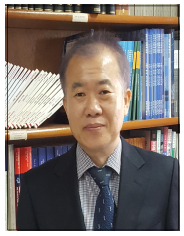
[24] S. Arlot & A. Celisse. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys, 4,* 40-79.

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