Information Systems Review
Vol. 17, No. 1 April 2015 http://dx.doi.org/10.14329/isr.2015.17.1.049

Support Vector Machine을 이용한 온라인 리뷰의 용어기반 감성분류모형

Terms Based Sentiment Classification for Online Review Using Support Vector Machine

이 태 원 (Taewon Lee) 부산대학교 경영학과 박사과정, 제1저자
홍 태 호 (Taeho Hong) 부산대학교 경영학과 교수, 교신저자

요 약

SNS의 확산으로 온라인 상점에서는 상품에 대한 주관적인 의견이 내포되어 있는 고객리뷰 정보가 빠르게 생성되고 확산되어 다른 고객들에게 큰 영향을 미치고 있다. 이와 더불어, 고객들의 긍정적 또는 부정적 의견을 분석하여 개선방안을 모색하려는 오피니언마이닝(opinion mining)이 주목 받고 있다. 고객리뷰에 내포된 감성정보를 가진 용어들은 감성분류를 하는데 가장 중요한 역할을 하기 때문에 영향력이 높은 용어를 선별하는 것이 가장 중요하다. 본 연구에서는 품사태깅을 이용하여 최적의 용어들을 선별하고 용어정보에 기반한 문서수준에서의 감성분류모형을 제안하고자 한다. 고객리뷰의 감성분류모형에 대표적인 기계학습기법인 SVM을 적용하고, SVM의 입력변수 선정과정에 품사태깅 방식과 용어추출기법을 다르게 조합하고 사용하여 긍정적/부정적 문서를 분류하였다. 본 연구에서 제안한 감성분류모형의 성과를 검증하기 위해 아마존(Amazon.com)의 영화와 도서에 대한 고객리뷰 80,000개를 수집하여 불필요한 용어들을 제거한 후 품사태깅을 통해 용어를 추출하였다. 추출된 용어는 문서빈도, TF-IDF, 정보획득량, 카이제곱 통계량의 값을 산출하여 값을 통해 용어들을 순위화하고, 각 상위 20개에 해당하는 최적의 용어를 선정한 후 SVM을 이용하였다. 제안된 감성분류모형을 통해 기존 연구에서 언급한 형용사만을 사용한 예측변수와 4품사를 사용한 예측변수에서의 실험결과를 통해 비교 분석하였다. 카이제곱 통계량 기반의 감성분류모형이 다른 모형보다 예측성과가 가장 우수하게 나타나는 것을 확인할 수 있었다. 본 연구에서 제안된 문서수준에서의 용어기반 감성분류모형을 이용함으로써 온라인 상점에서의 서비스 개선과 경쟁력 확보에 많은 도움이 될 것으로 기대된다.

키워드 : 오피니언마이닝, 감성분류, SVM

† 이 논문(저서)은 2014년 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임(NRF-2014S1A5A2A01017413).
2014년도 한국경영정보학회/한국정보기술응용학회 공동춘계학술대회 우수논문상을 수상한 논문임.

Ⅰ. 서 론

정보기술과 인터넷의 급속한 발달과 더불어 시간적, 공간적 제약에 구애 받지 않고 언제 어디서나 사용할 수 있는 스마트폰의 확산은 사용자들간의 커뮤니케이션을 촉진하여 소셜 네트워크 서비스(Social Network Service: SNS)의 폭발적 확산을 초래했다. 이와 더불어 온라인 상점에서의 고객리뷰가 SNS를 통해서 더 쉽고 빠르게 공유가 되면서 고객리뷰의 내용을 파악하는 것이 온라인 상점의 판매자는 물론 고객에게도 매우 중요한 이슈가 되었다. 온라인 상점들은 치열한 경쟁환경에서 경쟁우위를 확보하기 위한 방안으로 고객과 관련된 방대한 정보를 수집하고 분석한 결과를 활용하는 적극적인 경영활동이 요구되고 있다. SNS를 통한 고객리뷰의 빠른 전파로 인해 온라인 상점이 상품에 대한 불만사항이나 서비스 등의 문제점을 고려하여 사전조사를 통해 개선점을 찾고, 고객의 긍정적인 의견에 대해서도 신속한 대응을 할 필요성이 매우 높아지고 있다. 이처럼 고객리뷰에 대한 분석이 요구되고 있는 환경에서 풍부한 데이터 축적으로 주관적인 성격을 지닌 정보를 분석하고 분류하는 오피니언마이닝(opinion mining)이 주목을 받고 있다(Li *et al*., 2009; Tang *et al*., 2009; Chen *et al*., 2012; Wang *et al*., 2014; 박경미 등, 2011).

오피니언마이닝과 관련된 연구는 주로 온라인 상점의 고객리뷰 또는 댓글뿐만 아니라, SNS상의 고객의 의견을 감성의 극성(긍정 또는 부정)으로 판단하고 분류하는데 초점을 맞추어 왔다(Hu and Liu, 2004; Pang and Lee, 2008; Moraes *et al*., 2013). 하지만, 수 많은 리뷰들 중에는 필요한 정보 외에도 의도적으로 작성된 내용뿐만 아니라 잘못된 정보 혹은 광고 등이 포함되어 있기 때문에 불필요한 리뷰들을 제거하는 과정을 필요로 한다. 정확한 고객의 감성분류(sentiment classification)를 위해서는 고객리뷰에서 고객의 의견만을 대상으로 오피니언마이닝을 수행하여야 한다. 특히, 리뷰에 내포된 용어들이 감성분류를 하는데 가장 중요한 역할을 하고, 용어를 추출하는 과정에서 문서 전체에 많은 영향을 미치기 때문에 감성분류를 하는데 효율적이면서 효과적인 용어를 선별하는 것이 가장 중요하다.

지금까지 대부분의 연구에서는 감성분류를 위한 용어추출에만 집중되어 선별적인 용어추출을 통한 감성분류에는 소홀한 경향이 있어왔다(Hu and Li, 2011). 감성정보를 내포하고 있는 용어들을 선별하기 위해 품사에 대한 정보가 가장 중요한 지표가 된다(Xia *et al*., 2011). Hatzivassiloglou and McKeown(1997)는 명사나 동사 등과 같은 다양한 품사들을 이용하여 감성 정보를 표현할 수 있음에도 불구하고 모델을 통해 극성을 분류하는데 오직 형용사만이 좋은 입력변수라고 주장하였다. 명사의 경우 상품에 대한 속성을 많이 나타내고, 형용사와 동사는 구매자에 대한 주관적인 의견을 내포하고 있으며, 부사의 경우 다양한 표현방법과 수식어로 분류가 되어 문서를 분류하게 된다. Moraes *et al*.(2013)은 불필요한 용어들을 제거하고 정보획득량(information gain)을 이용하여 5,000개에 해당하는 많은 용어들을 추출한 후 SVM과 ANN에 적용하여 문서수준에서의 감성분류에 대한 연구를 진행하였다. 하지만 이와 같이 너무 많은 용어를 추출하여 입력변수로 사용할 경우 결측치의 발생가능성이 크고, 중복성이 높아 감성정보가 내포되어 있는지에 대한 여부를 알 수 없으며, 비효율적으로 너무 많은 용어를 입력변수로 사용한다는 단점을 가지고 있다. 이러한 단점을 보완하기 위해 본 연구에서는 4품사(형용사, 부사, 동사, 명사)에 해당하는 용어에 태깅하여 최적의 용어들을 선별하는 과정을 거쳐 문서수준에서의 감성분류에 대한 연구를 진행하고자 한다. 기존 연구처럼 형용사만을 사용한 입력변수에서의 우수성을 검증하기 위한 감성분류모형을 SVM을 이용하여 개발한다. 또한, 용어추출과정에서는 문서빈도(document frequency), TF-IDF(term frequency-inverse document frequency),

정보획득량(information gain), 카이제곱 통계량(chi-squared statistic)을 이용하여 추출된 용어 중 최적의 용어들만으로 감성분류모형에 적용함으로써 기존의 너무 많은 입력변수로 인한 모형의 복잡성과 문제점들을 해결하고자 한다.

본 논문의 구성은 다음과 같다. 제 Ⅱ장에서는 관련연구를 제시하고, 제 Ⅲ장에서는 문서수준에서 용어기반의 감성분류모형을 제시한다. 제 Ⅳ장에서는 제안된 모형에 대한 실험결과를 제시한다. 마지막으로 제 Ⅴ장에서는 결론을 맺는다.

Ⅱ. 관련연구

2.1 감성분류를 위한 오피니언마이닝

오피니언마이닝(opinion mining)은 사용자가 다양한 매체를 통해 표출한 의견을 추출, 분류, 이해하는 과정을 의미한다(Liu, 2012). 즉, 데이터마이닝의 문서분류 기술에서 발전된 오피니언마이닝은 특정 문서를 직접 작성한 사람의 감성을 추출해 내는 기술로써, 최근에 들어와서 소셜 네트워크 분석을 하는데 가장 큰 관심연구 분야로 성장하고 있다(Pang and Lee, 2008; Rao *et al.*, 2014; 장재영, 2009). 일반적으로 오피니언마이닝에서는 문서수준의 분석과 구나 문장수준의 분석으로 이루어진다. 문서수준의 분석은 전체 문서에 대한 의견을 종합하여 긍정적 혹은 부정적 성향을 판단하는 것으로써, 문서 내의 특정 용어들을 통해 감성을 표현하는 빈도수에 따라 판별해 내는 것이고, 구나 문장수준의 분석은 문서를 문장 단위로 나누어 개별적인 문장들을 통해 의견을 파악하고 감성을 포함한 문장을 추출하여 다양한 용어들로 분석해 내는 것이다(강대국, 박용태, 2012). 온라인 상에서의 오피니언마이닝을 이용하여 사람들이 특정 대상에 대한 긍정적인 견해와 부정적인 견해를 작성한 문서를 분석하고 대중의 관심이 실시간으로 어떻게 반영되고 변하는지에 대해 정보를 빠르게 분석해주는 많은 연구들이 진행되고 있다(O'Leary, 2011; Hu and Liu, 2004). 오피니언마이닝은 텍스트마이닝의 한 분야로 텍스트 문서를 이용하여 고객의 의견을 수집하고 분석하는 텍스트마이닝 기법을 활용하여 감성분류를 위한 방법으로 설명되고 있다(강범일 등, 2013; 김승우, 김남규, 2014).

Yang and Pedersen(1997)은 문서수준에서의 감성분류를 위한 연구로 문서빈도(document frequency), 정보획득량(information gain), 상호정보량(mutual information), 카이제곱 통계량(chi-squared statistic), 용어의 강도(term strength) 등을 이용하여 오프라인 문서를 기반으로 k-최근접이웃 기법과 회귀분석 기법 중 선형최소자승기법으로 성능 평가를 하였다. 이 연구에서는 정보획득량, 문서빈도, 카이제곱 통계량에서 강한 상관관계가 있다는 것을 실험결과에서 발견하였고, 수행속도 측면에서는 다른 방법들에 비해 문서빈도의 방법이 더 효율적이라고 주장하였다. Moraes *et al.*(2013)은 영화리뷰와 온라인 상품리뷰들을 대상으로 문서수준에서 포함된 용어들을 stemmer 알고리즘을 이용하여 불필요한 용어들을 제거하고 정보획득량을 이용하여 많은 용어들을 추출하여 SVM과 ANN에 적용하여 실증연구를 진행하였다. 이 연구에서는 50개에서 5,000개에 해당하는 너무 많은 용어들을 SVM과 ANN의 입력변수로 감성정보가 내포되어있는 용어들이 전체 문서에 잘 반영되었는지 혹은 유의한 용어들이 사용되었는지에 대한 논의는 없었다. 이와 같이 기계학습에서는 입력변수로 사용될 용어의 수가 많을수록 결측치의 발생가능성이 크고, 예측모형의 오류가 발생할 여지가 높게 된다는 단점을 가지고 있다. Hu and Li(2011)는 영화리뷰 문서와 전자제품에 대한 리뷰 문서를 이용하여 문서수준에서의 감성분류에 관한 연구를 진행하였다. 이 연구에서는 Topical Terms Description Model (TTDM) 모형을 제안하였으며, 임의로 문장을 선정하여 용어를 추출하는데 영화리뷰에서는 30개, 전자제품리뷰에서는 37개를 추출하였다. 평균 정

확성을 통해 영화리뷰에서는 상위 10개, 전자제품리뷰에서는 상위 23개에 해당하는 해당하는 용어들을 선별하여 모형을 검증하였다.

2.2 용어정보추출 기법

고객리뷰는 수많은 용어들과 고객들의 의견이 반영되어 있어 고객이 직접 느끼고 있는 감성과 감정, 그리고 의견까지 모든 주관적인 정보를 내포하고 있다. 따라서 고객리뷰에 내포되어있는 용어들을 추출하고, 용어에 대한 정보를 추출하는 것이 가장 중요한 핵심 요소가 된다.

2.2.1 문서빈도와 TF-IDF

문서빈도(document frequency)는 전체 문서들 중에서 용어가 발생한 문서의 수를 나타낸 것으로 일정 빈도 이상의 문서에 발생한 용어들을 내포하고 있는 문서의 비율을 나타낸다. 또한, 가장 간단하고 계산량이 적은 방법으로 알려져 있으며 빈도수가 낮은 용어일수록 문서분류에 기여하지 못한다는 가정을 기반으로 저빈도 및 불용어를 제거함으로써 문서 범주화의 정확성을 높일 수 있다(Li *et al*., 2009; Yang and Pedersen, 1997). 전체문서 N과 용어 i를 가진 문서빈도(DF)는 식 (1)과 같다.

$$DFi = \frac{ni}{N} \tag{1}$$

예를들어, 전체 문서의 수가 2,000개이고 bad를 포함한 문서의 수가 1,395개라 가정할 때, 문서빈도의 값은 1,395/2,000 = 0.698이 되므로 bad의 DF 값은 0.698이 되는 것이다. 이처럼 용어를 가진 문서들의 수와 전체 문서의 개수를 통해 DF 값을 산출 할 수 있다. 불용어의 경우 DF 값이 높게 산출되기 때문에 분류하는데 있어 정보력이 약할 뿐만 아니라 기여도가 크지 않게 된다는 문제점이 발생하게 된다.

TF-IDF(term frequency-inverse document frequency)는 개별적인 문서에서 용어에 대한 중요도를 표현할 수 있는 방법으로 문서빈도와 유사하며 간단하면서도 성능이 우수하기 때문에 많이 사용되는 기법이다. TF-IDF는 용어의 가중치를 부여하여 계산하는 방법으로 알려져 있으며 구와 절 수준에서의 분석에서 많이 이루어지고 있다(Han and Kamber, 2011). 전체 문서 N, 용어 t, 문서 d를 가진 TF-IDF의 계산방법은 식 (2)와 같다.

$$TF-IDF_{t,d} = TF_{t,d} \times IDF_t,$$
$$where$$
$$IDF_t = \log \frac{N}{DF_t} \tag{2}$$

2.2.2 정보획득량

정보획득량(information gain)을 이용하는 계산방법은 문서빈도의 방법보다 용어정보추출 과정에서 보다 효과적인 방법으로 오피니언마이닝 연구에서 가장 많이 사용되는 기법 중 하나이다. 문서에서의 발생빈도뿐만 아니라 발생하지 않은 빈도까지 고려하여 각 범주에서의 용어 정보량을 계산하는 방법이다. 전체 문서들의 집합을 하부집합으로 분리하고 특정 용어를 사용하기 전과 후에 달라지는 엔트로피를 계산한다. 알아내고자 하는 대상의 확률을 나누어 특정 용어에 대한 정보획득량을 구한 다음에 표준화된 획득량(normalizing gain)을 얻게된다(Abbasi *et al*., 2011; Moraes *et al*., 2013). 용어 t에 대한 정보획득량(IG)을 구하기 위해서는 먼저 전체 문서의 엔트로피 K를 구한다.

$$E(K) = \sum_{i=1}^{m} P_i \log_m \left(\frac{1}{P_i} \right) = -\sum_{i=1}^{m} P_i \log_m P_i \tag{3}$$

식 (3)에서 Pi는 알아내고자 하는 대상의 확률이고 m은 범주의 집합이다. 가령, 2,000개의 문서 중에서 긍정문서 800개, 부정문서 700개, 긍정문서도 부정문서도 아닌 문서가 500개라면 m은 긍정, 부정, 중립에 해당하는 집합으로 3이 된다. 확률을 구하여 식 (3)에 적용한 후 합을 계산하면 전

체 문서의 엔트로피 E(K)는 0.9835가 된다. 다음으로 용어에 대한 정보값(information value)를 계산하기 위해 문서에 용어가 발생하지 않으면 0, 발생하면 1로 정하여 두 그룹으로 나누고, 문서가 긍정인지 부정인지에 대한 Y값을 찾아 해당하는 용어의 정보값을 계산한다. <표 1>에서처럼 10개의 문서에 12개의 용어가 범주형 변수들로 표현되어 있다고 가정한다.

〈표 1〉 문서와 용어간의 빈도

	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10	T11	T12	Y
D1	0	1	1	1	0	1	1	1	1	1	0	0	0
D2	0	1	0	0	1	1	0	0	0	1	1	0	0
D3	1	1	1	1	1	1	0	0	0	1	0	1	1
D4	1	1	1	0	1	0	0	0	0	1	0	1	0
D5	1	1	1	1	1	1	1	1	1	0	1	1	1
D6	1	1	1	1	1	1	1	0	0	1	0	1	1
D7	1	0	1	0	0	1	0	0	0	0	1	1	0
D8	1	0	1	1	0	1	1	1	1	0	0	0	1
D9	1	0	1	0	1	0	1	0	0	0	0	1	0
D10	1	0	1	1	1	1	0	1	0	0	1	1	1

Ti: i번째 용어, Di: i번째 문서, Y: 긍정(1)/부정(0)
i = 1, 2, 3, …, n.

〈표 2〉 T12에 대한 두 범주

T12의 발생여부	전체용어 발생여부	긍정(1)/부정(0)
발생된 그룹	111111000101	1
	111010000101	0
	111111111011	1
	111111100101	1
	101001000011	0
	101010100001	0
	101111010011	1
발생하지 않은 그룹	011101111100	0
	010011000110	0
	101101111000	1

<표 2>에서처럼 12번째 용어가 발생한 그룹에서 긍정으로 분류한 문서는 1개, 부정으로 분류한 문서는 2개이다. 또한, 용어가 발생하지 않은 그룹에서 긍정으로 분류한 문서는 4개, 부정으로 분류한 문서는 3개이다.

$$E(S0) = -\left(\frac{2}{3}\right)\log_2\left(\frac{2}{3}\right) - \left(\frac{1}{3}\right)\log_2\left(\frac{1}{3}\right) \quad (4)$$

$$E(S1) = -\left(\frac{3}{7}\right)\log_2\left(\frac{3}{7}\right) - \left(\frac{4}{7}\right)\log_2\left(\frac{4}{7}\right) \quad (5)$$

E(S0) = 0.9183, E(S1) = 0.9852가 된다. 따라서, 용어에 대한 정보값은 E(Tn) = {E(S0)×(용어가 발생하지 않은 그룹에서의 문서수)/(총 문서수)}+{E(S1)×(용어가 발생한 그룹에서의 문서수)/(총 문서수)}를 이용하여 계산하면 T12에 대한 정보값은 0.9651이 된다. 표준화된 정보획득량을 산출하기 위해 gain값을 계산해야 한다. 즉, 전체 문서에서 발생한 엔트로피에서 해당용어의 정보값의 차이를 나타낸 것으로, T12에 대한 gain값은 0.0184가 되는 것이다. 마지막으로, T12의 표준화된 정보획득량의 값은 해당 용어의 정보값을 엔트로피로 나눈 값이므로 0.0187이 된다.

2.2.3 카이제곱 통계량

카이제곱 통계량(chi-squared statistic)의 계산방법은 교차분석 기법 중 가장 많이 사용되고 범주형 변수들(전체 문서에서 발생한 용어의 빈도수가 최소 한번 이상일 경우 1, 발생하지 않을 경우 0) 간의 연관성을 분석하는 것으로 용어 t_i와 범주 c_j간의 중요도를 구하는 기법이기도 하다. 카이제곱 통계량은 상호정보량과 비슷하지만 차이점은 카이제곱 통계량이 표준화된 값이기 때문에 용어추출에 있어서 훨씬 더 좋은 성능을 보인다는 것이다. 동일한 성질을 가지고 있는 범위에서의 범주 c_j와 용어 t_i간의 카이제곱 통계량(χ^2)은 식 (6)과 같이 정의될 수 있다(Li *et al.*, 2009; Yang and Pedersen, 1997).

〈표 3〉 용어(t_i)와 범주(c_j)간의 문서빈도

	범주 c_j가 발생된 문서	범주 c_j가 발생 되지않는 문서
용어 t_i를 가진 문서	A	B
용어 t_i를 가지지 않은 문서	C	D

$$\chi^2 = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)} \quad (6)$$

Where,
N = A+B+C+D (N: 전체 문서의 수)
i = 1, 2, …, n (n: 전체 용어의 수)
j = 1, 2, …, m (m: 전체 범주의 수)

예를들어, <표 1>에서 T12에 대한 카이제곱 통계량의 값을 계산하면 <표 4>처럼 나타낼 수 있다.

〈표 4〉 T12에 대한 용어와 범주간의 문서빈도

	범주 c_j가 발생된 문서	범주 c_j가 발생 되지않는 문서
용어 t_i를 가진 문서	4	1
용어 t_i를 가지지 않은 문서	3	2

전체 문서의 수 N은 10이 되고, T12의 카이제곱 통계량의 값을 식 (6)과 같이 계산하면 0.4762가 된다.

2.3 SVM을 활용한 감성분류

SVM(Support Vector Machine)은 Vapnik(1995)에 의해 개발된 기계학습기법으로 학습이론에 기반하여 이진분류의 문제를 해결하기 위해 만들어졌으며, 고차원적인 문제(high-dimensional problem)를 해결할 수 있기 때문에 많은 연구자들에게 관심을 보이고 있다. 비선형문제를 고차원에서 특정공간의 선형문제로 해결하기 위해 데이터를 서로 다른 두 개의 클래스로 분류하여 기준이 되는 최적분리경계면(maximum margin hyperplane)을 찾아 가장 가까운 점(support vector)과의 거리를 최대화하여 커널함수를 통해 데이터를 고차원의 특징공간으로 사상시킴으로써 비선형일 경우 선형문제로 전환하여 분리하는 방법으로 알려져 있다. 특히, 분류 및 회귀문제를 해결할 수 있는 능력을 가지고 있으며, 다른 분류기법들과 비교하여 우수한 성능을 보인다고 알려져 있다(Tay and Cao, 2001). SVM은 사용하는 커널함수와 파라미터(C, ɣ)의 설정값에 따라 모형의 전체성과가 달라지게 되고, 결과값의 차이가 다르게 나타나기 때문에 학습용 데이터를 통해 최적의 파라미터값을 산출하여 검증용 데이터에서의 예측성과를 알 수 있다. 일반적으로 사용되는 커널함수로 선형커널(Linear kernel), 다항식 커널(Polynomial kernel), 가우시안 RBF 커널(Gaussian Radial Basis Function kernel)과 시그모이드 커널(Sigmoid kernel) 등이 있으며, SVM 알고리즘의 구체적인 계산과정에 대한 설명은 다음과 같다.

학습 데이터를 이용하여 분류문제를 해결하기 위해 함수 f: Rn →{±1}을 추정하도록 하며, 두 개의 클래스 중 A는 $x_i \in R^n$, $y_i = +1$로, $x_i \in R^n$, $y_i = -1$로 표시한다. x는 입력벡터, w와 b는 분리경계면을 결정하는 모수로, w는 가중치 벡터이고, b는 바이어스이다.

$$y_i(w \cdot x_i + b) \geq 1, \ \forall x_i \in A \cup B \quad (7)$$

최대 분리경계면을 가지고 데이터를 최적화문제를 해결하기 위한 최적 의사결정 분리경계면은 식 (7)과 같다. a_i와 b는 분리경계면을 결정하는 모수이고, x는 학습 데이터, xi는 support vector를 나타낸다.

$$f(x, \alpha_i, b) = \Sigma y_i \alpha_i (x \cdot x_i) + b \quad (8)$$

학습 데이터가 비선형일 때 입력변수를 고차원의 특징공간으로 이동시켜 선형문제로 근사 시킬 수 있으며, 비선형문제에서는 식 (8)과 같이 나타낸다. K(x · xi)를 커널함수라 하고 데이터를 고차원공간으로 사상시킴으로써 특징공간 내에 선형으로 분리 가능한 입력데이터셋을 만든다.

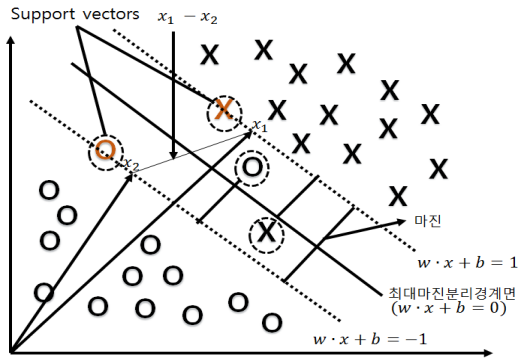
$$f(x, \alpha_i, b) = \Sigma y_i \alpha_i K(x \cdot x_i) + b \quad (9)$$

식 (10)은 다항식 커널을 나타내며, 식 (11)은 가우시안 RBF 커널로 나타낸 것이다.

$$K(x, x_i) = (x \cdot x_i + 1)^d \quad (10)$$

$$K(x, x_i) = \exp\left(-\frac{1}{\sigma^2}(x - x_i)^2\right) \quad (11)$$

오피니언마이닝 연구에서는 SVM을 활용하여 개선된 알고리즘을 제시하거나 트위터, 영화, 상품, 뉴스, 언어 등 온라인상에서 발생하는 수많은 데이터를 사용하여 감성분류 및 분석을 한다(홍초희, 김학수, 2012; Moraes *et al.*, 2013). 또한, 다른 분류기법들과 비교하여 분류문제를 해결하는데 우수한 성능을 보이고 있어 많은 연구에 사용되고 있다(임좌상, 김진만, 2014; Felipe *et al.*, 2014; Silva *et al.*, 2014).

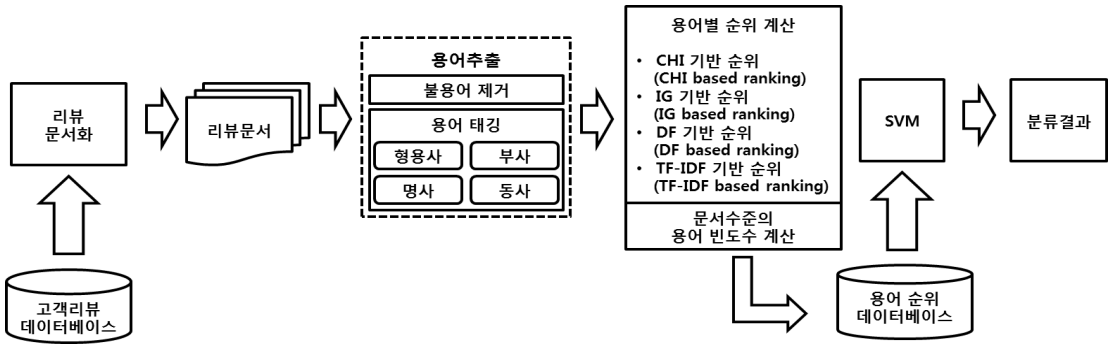


〈그림 1〉 SVM의 의사결정 경계와 마진(Vapnik, 1995)

Ⅲ. 연구 프레임워크

본 연구에서는 품사태깅을 이용하여 최적의 용어를 선정하는 과정을 통해 용어기반의 감성분류모형을 <그림 2>와 같이 개발하였다. 연구 프레임워크에 대한 구체적인 설명은 다음과 같다.

1) 고객리뷰 데이터베이스를 구축하여 아마존에서 영화와 도서에 대한 고객리뷰를 수집하기 위해 크롤링한 후 저장한다.
2) 저장된 리뷰를 이용하여 긍정적인 리뷰와 부정적인 리뷰로 분류하여 각 리뷰를 문서화한다.
3) 전체 문서에서 생성된 용어들을 모두 추출한 후 불용어(stopword)를 제거한다. 높은 빈도수를 가지고 있으면서 정보력이 낮은 불용어들이 과



〈그림 2〉 문서수준에서 용어기반의 감성분류모형을 위한 연구프레임워크

반수이기 때문에 제거의 대상이 된다. 불용어는 전체 문서에서 공통적으로 빈도수가 많으면서 특별하게 정보를 제공하지 못하는 것으로 주로 사용되는 관사(a, an, the 등), 접속사(that, and, when 등), 대명사(I, you 등), Be 동사(is, are 등) 등이 있다. 이를 제거함으로써 영향력이 높은 용어들만 선별할 수 있게 된다.

4) 불용어를 제외한 모든 용어들에 대해 4품사(형용사, 부사, 동사, 명사)에 대한 정보를 태깅한다. 품사에 대한 정보는 감성표현에 있어 가장 중요한 지표가 된다. 수많은 용어들을 추출한 후 품사태깅을 하지 않을 경우 용어 선별시 많은 수작업이 필요하며, 전문가의 견해를 통해 선정된 용어들을 사용하여야 하기 때문에 비용적인 측면과 시간적인 측면에서 많은 소비가 이루어지게 된다. 하지만 품사를 태깅함으로써 적은비용과 효율적인 시간을 활용할 수 있을 뿐만아니라 용어에 대한 성질을 정확히 판단할 수 있고, 감성정보를 내포하고 있는 품사에 대해서도 쉽게 선별할 수 있게 된다. 이와 같이 정보력과 영향력이 높은 용어들을 선별하기 위해 문법 구조 분석 도구인 Stanford POS phaser 프로그램을 이용하여 모든 용어들에 대해 태깅한다(Xia *et al.*, 2011).

5) 최적의 입력변수로 사용될 용어들을 추출하기 위해 태깅 정보로 용어 1,000개를 추출한다. 추출한 용어들을 이용하여 전체 문서에 대한 빈도수를 측정한 후 문서빈도, TF-IDF, 정보획득량, 카이제곱 통계량의 값들을 산출한다. 4가지 방법으로 산출한 용어들을 순위화하여 각각의 상위 20개에 해당하는 최적의 용어들을 선정한다. Hu and Li(2011)의 연구에서처럼 상위에 해당하는 용어들의 수를 이용하여 정확성에 대한 평균값을 측정하여 상위 20개의 용어를 사용하였을 때 가장 효과적인 것으로 확인하였다. Shmueli *et al.*(2010)은 입력변수들의 수가 너무 많을 때 발생할 수 있는 문제점들로 결측치의 증가 및 모형의 관리 측면에서 지적하였으며, 입력변수에 대한 파라미터의 개수가 적은 모형에서 입력변수의 영향력을 더 잘 이해할 수 있다고 하였다. 또한, 상관관계가 존재하는 변수들이 나타날 가능성이 매우 높기 때문에 이를 제거한다면 예측값의 분산이 줄어듦에 따라 예측력이 향상되기 때문에 좋은 모델이 된다고 하였다(Friedman, 1997; Geman *et al.*, 1992; Hastie *et al.*, 2008; Shmueli and Koppius, 2011).

6) 상위 20개에 해당하는 최적의 용어를 기반으로 전체 문서간의 빈도수를 계산한다. 최적의 용어를 선정하여 형용사만을 사용한 입력변수와 4품사를 사용한 입력변수와의 비교를 위해 전체 문서를 대상으로 용어 빈도수를 계산한다. 영화리뷰와 도서리뷰에서 형용사와 4품사에 대한 입력변수를 사용하게 되면 총 16개의 모형이 생성된다.

7) 16개의 모형을 저장하기 위해 용어순위 데이터베이스를 구축한다.

8) SVM을 이용하여 용어기반의 감성분류모형에 대한 결과를 알아본다.

Ⅳ. 실험 및 분석

4.1 데이터

아마존(Amazon.com)의 영화와 도서판매를 위한 고객 리뷰를 1점에서 5점의 형태로 제공되고 있으며 상품별로 매우 풍부한 고객리뷰를 보유하고 있다. 실제로 관람한 고객들과 책을 직접 읽은 고객들에 대한 주관적인 정보 모두 고객의 감성에 대한 정보가 다양하게 내포되어 있다는 장점을 가지고 있다. 본 연구에서는 영화와 도서에 해당하는 고객리뷰 데이터를 수집하였다. 데이터를 수집하는 과정에서 별점 4점 이상에 해당하는 문서와 2점 이하에 해당하는 문서를 각각 긍정적인 문서와 부정적인 문서로 수집하였다. 별점 3

점은 중립에 해당함으로 제외하였다. 또한, 영화의 경우 한 편당 긍정적 리뷰 100개, 부정적 리뷰 100개를 수집하였고, 도서 역시 한 권당 긍정적 리뷰 100개, 부정적 리뷰 100개를 수집하여 분류하였다. 최종적으로 본 연구에서는 영화리뷰 40,000개(긍정적 리뷰 20,000개, 부정적 리뷰 20,000개)와 도서리뷰 40,000개(긍정적 리뷰 20,000개, 부정적 리뷰 20,000개)로 총 80,000개의 데이터를 사용하였다.

4.2 실험 결과

본 연구에서 제안된 용어기반의 감성분류모형을 평가하기 위해 실험을 실시하였으며, 통계 소프트웨어 'R'을 사용하였다. 영화리뷰와 도서리뷰 모두 동일한 방법으로 32,000개의 학습용 데이터와 8,000개의 검증용 데이터로 나누어 실험하였다. 형용사만을 사용한 입력변수와 4품사(형용사, 부사, 동사, 명사)를 사용한 입력변수간의 비교를 위해 문서빈도, TF-IDF, 정보획득량, 카이제곱 통계량을 기반으로 추출한 용어들 중 상위 20개를 선정하였다.

<표 5>에서처럼 영화리뷰를 이용하여 4품사에 해당하는 1,000개의 용어 중 형용사에 해당하는 용어는 총 184개로 나타났다. 4품사를 사용하여 선정한 20개의 용어 중 문서빈도, 정보획득량, TF-IDF에서는 모두 2개의 형용사가 포함되어 있었고, 카이제곱 통계량에서는 12개의 형용사를 포함하고 있었다.

도서리뷰를 이용하여 4품사에 해당하는 1,000개의 용어 중 형용사에 해당하는 용어는 총 196개로 나타났다. 4품사를 사용하여 선정한 20개의 용어 중 문서빈도에서는 4개, 정보획득량에서는 5개, 카이제곱 통계량에서는 6개, TF-IDF에서는 3개의 형용사가 포함되어 있었다.

선정된 용어들을 기반으로 문서수준에서의 빈도수를 계산하여 총 16개의 모형에 대한 결과를 알아보기 위해 SVM에 적용하였다. 본 실험에서는 학습용 데이터를 통해 최적의 파라미터값을 산출하기 위해서 가우시안 RBF 커널함수를 사용하였으며, 파라미터 C는 C = {1, 5, 10, 15, 20, 25, 40, 60, 75, 100}으로 설정하였고, γ값은 γ = {0.001, 0.005, 0.01, 0.02, 0.03, 0.05, 0.1, 0.15, 0.2, 1}로 설정하여 격자 탐색(grid search)을 통해 파라미터를 선정하여 실험하였다.

〈표 5〉 영화리뷰에서 형용사와 4품사를 사용한 상위 20개의 입력변수

	형용사	4품사
DF_기반	great, bad, most, funny, real, right, big, last, boring, different, second, amazing, excellent, disappoint, favorite, video, classic, top, entire, perfect	star, people, movie, not, film, time, great, better, best, action, too, bad, plot, character, acting, say, effects, back, end, want
IG_기반	great, bad, most, boring, funny, real, stupid, amazing, excellent, right, big, disappointed, awful, favorite, last, perfect, awesome, christ, different, wonderful	people, movie, not, great, film, time, has, had, waste, best, plot, better, too, action, character, boring, say, effects, back, want
CHI_기반	great, bad, boring, awful, amazing, excellent, stupid, awesome, favorite, perfect, wonderful, fantastic, beautiful, disappoint, classic, outstanding, pointless, stunning, slow, dumb	great, waste, bad, boring, best, not, awful, amazing, excellent, enjoyed, stupid, people, awesome, action, favorite, perfect, wonderful, fun, fantastic, family
TFIDF_기반	great, bad, most, funny, real, right, big, last, video, boring, different, second, excellent, amazing, classic, favorite, comic, perfect, disappoint, top	film, not, movie, has, great, time, had, book, action, did, character, bad, made, best, most, also, better, version, plot, people

〈표 6〉 도서리뷰에서 형용사와 4품사를 사용한 상위 20개의 입력변수

	형용사	4품사
DF _기반	good, much, first, other, many, great, written, best, novel, little, new, long, old, real, interesting, right, bad, last, hard, different	stars, book, people, read, not, story, time, good, much, first, other, really, way, characters, many, life, well, great, love, think
IG _기반	good, much, first, other, many, great, written, best, bad, novel, little, new, long, interesting, old, real, right, last, disappointed, hard	stars, book, people, read, not, story, time, good, much, really, first, other, life, way, many, great, love, well, think, know
CHI _기반	great, wonderful, amazing, best, worst, disappoint, excellent, bad, poor, stupid, ridiculous, easy, awful, beautiful, terrible, favorite, predictable, human, annoying,	boring, waste, money, read, life, great, love, wonderful, amazing, poorly, loved, well, best, recommend, worst, highly, disappointing, world, heart, bad
TFIDF _기반	other, good, first, much, many, novel, great, written, best, little, new, old, long, real, bad, right, interesting, last, different, hard	not, story, read, book, time, life, very, characters, reading, other, good, first, really, love, much, only, many, way, novel, world

〈표 7〉 최적의 파라미터값

	파라미터(C, γ)	
	Movie	Book
SVM_DF	C = 10, γ = 0.1	C = 10, γ = 0.1
SVM_TF-IDF	C = 10, γ = 0.1	C = 10, γ = 0.1
SVM_IG	C = 1, γ = 0.02	C = 15, γ = 0.03
SVM_CHI	C = 5, γ = 0.001	C = 20, γ = 0.01

각각의 기법들을 통해 산출된 최적의 파라미터값에 대한 결과는 <표 7>과 같이 나타났다. 파라미터의 값은 4품사를 사용한 입력변수에서 선정되었으며, 이를 기준으로 형용사를 사용한 모형에 적용하여 실험하였다. 선정된 파라미터의 값으로 검증용 데이터에서의 예측성과를 알아보기 위해 정오분류표를 이용하여 정확성(accuracy), 재현성(recall), 정밀성(precision)에 대한 결과값을 알아보았다.

〈표 8〉 정오분류표

		예측범주	
		긍정	부정
실제 범주	긍정	TP(True Possitive)	FN(False Negative)
	부정	FP(False Possitive)	TN(True Negative)

$$\text{정확성}(accuracy) = \frac{TP+TN}{TP+TN+FP+FN} \quad (10)$$

$$\text{재현성}(recall) = \frac{TP}{TP+FN} \quad (11)$$

$$\text{정밀성}(precision) = \frac{TP}{TP+FP} \quad (12)$$

정오분류표는 실제 범주와 분석된 모형에 의해 분류된 예측 범주의 관계를 나타내는 표로서 정확성은 감성분류의 결과가 얼마나 정확한가를 나타내고, 높을수록 성능이 좋은 것으로 판단할 수 있다. 정밀성은 예측에 대해 얼마나 정확한지에 대해 나타낼 수 있으며, 재현성은 실제로 존재하는 감성분류 결과 중 실제 값의 비율로 나타낼 수 있다. 영화리뷰에서 형용사를 사용한 상위 20개의 입력변수를 이용하여 용어기반의 감성분류모형에 적용한 결과를 알아보았다.

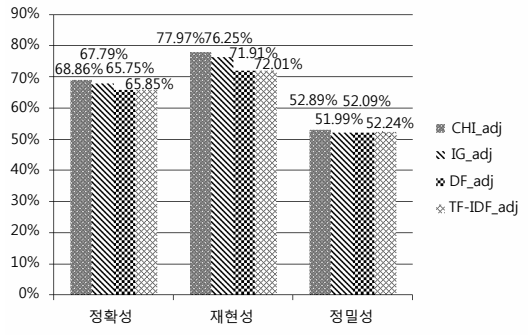
카이제곱 통계량 기반의 감성분류모형에서 예측 정확성은 다른 감성분류모형에 비해 68.86%로 가장 높게 나타내었으며, 정보획득량 기반의 감성분류모형에서는 67.79%, 문서빈도 기반의 감성분류모형에서는 65.5%로 가장 낮은 예측 정확성을 보이고 있었다. TF-IDF 기반의 감성분류모형에서의 결과에서는 65.85%로 문서빈도 기반의 감성분류모형과 거의 비슷하게 나타나는 것을 확

인할 수 있었다. 정밀성에 대한 결과는 전체적으로 다소 비슷한 양상을 보이고 있었다.

〈표 9〉 영화리뷰에서 형용사를 사용한 결과

(단위: %)

		SVM_CHI	SVM_IG	SVM_DF	SVM_TF-IDF
학습용 데이터	정확성	68.81	68.36	69.38	69.34
	재현성	77.57	77.03	77.81	77.60
	정밀성	52.85	52.23	54.16	54.31
검증용 데이터	정확성	**68.86**	**67.79**	**65.75**	**65.85**
	재현성	77.97	76.25	71.91	72.01
	정밀성	52.89	51.99	52.09	52.24



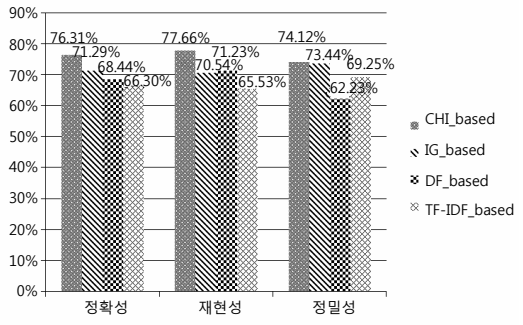
〈그림 3〉 영화리뷰에서 형용사를 사용한 감성 분류모형의 결과 비교

영화리뷰에서 4품사를 이용하여 실험한 결과 <표 10>에서처럼 카이제곱 통계량 기반의 감성분류모형에서 76.31%의 예측 정확성으로 가장 높게 나타났고, 정보획득량 기반의 감성분류모형에서는 71.29%, 문서빈도 기반에서는 68.44%의 예측 정확성으로 나타났으며, TF-IDF 기반의 감성분류모형의 예측 정확성은 66.30%로 가장 낮게 나타나는 것을 확인할 수 있었다. 재현성과 정밀성에 대한 결과 역시 카이제곱 통계량 기반의 감성분류모형에서 가장 높게 나타났으며, 형용사를 이용한 감성분류모형의 결과보다 모두 높게 나타나는 것을 확인할 수 있었다.

〈표 10〉 영화리뷰에서 4품사를 사용한 결과

(단위: %)

		SVM_CHI	SVM_IG	SVM_DF	SVM_TF-IDF
학습용 데이터	정확성	75.60	72.14	70.24	77.33
	재현성	77.23	71.37	73.04	76.30
	정밀성	72.56	73.87	64.09	79.22
검증용 데이터	정확성	**76.31**	**71.29**	**68.44**	**66.30**
	재현성	77.66	70.54	71.23	65.53
	정밀성	74.12	73.44	62.23	69.25



〈그림 4〉 영화리뷰에서 4품사를 사용한 감성 분류모형 결과 비교

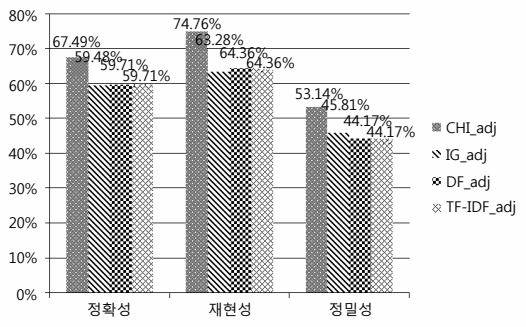
〈표 11〉 도서리뷰에서 형용사를 사용한 결과

(단위: %)

		SVM_CHI	SVM_IG	SVM_DF	SVM_TF-IDF
학습용 데이터	정확성	67.97	67.48	62.37	62.37
	재현성	75.61	73.78	67.49	67.49
	정밀성	52.96	54.16	47.58	47.58
검증용 데이터	정확성	**67.49**	**59.48**	**59.71**	**59.71**
	재현성	74.76	63.28	64.36	64.36
	정밀성	53.14	45.81	44.17	44.17

도서리뷰에서는 영화리뷰에서의 결과와 동일하게 카이제곱 통계량 기반의 감성분류모형에서 67.49%의 예측 정확성과 재현성, 정밀성 모두 가장 높게 나타났고, 정보획득량, 문서빈도, TF-IDF

기반의 감성분류모형 모두 60% 이하의 예측 정확성을 나타내고 있었다. 재현성과 정밀성 또한 영화리뷰와 동일하게 비슷한 결과로 나타났다.



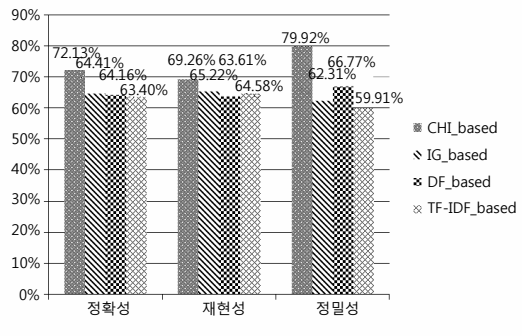
〈그림 5〉 도서리뷰에서 형용사를 사용한 감성 분류모형의 결과 비교

도서리뷰에서도 역시 <표 12>에서처럼 카이제곱 통계량 기반의 감성분류모형에서 72.13%의 예측정확성으로 가장 높게 나타났고, 정보획득량 기반의 감성분류모형에서의 예측 정확성은 64.41%, 문서빈도 기반에서는 64.16%의 예측 정확성으로 나타났으며, TF-IDF 기반의 감성분류모형에서는 63.40%의 예측 정확성으로 가장 낮게 나타난 것을 확인할 수 있었다. 도서리뷰 역시 형용사를 이용한 용어기반의 감성분류모형에 대한 실험결과보다 예측 정확성이 높게 측정된 것을 확인할 수 있었다.

〈표 12〉 도서리뷰문서에서 4품사를 사용한 결과

(단위: %)

		SVM_CHI	SVM_IG	SVM_DF	SVM_TF-IDF
학습용 데이터	정확성	71.84	71.80	65.93	64.20
	재현성	68.68	73.02	65.36	65.38
	정밀성	80.21	69.07	67.66	60.24
검증용 데이터	정확성	**72.13**	**64.41**	**64.16**	**63.40**
	재현성	69.26	65.22	63.61	64.58
	정밀성	79.92	62.31	66.77	59.91



〈그림 6〉 도서리뷰에서 4품사를 사용한 감성 분류모형 결과 비교

Ⅴ. 결론 및 연구의 한계

본 연구에서는 품사태깅을 이용하여 최적의 용어를 선정하는 과정을 통해 용어기반의 감성분류모형을 개발하였다. 영화리뷰에서 형용사에 해당하는 용어들을 입력변수로 사용한 실험결과에서 4품사를 이용한 예측 정확성에 대한 결과가 더 우수하다는 것을 알 수 있었다. 도서리뷰 역시 영화리뷰와 동일한 결과를 나타내고 있었다. 카이제곱 통계량 기반의 감성분류모형이 전체 실험결과에서 예측 정확성이 가장 우수하게 나타나는 것을 확인할 수 있었고, 재현성과 정밀성 또한 가장 높게 나타났다. 감성분류모형에서 예측 성과가 달라지는 것은 입력변수로 사용될 최적의 용어를 선정하는 과정에서 품사에 따라 용어의 영향력이 달라지기 때문이다.

본 연구의 실험을 통해 Hatzivassiloglou and McKeown(1997)의 단점을 보완하여 다양한 품사들로 입력변수의 우수성을 검증하였다. Moraes *et al.*(2013)의 많은 용어의 사용에 대한 단점을 보완하여 최적의 용어를 선정하여 문서빈도, TF-IDF, 정보획득량, 카이제곱 통계량 기반의 감성분류모형으로 비교하여 우수성을 검증하였다. 또한, 정보력이 높은 용어들에 대해 문서간의 연관성과 관계성이 높아 감성정보가 잘 반영되고 유의한 용어들의 사용여부에 대한 점도 최적의 용

어 선정을 통해 검증하였다.

본 실험에서 가장 우수하게 나타난 카이제곱 통계량 기반의 감성분류모형을 이용한다면 오피니언마이닝 연구에 좋은 예측성과를 낼 수 있을 뿐만 아니라 선별적인 용어추출을 이용한 연구의 실험과정을 통해 구축한 감성용어 데이터베이스를 재활용할 수 있다는 장점을 가진다. 또한, 본 연구에서 제안된 감성분류모형을 이용함으로써 온라인 상점에서의 서비스 개선과 품질 향상에 많은 도움이 될 것이다.

최적의 용어를 사용하여 감성분류모형에 대한 결과를 분석하였지만, 이를 통해 전체 리뷰를 평가하기에는 한계가 있다. 악의를 가지고 있는 리뷰나 잘못된 정보를 내포하고 있는 리뷰는 고객의 의도를 정확히 파악할 수 없기 때문에 감성분류를 위해 정확한 극성을 판별하기 어렵다는 것이다. 또한, 온라인상에서 자주 등장하는 신조어가 내포되어 있는 문서 역시 극성을 판별하기 어렵기 때문이다. 향후 연구에서는 문서에 대한 극성을 판별할 수 있는 방안을 모색하고, 다양한 기법들을 적용하여 예측성과를 향상시킬 수 있는 연구진행이 되어야 한다.

참 고 문 헌

강대국, 박용태, “리뷰 기반의 모바일 서비스 고객 요구사항 특성 분석”, 한국경영과학회 추계학술대회/방위사업청 무기체계 시험평가 세미나 논문집, 2012, pp. 945-951.

강범일, 송민, 조화순, “토픽 모델링을 이용한 신문 자료의 오피니언 마이닝에 관한 연구”, 한국문헌정보학회지, 제47권, 제4호, 2013, pp. 315-334.

김승우, 김남규, “오피니언 분류의 감성사전 활용 효과에 대한 연구”, 지능정보연구, 제20권, 제1호, 2014, pp. 133-148.

박경미, 박호건, 김형곤, 고희동, “SNS에서 오피니언마이닝 연구”, 정보과학회지, 제29권, 제11호, 2011, pp. 54-60.

임좌상, 김진만, “한국어 트위터의 감정 분류를 위한 기계학습의 실증적 비교”, 멀티미디어학회논문지, 제17권, 제2호, 2014, pp. 232-239.

장재영, “온라인 쇼핑몰의 상품평 자동분류를 위한 감성분석 알고리즘”, 한국전자거래학회지, 제14권, 제4호, 2009, pp. 19-32.

홍초희, 김학수, “트윗 감정 분류를 위한 다양한 기계학습 자질에 대한 비교 연구”, 한국콘텐츠학회논문지, 제12권, 제12호, 2012, pp. 471-478.

Abbasi, A., S. France, Z. Zhang, and H. Chen, “Selecting attributes for sentiment classification using feature relation networks”, *IEEE Transactions on Knowledge and Data Engineering*, Vol.23, 2011, pp. 447-462.

Chen, L., L. Qi, and F. Wang, “Comparison of feature-level learning methods for mining online consumer reviews”, *Expert Systems with Applications*, Vol.39, 2012, pp. 9588-9601.

Felipe, B., M, M. Mendoza, and B. Poblete, “Meta-level sentiment models for big social data analysis”, *Knowledge-Based Systems*, Vol.69, 2014, pp. 86-99.

Friedman, J. H., “On Bias, Variance, 0/1-Loss, and the Curse-of-Dimensionality”, *Data Mining and Knowledge Discovery*, Vol.1, 1997, pp. 55-77.

Geman, S., E. Bienenstock, and R. Doursat, “Neural Networks and the Bias/Variance Dilemma”, *Neural Computation*, Vol.4, 1992, pp. 1-58.

Han, J. and M. Kamber, Data Mining: Concepts and Techniques, 3rd edition, Morgan Kaufmann Publishers, 2011.

Hastie, T., R. Tibshirani, and J. H. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction (2nd ed.), New York: Springer, 2008.

Hatzivassiloglou, V. and K. McKeown, “Predicting

the semantic orientation of adjectives", *In Proceddings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL*, 1997, pp. 174-181.

Hu, M. and B. Liu, "Mining Opinion Features in Custmer Reviews", *Proceedings of the 19th national conference on Artifical intelligence*, 2004, pp. 755-760.

Hu, Y. and W. Li, "Document sentiment classification by exploring description model of topical terms", *Computer Speech and Language*, Vol. 25, 2011, pp. 386-403.

Li, S., R. Xia, C. Zong, and C. R. Huang, "A Framework of Feature Selection Methods for Text Categorization", *Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, 2009, pp. 692-700.

Liu, B., Sentiment Analysis and Opinion Mining, Morgan and Claypool Publishers, 2012.

Moraes, R., J. Valiati, and W. P. Gaviao Neto, "Document-level sentiment classification: An empirical comparison between SVM and ANN", *Expert Systems with Applications*, Vol.40, 2013, pp. 621-633.

O'Leary, D. E., "Blog mining-review and extensions: 'From each according to his opinion'", *Decision Support Systems*, Vol.51, 2011, pp. 821-830.

Pang, B. and L. Lee, "Opinion mining and sentiment analysis", *Foundations and Trends in Information Retrieval*, Vol.2, No.1/2, 2008, pp. 1-135.

Rao, Y., Q. Li, X. Mao, and L. Wenyin, "Sentiment topic models for social emotion mining", *Information Sciences*, Vol.266, 2014, pp. 90-100.

Shmueli, G. and O. Koppius, "Predictive Analytics in Information Systems Research", *MIS Quarterly*, Vol.35, No.3, 2011, pp. 553-572.

Shmueli, G., N. R. Patel, and P. C. Bruce, Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner, New Jersey, Wiley, Vol.2, 2010.

Silva, N., E. Hruschka, and E. Hruschka Jr., "Tweet sentiment analysis with classifier ensembles", *Decision Support Systems*, Vol.66, 2014, pp. 170-179.

Tang, H., S. Tan, and X. Cheng, "A survey on sentiment detection of reviews", *Expert Systems with Applications*, Vol.36, 2009, pp. 10760-10773.

Tay, F. E. H. and L. Cao, "Application of support vector machines in financial time series forecasting", *Omega*, Vol.29, 2001, pp. 309-317.

Vapnik, V., *The Nature of Statistical Learning Theory*, Springer, 1995.

Wang, G., J. Sun, J. Ma, K. Xu, and J. Gu, "Sentiment classification: The contribution of ensemble learning", *Decision Support Systems*, Vol. 57, 2014, pp. 77-93.

Xia, R., C. Zong, and S. Li, "Ensemble of feature sets and classification algorithms for sentiment classification", *Information Sciences*, Vol.181, 2011, pp. 1138-1152.

Yang, Y. and J. Pedersen, "A comparative study on feature selection in text categorization", *In Proceedings of ICML-97, the 14th International Conference on Machine Learning*, 1997, pp. 412-420.

Information Systems Review
Volume 17 Number 1
April 2015

Terms Based Sentiment Classification for Online Review Using Support Vector Machine

Taewon Lee* · Taeho Hong**

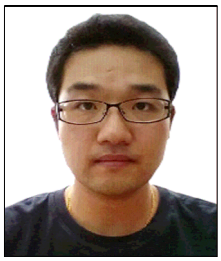
Abstract

Customer reviews which include subjective opinions for the product or service in online store have been generated rapidly and their influence on customers has become immense due to the widespread usage of SNS. In addition, a number of studies have focused on opinion mining to analyze the positive and negative opinions and get a better solution for customer support and sales. It is very important to select the key terms which reflected the customers' sentiment on the reviews for opinion mining. We proposed a document-level terms-based sentiment classification model by select in the optimal terms with part of speech tag. SVMs (Support vector machines) are utilized to build a predictor for opinion mining and we used the combination of POS tag and four terms extraction methods for the feature selection of SVM. To validate the proposed opinion mining model, we applied it to the customer reviews on Amazon. We eliminated the unmeaning terms known as the stopwords and extracted the useful terms by using part of speech tagging approach after crawling 80,000 reviews. The extracted terms gained from document frequency, TF-IDF, information gain, chi-squared statistic were ranked and 20 ranked terms were used to the feature of SVM model. Our experimental results show that the performance of SVM model with four POS tags is superior to the benchmarked model, which are built by extracting only adjective terms. In addition, the SVM model based on Chi-squared statistic for opinion mining shows the most superior performance among SVM models with 4 different kinds of terms extraction method. Our proposed opinion mining model is expected to improve customer service and gain competitive advantage in online store.

Keywords: Opinion Mining, Sentiment Classification, Support Vector Machine

* Doctoral Candidate, College of Business Administration, Pusan National University
** Professor, College of Business Administration, Pusan National University

◐ 저 자 소 개 ◑



이 태 원 (twanny@pusan.ac.kr)

동국대학교 컴퓨터학과에서 학사를 마쳤으며, 영남대학교 대학원 컴퓨터공학과에서 석사학위를 취득하였다. 현재 부산대학교 대학원 경영학과 박사과정에 재학중이며 경영정보 생산관리전공을 하고 있다. 주요 연구분야는 데이터마이닝, 오피니언마이닝, CRM, 소셜 네트워크 분석 등이다.

홍 태 호 (hongth@pusan.ac.kr)

현재 부산대학교 경영대학 교수로 재직하고 있다. KAIST에서 산업공학사를 취득하였고 경영정보시스템을 전공하여 공학석사와 박사를 취득하였다. 딜로이트 컨설팅에서 컨설턴트로 재직했으며, 주요 관심분야는 CRM, Data Mining, Business Intelligence, 지식경영 등이다. 주요 논문을 *Expert Systems*, *Expert Systems with Applications*, *Asia Pacific Journal of Information Systems*, 정보시스템연구 등에 게재하였다.

논문접수일 : 2014년 10월 15일 **게재확정일 : 2015년 02월 23일**
1차 수정일 : 2015년 01월 18일