효과적인 상품평 감정 분류를 위한 어휘 자질의 순차적 사용 방법

신준수○, 김학수
강원대학교 컴퓨터정보통신전공
nlpsjs@kangwon.ac.kr, nlpdrkim@kangwon.ac.kr

A method to sequentially use lexical features for effective sentiment categorization of Korean Customer Reviews

Jun-Soo Shin○, Harksoo Kim
Program of Computer and Communications Engineering,
Kangwon National University

요 약

인터넷이 크게 발전하면서 현재는 인터넷으로 쉽게 쇼핑을 할 수 있다. 이 때 물건의 구입에 큰 영향력을 미치는 것이 바로 그 물건의 상품평이다. 하지만 실제로 수많은 상품평을 사용자가 일일이 확인하고 판단하는 데에는 많은 시간이 소모된다. 이러한 문제점을 해결하기 위해서 본 논문에서는 상품평 문장을 일반, 긍정, 부정의 세 단계로 나누는 시스템을 제안한다. 감정을 판단하는데 중요한 역할을 하는 품사에 따라 우선순위를 달리하여 자질을 추출한다. 추출된 자질을 사용하여 Paul Graham을 사용하여 가중치를 계산하고 기계학습을 한다. 실험은 일반과 감정(긍정, 부정)으로 분류하는 실험과 긍정과 부정으로 분류하는 실험을 하였다. 실험 결과 품사에 우선순위를 사용하여 만든 시스템이 기본 시스템보다 더 적은 자질을 사용하고 더 높은 성능을 보였다.

주제어: 감정 분류, 상품평, 문서 분류

1. 서 론

인터넷 쇼핑몰에서 상품평은 사용자의 구매의사 결정에 큰 역할을 한다. 인터넷 쇼핑몰을 이용하는 사용자들이 증가함에 따라 작성되는 상품평도 계속해서 증가하고 있다. 이렇게 작성되는 수많은 상품평을 개인 사용자가 일일이 다 읽어보는데는 상당한 시간이 소비된다. 제품을 생산한 기업의 입장에서도 자사가 출시한 제품에 대한 소비자의 생각을 쉽게 파악할 수 있는 방법 중 하나가 상품평을 읽는 것이다. 하지만 상품평을 일일이 확인하기에는 많은 어려움이 있다. 인터넷 쇼핑몰에서는 좋은 상품들을 사용자에게 추천해주기 위해 신뢰도나 만족도 등의 조사 결과를 사용하고 있다 그러나 이러한 조사 결과를 실제 상품평의 내용과 맞지 않는 경우도 상당 수 존재하고 있다. 그렇기 때문에 상품평에 대한 신뢰도와 만족도는 조사 결과가 아닌 실제 상품평의 내용에 기반을 두어야 한다. 상품평의 특징을 살펴보면 긍정적인 문장과 부정적인 문장 그리고 일반적인 문장의 크게 세 가지의 범주로 나타난다. 그리고 세 가지 범주의 문장들이 하나의 상품평에 나타나기 때문에 상품평 자체를 하나의 범주에 할당하는 것은 바람직하지 않다.

본 논문에서는 이러한 문제점들을 해결하기 위해서 상품평의 문장에 대해서 감정을 분류하는 시스템을 구현한다. 기존의 연구에서는 문장에서 나타나는 용언과 체언을 모두 자질로 선택한다. 이러한 방법은 감정 분석의 관점에서 의미 없는 단어가 한 범주에서 집중적으로 나오게 되면 감정 분석에서 의미가 없는 단어임에도 불구하고 감정을 판단하는 중요한 자질로 작용한다 예를 들어 “배송이 빨라요” 와 같은 문장과 “고장이 빨리나요” 와 같은 문장이 있다. “배송이 빨라요” 와 같이 배송에 관한 긍정적인 문장의 빈도수가 높으면 ‘배송’ 이 긍정에서 중요한 자질로 사용이 된다 그러나 실제로는 ‘배송’ 과 ‘빠르’ 가 함께 출현하였기 때문에 전체적으로 긍정인 문장으로 판단된다. 반면에 “고장이 빨리나요” 와 같이 ‘고장’ 과 ‘빠르’ 가 함께 나타난다. 이 경우에는 ‘고장’ 이 문장을 부정으로 판단하는데 중요한 자질처럼 보이지만 실제로는 ‘고장’ 과 ‘빠르’ 가 함께 나타나기 때문에 부정으로 판단된다. ‘고장’ 과 ‘빠르’ 의 경우에는 ‘빠르’ 가 부정적인 의미로 사용되었고, ‘배송’ 과 ‘빠르’ 의 경우에는 ‘빠르’ 가 긍정적인 의미로 사용되었다. 즉, 어느 한 가지 형태소도 주변의 단어들로 인해서 그 의미가 달라질 수 있다. 이러한 문제점을 해결하기 위해서 본 논문에서는 자질 추출 단계에서 감정을 판단하는데 중요한 자질인 용언과 용언의 주변 체언들을 함께 사용한다 또한 용언이 없는 문장의 경우에는 체언만을 사용한다 예를 들어 “배송이 빨라요” 와 같은 문장에서는 ‘배송’ 과 ‘빠르’ 를 자질로 추출한다. 용언이 없는 문장인 “매우 친절하네요.” 와 같은 문장에서는 ‘친절’ 만을 자질로 추출한다. 이렇게 추출된 자질의 가중치를 계산하여 기계학습을 한다 “어제 주문했는데 배송도 빠르고 매우 친절하시네요” 와 같은 실험 문장에서는 앞에서 학습을 위해 추출한 자질인 ‘배송’, ‘빠르’, ‘친절’ 만을 사용하여 실험 문장의 감정 범주를

판단하게 된다.

본 논문의 구성은 다음과 같다. 2절에서 감정 분류와 관련된 연구를 소개하고 3절에서는 본 논문에서 제안하는 자질 추출 방법을 사용한 시스템에 대해서 소개하고 4절에서는 제안한 시스템의 성능을 살펴본다. 마지막으로 결론 및 향후과제를 기술한다.

2. 관련연구

상품평의 감정을 분석하는 연구는 문서를 분류하는 방법과 크게 다르지 않다. 형태소 분석기를 사용하여 자질을 추출하고 선택하여 Naive Bayes, Machine Learning 등을 이용한다. 이 때 사용되는 자질 추출 방법과, 자질 선택 방법, 분류 방법 등에 따라서 시스템의 성능이 크게 좌우된다.

Topic Signature를 이용한 댓글 분류 시스템에서 댓글은 띄어쓰기와 정확한 단어의 사용이 이루어지지 않는다는 특징이 있기 때문에 자질 추출 방법에서 음절 n-gram 방법을 사용한다. 자질 선택 방법에서는 Topic Signature를 사용하여 일정 확률 이상의 자질만을 사용한다.[3] 악성 댓글을 판단하는 자질과 상품평의 감정을 판단하는 자질은 차이가 있기 때문에 실질적으로 상품평의 감정 시스템에 적용할 수 없다.

감정 자질을 이용한 한국어 문장 및 문서 감정 분류 연구에서는 영어단어 시소러스 정보를 이용하여 감정을 나타내는 자질들을 사용한다. 이 때 자질의 가중치 계산은 TF-ISF 기법을 사용한다.[4] 영어단어 시소러스의 정보를 사용하였기 때문에 상품평의 감정 분류 시스템에서는 이 자질을 그대로 사용 할 수 없다. 예를 들어 “배송이 빨라요” 와 같은 문장에서 ‘배송’ 과 ‘빠르’ 를 보고 긍정으로 판단하여야 하지만 영어 단어 시소러스에서는 ‘배송’, ‘빠르’ 에 대한 감정 자질을 출현하지 않는다.

이러한 이유에서 기존의 연구들은 상품평의 감정 분석 시스템에 적용할 수 없다. 본 논문에서는 상품평의 감정 판단에 중요한 자질을 추출하는 방법을 제시하고 추출된 자질의 가중치를 계산하는 방법을 제시한다.

3. 상품평 분석 시스템

상품평을 분석하는 시스템의 순서도는 그림1과 같다.

본 논문에서 제안한 시스템은 크게 학습 과정과 감정 분류 과정으로 나뉜다. 학습 과정은 다음과 같다. 형태소 분석 단계를 거친 후에 자질을 추출한다. 추출된 자질로 테이블을 생성하여 가중치를 계산한다. 다시 형태소 분석된 문장에서 체언과 용언을 추출하고 테이블 내에 존재하는 자질들만을 사용하여 지지 벡터 기계 학습을 하여 모델을 생성한다. 감정 분류 과정에서는 형태소 분석

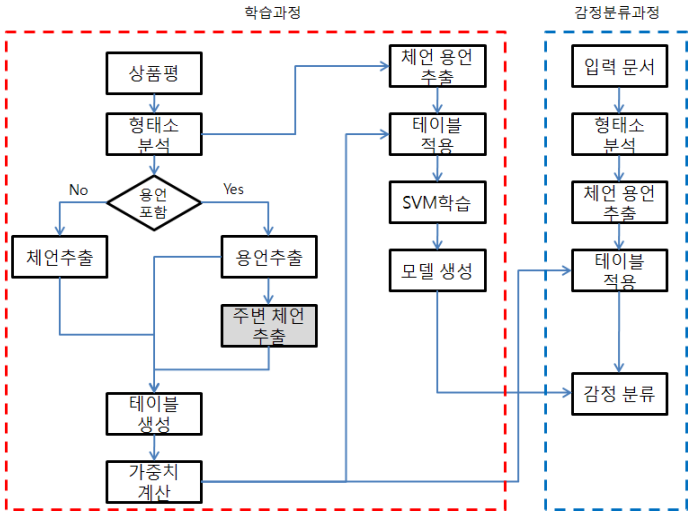

그림 1. 제안 시스템의 전체 순서도

단계에서 자질을 추출하고 테이블 내에 존재하는 자질만을 사용하여 문장을 분류한다.

3.1 자질 추출

형태소 분석 단계에서 자질을 추출하여 테이블을 생성한다. 이 때 용언과 체언을 동시에 모두 추출하지 않는다. 문장 내에 용언이 포함되어 있는 경우에는 용언과 용언의 주변 체언을 추출하여 자질로 사용하고 용언이 포함되어 있지 않은 문장에서는 체언만을 추출하여 자질로 사용한다.

휴일에 주문했는데 배송이 빠르네요.
생각보다 고장이 빨리나요.
배송에 감사드립니다.

그림 2. 세탁기 상품평의 예

위의 그림 2와 같은 상품평에서 용언과 체언을 동시에 추출하게 되면 ‘휴일’, ‘주문’, ‘생각’ 등과 같이 상대적으로 ‘배송’, ‘빠르’, ‘감사’, ‘고장’ 에 비해서 감정을 판단하는데 중요하지 않은 자질도 추출된다.

본 논문에서 제시한 자질 추출 방법을 사용하게 되면 첫 번째 문장에서 ‘배송’, ‘빠르’ 가 추출되고 두 번째 문장에서는 ‘고장’, ‘빠르’ 가 추출되고, 세 번째 문장에서 ‘배송’, ‘감사’ 가 추출된다. 본 논문에서 제시하는 자질 추출 방법을 사용하게 되면 앞의 모든 용언과 체언을 동시에 추출하는 방법에 비해서 ‘휴일’, ‘주문’, ‘생각’ 등의 중요도가 상대적으로 작아지거나 아예 자질로 추출되지 않으며, ‘배송’, ‘빠르’, ‘고장’, ‘감사’ 의 자질 중요도가 더 커진다.

그림 3의 다른 세탁기 상품평의 경우, 용언의 주변 체언을 포함하지 않고 용언만 추출하는 방법을 사용했을 때 첫 번째 문장에서 ‘해주시’ 라는 용언이 추출된다. ‘해주시’ 는 감정을 판단하는데 중요한 자질이 아니다. ‘친절’ 이 문장을 긍정으로 판단하는데 중요한 자질로 사용된다. 이 때 본 논문에서 제시한 용언의 주변 체언 정보를

함께 이용하게 되면 첫 번째 문장에서 ‘친절’, ‘해주시’의 자질이 사용된다. 두 번째 문장에서는 먼저 ‘좋’이 추출된다. ‘좋’은 감정을 판단하는데 중요한 자질로 사용된다. 그러나 ‘저렴’이라는 체언 역시 중요한 자질이다. 이 문장 역시 용언의 주변 체언을 함께 추출하는 방법을 사용하게 되면 ‘저렴’, ‘좋’이 추출된다.

배송도 친절하게 잘 해주셨고요. 배송도 친절하고 무엇보다 물건도 저렴해서 좋고요.

그림3. 세탁기 상품평의 예

위와 같은 이유에서 먼저 용언을 추출하고 용언의 주변 체언을 함께 추출하여 자질로 사용한다. 아래 [표 1]은 그림 2와 그림 3의 세탁기 상품평의 예에서 용언과 체언을 동시에 추출하는 방법의 자질 테이블이고 [표 2]는 본 논문에서 제안하는 자질 추출 방법을 사용하였을 때 그림 2와 그림 3의 세탁기 상품평의 예에서 추출되는 자질 테이블이다.

표 1. 용언과 체언을 모두 추출

자질	빈도수
휴일	1
주문	1
배송	4
빠르	2
생각	1
고장	1
감사	1
친절	2
해주시	1
물건	1
저렴	1
좋	1

표2. 본 논문에서 제안하는 방법

자질	빈도수
배송	2
빠르	2
고장	1
감사	1
친절	1
해주시	1
저렴	1
좋	1

3.2 자질 가중치 계산

본 논문에서는 추출된 자질의 가중치를 계산하는 방법으로 식 (1)을 사용한다.[9] $W(w)$는 자질의 가중치를 나타낸다. tf_1과 tf_2는 각 범주에서 자질 w의 빈도수를 나타내고, C_1과 C_2는 각 범주의 총 자질의 수를 나타낸다. 자질 w의 특정 범주 내에서의 빈도수 tf는 특정 범주의 총 자질의 수로 정규화하여 확률 값을 계산한다.

$$W(w) = \frac{\frac{tf_1}{C_1}}{\frac{tf_1}{C_1} + \frac{tf_2}{C_2}} \quad (1)$$

3.3 SVM 학습

SVM은 이진 분류에서 높은 성능을 보이고 있는 기계학습 방법이다. 본 논문에서는 먼저 감정(긍정, 부정)과 일반 문장을 분류하는 시스템을 만들고, 긍정과 부정 문장을 분류하는 시스템을 만든다. 두 시스템 모두 이진 분류로 볼 수 있기 때문에 SVM을 사용하여 모델을 생성하고 문장의 감정을 분류한다

식 (1)을 사용하여 자질의 가중치를 계산하고 테이블을 만든다. 그리고 테이블에 있는 자질들만을 사용하여 학습을 한다. 예를 들어 그림 3의 두 번째 문장인 “배송도 친절하고 무엇보다 물건도 저렴해서 좋고요”에서 용언과 체언을 모두 추출한다 ‘배송’, ‘친절’, ‘무엇’, ‘물건’, ‘저렴’, ‘좋’이 추출되고 이 중 [표 2]의 테이블에 있는 자질 ‘배송’, ‘친절’, ‘저렴’, ‘좋’만을 사용하여 학습을 한다.

4. 실험

4.1 실험 방법

실험에 사용된 문장은 가격비교 사이트에서 수집하였으며 철자 오류 및 띄어쓰기를 수작업으로 보정하였다[8] 수집 된 문장은 긍정 2585문장과 부정 679문장 일반 840문장의 총 4104 문장이다.

실험 문장에서 학습 단계와 같은 방법으로 모든 용언과 체언을 추출한다. 추출된 용언과 체언 중 테이블에 존재하는 자질만을 사용하여 SVM 모델을 적용하고 분류한다. 실험은 10-fold cross validation 방법을 사용하였으며 성능 평가 방법으로는 정보 검색 분야에서 일반적으로 사용되는 정확율, 재현율을 측정하고 이를 하나의 값으로 표현하는 $F_1 - measure$를 사용하였다. 정확율, 재현율, $F_1 - measure$에 대한 식은 (2), (3), (4)와 같다.

$$\text{정확율} = \frac{\text{제대로 판단된 감정 수}}{\text{시스템이 판단한 감정 수}} \quad (2)$$

$$\text{재현율} = \frac{\text{제대로 판단된 감정 수}}{\text{정답 감정의 수}} \quad (3)$$

$$F_1 - measure = \frac{2 \times 재현율 \times 정확율}{재현율+정확율} \quad (4)$$

4.2 실험 결과

[표 3]은 용언과 체언을 동시에 모두 자질로 사용하는 시스템(1), 용언을 추출하고 용언이 없는 경우에 체언을 추출하는 시스템(2), 본 논문에서 제안한 용언을 추출하고 용언이 있는 경우에 주변의 체언을 함께 추출하고 용언이 없는 경우에는 체언을 추출하는 시스템(3)의 $F_1 - measure$ 실험 결과이다.

표3 실험 결과 (단위 : %)

시스템	자질의 수	일반/감정 분류		부정/긍정 분류	
(1)	4343	일반	71.2	부정	83.64
		감정	67.18	긍정	80.83
		평균	69.19	평균	82.24
(2)	1607	일반	68.34	부정	81.95
		감정	67.36	긍정	81.24
		평균	67.85	평균	81.6
(3)	3359	일반	71.1	부정	84.21
		감정	68.08	긍정	82.58
		평균	69.59	평균	83.4

실험 결과에서 알 수 있듯이 본 논문에서 제안한 시스템(3)이 용언과 체언을 동시에 모두 자질로 사용하는 시스템의 77.34%의 자질만 사용하고 평균적으로 더 높은 성능을 보였다. 일반과 감정을 분류하는 실험에서는 $F_1 - measure$ 값이 0.4%로 큰 차이는 없었지만 부정과 긍정을 분류하는 실험에서는 1.16%의 차이를 보였다. 용언을 추출하고 용언이 없는 경우에 체언을 추출하는 시스템(2)은 용언과 체언을 동시에 모두 자질로 사용하는 시스템의 37.02%의 자질만을 사용하고도 성능차이가 크게 나지 않았다.

5. 결론 및 향후과제

본 연구에서는 자질 추출 단계에서 용언에 우선순위를 두고 용언 주변의 체언들을 함께 포함하여 자질을 추출하는 방법을 제안하였다. 일반적으로 사용되는 용언과 체언을 동시에 자질로 사용하는 방법보다 더 적은 자질을 사용하고 더 높은 성능을 보이는 것을 알 수 있었다. 자질을 추출하는데 우선순위를 두고 추출된 자질의 주변 정보들을 포함하게 되면 더 높은 성능을 낸다는 것을 확인할 수 있었다. 실험 데이터의 양이 적어서 처리 속도 및 용량에 대해서는 실험을 하지 못하였지만 데이터가 많아질수록 속도 및 용량의 차이가 날 것으로 예상된다. 본 논문에서는 철자 오류와 띄어쓰기를 수작업으로 수정하였지만 앞으로 철자 오류 및 띄어쓰기에 영향을 받지 않고 자질을 추출할 수 있는 연구를 진행할 예정이다. 또한 본 논문에서 제시한 자질 추출 방법을 응용하여 문서 분류, 악성 댓글 분류, 문서 요약 등에도 적용 할 계획이다. 그리고 본 연구에서는 상품평을 분석하는데 용언의 주변 체언만을 살펴봤는데 추가적으로 용언을 기준으로 슬라이딩 윈도우를 사용하는 방법을 적용해보고 이를 실험할 예정이다. ㅂ3

참고문헌

[1] 명재석, 이동주, 이상구, "반자동으로 구축 된 의미 사전을 이용한 한국어 상품평 분석 시스템", 정보과학회 논문지 제 35권 제 6 호, 2008

[2] 배원식, 한요섭, 차정원, "Topic Signature와 동시 출현 단어 쌍을 이용한 문서 범주화",한국컴퓨터종합학술대회 논문집, 2008

[3] 배민영, 차정원 "Topic signature와 n-gram을 이용한 댓글 분류 시스템", 한글 및 한국어 정보처리 학술대회, 2008

[4] 황재원, 고영중, "감정 자질을 이용한 한국어 문장 및 문서 감정 분류 시스템", 정보과학회논문지 제 14권 제 3호, 2008

[5] 신준수, 이주호, 김학수, "CRFs를 이용한 한국어 상품평의 감정 분류". 한글 및 한국어 정보처리 학술대회, 2008

[6] 고영중, 박진우, 서정연, "문장 중요도를 이용한 자동 문서 범주화", 정보과학회논문지 제 29권 제 6호, 2002

[7] Anindya Chose, Panagiotis G. Iperotis and Arun Sundararajan, "Opinion Mining Using Econometrics: A Case Study on Reputation System", Proceedings of ACL2007

[8] 비비, http://www.bb.co.kr

[9] Paul Graham Homepage, www.paulgraham.com