ISSN 2383-630X(Print) / ISSN 2383-6296(Online)
Journal of KIISE, Vol. 42, No. 4, pp. 512-521, 2015. 4
http://dx.doi.org/10.5626/JOK.2015.42.4.512

인터넷 감정기호를 이용한 긍정/부정 말뭉치 구축 및 감정분류 자동화

(Automatic Construction of a Negative/positive Corpus and Emotional Classification using the Internet Emotional Sign)

장 경 애 † (Kyoungae Jang)　박 상 현 †† (Sanghyun Park)　김 우 제 ††† (Woo-Je Kim)

요 약 네티즌은 인터넷을 통해서 상품을 구매하고 상품에 대한 감정을 긍정 혹은 부정으로 상품평에 표현한다. 상품평에 대한 분석은 잠재적 소비자뿐만 아니라 기업의 의사결정에 중요한 자료가 된다. 따라서 인터넷의 대량 리뷰에서 의미 있는 정보를 분석하여 의견을 도출하는 오피니언 마이닝 기술의 중요성이 증대되고 있다. 기존의 연구는 대부분이 영어를 기반으로 진행되었고 아직 한글에 대한 상품평 분석은 활발히 이루어 지지 않고 있다. 또한 한글은 영어와 달라 꾸미는 말과 어미가 복잡한 특성을 갖고 있다. 그리고 기존의 연구는 통계적 기법, 사전 기법, 기계학습 기법 등을 사용하여 연구되었으나 인터넷 언어의 특성을 감안하지는 못하였다. 본 연구에서는 감정이 포함된 인터넷 언어의 특성을 분석하여 감정분석의 정확률을 높이는 감정분류 방법을 제안한다. 이를 통해 데이터에 독립적인 인터넷 감정기호를 이용해서 자동으로 긍정 및 부정 상품평을 분류할 수 있었고 높은 정확률, 재현율, Coverage 결과를 통해서 제안 알고리즘의 유효성을 확인할 수 있었다.

키워드: 오피니언 마이닝, 텍스트 마이닝, 상품평, 기계학습, 분류기법, 긍정/부정 단어사전

Abstract Internet users purchase goods on the Internet and express their positive or negative emotions of the goods in product reviews. Analysis of the product reviews become critical data to both potential consumers and to the decision making of enterprises. Therefore, the importance of opinion mining techniques which derive opinions by analyzing meaningful data from large numbers of Internet reviews. Existing studies were mostly based on comments written in English, yet analysis in Korean has not actively been done. Unlike English, Korean has characteristics of complex adjectives and suffixes. Existing studies did not consider the characteristics of the Internet language. This study proposes an emotional classification method which increases the accuracy of emotional classification by analyzing the characteristics of the Internet language connoting feelings. We can classify positive and negative comments about products automatically using the Internet emoticon. Also we can check the validity of the proposed algorithm through the result of high precision, recall and coverage for the evaluation of this method.

Keywords: opinion mining, text mining, product review, machine learning, Classification, positive/negative dictionary

† 정 회 원 : 서울과학기술대학교 IT정책대학원 산업정보시스템
jkalove@hanmail.net
†† 종신회원 : 연세대학교 컴퓨터과학과 교수
sanghyun@cs.yonsei.ac.kr
††† 비 회 원 : 서울과학기술대학교 글로벌융합산업공학과 교수
wjkim@seoultech.ac.kr
(Corresponding author임)
논문접수 : 2014년 10월 13일
(Received 13 October 2014)
논문수정 : 2015년 1월 1일
(Revised 1 January 2015)
심사완료 : 2015년 2월 11일
(Accepted 11 February 2015)

Copyright©2015 한국정보과학회 : 개인 목적이나 교육 목적인 경우, 이 저작물의 전체 또는 일부에 대한 복사본 혹은 디지털 사본의 제작을 허가합니다. 이 때, 사본은 상업적 수단으로 사용할 수 없으며 첫 페이지에 본 문구와 출처를 반드시 명시해야 합니다. 이 외의 목적으로 복제, 배포, 출판, 전송 등 모든 유형의 사용행위를 하는 경우에 대하여는 사전에 허가를 얻고 비용을 지불해야 합니다.
정보과학회논문지 제42권 제4호(2015. 4)

1. 서 론

1.1 연구 배경 및 목적

인터넷을 통한 전자상거래는 최근 스마트 디바이스의 확산으로 시공간의 제약 없이 인터넷을 이용할 수 있게 되면서 더욱 활발하게 이루어지고 있다. 방송통신위원회와 한국인터넷진흥원(KISA)이 발표한 '2012년 인터넷이용실태조사[1]'에 따르면 만3세 이상 인구의 인터넷이용률은 78.4%로, 인터넷 이용자 수는 전년보다 94만 명 증가한 3천 8백 12만 명으로 조사되었다. 인터넷을 이용하는 용도로는 '자료 및 정보 획득'이 92.3%로 가장 높았고, '상품이나 서비스를 구매 및 판매'하는 경우도 57.7%로 절반 이상으로 높게 조사되었다[1].

인터넷을 이용한 전자상거래는 일반적인 일이 되었고, 인터넷을 통해 공유되는 상품평은 소비자의 구매활동에 직접적인 영향을 미치게 되었다. 동일한 상품에 대해서 상품평을 읽은 소비자들은 읽기 전보다 상품의 신뢰, 기대치, 구매의도 등이 변화되어 실제 구매활동에 영향을 받는다것이 연구결과 밝혀졌다[2].

상품평에 대한 분석은 잠재적 소비자의 행동뿐만 아니라 기업의 의사결정에 중요한 자료가 된다. 그러나 지속적으로 늘어나고 다양한 사이트로 분산되어 있는 방대한 상품평 및 리뷰 데이터를 수작업으로 분석하기는 어렵다. 따라서 인터넷의 대량 리뷰에서 유의미한 정보를 분석하여 의견을 유추해주는 오피니언 마이닝(Opinion Mining) 기술의 중요성은 증대되고 있다.

기존 연구의 대부분이 영어를 기반한 연구로 진행되어 아직 한글에 대한 상품평 분석은 활발히 이루어지고 있지 않았으며 활용할 만한 자료가 존재하지 않았다[4,5]. 한글은 꾸미는 말과 어미가 다양하여 영어와 달리 분석이 어려우며, 네티즌은 인터넷쇼핑몰이나 SNS공간에서 정제된 한글이 아니라 인터넷 언어를 사용하고 있다. 인터넷 언어를 고려하지 않은 상품평 분석은 정확도가 떨어지고[6,11], 기존 구매자의 정확한 의견을 파악할 수 없다. 따라서 본 논문에서는 네티즌이 사용하는 인터넷 언어의 특성을 분석하고 상품평에 잠재된 의미를 파악하여 감정분석의 정확도를 높이고 감정분류의 성능을 향상시키는 방법을 제안한다.

1.2 연구 범위 및 방법

한국 네티즌이 감정을 표현하는 이모티콘과 한글의 초성, 특수기호로 인터넷 감정기호사전을 구축하고 이를 통해 긍정/부정 말뭉치와 상품평의 감정분류를 자동으로 수행할 수 있는 분류방법을 제안한다. 분석데이터는 노트북, MP3, 모니터의 상품평을 수집하고 분석하였다. 그 이유는 20대와 30대가 인터넷 상품구매 및 판매의 비율이 가장 높은 연령대이므로, 이 연령대의 관심 영역에 해당하는 최신기기를 선정하였다. 연구를 위하여 Python 2.7, 루씬 형태소분석기, Weka 3.6, Java와 Visual Basic을 이용한 메크로를 구현하여 실험을 진행하였다.

본 연구는 총 5장으로 구성되며 1장은 연구의 배경 및 목적을 소개하고, 2장에서는 관련 선행연구를 검토하여 문제점 및 개선방향을 도출하고, 3장에서는 본 연구 방법을 설계하고, 4장에서는 연구의 결과를 제시한다. 5장에서는 결론과 향후 추가 연구과제에 대해서 논의한다.

2. 이론적 배경과 선행연구의 고찰

2.1 오피니언 마이닝

오피니언 마이닝은 텍스트 데이터에서 긍정(positive), 부정(negative)의 의견을 판단하고 활용하는 목적으로 사용된다. 네티즌이 인터넷 SNS, 전자상거래 등을 통해서 인터넷에 댓글, 리뷰, 상품평을 남기면서 대량의 인터넷데이터에 숨은 감정을 분석하기위해 오피니언 마이닝이 활용되고 있다. 또한 전자상거래 분야에서는 소비자가 상품을 눈으로 직접 확인할 수 없는 특성 때문에 상품평 감정 분석에 대한 연구로 오피니언 마이닝을 활용하기 시작했다. 오피니언 마이닝은 게시글이나 특정 주제에 따른 네티즌의 특성을 파악하고 문장에 잠재한 의미를 도출하여 긍정인지 부정인지 감정을 파악하는 연구분야이다.

오피니언 마이닝은 일반적으로 데이터를 수집하여 특징을 추출하고, 감정을 분석하고 요약 및 표현하는 과정으로 진행된다[4-7]. 먼저 특징 추출과정에서는 분석하고자 하는 상품의 특징에 해당하는 속성정보와 감정정보를 추출한다. 수집된 데이터의 전처리 과정을 거치고 문장의 형태소를 분석하여 파싱한다. 전처리 과정에서 객관적인 문장과 주관적인 문장을 분리하여 객관적인 문장은 분석에서 제외한다. 그리고 감정 분석과정에서 단어 및 문장의 감정을 통계기법, 기계학습, 자연어처리 등 다양한 기법을 활용하여 긍정과 부정으로 분류한다.

본 논문은 특징추출과 감정분석 단계를 인터넷 감정기호사전 구축단계, 긍정/부정 말뭉치 구축 단계, 감정기호와 긍정/부정 말뭉치를 통한 극성 통합분류 단계로 전체 3단계 사이클(Cycle)을 구성하여 네티즌 상품평의 정확도를 높이는 방법을 제안한다.

2.2 선행연구 분석

사용자의 상품이 긍정적인지 부정적인지 판단하는 극성판단을 위한 오피니언 마이닝에 사용되는 기법은 자연언어처리 기법과 정량적 실험기반으로 나눌 수 있다. 자연언어 처리 기법은 어휘에서 품사의 관계 분석 및 품사의 의미방향 등을 통해 극성을 판단한다[5,11,14,15]. [5]에서는 자연어처리 기법을 활용하여 후보어휘를 추천하여 의미사전을 구축하였으며, [11]에서는 형용사, 동사,

부사의 각 문장에서 성질과 방향을 이용해서 극성을 판단하는 알고리즘을 제안하였다. 또한 기 구축된 긍정/부정 말뭉치 사전을 활용하거나 외국의 경우는 워드넷(WordNet)을 활용하여 연구도 시도되었다[17,18]. 이 연구에서는 워드넷을 활용하여 유의어와 반의어에 의한 분석으로 긍정, 부정의 어휘를 파악하고 센티 워드넷(SentiWordNet)으로 감정을 수치화하는 작업을 수행하였다.

최근에는 데이터의 정량적인 실험을 통한 통계기반 기법과 기계학습법을 통한 극성을 분류하는 연구가 진행되고 있다. 통계기반 기법의 대표적인 방법으로 단어의 출현빈도에 따른 PMI(Pointwise Mutual Information) 방법을 사용한다[7,19]. SVM(Support Vector Machine), CRFs(Conditional Random Fields), 네이브베이즈 등을 활용한 기계학습 기법으로 극성을 분류하기도 한다[16, 20,21,24,25]. 또한 자연어처리 기법과 통계적, 기계학습법을 혼합하여 분석을 진행하기도 한다[6,13,16,21].

선행연구방법을 살펴보면, 감정 어휘를 효율적으로 추출하기 위해서는 초/중성 음운패턴을 사용하거나[6], 한글어 구문 패턴을 정의하여 감정단어를 추출하는 방법[13]이 있었다. 또한 상품리뷰의 순위를 지정하는 방법으로는 출현빈도, 인트로피, 근접도 등의 알고리즘을 활용한 연구[22], 의미사전을 구축하고 가중치를 부여하여 순위를 지정한 연구[5] 등이 있었다. 그리고 분석된 어휘를 긍정/부정 감정으로 분류하는 사전을 자동으로 구축하는 방법을 제안한 연구도 시도되었다[5,7,8].

2.3 선행연구와의 차별성

그러나 오피니언 마이닝을 통한 상품평 감정분석에 대한 선행연구에는 한계점이 존재한다. 첫째, 전문가들이 수작업으로 긍정/부정 말뭉치를 구축하였는데, 이는 정확도는 향상될 수 있으나 샘플링에 의한 분석이 아니라 대량의 의견분석에서는 상당한 시간과 노력이 소요되므로 어려운 작업이 된다. 이를 위하여 별점 및 리뷰점수로 분류하는 경우가 존재하였으나 다각적인 연구는 부족하였다.

둘째, 대부분 영어를 기반으로 오피니언 마이닝이 연구가 선행되었다. 한국어와 영어의 문장구조가 상이하여 한국어를 영어로 번역한다고 하더라도 그 의미가 똑같지 않아 선행연구를 활용하기 어려우며, 워드넷(WordNet)은 통한 연구 또한 한글에 그대로 적용하기는 한계가 존재한다.

셋째, 기계학습 기법을 적용한 연구는 학습한 어휘가 동일한 도메인에서는 좋은 결과를 나타낼 수 있으나 다른 도메인에서는 상이한 결과가 나타날 수 있어 활용도가 저하된다. 따라서 데이터분석에서 도메인, 토픽, 시간의 의존성을 줄이는 노력이 필요하다[20]. 일부에서 데이터 도메인의 의존성을 줄이기 위해 공통 도메인단어를 분류하여 활용하였으나 복잡도가 증가하였다[7,8,15].

따라서 본 연구는 인터넷 언어에서 즐거움, 행복, 슬픔, 분노, 실망 등의 감정을 추출하여 극성분석에 효율성을 높이고자 하였다. 이를 위하여 인터넷 언어에서 감정이 표현되는 이모티콘과 초성을 기반으로 긍정/부정 인터넷 감정기호사전을 구축하고 이를 통해서 감정기호가 포함된 상품평의 극성을 분류한다. 또한 분류된 극성 상품평을 활용하여 긍정/부정 말뭉치를 자동 구축하고 이를 활용하여 전체 상품평의 극성을 분류하는 방법으로 상품평의 극성분석의 정확도 및 커버리지를 높였다.

3. 연구설계

3.1 제안 메소드의 개요

본 연구에서는 상품평의 감정을 분석하여 극성을 판별하는 3단계 방법을 제안한다. 1단계는 인터넷 감정언어를 기반으로 상품평의 제한 범위에서 극성분류를 하는 단계이다. 네티즌이 감정을 표현하는 이모티콘, 한글초성, 특수기호를 분석하고 긍정/부정으로 분류하여 인터넷 감정기호사전을 구축한다. 이를 기반으로 감정기호가 포함된 상품평의 극성분류를 수행한다. 2단계는 1단계에 도출된 감정분류 상품평을 기준으로 긍정/부정 단어를 도출해서 자동으로 말뭉치를 구축하는 단계이다. 1단계에서 분류된 긍정/부정 상품평 그룹을 기반으로 형태소 분석을 통해 속성단어와 감정단어를 도출하여 한

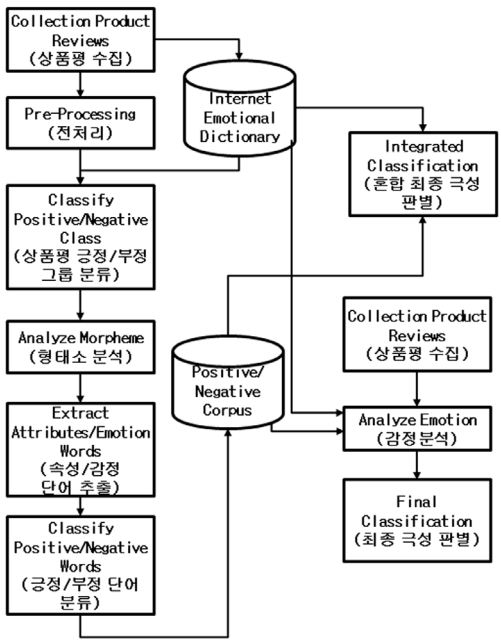

그림 1 연구절차
Fig. 1 Process of Research

글 긍정/부정 말뭉치 사전을 구축한다. 3단계는 인터넷 감정기호사전과 긍정/부정 단어사전으로 전체 범위의 상품평을 감정 분석하여 상품평의 최종 극성을 판별하는 자동화 단계로 이루어진다. 본 연구에서는 2개의 사전을 구축하게 되는데 1단계의 인터넷 감정기호사전은 수작업으로 수행되며, 2단계 및 3단계에서 활용되는 긍정/부정 말뭉치는 자동으로 구축된다.

단계별 극성분류는 정확률과 재현율을 기준으로 비교 검증한다. 정확률은 전체건수 대비 적합한 건수를 찾은 비율이고 재현율은 전체적합건수 대비 적합건수를 찾은 비율을 의미한다. 즉 정확률은 검색된 건수들이 얼마나 적합한지를 나타내는 척도로 부합된 값을 정확히 도출한 것에 대한 검증, 재현율은 적합건수를 얼마나 많이 검색했는지를 의미하여 적합한건수를 검색하는 능력을 검증이다.

- Precision(정확률) = 검색된 적합건수/(검색된 적합건수 + 검색된 부적합건수) * 100
- Recall(재현율) = 검색된 적합건수/(검색된 적합건수 + 검색되지 않은 적합건수) * 100

3.2 감정기호를 이용한 상품평 극성 분류

기존 연구에서는 긍정/부정 사전을 상품평의 어휘에 한정하여 자연어 분석 혹은 통계적 기법 등을 사용하였다. 그러나 사전구성에서 수작업의 문제[5,7,8], 상품 속성과 문맥에 따라 감정어의 극성이 다르게 분류되는 도메인에 대한 의존성의 약점이 발견되었다[20].

이러한 문제의 대안을 제시하기 위해 본 연구에서는 네티즌의 감정을 소실 없이 분석하기 위하여 인터넷 감정기호를 활용한 사전을 구축한다. 그리고 인터넷 쇼핑몰에서 노트북, MP3, 모니터의 상품평을 수집하여 상품평의 긍정과 부정으로 분류한다.

상품평은 긍정 문장과 부정 문장이 혼합되어 있으므로 인터넷 감정기호사전을 이용한 극성분류 시 문장단위 극성을 판단하는 SO-PMI(Semantic Orientation from Point-wise Mutual Information) 방식을 적용한다. 본 연구에서는 SO-PMI를 그대로 적용하지 않고 감정기호에 건수를 적용한 SO-PMI를 적용한다. 각 감정기호에 동일한 감정기호가 중첩될 경우 건수를 곱한 긍정 단어 집합에서 부정단어 집합의 차를 구하여 양수이면 긍정적이고 음수이면 부정적으로 최종 감정의 극성을 판별한다. 수식으로 나타내면 다음과 같다.

$$SCORE(rs) = \sum_{p \in P(RS)} SCORE(p) - \sum_{n \in N(RS)} SCORE(n)$$

$$= \sum_{p \in P(RS)} EMOT(p) CNT(p) - \sum_{n \in N(RS)} SCORE(n) CNT(n)$$

IF (SCORE < 0 ; Positive Reviews; else Negative Reviews)

3.3 긍정/부정 말뭉치 자동구축

상위에 사용된 인터넷 감정기호사전을 이용한 분류는 인터넷 감정기호가 포함되지 않을 상품평에는 적용할 수 없다. 따라서 인터넷 감정기호를 이용해서 분류된 긍정/부정 상품평에서 유의미한 단어를 추출하여 전체적 상품평 분석에 활용할 수 있다. 본 연구에서는 인터넷 감정기호사전에서 분류된 긍정문장과 부정문장을 형태소분석하여 긍정/부정단어를 추출하여 말뭉치 사전을 자동으로 구축한다. 이 방법은 기존에 수작업으로 진행되는 것보다 효율적인 결과를 도출할 수 있었다.

형태소분석은 Java기반의 오픈소스 '루씬 형태소분석기'를 활용한다. 형태소 분석기를 활용하면 표 1과 같이 단어를 파싱하여 품사를 구분하여 결과를 도출할 수 있다.

분류된 형태소에서 상품평의 감정을 분류하기 위하여 속성단어와 감정단어를 추출한다. 이들은 문장에서 명사형이나 명사구절의 형태소로 나타나고 문장에 주어 역할을 한다. 그리고 감정단어는 상품 속성에 의견을 부여하여 형용사의 형태소로 나타나고 이는 문장에서 서술어 혹은 주격보어, 목적격보어로 나타난다. 형태소 분석 결과에서 명사(N)과 동사(V)를 추출하여 긍정/부정 분류에 활용한다.

추출한 단어를 긍정/부정으로 자동분류하고, 속성단어와 감정단어로 분류하여 출현빈도를 계산한다. 출현빈도가 높은 단어는 상품평의 감정분석에 영향도가 높다고 판단할 수 있다. 그리고 긍정문장과 부정문장에서 동시에 추출된 단어가 존재할 수 있으므로 이들의 출현빈도를 기준으로 긍정 혹은 부정 단어로 재분류하여 긍정/부정 말뭉치를 구축한다.

표 1 상품평 형태소 분석 예시
Table 1 Reviews of Morphological Analysis Result

Ex)	아, 정말 너무 이뻐요 ㅋㅋ노란색 완전 사랑스러운데요??? ㅋㅋㅋㅋ 배송도 빨랐고
Result	아(Z)-> <100> 정말(Z)-> <100> 너무(Z)-> <100> 이쁘(V),어요(e)-> <100> 노란색(N)-> <30> 완전(N)-> <100> 사랑스러운데(N),요(j)-> 사랑/ 스러운데/ <70> 배송(N),도(j)-> <30> 배송도(N)-> <30> 빠르(V),었(f),고(e)-> <100>

3.4 최종 상품평 극성 자동판별

상위 1단계와 2단계에서 구축된 인터넷 감정기호사전과 긍정/부정 말뭉치를 기반으로 통합된 상품평의 감정을 분석하여 자동으로 극성을 판별한다. 또한 상품평을 인터넷 감정기호와 긍정/부정 말뭉치를 기반으로 실험

데이터를 생성하여 기계학습을 통해 예측 실험을 하고 제안 알고리즘과 비교 실험한다.

4. 실험 결과

4.1 인터넷 감정기호 기반 상품평 극성 분류

인터넷 언어를 이용한 상품평의 극성 분류를 위해서 먼저 Wikipedia에 명시된 '대한민국 인터넷 신조어'를 참조하여 인터넷 감정기호사전을 구축하였다. 인터넷 신조어 중에서 감정을 표현하는 이모티콘, 특수기호, 상품평에 해당하는 출현빈도 및 감정을 내포하고 있는 한글 초성 43개를 추출하여 긍정/부정으로 수동으로 분류하고 사전을 표 2와 같이 구축하였다.

인터넷 상품평 MP3, 노트북, 모니터 각 400건씩 1200개의 상품평을 무작위 추출하여 인터넷 감정기호의 포함여부를 표시한 실험 데이터를 생성하였다. 실험 데이터에 인터넷 감정기호사전의 누적건수 SO-PMI기법을 적용하여 상품평의 극성을 분류하였다. 또한 실험에 사용된 1200개의 상품평 데이터는 의미판단을 위하여 수작업으로 관찰하여 긍정/부정 판단을 하고 태깅을 달았다. 이때 3명의 연구인력을 투입하여 의미를 파악하고 2인 이상이 동일한 판단을 한 상품평을 시험 데이터로 사용하였다. 실제 판단한 긍정/부정 판단과 감정기호사전 기반의 제안 알고리즘으로 판단된 결과를 비교하였다. 실제 상품평의 극성분류가 감정기호를 기반으로 정확하게 분류되었는지를 평가하기 위하여 정확률(Precsion)과 재현율(Recall), 분포도(Coverage)를 산출하였다.

표 2 인터넷 감정기호 사전 구축 사례
Table 2 The Dictionary based on the Internet Emotional Sign

Class	Sign	Word Class	Meaning
Positive	ㅇㅋ	Verb(동사)	허락하다
	ㅋㅋ	Onomatopoeia (의성어)	킥킥, 크크크
	ㅎㅎ	Onomatopoeia (의성어)	호호호, 하하하
	ㅅㄱ	Verb(동사)	수고하세요
	^^	Mimetic Word (의태어)	웃는표정
Negative	T-T	Mimetic Word (의태어)	우는표정
	ㅠㅠ	Mimetic Word (의태어)	슬프게 우는표정
	OTL	Mimetic Word (의태어)	좌절하는 모습
	-_-;;	Mimetic Word (의태어)	경직된 표정
	ㄷㄷ	Mimetic Word (의태어)	떨리는 모습

표 3 인터넷 감정기호 기반 극성분류 실험결과
Table 3 Results of Classification based on the Internet Emotional Sign (Note: P.:Positive, N.:Negative, Pre.:Precision, Re.:Recall, Dst.:Distribution)

Data	Real (실제)	Exam (실험)	Fit	Pre.	Re.	Dst.
MP3	371	243	206	0.85	0.56	0.65
Laptop	368	194	171	0.88	0.47	0.53
Monitor	381	194	176	0.91	0.46	0.51
Average				0.88	0.49	0.56

표 4 제안 알고리즘과 기계학습 비교실험 결과
Table 4 Machine Learning Algorithm and Compares Results

Class	Precision	Recall
Internet Emotional Dictionary	0.879	0.494
Machine Learning	0.850	0.880

실험 결과는 표 3과 같이 정확률은 평균 88%로 높게 나타났으나, 재현율은 평균 49%로 낮게 나타났다. 이 결과는 감정기호가 상품평에 분포하는 분포도가 평균 56% 수준이기 때문에 낮은 재현율이 나타난 것으로 판단된다.

추가적으로 상대평가를 위하여 제안 알고리즘이 기계학습의 예측율과 어떤 차이를 나타내는지 비교실험을 실시하였다. 인터넷 감정기호로 감정분류한 데이터에Naive-Bayes 알고리즘, SMO알고리즘, J48알고리즘, IBK 알고리즘을 사용하여 분류하였다. 준비된 상품평 데이터 중 MP3, 노트북, 모니터 각각 100개씩 추출하여 학습데이터로 활용하여 실험하였다. 그리고 제안 알고리즘과 비교하기 위하여 상위에 사용된 1200개의 동일한 상품평을 실험 데이터로 준비하여 실험을 실시하였다. 실험은 10배 교차검증(10-fold-cross validation)으로 진행하였고 정확률과 재현율로 극성분류의 정확도를 비교하였다.

실험결과는 표 4와 같이 NaiveBayes, SMO, J48, IBK 알고리즘에서 평균 정확률 85%, 재현율 88%를 예측했다.

인터넷 감정기호사전 기반 극성 분류를 대표적인 분류 기계학습 알고리즘과 비교한 결과 정확률은 88%로 더 높게 나타났으나, 재현율에는 차이가 있는 것으로 확인되었다. 그 이유는 기계학습에서는 감정기호가 미포함된 데이터를 하나의 패턴으로 인식해서 분류하였기 때문이고, 제안 알고리즘에서는 건수별 오류로 분류하였기 때문으로 사료된다.

4.2 긍정/부정 말뭉치 자동구축

무작위 추출된 상품평에서 인터넷 감정기호의 분포도를 살펴본 결과 60%로 전체 인터넷 상품평을 커버하기에는 부족하였다. 따라서 본 연구에서는 나머지 40%의

표 5 형태소 분석 후 후보단어 추출 결과
Table 5 After Morphological Analysis Candidate Word Extraction Results

Data	Total	Positive	Negative
MP3	3,045건	2,530건	515건
Laptop	2,599건	1,959건	640건
Monitor	548건	441건	107건
Total	6,192건	4,930건	1,262건

표 6 MP3 추출 속성/감정단어(상위 5개)
Table 6 Extract MP3 Attributes/Emotion Words(Top 5)

구분	Attribute		Emotion	
	Word	Count	Word	Count
Positive	가격(N)	1060	좋(V)	1185
	배송(N)	924	만족(V)	489
	이어폰(N)	751	맘(N)	348
	음질(N)	582	괜찮(V)	339
	디자인(N)	561	추천(N)	298
Negative	가격(N)	172	없(V)	138
	이어폰(N)	142	않(V)	123
	배송(N)	118	아니(V)	82
	음질(N)	112	추천(N)	67
	동영상(N)	96	만족(V)	54

표 7 노트북 추출 속성/감정단어(상위 5개)
Table 7 Extract Laptop Attributes/Emotion Words(Top 5)

Class	Positive		Negative	
	Word	Count	Word	Count
Positive	가격(N)	103	좋(V)	220
	배송(N)	75	괜찮(V)	99
	성능(N)	74	어서(e)	97
	인터넷(N)	39	많(N)	94
	화면(N)	33	만족(N)	86
Negative	가격(N)	34	없(V)	72
	배송(N)	24	아니(V)	39
	인터넷(N)	22	않(V)	36
	화면(N)	22	많(N)	33
	성능(N)	22	문제(N)	28

상품평에 대한 극성분류를 수행하기 위해서 상위에 실험된 인터넷 감정기호사전을 활용하여 의미 있는 단어를 도출하였다. 먼저 인터넷 감정기호사전으로 긍정/부정 문장 그룹을 기준으로 형태소 분석한 결과 긍정 형태소 176,176건, 부정 형태소 46,994건이 분류되었다. 그리고 형태소 분석결과를 기준으로 상품의 특성을 나타내는 속성단어와 상품의 감정을 나타내는 감정단어를 표 5와 같이 추출하였다.

표 8 모니터 추출 속성/감정단어(상위 5개)
Table 8 Extract Monitor Attributes/Emotion Words(Top 5)

Class	Attribute		Emotion	
	Word	Count	Word	Count
Positive	가격(N)	885	좋(V)	1647
	배송(N)	555	많이(N)	954
	화면(N)	483	만족(N)	466
	화질(N)	332	사용(N)	461
	디자인(N)	295	구입(N)	386
Negative	가격(N)	170	없(V)	134
	화면(N)	104	만족(N)	114
	화질(N)	81	않(V)	73
	배송(N)	74	하지(N)	52
	디자인(N)	48	아니(V)	51

표 9 공통 긍정/부정 말뭉치 구축(상위 5개)
Table 9 Extract Positive/Negative Corpus(Top 5)

Rank	Positive Word	Negative Word
1	좋(V)	없(V)
2	만족(V)	않(V)
3	맘(N)	아니(V)
4	괜찮(V)	불편(N)
5	추천(N)	모르(V)

형태소 분석결과에서 MP3, 노트북, 모니터에서 추출빈도 기준 상위 20개의 긍정/부정단어를 추출하였다. 감정단어 및 속성단어를 추출한 결과, 속성단어는 10%인 데이터 도메인에 한정된 단어 이외 90%는 도메인과 무관하게 일치하였고, 감정단어 또한 90%이상 일치하는 결과를 나타냈다. 감정단어 중 긍정과 부정 문장에서 출현 빈도를 비교하여 높은 빈도의 단어를 긍정과 부정으로 분류하였다.

이를 기준으로 공통으로 사용되는 긍정/부정 단어 말뭉치를 표 9와 같이 긍정 44개 부정 21개의 단어를 구축하였다.

또한 공통으로 출현되는 속성단어와 감정단어를 제외하고 데이터 도메인에 특수하게 나타난 속성단어를 추출하여 관련된 감정단어를 단어사전에 추가하였다.

감정기호를 기반으로 긍정/부정단어를 추출한 말뭉치로 상품평의 극성분류 실험을 실시한 결과 표 11과 같이 정확률 86%로 비슷하게 나타났으나, 재현율 86%로 상당히 높은 값으로 나타났다. 이렇게 재현율이 높은 이유는 긍정/부정 단어가 실험 데이터에 나타난 분포도가 99%로 높은 이유라고 사료된다.

이 결과를 제안 알고리즘과 비교 검증해보면, 정확률과 재현율 모두 제안하는 알고리즘이 기계학습 알고리즘

표 10 도메인 기반 긍정/부정 말뭉치 사례
Table 10 Extract Positive/Negative Corpus based on the Business Domain

Data	Attribute	Positive	Negative
MP3	이어폰(N)	잘들리기(N)/ 괜찮(N)/ 무통증(N)	안나오네(N)/ 안들리네(N)/ 다르(V)
MP3	음질(N)	괜찮(N)/깨끗(N)/ 고음질(N)/깔끔(N)	나쁜(N)/불만(N)/ 떨어지(V)
Laptop	인터넷(N)	빠르(V)/좋(V)/ 쓸만(N)	느리(V)/꺼지(V)/ 별루(N)
Laptop	화면(N)	깔끔(N)/크고(N)/ 크(V)	떨어져(V)/잡티(N)/ 눈부셔(N)
Laptop	마우스(N)	좋(V)/잘되(V)/ 쓸만(N)	불량(N)/안(N)/ 없(V)
Monitor	화면(N)	좋(V)/밝은(N)/ 환(N)/크(V)	떨리(V)/꺼지(V)/ 안뜨네(N)

표 11 긍정/부정 말뭉치 기반 극성분류 실험결과
Table 11 Classification Results based on the Positive/Negative Dictionary (Note: P.:Positive, N.:Negative, Pre.:Precision, Re.:Recall, Dst.:Distribution)

Data	Real (실제)	Exam (실험)	Fit	Pre.	Re.	Dst.
MP3	371	370	312	0.84	0.84	0.99
Laptop	368	365	303	0.83	0.82	0.99
Monitor	381	380	351	0.92	0.92	0.99
Average				0.87	0.86	0.99

표 12 긍정/부정 단어사전 기반 상품평 극성분류 비교 실험 결과
Table 12 Classification Results based on the Positive/Negative and Emotional Dictionary

Class	Precision	Recall
Emotion Dictionary (Step 1)	0.878	0.494
Positive/Negative Dictionary (Step 2)	0.865	0.862

표 13 제안 알고리즘과 기계학습 비교실험 결과
Table 13 Classification Compares Results based on the Positive/Negative and Emotional Dictionary

Class	Precision	Recall
Positive/Negative Dictionary	0.865	0.862
Machine Learning	0.830	0.861

보다 높게 나타났다. 이 결과로 인터넷 감정기호 사전기반 실험으로 정확도가 높은 긍정/부정 단어가 추출되었음을 확인할 수 있었다.

4.3 상품평의 최종극성 자동판단

인터넷 감정기호사전을 구축하여 상품평의 감정을 분류한 결과 감정기호가 포함되지 않은 상품평에 대해 감정을 분류할 수 없는 이슈가 발생되었다. 따라서 감정기호로 분류된 긍정/부정 문장에서 단어를 추출하여 긍정/부정 단어사전을 구축하였다. 본 장에서는 인터넷 감정기호와 긍정/부정 말뭉치를 통합하여 상품평의 감정을 분류하는 실험을 실시하였다.

먼저 상위에서 사용되었던 MP3, 노트북, 모니터의 1200개 상품평을 동일하게 실험 데이터로 구성하여 상품평 감정분류를 실험한 결과 표 14와 같이 제안 알고리즘의 분포도(Coverage)는 99.9%이고 정확률과 재현율은 88%를 나타냈다. 전체 실험 건수 대비 공통으로 평가된 건수의 비율은 93%을 나타냈다.

그림 2와 같이 1단계 실험의 인터넷 감정기호사전, 2단계 실험의 긍정/부정 단어사전, 3단계 실험의 혼합방식의 실험결과를 비교하면 표 15와 같이 3단계 혼합방식이 정확률 및 재현율, 분포도가 높게 모두 나타났다.

다음은 검증을 위해서 인터넷 쇼핑몰에서 MP3, 노트북, 모니터의 상품평의 신규 실험 데이터 600건을 추가로 추

표 14 통합 극성분류 실험결과
Table 14 Integrated Classification Results (Note: P.:Positive, N.:Negative, Pre.:Precision, Re.:Recall, Dst.:Distribution)

Data	Real (실제)	Exam (실험)	Fit	Pre.	Re.	Dst.
MP3	371	371	320	0.86	0.84	0.99
Laptop	368	367	310	0.84	0.82	0.99
Monitor	381	381	358	0.93	0.92	0.99
Average				0.88	0.86	0.99

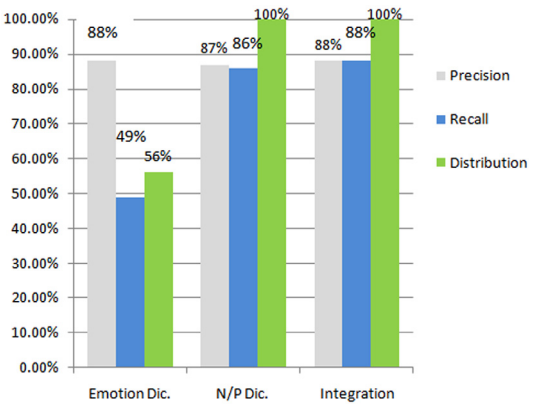

그림 2 단계별 제안 알고리즘 비교 그래프
Fig. 2 Step-by-Step Algorithm Comparison Graph (Note: P.:Positive, N.:Negative, Dic.:Dictionary)

표 15 최종 극성 판단결과
Table 15 Final Classification Results (Note: P.:Positive, N.:Negative, Pre.:Precision, Re.:Recall, Dst.:Distribution)

Data	Real (실제)	Exam (실험)	Fit	Pre.	Re.	Dst.
1200	1120	1119	989	0.88	0.88	0.99
600	600	600	545	0.91	0.91	1.00
Average				0.90	0.90	0.99

출하고 실험자와 2명이 태깅한 데이터를 제안하는 알고리즘과 기존 분류알고리즘을 적용하여 실험을 실시하였다. 기 구축된 인터넷 감정기호사전과 긍정/부정 단어사전을 기반으로 상품평의 극성을 분류한 결과 표 15와 같이 정확률 91%, 재현율 91%, 분포도 99%의 결과를 나타냈다.

5. 결론 및 향후 연구

5.1 연구의 요약 및 결론

본 연구에서는 한국 인터넷 언어의 특성을 분석하여 감정기호를 보존하면서 상품평의 극성 분류의 정확도를 높이는 것을 목표로 하였다. 이를 위하여 3단계 분석 방법을 제안하였으며, 1단계는 인터넷 수동 감정기호사전 구축 및 극성 분류 단계, 2단계는 긍정/부정 말뭉치 자동구축 및 극성 분류 단계, 3단계는 감정기호사전과 긍정/부정 말뭉치 사전을 통합하여 상품평 자동극성 분류하는 방법을 제안하였다.

먼저 한국 네티즌의 인터넷 언어에서 이모티콘, 특수기호, 한글초성의 감정기호를 추출하여 인터넷 감정기호사전을 구축하였다. 이 제안 알고리즘은 간단하면서도 데이터 도메인이나 토픽, 시간에 독립적인 방법으로 의미가 크다. 다음은 인터넷 감정기호사전 기반으로 긍정/부정 문장에서 단어를 추출하여 말뭉치를 구축하였다. 구축한 사전의 단어를 기반으로 상품평의 극성을 분류한 결과 분포도(Coverage)는 전체를 수용할 수 있는 99%의 우수한 결과를 도출하였고 기계학습 알고리즘 대비 높은 정확도를 얻을 수 있었다. 마지막으로 감정기호사전과 긍정/부정 말뭉치 사전을 통합하여 상품평 극성분류를 실험한 결과 상품평에 실험 데이터의 분포도(Coverage)는 99%로 나타났고, 정확률 및 재현율은 다른 알고리즘 대비 높은 성능을 보였다.

따라서 본 연구는 한국어의 초성과 특수기호가 포함된 감정기호를 기반으로 감정을 분류한 것과 인터넷 감정기호 포함여부와 데이터의 도메인에 무관한 극성분류 알고리즘을 제안하고 구현하여 결과를 검증한 것에 의미가 있다고 할 수 있다.

5.2 향후 연구

인터넷 감정기호를 이용한 극성분류에서는 정확률이 긍정/부정분류가 균등하게 나타났으나, 긍정/부정 단어를 기반으로 상품평의 극성분류에서는 부정 단어에 대한 정확률이 떨어지는 결과가 나타났다. 이는 부정 상품평을 제시할 경우 네티즌들이 긍정문장을 먼저 기술하고 부정문장을 사용하는 경우가 많았기 때문으로 사료된다. 따라서 향후에는 긍정/부정 말뭉치 사전을 구축할 때 언어의 문맥 기준으로 접속사, 부사 등 자연어에 대한 언어학적 접근에 대한 추가 연구가 필요하다. 또한 무작위 추출한 상품평에서 부정문장의 비율이 긍정문장에 비해 적어 부정문장에 대한 분석을 상세히 진행하지 못한 아쉬움이 존재하여 부정문장의 데이터를 추가 확보하여 부정문장에 대한 상세연구가 필요하다.

그리고 인터넷 감정기호는 강한 긍정과 부정을 표현하고 있으므로 상세 감정을 분류하기에 적합하지 않았다. 따라서 인터넷 감정기호에 추가적인 분석기법을 혼합하여 상품평의 감정을 긍정, 부정으로 극성분류에 국한하지 않고 매우 긍정, 긍정, 중립, 부정, 매우 부정으로 분류하여 감정을 상세분류하는 연구가 요구된다.

References

[1] KISA, "Internet Use Survey 2012 Survey," *KISA*, pp. 23-37, 2012.

[2] Kook Yong Lee and Seung Woon Kim, "The Impact of Online Reviews in Purchasing Decision Making," *Academy of customer satisfaction management*, Vol. 14, No. 3, pp. 85-102, 2012.

[3] Eun Ah Seo, *Speaking as a writing or linguistic analysis of Quote, Reply, good reply, bad reply, ID and emoticons*, Communication Books, Seoul, 2007.

[4] Kyungmi Park, Hogun Park, Hyunggun Kim and Heedong Ko, "Opinion mining research in SNS," *Journal of KIISE*, Vol. 29, No. 11, pp. 54-60, 2011.

[5] Jaeseok Myung, Dongjoo Lee and Sang-goo Lee, "A Korean Product Review Analysis System Using a Semi-Automatically Constructed Semantic Dictionary," *Journal of KIISE: Software and Application*, Vol. 35, No. 6, pp. 392-403, 2008.

[6] Junsoo Shin, Harksoo Kim, "A Robust Pattern-based Feature Extraction Method for Sentiment Categorization of Korean Customer Reviews," *Journal of KIISE:Software and Application*, Vol. 37, No. 12, pp. 946-950, 2010.

[7] Jung-yeon Yang, Jaeseok Myung and Sang-goo Lee, "A Sentiment Classification Method Using Context Information in Product Review Summarization," *Journal of KIISE:Database*, Vol. 36, No. 4, pp. 254-262, 2009.

[8] Jongseok Song and Soowon Lee, "Automatic Construction of Positive/Negative Feature-Predicate Dictionary for Polarity Classification of Product Reviews," *Journal of KIISE:Software and Application*, Vol. 38, No. 3, pp. 157-168, 2011.

[9] Likun Qiu, WeishiZhang, Changjian Hu,KaiZhao. "SELC:A Self-Supervised Model for Sentiment Classification," *Conference on Information and Knowledge Management, Proc. of the 18th ACM Conferenceon Information and Knowledge Management*, Hong Kong, China, 929-936, 2009.
[10] Hu, M. and Liu, B., "Mining and summarizing customer reviews," *Proc. of the 10th ACM SIGKDD Conf.*, pp. 168-177, 2004.
[11] Jae-Young Chang, "A Sentiment Analysis Algorithm for Automatic Product Reviews Classification in On-Line Shopping Mall," *The Journal of Society for e-Business Studies*, Vol. 14, No. 4, pp. 19-33, 2009.
[12] Gi young Kim, Haiin Lee, Suhwan Yook and Woojin Paik, "Customer Preference Identification System using Natural Language Processing-based Analysis and Automatic Classification of Product Reviews," *Korea Society for Information Management*, Vol. 16, pp. 65-70, 2009.
[13] Hanhoon Kang, Seong Joon Yoo and Dongil Han, "Automatic Extraction of Opinion Words from Korean Product Reviews Using the k-Structure," *Journal of KIISE:Software and Application*, Vol. 37, No. 6, pp. 470-479, 2010.
[14] Xiaowen Ding, Bing Liu., "The Utility of Linguistic Rules in Opinion Mining," *Proc. of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 811-812, 2007.
[15] Theresa Wilson, Janyce Wiebe and Paul Hoffmann. "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis," *HLT/EMNLP*, pp. 347-354, 2005.
[16] Alexander Pak and Patrick Paroubek, "Twitter based system: Using Twitter for Disambiguating Sentiment Ambiguous Adjectives," *Proc. of International Workshop of Semantic Evaluations*, 2010.
[17] Courses, E., and Surveys, T., "Using SentiWordNet for multilingual sentiment analysis," Data Engineering Workshop ICDEW, 2008.
[18] Pavel Smrž, "Using WordNet for Opinion Mining," *Proc. of the International WordNet Conference 2006*, pp. 333-335, 2006.
[19] Peter D. Turney, "Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews," *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, pp. 417-424, 2002.
[20] Jonathon Read, "Using emoticons to reduce dependency in machine learning techniques for sentiment classification," *In ACL, The Association for Computer Linguistics*, 2005.
[21] Alexander Pak and Patrick Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," *Proc. of theEuropean Language Resources Association (ELRA)*, 2010.
[22] Hongjune Yune, Han-joon Kim and Jae-Young Jang, "An Efficient Search Method of Product Reviews Using Opinion Mining Techniques," *KIISE Transactions on Computing Practices*, Vol. 16, No. 2, pp. 222-226, 2010.
[23] H. Nishikawa, T. Hasegawa, Y. Matsuo, and G. Kikui, "Opinion summarization with integer linear programming formulation for sentence extraction and ordering," *In COLING*, 2010.
[24] K. Dave, S. Lawrence, D. Pennock, "Mining the Peanut Gallery: Opinion Extraction and Semantic Classification of Product Reviews," *Proc. of the 12th Intl. World Wide Web Conference (WWW '03)*, pp. 512-528, 2003.
[25] Qiang Ye, Ziqiong Zhang and Rob Law, "Sentiment classification of online reviews to travel destination by supervised machine learning approaches," *Expert Systems with Applications, Elsevier*, pp. 1-9, 2008.
[26] P. Turney and M. Littman, "Measuring praise and criticism: Inference of semantic orientation from association," *Proc. of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics*, pp. 417-424, 2002.
[27] Minqing Hu and Bing Liu, "Mining and Summarizing Customer Reviews," KDD'04, Seattle, Washington, USA, 2004.
[28] V. Vapnik, "Estimation of Dependences Based on Empirical Data," Springer-Verlag, 1982.
[29] J.C Platt, "Fast training of support vector machines using sequential minimal optimization. Advances in kernel methods: support vector learning," MIT Press, Cambridge, MA, 1999.
[30] G. H. John, P. Langley, "Estimating Continuous Distributions in Bayesian Classifiers," *Uncertainty in Artificial Intelligence*, Vol. 11, pp. 338-345, 1995.
[31] J.R. Quinlan, C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, San Mateo: CA, 1993.
[32] Internet shopping mall page; http://www.auction.co.kr
[33] Internet shopping mall page; http://www.gmarket.co.kr

장 경 애
1996년 대구대학교 문헌정보학과(문헌정보학사). 2014년 연세대학교 컴퓨터공학과(공학석사). 2014년~현재 서울과학기술대학교 IT정책대학원 산업정보시스템 박사과정. 관심분야는 데이터 품질, 데이터 분석, 인공지능, 최적화, 소프트웨어 품질 등

박 상 현

1989년 서울대학교 컴퓨터공학과(공학사). 1991년 서울대학교 컴퓨터공학과(공학석사). 2001년 UCLA대학교 전산학과(공학박사). 1991년~1996년 대우통신 연구원. 2001년~2002년 IBM T. J. Watson Research Center Post-Doctoral Fellow 2002년~2003년 포항공과대학교 컴퓨터공학과 조교수. 2003년~2006년 연세대학교 컴퓨터과학과 조교수. 2006년~2011년 연세대학교 컴퓨터과학과 부교수. 2011년~현재 연세대학교 컴퓨터과학과 정교수. 관심분야는 데이터베이스, 데이터 마이닝, 바이오인포매틱스, 적응적 저장장치 시스템

김 우 제

1986년 서울대학교 산업공학과(공학사) 1988년 서울대학교 산업공학과(공학석사) 1994년 서울대학교 산업공학과(공학박사) 2003년~현재 서울과학기술대학교 글로벌융합산업공학과 교수. 1988년 4월~1991년 2월 동양경제연구소 연구원. 1999년~2001년 University of Michigan Visiting Scholar. 관심분야는 IT서비스, 소프트웨어공학, 최적화, 스마트그리드 등