

오피니언 마이닝을 활용한 블로그의 극성 분류 기법

이종혁*, 김원상*, 박제원**, 최재현**

요약

기존의 감정분석을 통한 극성 분류는 주로 평점을 기반으로 하는 상품평을 기준으로 문장규칙을 이용하여 분석해왔다. 이러한 분석방법은 평점이 없는 블로그 같은 경우 적용되기 어려움 점이 있고 댓글 아르바이트나 관리자에 의해 상품평이 조작될 가능성이 있어서 상품평 만으로는 상품, 매장에 대한 의견을 파악하기에는 어려움이 있다. 이러한 문제점을 고려할 때 개인들의 솔직한 의견이 담겨 있는 블로그를 분석하여 극성을 분류하면 상품, 매장에 대한 올바른 이해가 가능하다. 본 논문은 도메인별로 블로그 글에 대한 고빈도 단어를 추출하여 주제어를 선정하고, 선정된 주제어를 기준으로 제안하는 감정분석 기법을 적용하여 블로그 글에 대한 극성을 분류한다. 감정분석 기법의 성능을 평가하기 위하여 정보 검색 분야에서 사용되는 측정지표 Precision, Recall, F-score를 사용하여 본 연구의 극성 분류기법의 유용성을 검증한다. 평가 결과 기존의 상품평을 문장규칙을 이용하여 분석하여 극성 분류를 하는 기법들에 비해서 제안한 감정분석 기법을 적용할 경우에 우수한 성능으로 극성 분류를 하는 것으로 나타났다.

키워드 : 감정단어, 감정방향, 오피니언 마이닝, 극성분류, 블로그

The Blog Polarity Classification Technique using Opinion Mining

Jong-Hyuk Lee*, Won-Sang Lee*, Jea-Won Park**, Jae-Hyun Choi**

Abstract

Previous polarity classification using sentiment analysis utilizes a sentence rule by product reviews based rating points. It is difficult to be applied to blogs which have not rating of product reviews and is possible to fabricate product reviews by comment part-timers and managers who use web site so it is not easy to understand a product and store reviews which are reliability. Considering to these problems, if we analyze blogs which have personal and frank opinions and classify polarity, it is possible to understand rightly opinions for the product, store. This paper suggests that we extract high frequency vocabularies in blogs by several domains and choose topic words. Then we apply a technique of sentiment analysis and classify polarity about contents of blogs. To evaluate performances of sentiment analysis, we utilize the measurement index that use Precision, Recall, F-Score in an information retrieval field. In a result of evaluation, using suggested sentiment analysis is the better performances to classify polarity than previous techniques of using the sentence rule based product reviews.

Keywords : Sentiment Vocabulary, Sentiment Orientation, Polarity Classification, Opinion Mining

1. 서론

최근에 사람들은 합리적인 소비활동을 위해서 물건을 살 때나 음식점을 이용할 때 인터넷을 검색하여 이미 전에 사용하거나 방문한 사람들의 생각이 담긴 블로그의 글을 읽는다. 온라인 쇼평몰이나 맛집 사이트에서 평점을 매기거나 후기를 남기는 경우가 많은데 댓글 아르바이트를 하는 사용자들 후기를 조작하거나 관리자가 부정적인 글은 지울 수 있기 때문에 정확한 평을 기대하기는 어렵다. 하지만 블로그의 경우 개인적인 생각과 의견을 담는 공간이기 때문에 사

※ 교신저자(Corresponding Author): Jae-Hyun Choi
접수일:2014년 07월 11일, 수정일:2014년 08월 30일
완료일:2014년 08월 31일

* 숭실대학교 SW특성화대학원 석사과정

Tel: +82-010-5168-0114, +82-010-7628-2615

email: lifesasimi@naver.com

** 숭실대학교 SW특성화대학원 교수

람들의 솔직한 감정을 알 수 있다. 사람들의 생각은 주관적이기 때문에 한 블로거의 글만을 읽는다면 정확한 판단을 내리기 어렵다. 따라서 정확한 판단을 위해서는 다수의 블로거의 글을 추출하여 오피니언 마이닝을 통해 상품, 매장에 대한 블로거들의 의견을 분류해야 한다.

극성에 따른 의견 분류를 위해서 감정사전을 구축하고, 문장 규칙을 설정하여 주제를 중심으로 글을 분석해야 한다.[1],[3] 그 중 문장 규칙을 정하는데 있어서 관형사규칙, 부정어 규칙만을 사용하여 구문을 분석할 경우 서술어가 갖는 의미적 강도 파악에 어려울 수 있다. 예를 들어 '핸드폰이 좋다'라는 표현과 '핸드폰이 정말 좋다'의 표현은 긍정이라고 분류는 가능하지만 긍정적인 분류 중에서도 얼마만큼의 강도의 긍정인지 파악이 어려울 수 있다. 또한 한국말의 특성상 문장이 길어질수록 서술어를 꾸며주는 형용사, 부사어가 많이 등장한다는 것을 볼 수 있다. 이는 주어와 서술어 사이의 거리가 증가하면 의미적 강도가 높아진다고 볼 수 있다. 따라서 문장규칙을 정하는데 있어서 관형사, 부정어 규칙 이외에도 의미적 강도를 파악할 수 있는 부사어를 고려할 필요가 있다.

본 연구는 오피니언 마이닝에서 자주 사용하는 평점이 매겨져 있는 상품평 이외에 상품, 매장에 대한 의견이 담겨 있는 블로그의 글을 대상으로 활용할 수 있는 감정분석 알고리즘을 제안한다. 제안방법은 먼저 수집된 블로그 글의 형태소를 분석하여 명사추출 후, 빈도를 세서 주제를 선정한다. 선정된 주제를 문장규칙으로 만든 감정 분석 알고리즘을 적용하여 글에 대한 극성을 분류한다. 제안하는 감정 분석 알고리즘에서는 다른 연구에서 주로 사용하는 관형사 규칙 이외에도 부정규칙, 감정의 의미 강도에 따른 깊이를 구분하는 부사규칙과 거리규칙을 사용하여 이전 연구들보다 더 정확하게 분류가 가능하다는 장점을 가지고 있다. 또 평점이 없는 글에서도 분류가 가능하기 때문에 제안하는 알고리즘은 더 많은 도메인에서도 활용할 수 있다는 이점이 있다.

논문의 구성은 다음과 같다. 2장에서는 관련연구에 대해 살펴보고, 3장에서는 본 논문에서 제안하는 분류기법을 제안한다. 4장에서는 제안한 분류기법에 대한 실험 결과를 분석한다. 마지막

으로 5장에서는 결론과 향후 연구에 대해 논의한다.

2. 관련연구

2.1 오피니언 마이닝

오피니언 마이닝(Opinion Mining)은 문서로부터 사용자가 나타내는 의견과 감정을 나타내는 패턴을 이용해 특정 대상에 대한 의견의 극성을 분류하는 기술이다. [8]은 블로그에 쓰인 글을 통하여 대중들의 주관적인 의견을 효율적으로 판단하기 위하여 오피니언 마이닝은 다른 기법에 비해 효율적이다. 문서에 담긴 감정방향을 분류하는 연구가 많이 진행 중인데, 국내에서는 [1],[3],[5],[16]는 문장의 패턴과 언어학적 규칙을 이용하여 분류하는 연구를 시도하였고 국외에는 [2],[9],[10],[11],[12],[13],[14]가 연구하였다. 그 중 [2]는 문장 간의 거리에 따른 의미적 강도를 연구하였는데 이 부분을 한글에 맞게 변형하여 본 논문의 감정분석 기법에 적용하였다.

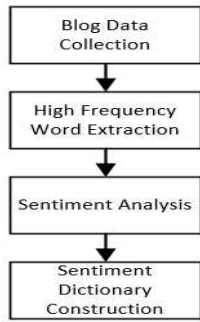
2.2 감정분석

[1]에서는 상품평을 기반으로 영화, 의류 등의 사용자의 감정을 분석할 수 있는 감정사전을 자동으로 구축하는 시스템을 구현하였다. [3]은 감정분석을 통해 추천할 영화를 랭킹으로 정렬하여 추천해준다. 이러한 연구들에서는 감정분석 방법은 평점이 담겨져 있는 글을 대상으로 극성을 분류하는 방법을 사용한다. 그렇기 때문에 평점이 없는 글에 대해서는 평가하기 어렵다. 또한 [1]에서는 문장규칙을 적용할 때 관형사 규칙만을 적용하여 의미적 강도를 파악하기 어려운 점이 있다. 본 논문에서는 평점이 없는 글에서도 극성분석을 통해 긍정적인 감정이 담겨 있는 글과 부정적인 글을 분류할 수 있는 방법을 연구한다. [2]는 문장 간의 거리에 가중치를 두어 의미의 강도를 파악하여 적용하였다. 하지만은 이는 영어를 기준으로 하였기 때문에 한글 어순에는 맞지 않는 부분이 있어 본 논문에서 한글 어순에 맞게 거리규칙을 변형하여 적용한다.

3. 제안하는 분류 기법

본 논문에서 제안하는 감정분석 분류 기법의 프로세스는 (그림 1)과 같다. 제안하는 기법은 기존기준 없이도 감정분석을 통해 긍정, 부정을 판별할 수 있다.

(그림 1) 분류기법 프로세스



(Figure 1) Classification Technique Process

이를 위한 분류 기법의 단계는 총 4단계로 구성된다. 첫 번째 단계는 네티즌들의 특정 도메인에 관한 개인적인 글이 담겨있는 블로그에서 정보를 수집하는 단계이다. 두 번째 단계는 도메인과 관련하여 수집된 블로그 글 중에서 주로 언급되는 특징이 담긴 단어를 추출하는 단계이다. 세 번째 단계는 글의 극성 판별을 위한 감정분석 알고리즘을 구축 및 적용 단계이다. 마지막 단계는 감정단어 사전을 구축하여 세 번째 단계에서 구축한 감정분석 알고리즘과 비교하여 글의 극성을 판별한다. 본 장에서는 총 4단계의 프로세스를 각 절에서 자세히 설명한다.

3.1 블로그 정보 수집

(그림 2)는 블로그 정보수집 프로세스를 나타낸다. 정보 수집 프로세스의 첫 번째 단계로 Google에서 제공하는 Custom Search API를 이용하여 국내 블로그 서비스 업체인 Naver, Daum, Tistory, Cyworld의 핸드폰, 영화, 음식점 총 3개의 도메인과 관련된 블로그 글을 추출하였다. 영화 도메인에서는 ‘트랜스포머’, 핸드폰 ‘갤럭시S3’, 음식점 도메인은 종류별로 10개로 선정하였다. 각각 1만개, 도합 총 14만개를 추출하였다.

(그림 2) 블로그 정보수집 프로세스

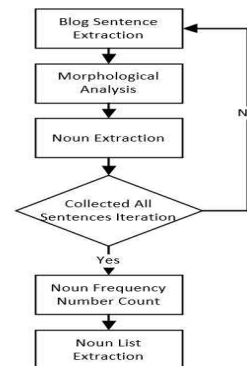


(Figure 2) Blog Data Extraction Process

총 14만개의 데이터를 2단계의 필터링을 거친다. Custom Search는 미리 선정한 키워드가 있는 경우, 모두 추출하기 때문에 필요하지 않거나 중요하지 않은 글 또한 추출할 수 있다. 불필요한 글을 제거하기 위해 첫 번째 단계로 글의 제목에 선정된 도메인 관련 키워드가 있을 경우에 필요한 글로 간주하여 추출한다. 두 번째 단계는 Jsoup API를 이용하여 http 파싱을 하여 평가가 있는 글만을 수집한다. 필터링 결과 14만개의 데이터 중 5619개의 데이터를 선별할 수 있었다.

3.2 고빈도 단어 추출

(그림 3) 고빈도 단어 추출 프로세스



(Figure 3) High Frequency Vocabulary Extraction Process

이번 장에서는 (그림 3)과 같이 고빈도 단어 추출 프로세스를 기술한다. 프로세스의 단계를 보면 블로그에 쓰인 문장들을 추출하여 형태소 분석을 통해 단어들 중 빈번하게 등장하는 단어를 뽑는다. 이 단어들은 글의 주제어로 생각할 수 있다. 따라서 이 단어들을 주제어 리스트로

간주하여 다음 단계에서 구축하는 언어문장규칙의 기준점으로 사용한다. 본 연구에서는 초고속 형태소 분석기 mach 1.0[15]을 사용하여 명사(NN)만을 추출하고, 글마다의 추출된 명사어가 빈도가 얼마나 되는지를 카운트한다. 추출결과는 <표 1>에 나타나있다.

<표 1> 고빈도 순으로 추출된 주제어 일부

Domain	Restaurant		Mobile	
Detail	Nolboo-Bossam		Galaxy S3	
Number	Word	Frequency	Word	Frequency
1	Bossam	171416	Mobile	9312641
2	Nolboo	144031	Galaxy	2074558
3	Menu	34623	Design	343321
4	Pig's feet	29434	Price	201891
5	Kimchi	26613	LCD	163970
6	Taste	23687	Quality of sound	87862
7	Brand	19644	Battery	77434
8	Price	17888	Screen	63111
9	Order	14097	Touch	45266
10	Food	8589	Camera	43680

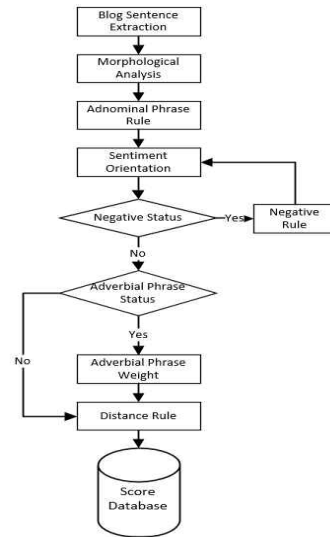
<Table 1> The part of the extracted Keyword ascending by High Frequency

3.3 감정분석 적용

이 장에서는 글의 감정을 분석하는데 있어서 평점을 기반을 두지 않고 문장의 규칙만을 이용하여 극성을 판별할 수 있는 기법을 (그림 4)에 나타냈다. 언어문장 규칙은 관형어 규칙, 부정어 규칙, 부사어 규칙, 거리 규칙 4가지 부분으로 나누어진다. 한글은 '핵말언어'로 중요한 의미가 뒤에 나오고 평서문에서 서술어가 가장 마지막에 나온다. '명동칼국수는 맛있다.'라는 문장에서 주어인 '명동칼국수'의 의미를 나타내는 '맛있다'라는 말이 가장 뒤에 나온다. 이처럼 한국어는 문장에서 중요한 말이 뒤에서 나타난다. '명동칼국수는 정말 맛있다.'라는 문장처럼 주어와 동사 사이에 감정을 나타내는 관형어와 부사가 추가될수록 문장이 길어지는 점을 볼 수 있다. 이러한 문장의 규칙에 맞게 본 논문에서는 부사가 들어가서 관형어를 꾸며주면서 문장 길이가 길어지면 감정의 깊이를 나타낸다고 볼 수 있다. 따라서 이 알고리즘에서 부사가중치와 거리가중

치를 두어 극성 분류이외에도 극성의 강도를 파악할 수 있도록 하였다.

(그림 4) 감정분석 프로세스



(Figure 4) Sentiment Analysis Process

식(1)은 본 논문에서 제안하는 감정분석 알고리즘이다. 식(1)에서 $AD+w_i$ 는 부사가중치 AD 와 감정서술어 w_i 의 합이다. 감정서술어 w_i 를 꾸며주는 부사어가 있을 경우 부사 가중치를 더 해준다. SO 는 Sentimental Orientation의 약자로 감정방향을 의미한다. 감정방향 값은 1과 -1 값으로 각각 긍정과 부정을 의미한다. 앞에서 w_i 는 감정서술어이지만 이 감정이 긍정의 의미인지 부정의 의미인지 구분되어 있지 않다. 이때 SO 를 적용하여 방향을 결정해준다. $dis(w_i, f)$ 는 감정서술어 w_i 와 주제어 f 값의 거리 값을 표현해준다. $dis(w_i, f)$ 의 값을 거리 가중치로 두어 $(AD+w_i) \cdot SO$ 값에 곱한 후 전체 총 값을 합한다. 전체 값은 블로그에 있는 총 문장을 의미하고 w_i 에서 i 값은 한 문장 안에 존재하는 감정서술어의 개수이다.

$$score(f) = \sum (AD+w_i) \cdot SO * dis(w_i, f) \quad (1)$$

AD : 부사 가중치

w_i : 단어

SO (Sentimental Orientation): 감정방향

(긍정: 1, 부정: -1)
 $dis(w_i, f)$:거리 가중치
 (감정단어와 주제어와의거리)
 f :주제어(고빈도 단어)

관형사 규칙은 문장에서 ‘형용사 + 어미’와 ‘동사 + 어미’의 문법구조에서 형용사와 동사만을 추출하여 적용하도록 한다. 식(1)에서 감정서술어 w_i 들을 (그림 5)의 알고리즘1에 적용하여 감정사건과 비교하여 감정방향을 결정한다.

(그림 5) 알고리즘1: 관형어, 부정규칙

```

for each sentence  $s_i$  that contains a
    set of features do
    if feature  $f_j$  is equal to component
    of sentence  $w_k$  that is verb or adjective
    then
        opinionOrientation = compare to
        Sentimental Dictionary;
    if feature  $f_j$  is equal to negative
    word then
        opinionOrientation
        =inverse * opinionOrientation;
    endif
    if feature  $w_{k-1}$  that is adverb exists
    then
        Algorithm2();
    endif
endif
endfor;
    
```

(Figure 5) Algorithm1:
 Adnominal Phrase, Negative Rule

부정어 규칙은 관형어 규칙에서 적용되어 감정방향이 결정되었지만 관형어 뒤에 부정어가 나오므로 의미가 변화는 예외의 경우에 적용된다. 한글에는 앞이 긍정이라도 뒤에 단어의 선어말어미 중 부정의 의미를 담고 있는 ‘안’, ‘않’, ‘지만’, ‘못’이 나오면 부정이 된다. 예를 들면 ‘점원이 친절하지 않다.’라는 문장에서 ‘친절’이라는 단어는 긍정의 의미를 가지고 있지만 ‘않다’라는 단어가 나오면서 문장전체가 부정의 의미로 변하게 된다. 따라서 앞이 긍정이지만 부정어가 나

오면 부정으로 처리하고, 앞이 부정이고 뒤 또한 부정이 되면 긍정으로 감정방향을 바꾸어 준다. 알고리즘1에서 부정어가 있는 경우 opinionOrientation에 inverse를 곱하여 감정방향을 바꾸어준다.

부사규칙은 감정서술어를 꾸며주는 부사어가 문장에 나올 경우 추출하여 가중치를 적용한다. 부사는 서술어의 뜻을 분명하게 만들어주는 품사이기 때문에 본 논문에서는 부사가 감정서술어의 감정의 강도를 깊게 나타낸다고 생각하여 부사에 가중치를 두었다.

(그림 6) 알고리즘2: 부사규칙

```

if feature  $w_{k-1}$  that is adverb exists
then
    if opinionOrientation is equal to 1
    then
        opinionOrientation += 1;
    endif
    if opinionOrientation is equal to -1
    then
        opinionOrientation -= 1;
    endif
endif
    
```

(Figure 6) Algorithm2:
 Adverbial Phrase Rule

알고리즘2 부사규칙은 감정의 방향을 나타내는 opinionOrientation이 1(긍정)인 경우에는 가중치 1를 가산하고 opinionOrientation이 -1(부정)인 경우 1을 감산해준다.

거리규칙은 식(1)의 감정서술어 w_i 와 주제어 f 사이의 거리를 가중치로 준다. 문장에서 감정단어와 고빈도 단어 추출을 통해 추출된 주제어 사이에 관형어와 부사가 존재할 경우 문자의 길이가 길어지며, 그만큼 감정의 깊이가 깊어진다 고 볼 수 있다. 따라서 그 깊이에 따라 곱하여 가중치를 주었다. 거리는 감정단어를 기준으로 주제어 사이의 단어 수로 식(2)로 정의하였다.

$$distance = w_i - f \tag{2}$$

w_i :단어
 f :주제어(고빈도 단어)

예를 들어 ‘명동칼국수는 정말 맛있다’라는 문장이 있으면 w_i 는 감정서술어인 맛있다가 되고 f 는 주제어인 ‘명동칼국수’가 된다. 그러므로 w_i 의 값은 첫 문장의 단어부터 ‘맛있다’까지의 단어 수가 되므로 $w_i=3$ 이고, f 의 값은 시작 단어가 되기 때문에 $f=1$ 이 되어 $w_i-f=2$, 즉 거리 가중치 값은 2가 된다.

3.4 감정사전 구축

감정사전은 3.3절에서 언어 문장규칙으로 추출한 감정서술어의 감정방향을 비교하기 위해 필수적이다. 감정사전의 구축은 동양권에서 감정을 설명할 때 가장 많이 인용되는 ‘회로애락(喜怒哀樂)’을 기준 ‘기쁨, 분노, 슬픔, 즐거움’의 동의어로 감정사전 데이터베이스를 구축하였다. 동의어 데이터는 [4],[6],[7]에서 가져오고, 도메인에 맞는 동의어를 일부 수기로 입력하였다.

<표 2> 감정사전 일부

Number	Sentiment Predicate	Sentiment Orientation
1	Cheerful	p(positive)
2	Violent Anger	n(negative)
3	Anguish	n(negative)
4	Good	p(positive)
5	Painful	n(negative)
6	Sympathetic	p(positive)
7	Pleased	p(positive)
8	Sharp	p(positive)
9	Aggressive	n(negative)
10	Embarrassed	n(negative)

<Table 2> The Part of Sentiment Dictionary

<표 2>는 전체 424개의 데이터 중 데이터베이스에 저장되어 있는 일부 데이터를 나타내었다. 감정서술어는 형용사형으로 입력하였고, 각각의 단어의 의미방향을 긍정이면 p로, 부정이면 n으로 표현하였다.

4. 실험 및 검증

4.1 테스트 데이터 및 평가 방법

본 연구에서는 Naver, Daum, Tistory, Cyworld 블로그에서 핸드폰, 영화, 그리고 음식

점 도메인의 데이터를 수집하였다. <표 3>은 도메인별로 추출한 블로그 데이터의 수이다. 테스트 데이터를 도메인별로 200개를 임의로 추출하고 본 논문의 분류기법을 적용하였다.

<표 3> 도메인별 수집한 블로그 수

Domain	Detail Domain	Number
Movie	Transformer	948
Mobile	Galaxy S3	482
Restaurant	Nolboo-Bossam	208
	Primo BacioBaci	445
	Mies container	434
	Seoga and Cook	485
	Abiko	238

<Table 3> The number of Blogs collected by Domain

평가방법은 분류기법을 통해 결정된 감정방향 결과와 실제 수작업으로 감정방향을 확인하였을 때의 결과를 비교하여 일치하는지를 판단하였다. 통계적 분류 평가척도인 식(3) Precision, Recall F-Score를 사용하였다. 식(3)에서 True Positive는 시스템이 감정방향을 긍정이라고 분류하였는데 실제로 사람이 분류하였을 때 긍정인 블로그의 수이고, False Negative는 시스템이 긍정이라고 분류하였는데 사람이 확인하였을 때 부정인 블로그의 수이다. False Positive는 시스템이 부정이라고 분류하였는데 실제로 사람이 확인하였을 때는 긍정인 블로그의 수를 나타낸다. 이를 기반으로 Precision, Recall, 계산하였다.

$$Precision = \frac{True\ Positive}{True\ Positive + False\ Positive} \quad (3)$$

$$Recall = \frac{True\ Positive}{True\ Positive + False\ Negative}$$

$$F-Score = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

4.2 제안 분류기법의 실험결과

본 연구에서 실험은 부사규칙만 적용하였을 때의 실험, 거리규칙만 적용하였을 때의 실험, 둘 다 적용한 본 연구의 분류기법을 사용한 실험 3가지로 진행하였고 기존 연구 [1],[2]의 실험과 비교하였다. 도메인 음식점은 3가지 실험을

모두 진행하였고 영화, 핸드폰은 제안하는 분류 기법만을 적용하여 실험하였다.

4.2.1 자체 실험 3가지:

부사규칙, 거리규칙, 제안하는 분류기법

음식점은 수집된 10가지 음식점 중 5가지 음식점을 임의로 선택하여 실험하였다. 선정된 음식점은 <표 3>에 음식점 세부 도메인으로 구성된다. 부사규칙 실험의 결과는 <표 4>에 보여준다. 식(1) 감정분석 알고리즘 중에서 알고리즘1의 관형사, 부정규칙을 고려한 기법의 성과와 추가적으로 알고리즘2 부사규칙을 적용하였을 경우 결과는 <표 4>와 같이 Precision은 약 0.28% 저하, Recall은 0.75% 향상, F-score은 0.09% 저하되었다.

<표 4> 부사규칙 성능

Detail Domain	Consideration of Adnominal & Negative Phrase Rule			Consideration of Adverbial Phrase Rule		
	Precision	Recall	F-Score	Precision	Recall	F-Score
Nolboo-Bossam	96.08	60.12	73.96	97.20	63.41	76.75
Primo BacioBaci	93.68	50.57	65.68	91.92	51.12	62.70
Mies container	96.36	58.24	72.60	96.55	60.87	74.67
Seoga and Cook	89.72	58.54	70.85	88.88	58.79	70.80
Abiko	94.64	60.92	74.12	94.44	57.95	71.83
Average	94.10	57.68	71.44	93.82	58.43	71.35

<Table 4> The Performance of Adverbial Phrase Rule

<표 5>는 위의 부사규칙 실험처럼 알고리즘1의 관형사, 부정규칙만을 적용한 성과와 식(2) 거리규칙을 적용하였을 경우의 성능을 비교한 결과이다. Precision은 그대로 유지되었고 Recall은 약 1.88% 향상, F-score은 0.48%로 향상되었다.

<표 5> 거리규칙 성능

Detail Domain	Consideration of Adnominal & Negative Phrase Rule			Consideration of Distance Rule		
	Precision	Recall	F-Score	Precision	Recall	F-Score
Nolboo-Bossam	96.08	60.12	73.96	97.20	62.58	75.84
Primo BacioBaci	93.68	50.57	65.68	91.92	55.37	69.50
Mies container	96.36	58.24	72.60	96.55	61.75	75.33
Seoga and Cook	89.72	58.54	70.85	88.88	60.98	72.99
Abiko	94.64	60.92	74.12	94.44	57.14	70.92
Average	94.10	57.68	71.44	94.10	59.56	72.92

<Table 5> The performance of Distance Rule

관형사, 부정규칙을 사용한 성과와 본 논문에서 제안하는 분류기법을 적용한 성능을 비교한 결과는 <표 6>에 나타났다. Precision은 1.14% Recall은 4.51%, F-score은 3.8% 향상되었다.

<표 6> 제안한 분류기법 성능

Detail Domain	Consideration of Adnominal & Negative Phrase Rule			Consideration of Suggested Classification Technique		
	Precision	Recall	F-Score	Precision	Recall	F-Score
Nolboo-Bossam	96.08	60.12	73.96	97.20	64.85	77.82
Primo BacioBaci	93.68	50.57	65.68	91.92	60.34	73.97
Mies container	96.36	58.24	72.60	96.55	62.70	76.07
Seoga and Cook	89.72	58.54	70.85	88.88	62.05	74.10
Abiko	94.64	60.92	74.12	94.44	61.02	74.23
Average	94.10	57.68	71.44	95.24	62.19	75.24

<Table 6> The performance of suggested Classification Technique

4.2.2 비교 실험

[1]은 도메인의 특징에 따라 감정방향을 상이하게 사용하는 것을 고려하여 극성을 분류하는 기법이다. 이 기법에서는 평점이 나타나 있는 상품평을 기준으로 서술어의 감정방향을 1차적으로 나누어 놓고 그 뒤에 감정별 서술어의 빈도

분석과 극성 비율 값을 계산하여 감정사전을 구축한 후, 서술어에 대한 품사 태깅 규칙을 사용하여 긍정, 부정을 분류한다. [2]는 상품평을 기준으로 영어로 된 문장을 거리규칙과 부정규칙을 적용하여 극성은 분류한다.

본 논문에서 제안하는 감정분석 알고리즘의 경우, 기본적으로 평점으로 감정방향이 결정되지 않는 블로그의 일반적인 글을 대상으로 하여 관형사, 부정, 부사, 거리 규칙을 적용하여 긍정, 부정을 분류하였다. <표 7>은 [1]과 본 분류 기법의 비교 실험의 결과이다. [1]의 논문의 실험 결과 중 긍정, 부정의 글의 비율정보를 미 고려한 데이터를 기준으로 비교하였다. 블로그의 글 작성 특성상 도메인 관련 긍정관련 글이 부정관련 글에 비해 많기 때문이다. <표 7>를 보면 본 논문의 분류기법이 상품평을 Precision은 15.97%, Recall은 2.91%, F-score는 8.64% 향상된 것을 볼 수 있다.

<표 7> [1]과 본 논문의 성능비교

Domain	Automatic Construction of Positive/Negative Feature-Predicate Dictionary for Polarity Classification of Product Reviews[1]			Suggested Classification Technique		
	Precision	Recall	F-Score	Precision	Recall	F-Score
Mobile	81.74	77.05	79.33	97.66	86.21	91.58
Movie	71.73	71.73	71.73	91.87	71.07	80.14
Average	78.79	75.73	77.22	94.76	78.64	85.86

<Table 7> The Performance Comparison suggested Classification Technique with [1]

<표 8>은 [2]와 본 분류 기법의 비교실험의 결과이다. [2]의 경우 영어를 기준으로 거리 규칙과 부정규칙을 적용하였기 때문에 한글을 기준으로 하는 본 감정분석 알고리즘의 거리, 부정 규칙과 상이한 점이 있다. <표 8>의 결과를 보면 Precision이 15.92%, Recall이 9.16%, F-Score이 12.25% 향상되었다.

<표 8> [2]와 본 논문의 성능비교

Domain	A Holistic Lexicon-Based Approach to Opinion Mining[2]			Suggested Classification Technique		
	Precision	Recall	F-Score	Precision	Recall	F-Score
Mobile	81.74	77.05	79.33	97.66	86.21	91.58

<Table 8> The Performance Comparison suggested Classification Technique with [2]

두 가지 비교 실험에서 볼 수 있듯이 제안하는 감정분석 알고리즘이 기존의 연구에서 사용되는 기법보다 약 10%의 향상된 성능을 나타낸 것으로 볼 때 기존의 분류기법보다 높은 성능을 가진다고 볼 수 있다.

5. 결론 및 향후 연구

본 연구에서는 기존연구에서 활용하였던 평점이 내포되어 있는 상품평 기반이 아닌 네티즌들의 일반적인 글이 담겨 있는 블로그를 기준으로 도메인별로 데이터를 수집하였다. 추출된 데이터 안에서 고빈도 단어를 분석하여 주제를 선정하였다. 선정된 주제를 기준으로 제안한 문장 규칙(관형사, 부정, 부사, 거리)을 적용하여 극성을 분류하였다. 제안하는 감정분석 알고리즘안의 세부 규칙인 관형사, 부정, 부사, 거리 규칙의 각각의 성능을 비교 실험을 하였다.

관형사, 부정 규칙을 고려한 성과와 이 기법에 추가적으로 부사규칙을 적용하였을 때의 성능 비교는 성능에 미비하게 저하된 결과가 나타났다. 하지만 거리 규칙을 추가적으로 사용한 기법의 성능은 전 보다는 성능이 향상되었다. 또한 최종적으로 4가지 규칙을 모두 적용한 경우는 관형사, 부정규칙 만을 사용한 결과보다 높게 향상되었다. 또한 기존의 다른 논문에서 연구한 2가지 실험과 비교실험을 하였는데 성능이 각각 8.64%, 12.25%의 높은 향상된 결과를 보여주었다.

향후 연구는 감정분석을 통한 극성 분류의 성능을 향상을 위해 감정단어 사전의 연구가 필요하다. 앞서 실험에서 간혹 특정 도메인의 경우 낮은 recall값이 나타났는데 이는 블로그의 특성

상 긍정적인 글이 부정적인 글보다 많아서 나온 결과이기도 하지만 감정사전에 담겨 있지 않은 경우에도 해당된다. 극성 분류의 높은 성능 향상을 위해서는 감정사전의 단어들의 긍정, 부정 분류에 대한 정확도를 높일 수 있는 연구가 필요하다. 또한 인터넷 용어나 정확하지 않은 문장, 관용적 표현 등의 규칙을 판단할 수 있는 문장 규칙에 대한 연구가 필요하다.

References

- [1] Jong-Seok Song, Soo-Won Lee, "Automatic Construction of Positive/Negative Feature-Predicate Dictionary for Polarity Classification of Product Reviews", The Korean Institute of Information Scientists and Engineers: Software and Application, Vol. 38, No.3, March 2011.
- [2] Xiaowen Ding, Bing Liu, Phips S. Yu, "A Holistic Lexicon-Based Approach to Opinion Mining", Conference on Web Search and Data Mining, 2008
- [3] Sung-Ho Oh, Shin-Jae Kang, "Movie Retrieval System by Analyzing Sentimental Keyword from User's Movie Reviews", Journal of Korea Academia-Industrial cooperation Society, Vol. 14, No. 3 pp. 1422-142, July 2013.
- [4] <http://korean.abcthesaurus.com/>
- [5] Cheol-Seong Lee, Dong-Hee Choi, Seong-Soon Kim Jaewoo Kang, "Classification and Analysis of Emotion in Korean Microblog Texts", The Korean Institute of Information Scientists and Engineers: Database, Vol. 40, No.3, June 2013.
- [6] In-Jo Park, "The analysis of Korean affective terms : listing affective terms and exploring dimensions in the affective terms", Seoul National University, Psychology College, Master's thesis, 2001
- [7] In-Jo Park, Kyung-Hwan Min, "Making a List of Korean Emotion Terms and Exploring Dimensions Underlying Them", Korean Journal of Social and Personality Psychology, Vol. 19, No. 1, pp. 109~129, 2005.
- [8] Hong-June Yune, Han-Joon Kim, Jae-Young Chang "An Efficient Search Method of Product Reviews using Opinion Mining Techniques", The Korean Institute of Information Scientists and Engineers: CPL and Letter", Vol. 16, No. 2, February 2010.
- [9] Pavel Smrž., "Using WordNet for Opinion Mining," Proc. of the International WordNet Conference 2006, pp.333-335, 2006.
- [10] Andrea Esuli, Fabrizio Sebastiani, "PageRanking WordNet Synsets: An Application to Opinion Mining," Proc. of the Annual Meeting of the Association for Computational Linguistics 2007, pp.424-431, 2007.
- [11] Hong-Gu Choi, Een-Jun Hwang, "Emotion-based Music Recommendation System based on Twitter Document Analysis", The Korean Institute of Information Scientists and Engineers: CPL and Letter", Vol. 18, No. 11, November 2012.
- [12] Minqing Hu and Bing Liu, "Mining and Summarizing Customer Reviews." KDD'05, Seattle, Washington, USA., Aug.2004.
- [13] J. Kamps M. Marx, R. Mokken, and M. de Rijke., "Using WordNet to Measure Semantic Orientations of Adjectives" Proceedings of LREC, 2004.
- [14] Hatzivassiloglou V., Mackeown K., "Predicting the Semantic Orientation of Adjectives." Proceedings of the 8th Conference on European chapter of the association for Computational Linguistics, pp.174-181, 1997
- [15] Kwang-Seob Shim, Jae-Hyung Yang, "MACH : A Supersonic Korean Morphological Analyzer", Proceedings of the 19th International Conference on Computational Linguistics (COLING-2002), pp.939-945, 2002
- [16] Kwang-Mo Ahn, Yun-Suk Kim, and Young-Hoon Kim, Young-Hoon Seo, "Sentiment Classification of Movie Reviews using Levenshtein Distance", Journal of Digital Contents Society Vol. 14, No. 4., pp.

581-587, Dec. 2013



이 중 혁

2013년 : 숭실대학교
컴퓨터학부 학사졸업
2014년 : 숭실대학교
SW특성화 대학원 재학

2013년~현 재: 숭실대학교 SW특성화대학원
석사과정
관심분야 : 소프트웨어공학, 빅데이터, 데이터마이닝

최 재 현



2004년 2월 : 숭실대학교
컴퓨터학부(석사)
2011년 8월 : 숭실대학교
컴퓨터학부(박사)

2004년~현 재 : 숭실대학교 SW특성화대학원 교수
관심분야 : SW공학, SW프레임워크, 서비스 엔지니어링, 클라우드, 데이터마이닝



김 원 상

2013년 : 숭실대학교
컴퓨터학부 학사 졸업
2014년 : 숭실대학교
SW특성화 대학원 재학

2013년~현 재: 숭실대학교 SW특성화대학원
석사과정
관심분야 : 모바일 컴퓨팅, 데이터 마이닝,
소프트웨어 공학



박 제 원

2004년 2월 : 숭실대학교
컴퓨터학부(석사)
2011년 8월 : 숭실대학교
컴퓨터학부(박사)

2004년~현 재 : 숭실대학교 SW특성화대학원 교수
관심분야 : 소프트웨어테스팅, 소프트웨어프로세스,
웹서비스, SOA/ESB, 프로젝트관리