Increasing Accuracy of Classifying Useful Reviews by Removing Neutral Terms

Minsik Lee (Undergraduate Student, Dept. of Business Administration, Catholic University of Korea)
Hong Joo Lee (Associate Professor, Dept. of Business Administration, Catholic University of Korea)

[교신저자 연락처]
◎ 이홍주
- 주소: 14662, 경기도 부천시 원미구 지봉로 43 가톨릭대학교 경영학전공
- 전화번호: 02-2164-4009, Fax 번호: 02-2164-4280
- E-mail주소: hongjoo@catholic.ac.kr
The purpose of this study is to find a way of deleting words for creating document–term matrix in text mining. Since there are many documents, the size of document–term matrix is so large. Thus, researchers delete terms in terms of sparsity since sparse words have little effects on classifications or predictions. In this study, we propose neutrality index to select words to be deleted. Many words still appear in both classifications and these words have little or negative effects on classification performances. Thus we deleted neutral words which are appeared in both classifications similarly. After deleting sparse words, we selected words to be deleted in terms of neutrality.

We tested our approach with Amazon.com’s review data from five different product categories and compared the classification performances in terms of precision, recall, and F–measure. By selecting deleted words with neutrality index, we can best classification performances.

**Key Words**: Neutrality, Term Remove, Customer Review, Classification, Usefulness index
Ⅰ. 서 론

전자상거래의 성숙으로 거의 모든 상품들이 온라인으로 판매되고 있으며, 판매 제품을 이미 구매하여 사용한 고객의 리뷰가 구매 의사결정에 중요한 영향을 미치고 있다 (Dellarocas, Gao, & Narayan, 2010). 다른 고객이 제품을 사용한 후 자신의 경험에 근거하여 작성한 제품리뷰가 더 많은 정보를 제공하고, 객관적이며, 신뢰할 만하다고 생각하고 있다 (Dellarocas, 2003).

제품에 대한 고객 리뷰가 많이 작성되고 있기 때문에 구매하려는 제품의 모든 리뷰를 읽고 제품의 장단점을 파악하는 것은 무척 힘든 일이 되었다 (David and Pinch, 2006; Liu et al., 2008). 따라서, 전자상거래 업체들과 연구자들은 텍스트 마이닝을 활용하여 리뷰들 중에서 유용한 리뷰들의 속성을 파악하거나 유용한 리뷰와 유용하지 않은 리뷰를 미리 분류하는 노력을 수행하고 있다(Cao, Duan & Gan, 2011; Mudambi & Schuff, 2010). 또한, 리뷰 자체가 가지고 있는 리뷰길이, 작성자, 사용 단어 등의 속성을 활용하여 리뷰의 유용성에 미치는 영향을 분석하여왔다 (이상재, 최준연, 최진호, 2014).

텍스트 마이닝을 활용한 문서 분류의 다양한 문제에서는 텍스트의 전처리 과정에서 다양한 이유로 불필요한 단어들을 제거하게 된다. 어근추출(Stemming)을 통해 한 단어의 유사 표현을 하나의 단어로 단순화하고, 기호나 숫자, 불용어 혹은 사용자의 상황에 따른 불필요한 단어들을 제거한 후에 문서-단어 매트릭스를 생성한다 (Choeh, Lee and Park, 2015; Pak and Paroubek, 2010). 문서의 수가 많아질수록 추출되는 단어가 무척 많기 때문에 문서-단어 매트릭스의 차원을 줄이기 위한 다양한 노력들이 시도되어왔다. 기본적으로 단어의 희소성(Sparsity)에 기반하여 문서의 수에 비해 출현빈도가 현저히 적은 단어들을 분석에서 제거하거나 단어의 정보기여도를 산출하여 정보 기여가 낮은 단어들은 제거하는 방식이 사용되어왔다(Naji, 2013; Perkins, 2014).

본 연구에서는 문서-단어 매트릭스에서 단어의 제거 기준으로 온라인 고객 리뷰가 유용한 지, 그렇지 않은지를 구분하는 문제에서 단어들이 유용 리뷰 집합과 유용하지 않은 리뷰집합에 중복하여 등장하는 정도를 측정한 중립도를 제시한다. 제시한 중립도를 희소성과 함께 분석에 활용하여 제거할 단어를 선정한 후에 각 분류 알고리즘의 성과를 비교하였다. 최적의 성과를 보이는 중립도를 찾았으며, 희소성과 중립도에 따라 단어를 선택적으로 제거하였다.

실험은 Amazon.com의 5개 제품 분야 고객 리뷰와 사용자들의 리뷰에 대한 평가를 활용하였다. Information gain (Zhang and Tran, 2011)과 Support Vector
Machines (박성철, 김선웅, 최홍식, 2014; 이현욱, 안현철, 2011; 홍의석, 2011) 모두 F-measure 기준으로 5개 제품 카테고리에서 중립도와 회소성을 함께 활용한 방안이 전체 단어를 활용하거나 회소성만을 기준으로 단어를 제거할 방안보다 높은 성과를 보였다. 본 논문의 2장에서는 실험에 활용된 자료에 대해 소개하며, 3장에서는 분류방안과 결과를 제시한다. 4장에서는 본 논문의 의의와 향후 연구방향에 대해 논하였다.

II. 자료

본 연구에서는 Amazon.com에서 판매되는 상품에 대한 고객리뷰들을 활용하여 연구를 진행하였다. <그림 1>은 Amazon.com에 게재된 고객 리뷰의 한 사례이다. Amazon.com에서는 고객이 상품에 대한 리뷰를 작성하면 다른 고객들이 작성된 리뷰에 대해서 유용했는지와 유용하지 않았는지를 하단의 투표 버튼을 사용하여 평가할 수 있도록 만들어 두었다. <그림 1> 리뷰의 경우 47개의 총 득표수 중에서 47개가 유용하다고 평가된 리뷰이다. 연구에서는 1999년부터 2014년까지 등장한 ‘Cellphone & Accessories’, ‘Movies & TV Program’, ‘Automotive’, ‘CDs & Vinyl’, ‘Clothing, Shoes & Jewelry’ 5가지 제품 카테고리에 속한 리뷰를 분석하였다 (McAuley, Targett, Shi, and van den Hengel, 2015; McAuley, Pandey, and Leskovec, 2015).

![그림 1] Amazon.com 리뷰 예시

<그림 2>는 본 연구의 분류 알고리즘을 나타낸 그림이다. 연구에서 첫 번째 단계는 필터링 과정을 하는 것이다. 5개의 카테고리에 속한 리뷰 데이터 중에서 유용한 득표
의 숫자와 유용하지 않은 득표를 더한 전체 득표의 수가 4개 이상인 리뷰로 한정한다. 추출된 데이터에서 제품 카테고리별로 유용하다고 판단되는 1,500개의 리뷰와 유용하지 않다고 판단되는 1,500개의 리뷰를 무작위로 추출하여 연구에 사용하였다. 본 연구에서 정의하는 유용한 리뷰와 유용하지 않은 리뷰는 유용득표 숫자가 전체 득표수의 60%를 초과하는 경우를 유용한 리뷰라고 정의하고 60%를 초과하지 못하는 경우 유용하지 않은 리뷰라고 정의했다 (Zhang and Tran, 2011).

본 연구에서 사용된 통계 프로그램은 R이다. 수집된 데이터를 전처리 하는 과정으로 R의 tm (Feinerer, Hornik, and Meyer, 2008) 패키지를 사용했으며 모델링 구축 과정에서는 e1071 (Meyer, 2015) 패키지를 사용했다. 전처리 과정에서 리뷰에 등장하는 숫자, 특수기호와 불용어(stopword)는 제거하였으며 어근 추출을 수행하였다. 5개의 데이터 종류에서 3,000개 리뷰의 단어들을 추출하여 문서-단어 매트릭스를 만들었다. ‘Cellphones & Accessories’ 데이터의 경우 10,420개, ‘Movies & TV program’ 데이터의 경우 21,770개, ‘Automotive’ 데이터의 경우 9,580개, ‘CDs & Vinyl’ 데이터의 경우 19,544개, ‘Clothing, Shoes & Jewerly’ 데이터의 경우 8,671개의 단어가 추출되었다 (표 1 참조).

생성된 문서-단어 매트릭스를 본 연구에서는 두 가지 ‘단어 제거’ 과정을 수행했다. 첫 번째 방법인 ‘Delete Sparse terms’는 희소한 단어들을 제거하는 것이다.

| Category                  | All terms | Remaining terms |
|---------------------------|-----------|-----------------|
| Cellphones & Accessories | 10,420    | 3,668           |
| Movies & TV program       | 21,770    | 8,290           |
| Automotive                | 9,580     | 3,810           |
| CDs & Vinyl               | 19,544    | 7,105           |
| Clothing, Shoes & Jewerly| 8,671     | 3,485           |

이방식의 목적은 오타로 인해 추출된 단어를 제거하거나 매우 적게 등장하는 단어를 제거하기 위한 것이다. 희소한 단어들을 제거하기 위해 설정한 절삭 값(threshold)은 0.1로써 문서 수 대비 단어의 등장 횟수가 0.1% 미만인 희소한 단어를 제거하였다.
희소성 기반의 단어들을 제거한 결과 ‘Cellphones & Accessories’ 데이터의 경우 6,752개, ‘Movies & TV program’ 데이터의 경우 13,480개, ‘Automotive’ 데이터의 경우 5,770개, ‘CDs & Vinyl’ 데이터의 경우 12,439개, ‘Clothing, Shoes & Jewelry’ 데이터의 경우 5,186개의 단어가 추출되었다 (표 2 참조). 많게는 66%에서 적게는 57%까지 희소한 단어가 감소하며 평균적으로 62% 단어가 감소하는 결과가 나왔다.

두 번째 방법인 ‘Delete Neutral terms’는 중립성을 기준으로 단어들을 제거하는 것이다. 이 방식의 목적은 분류에 영향을 주는 단어가 아닌 데이터의 종류별 속성에 따라서 자주 등장하는 단어들을 제거하기 위한 것이다. <그림 4>는 ‘Movies & TV program’ 데이터의 c, m으로 시작하는 중립 단어, <그림 5>는 ‘Cellphones & Accessories’ 데이터의 c, m으로 시작하는 중립 단어이다. <그림 4>에서는 ‘Movies & TV program’의 리뷰 데이터기 때문에 영화와 관련된 ‘movi’, ‘moviethi’의 단어가 등장했다. 반면에 <그림 5>는 ‘Cellphones & Accessories’의 리뷰 데이터에서 휴대폰과 관련된 ‘cell’, ‘cellpho’의 단어가 등장했다.
<그림 4> Movies & TV Program의 중립 단어

| acdog | cat | caveat | chiron | chronolog | celelegu | constant | crisis | mar | mer | more |
|------|-----|--------|--------|-----------|----------|----------|--------|-----|----|-----|
| califou | celeb | chucki | columbia | construct | cristina | marathon | metal | most |
| calib | centari | dung | comb | consumpt | crocodili | marc | michael | motion |
| california | certain | chunk | come | contain | crux | marguerit | middli | motorcycl |
| call | certifi | church | comedi | contempori | crunch | marian | midway | mountain |
| came | chair | ca | comedian | content | csi | marilla | mild | movi |
| camel | chaplin | cinema | comet | controversi | cuba | marilyn | million | moviethi |
| cameraworl | chappell | circl | comic | conveni | cue | marion | mindi | much |
| camouflag | charact | circuit | command | cool | curious | market | mini | med |
| cancer | charad | civilian | complet | coppola | curtain | marvel | mininelli | multfacet |
| carn | charisma | clerk | completist | copyright | custodi | mass | mirror | multisysten |
| cannot | chari | clow | comprehen | core | cute | may | mirror | multivers |
| canva | chase | clay | conceiv | comi | cynic | mayb | misfit | munich |
| capit | cheaper | clearanc | concentr | cicon | macabr | mcaidan | mishnish | muki |
| cardin | check | cleer | concept | correct | maden | mccartney | mission | most |
| care | cheeki | chich | condernn | could | makeup | mcklinney | misunderstit | mutat |
| caro | child | clichaceut | confes | counter | male | mcqueen | mitchum | mite |
| carolina | chock | climar | conflict | coupl | malibu | mcsheane | mom | myer |
| cartoon | choic | clockwork | confus | coward | malign | neart | momentun | |
| cary | choppli | closeup | congressm | cowork | manag | measur | monkey |
| casino | chris | cowen | connect | crank | manifest | necha | monioagu | |
| casper | christ | cluchi | conquer | crush | mann | media | monotoni | |
| cassidi | christensen | cocknach | conquest | crass | marnequin | medoocr | monster |
| castle | christian | codi | consequ | creatur | manor | megatron | montag |

<그림 5> Cellphones & Accessories의 중립 단어

### III. 분류 방안 및 결과

1. 중립도 기반 단어 제거 방안

중립도 기반 단어 제거는 기본적으로 희소성에 기반하여 희소한 단어를 제거한 후에 중립도를 기반으로 두 집합에 모두 속하는 단어를 제거하는 방안이다. 본 연구에
서 제안하는 중립도를 구하는 식은 (1)과 같다. \( HT_i \)는 단어 \( i \)가 등장한 유용한 리뷰의 수이며, \( H\overline{T_i} \)는 단어 \( i \)가 등장한 유용하지 않은 리뷰의 수이다. 즉, 한 단어가 유용한 리뷰 집합과 유용하지 않은 리뷰 집합에 모두 등장하는 정도를 중립도(Neutrality Index, \( N_i \))라고 하고 다음과 같이 정의한다. 리뷰가 유용 리뷰인지 유용하지 않은 리뷰인지에 관계없이 전체 투표수에서 유용하다는 투표수가 60%이상인 경우를 유용한 리뷰로 설정하였다.

\[
N_i = \begin{cases} 
\frac{HT_i}{HT_i} & (HT_i > HT_i) \\
\frac{H\overline{T_i}}{HT_i} & \text{otherwise}
\end{cases}
\quad (1)
\]

모든 단어들의 중립도를 구하여 일정 절삭 값(threshold)에 포함하는 단어들을 중립 단어라고 판단하여 제거하는 과정을 거쳤다. 최적의 절삭 값을 구하는데 있어서 각 데이터와 알고리즘 별로 최적의 중립 단어 제거 지점을 최적 중립도(Best Neutrality Index, BNI)를 찾는 과정을 진행했다. (<그림 6>, <그림 7>, <표 2>, <표 3>)
| Cellphones & Accessories | Movies & TV program |
|-------------------------|---------------------|
| ![Image](89x490 to 261x611) | ![Image](266x484 to 442x609) |
| ![Image](85x335 to 256x456) | ![Image](268x332 to 446x458) |
| ![Image](84x181 to 261x306) | ![Image](255x44) |

> <그림 6> Information Gain 최적 중립도 (단위: F-measure)
| Category                        | Best Neutrality Index | Remaining terms |
|--------------------------------|-----------------------|-----------------|
| Cellphones & Accessories       | 0.4                   | 1,643           |
| Movies & TV program            | 0.5                   | 4,244           |
| Automotive                     | 0.6                   | 2,390           |
| CDs & Vinyl                    | 0.4                   | 3,301           |
| Clothing, Shoes & Jewelry      | 0.5                   | 1,633           |

<표 2> Information gain 최적 중립도

<그림 6>은 Information Gain 알고리즘을 사용하여 찾아낸 최적 중립도이다. 그래프의 X축은 중립도로 오른쪽으로 갈수록 중립 단어 선정 범위가 넓어져 제거되는 중립 단어가 많아진다. Y축은 F값으로 위로 상승할수록 값이 올라갈을 나타낸다. Information Gain을 적용한 모든 그래프에서 단어를 줄일수록 F값이 상승하는 모습을 보이다가 다시 감소하는 형태를 보인다. F값이 가장 높은 지점은 최적 중립도라고 판단하여 중립 단어들을 제거한 결과 ‘Cellphones & Accessories’ 데이터의 경우 2,025개, ‘Movies & TV program’ 데이터의 경우 4,046개, ‘Automotive’ 데이터의 경우 1,420개, ‘CDs & Vinyl’ 데이터의 경우 3,804개, ‘Clothing, Shoes & Jewelry’ 데이터의 경우 1,852개의 단어가 추가로 제거되었다. 적게는 37%에서 많게는 55% 단어 제거를 했다.

| Category                        | Best Neutrality Index | Remaining terms |
|--------------------------------|-----------------------|-----------------|
| Cellphones & Accessories       | 0.7                   | 2,881           |
| Movies & TV program            | 0.5                   | 4,223           |
| Automotive                     | 0.5                   | 1,711           |
| CDs & Vinyl                    | 0.8                   | 5,821           |
| Clothing, Shoes & Jewelry      | 0.7                   | 2,883           |

<표 3> SVM 최적 중립도
| Cellphones & Accessories | Movies & TV program |
|--------------------------|---------------------|

| Automotive               | CDs & Vinyl         |

| Clothing, Shoes & Jewerly|                       |

<그림 7> SVM 최적 중립도 (단위: F-measure)

<그림 7>은 <그림 6>과 동일한 방식으로 SVM의 알고리즘을 사용하여 찾아낸 최적 중립도를 찾아내었다. SVM의 방식도 모든 그래프에서 단어를 줄일수록 F-값이 상승하는 모습을 보이며 다시 감소하는 형태를 보인다. ‘Cellphones & Accessories’ 데이터의 경우 787개, ‘Movies & TV program’ 데이터의 경우 4,067개, ‘Automotive’ 데이터의 경우 2,099개, ‘CDs & Vinyl’ 데이터의 경우 1,284개,
Clothing, Shoes & Jewerly’ 데이터의 경우 602개의 단어가 추가로 제거되었다. 적게는 17%에서 많게는 55% 추가 단어 제거를 하였다.

같은 데이터에서 2개 알고리즘을 사용하여 최적 중립도 지점을 확인해본 결과 알고리즘에 따라서 최적 중립도가 다른 것을 확인했다. 예를 들어 동일한 ‘Cellphones & Accessories’ 데이터에서 Information Gain의 최적 중립도는 0.4, SVM의 최적 중립도는 0.7이 나온다. 또한 같은 알고리즘에서 5개 제품 카테고리 데이터를 사용하여 최적중립도 지점을 확인해본 결과 찾아낸 결과 최적 중립도가 다른 것을 확인했다. 예를 들어 동일한 Information Gain의 알고리즘에서 찾아낸 5개 데이터 집합의 최적 중립도는 0.4(Cellphones & Accessories, CDs & Vinyl), 0.5(Movies & TV program, Clothing, Shoes & Jewerly), 0.6(Automotive)이었다.

2. 분류 결과

축소된 데이터를 SVM, Information Gain 알고리즘에 적용하여 분류 결과의 Recall, Precision, F-값을 측정하고, 전체 단어를 사용한 경우와 회소성만을 기준으로 단어를 제거한 경우와 분류 성과를 비교하였다.

표 4는 3000개의 리뷰를 무작위로 70%는 학습 집합으로 30%는 테스트 집합으로 나누어 30번 반복 수행을 실시한 결과다. 제시된 수치는 각 알고리즘을 통한 분류 결과의 Recall, Precision, F-값이다. All terms은 추출된 모든 단어를 바탕으로 예측한 결과값이다. Sparsity는 회소성을 기준으로 단어를 제거한 후에 알고리즘에 적용한 결과이다. Neutrality + Sparsity는 회소성에 근거하여 단어를 제거한 후 중립성에 기반을 두어 추가적으로 단어를 제거한 집합이 알고리즘에 적용된 결과이다.

전체 단어를 사용하는 것에 비해서 회소성을 기반으로 단어를 제거한 후에 활용하는 것이 SVM 알고리즘에서는 더 좋은 성과(F-measure)를 보였지만, Information Gain 알고리즘에서는 그렇지 않았다. 값이 근사하기는 하지만 단어별 중요도를 활용하는 Information Gain 알고리즘에서는 전체 데이터를 활용하는 경우가 회소성 기준으로 단어를 제거한 후에 활용하는 경우보다 분류 정확도가 높았다. 하지만 본 연구에서 제안하는 회소성과 중립성을 고려한 단어를 제거하는 방법은 제품 카테고리와 알고리즘과 관계없이 모두 가장 높은 F-값을 보였다.
<표 4> 분류 정확도

| Category       | Methods                      | Information Gain                  | SVM                |
|---------------|------------------------------|-----------------------------------|--------------------|
|               |                              | Precision | Recall | F -measure | Precision | Recall | F -measure |
| Cellphone & Accessories | All terms        | 0.5053    | 0.9964 | 0.6705     | 0.7074    | 0.3602 | 0.4769     |
|               | Sparsity             | 0.5047    | 0.9972 | 0.6702     | 0.6414    | 0.6159 | 0.6282     |
|               | Neutrality + Sparsity | 0.5397    | 0.9344 | 0.6841     | 0.631     | 0.6264 | 0.6285     |
|               | F statistics (p-value) | 961.585   | 466.081 | 112.723 (0.000) | 171.090 (0.000) | 171.090 (0.000) | 549.357 (0.000) |
| Movies & TV Program | All terms | 0.5075    | 0.9967 | 0.6705     | 0.6758    | 0.3737 | 0.4810     |
|               | Sparsity             | 0.5067    | 0.9970 | 0.6702     | 0.64      | 0.5690 | 0.6022     |
|               | Neutrality + Sparsity | 0.5327    | 0.969  | 0.6841     | 0.6288    | 0.6064 | 0.6172     |
|               | F statistics (p-value) | 411.928   | 372.658 | 112.723 (0.000) | 66.3326 (0.000) | 1179.80 (0.000) | 699.954 (0.000) |
| Automotive    | All terms | 0.5115    | 0.9871 | 0.6739     | 0.6389    | 0.3486 | 0.4506     |
|               | Sparsity             | 0.5108    | 0.9889 | 0.6736     | 0.6152    | 0.5883 | 0.6013     |
|               | Neutrality + Sparsity | 0.5248    | 0.964  | 0.6793     | 0.5956    | 0.6183 | 0.6066     |
|               | F statistics (p-value) | 108.717   | 76.9284 | 21.4104 (0.000) | 39.7933 (0.000) | 1249.79 (0.000) | 610.499 (0.000) |
| CDs & Vinyl   | All terms | 0.51      | 0.9950 | 0.6745     | 0.676     | 0.3972 | 0.5002     |
|               | Sparsity             | 0.5088    | 0.9965 | 0.6740     | 0.6452    | 0.6607 | 0.6527     |
|               | Neutrality + Sparsity | 0.5580    | 0.9459 | 0.7024     | 0.6446    | 0.6695 | 0.6566     |
|               | F statistics (p-value) | 692.332   | 490.331 | 282.266 (0.000) | 38.5171 (0.000) | 1217.64 (0.000) | 716.847 (0.000) |
| Clothing, Shoes & Jewelry | All terms | 0.536     | 0.9703 | 0.6905     | 0.6489    | 0.3461 | 0.4510     |
|               | Sparsity             | 0.5336    | 0.9716 | 0.6888     | 0.6194    | 0.5859 | 0.6020     |
|               | Neutrality + Sparsity | 0.572     | 0.8844 | 0.6946     | 0.6172    | 0.6019 | 0.6093     |
|               | F statistics (p-value) | 234.792   | 545.061 | 5.6308 (0.005) | 23.9686 (0.000) | 1175.00 (0.000) | 653.872 (0.000) |

IV. 결론

데이터 집합에 따라 정확도 개선 정도가 상이하며, F-measure 기준으로는 두 알고리즘에서 모두 희소성과 중립도에 기반하여 단어를 제거하는 방안이 더 성과가 높았
다. 하지만 Information Gain 알고리즘에서는 Recall 기준으로는 5개 제품 카테고리 데이터에서 언제나 회소성을 기준으로 단어를 제거하는 방안의 성과가 높았으며, SVM에서는 전체 단어를 활용하는 방안이 Precision 기준으로 성과가 더 높았다. 따라서, 활용하는 알고리즘과 분석 목적에 따라서 단어 제거 방안을 고려하는 것이 필요하다.

회소성을 기준으로 단어를 제거한 후에 중립도를 기준으로 10% 정도의 추가 단어 제거를 통해서 성과를 개선할 수 있다는 것을 확인하였다. 추가적으로 더 많은 데이터 집합과 회소성, 중립성 기준의 적용을 통해 적합한 활용기준을 정하는 것이 필요하다.

참고문헌

박성철, 김선웅, 최홍식, “SVM을 이용한 시스템트레이딩전략의 선택모형”, 제20권, 제2호, 2014, pp. 59-71.
이상재, 최준연, 최진호, “온라인 리뷰의 경제적 효과, 유용성과 유용성 투표수에 영향을 주는 결정요인,” 한국IT서비스학회지, 제13권, 제1호, 2014, pp. 43-55.
이현욱, 안현철 “비정상 오류비용을 고려한 분류기준값 최적화와 SVM에 기반한 지능형 침입탐지모형”, 지능정보연구, 제17권, 제4호, 2011, pp. 157-173.
홍의석, “Support Vector Machine을 이용한 초기 소프트웨어 품질 예측”, 한국IT서비스학회지, 제10권, 제12호, 2011, pp. 235-245.
Cao, Q., Duan, W., and Gan, Q., “Exploring determinants of voting for the ‘helpfulness’ online userreviews: A text mining approach,” Decision Support Systems, Vol. 50, No. 2, 2011, pp. 511-521.
Choeh, J. Y., Lee, H. J. and Park, S. J., “A Personalized Approach for Recommending Useful Product Reviews Basedon Information Gain,” KSII Transactions on Internet and Information Systems, Vol. 9, No. 5, 2015, pp. 1702-1716.
David, S., and Pinch, T., "Six Degrees of Reputation: The Use and Abuse of Online Review and Recommendation Systems," First Monday, Vol. 11, No. 3, 2006, Available at http://dx.doi.org/10.5210/fm.v11i3.1315.
Dellarocas, C., “The Digitization of Word of Mouth: Promise and Challenges of Online Feedback Mechanisms”, Management Science, Vol. 49, No.10,
Dellarocas, C., Gao, G., and Narayan, R., “Are consumers more likely to contribute online reviews for hit or niche products?”, Journal of Management Information Systems, Vol. 27, No. 2, 2010, pp. 127–157.

Feinerer, I., Hornik, K. and Meyer, D., “Text Mining Infrastructure in R,” Journal of Statistical Software, vol. 25, no.5, 2008, pp. 1–54.

Liu, Y., Huang, X., An, A., and Yu, X., “Modeling and Predicting the Helpfulness of Online Reviews”, Proceedings of the Eighth IEEE International Conference on Data Mining, 2008, pp. 443–452.

McAuley, J., Targett, C., Shi, J., and van den Hengel, A., “Image-based recommendations on styles and substitutes”, SIGIR, 2015

McAuley, J., Pandey, R., and Leskovec, J., “Inferring networks of substitutable and complementary products”, Knowledge Discovery and Data Mining, 2015

Naji, I., “10 Tips to Improve your Text Classification Algorithm Accuracy and Performance,” accessed at http://thinknook.com/10-ways-to-improve-your-classification-algorithm-performance-2013-01-21/

Pak, A. and Paroubek, P., “Twitter as a Corpus for Sentiment Analysis and Opinion Mining”, LREc, Vol. 10, 2010

Perkins, J., Python 3 Text Processing with NLTK 3 Cookbook, Packt Publishing, 2014

Zhang, R. and Tran, T., “An Information gain-based approach for recommending useful product reviews,” Knowledge and Information Systems, vol. 26, no. 3, 2011, pp. 419–434.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F., “e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071)”, TUWien. R package version 1.6–7. https://CRAN.R-project.org/package=e1071, 2015.