# 나이브 베이즈 기반 소셜 미디어 상의 신조어 감성 판별 기법

## (Sensitivity Identification Method for New Words of Social Media based on Naive Bayes Classification)

김정인*, 박상진**, 김형주***, 최준호****, 김한일*****, 김판구******

(Jeong In Kim, Sang Jin Park, Hyoung Ju Kim, Jun Ho Choi, Han Il Kim, Pan Koo Kim)

**요 약**

인터넷의 발달과 스마트폰의 보급으로 인하여 그에 따른 소셜 미디어 문화가 형성됨에 따라 PC통신부터 지금까지 소셜 미디어 신조어가 그 문화로 자리 잡아가고 있다. 소셜 미디어의 등장과 사람들의 가교역할을 해주는 스마트폰의 보급화로 신조어가 생기고 빈번하게 사용되고 있는 추세이다. 신조어의 사용은 다양한 문자 제한 메신저의 문제점을 해결하고 짧은 문장을 사용하여 데이터를 줄이는 등 많은 장점을 가지고 있다. 그러나 신조어에는 사전적인 의미가 없으므로 데이터 마이닝 기술이나 빅데이터와 같은 연구에서 사용되는 알고리즘의 성능 저하와 연구에 제약사항이 발생한다. 따라서 본 논문에서는 웹 크롤링을 통해 텍스트 데이터를 추출하고, 텍스트 마이닝과 오피니언 마이닝을 통해 의미부여 및 단어들에 대한 감정적 분류를 통한 문장의 오피니언 파악을 진행하고자 한다. 실험은 다음과 같이 3단계로 진행하였다. 첫째, 소셜 미디어에서 새로운 단어를 수집하여 수집된 단어는 긍정적이고 부정적인 학습을 받게 하였다. 둘째, 표준 문서를 사용하여 감정적 가치를 도출하고 검증하기 위해 TF-IDF를 사용하여 데이터의 감정적 가치를 측정하기 위해 명사 빈도수를 측정한다. 신조어와 마찬가지로 분류된 감정적 가치가 적용되어 감정이 표준 언어 문서로 분류되는지 확인하였다. 마지막으로, 새로 합성된 단어와 표준 감정적 가치의 조합을 사용하여 장비 기술의 비교분석을 수행하였다.

■ 중심어 : 신조어 ; 감성 판별 ; 나이브 베이즈 분류 ; 텍스트 마이닝 ; 오피니언 마이닝

**Abstract**

From PC communication to the development of the internet, a new term has been coined on the social media, and the social media culture has been formed due to the spread of smart phones, and the newly coined word is becoming a culture. With the advent of social networking sites and smart phones serving as a bridge, the number of data has increased in real time. The use of new words can have many advantages, including the use of short sentences to solve the problems of various letter-limited messengers and reduce data. However, new words do not have a dictionary meaning and there are limitations and degradation of algorithms such as data mining. Therefore, in this paper, the opinion of the document is confirmed by collecting data through web crawling and extracting new words contained within the text data and establishing an emotional classification. The progress of the experiment is divided into three categories. First, a word collected by collecting a new word on the social media is subjected to learned of affirmative and negative. Next, to derive and verify emotional values using standard documents, TF-IDF is used to score noun sensibilities to enter the emotional values of the data. As with the new words, the classified emotional values are applied to verify that the emotions are classified in standard language documents. Finally, a combination of the newly coined words and standard emotional values is used to perform a comparative analysis of the technology of the instrument.

■ keywords : New Words ; Sensitivity Identification ; Naive Bayes Classification ; Text Mining ; Opinion Mining

## Ⅰ. 서 론

인터넷의 발달과 스마트폰의 보급화로 소셜 미디어 문화가 형성되면서 사용자들 또한 소셜 미디어 문화에 발맞춰 나아가고 있다. 이러한 소셜 미디어 문화에서 가장 중요한 건 소통이다. 인터넷 보급화 초기 시절 PC 통신 게시판이나 채팅에서는 특수문자를

* 정회원, 조선대학교 컴퓨터공학과 박사후 연구원
** 정회원, 조선대학교 소프트웨어융합공학과 석사
*** 정회원, 조선대학교 컴퓨터공학과 박사과정
**** 정회원, 조선대학교 자유전공학부 부교수
***** 정회원, 제주대학교 컴퓨터교육과 교수
****** 정회원, 조선대학교 컴퓨터공학과 교수

이 논문은 2019년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2019S1A5A2A03049825).

접수일자 : 2020년 03월 09일
게재확정일 : 2020년 03월 17일

교신저자 : 김판구 e-mail : pkkim@chosun.ac.kr

사용하여 사용자의 감정이나 느낀 점 등을 나타내어 소통하는 것이 전부였으나 소셜 미디어의 등장과 사용자들의 가교역할을 해주는 스마트폰의 보급화로 신조어가 생기고 빈번하게 사용하고 있는 추세다. 신조어의 등장은 소셜 미디어 주 사용층인 10대와 20대 사용자들이 많이 사용하면서 다양한 신조어가 등장하고 사용자들의 이용을 통해서 또 다른 신조어를 탄생시키고 발전되어지고 있다. 신조어의 사용은 여러 장단점을 가져올 수 있는데 가장 큰 장점은 빠른 의미전달이다. 소셜 미디어가 지닌 문제점인 글자 수 제한이라는 것을 신조어 사용으로 긴 문장을 짧은 문장으로 또는 긴 단어를 짧은 단어로 함축적인 의미전달이 가능하여 글자 수 제한 문제점 해소가 가능하다. 하지만 데이터 마이닝(Data Mining) 기술이나 빅 데이터(Big Data) 같은 연구에서는 사전적인 의미를 갖고 있지 않아 알고리즘의 성능 저하와 연구에 제약사항이 발생한다. 본 논문에서는 소셜 미디어 신조어 분석을 통해 신조어에 대한 감성 판별 연구를 진행하고자 한다. 위키피디아(Wikipedia)에 구축되어 있는 소셜 미디어 신조어 목록의 단어들과 문서에서 신조어와 동시에 사용하는 표준어 단어들을 수집하여 의미를 통한 긍정 값과 부정 값의 분류를 진행한다. 분류를 통해 추출된 긍정, 부정 단어들은 파이썬(Python)을 통해 학습시킨 후 구축된 데이터를 이용하여 데이터 마이닝 기법의 성능을 향상시키는 방법을 제안하고 기존의 연구와 비교 실험을 통해 성능평가를 진행한다.

## Ⅱ. 관련 연구

### 1. 신조어 형성 원리

소셜 미디어 문화 중 하나인 신조어는 소셜 미디어 내에서는 물론이고 상당수가 현실 세계에서도 사용한다. 신조어들은 대부분 유행에 예민하여 빠르게 확산되어 타올랐다가 식어버리는 취약부분이 존재하여 단어의 의미 이외에 구조나 탄생에 대해서는 의의가 필요하지 않다. 하지만 신조어의 형성과 탄생은 흥미로운 부분이고, 단어가 유행하는 기간에는 단어에 대한 평가가 필요하다.

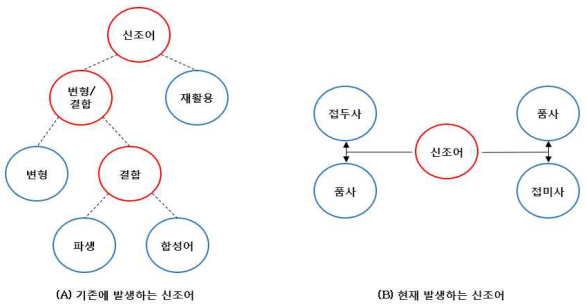


그림 1. 신조어 형성과정

신조어는 '동사+형용사', '형용사+명사', '명사+명사'나 반대의 조합으로 이루어져 있거나 긴 단어를 줄여 줄임말 형태의 단어로 불규칙적인 형태로 구성되어 있다[1, 2]. 또한, 해당 시대의 유행하는 어떤 사물이나 사람의 형태를 가져와 탄생하는 단어도 존재한다. 보통의 경우 유행하는 단어들은 특정 커뮤니티에서 사용하다가 점차 퍼져나가 오랫동안 사용자들에 의해 쓰이게 되면 관용어처럼 탄생한다. 이렇게 탄생한 신조어는 진화하거나 변형시켜 또 다시 재탄생하게 된다.

### 2. 텍스트 마이닝(Text Mining)

데이터 마이닝 기술 중 하나인 텍스트 마이닝은 자연어 처리와 정보 추출 등의 분야를 연구하는데 유용한 기술 중 하나이다. 소셜 미디어에서 흔히 찾을 수 있는 데이터들은 구조가 완전하지 않는 형태로 구성되어 있고 가공되지 않은 데이터로 그 안에서 불분명한 형태 안에 필요한 키워드 추출하는 작업은 중요하다[3].

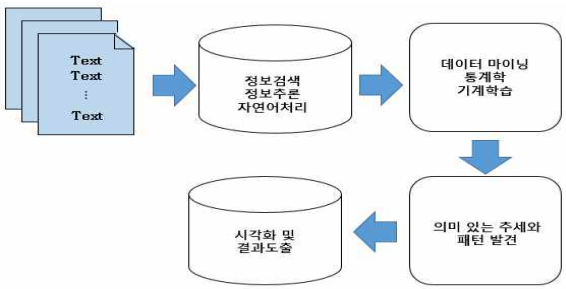


그림 2. 텍스트 마이닝 과정

텍스트 마이닝 과정은 대규모 문서나 유저를 통해 실시간으로 생성되거나 비정형적 데이터 안에서 텍스트의 특징 정보를 추출하여 키워드 형태로 표현하고 텍스트 간의 유사도를 확인하여 군집화하고, 정형화된 데이터는 새로운 정보를 생성하고 찾고자 하는 패턴이나 키워드를 찾는 과정으로 진행한다[4]. 소셜 미디어 상의 신조어 추출은 텍스트 마이닝에서 이용하는 TF-IDF를 이용하여 추출한다[5]. 추출된 신조어는 단순히 의미만을 갖고 있는 것이 아니기 때문에 위키피디아에서 제공하는 '대한민국 인터넷 신조어' 목록을 통해 임의로 긍정, 부정 분류를 진행하고 분류된 데이터는 오피니언 마이닝을 통해 학습하여 사용한다[6].

### 3. 오피니언 마이닝(Opinion Mining)

오피니언 마이닝은 자연 언어 처리, 텍스트 분석 등을 사용해

정서적인 상태와 주관적인 정보를 체계적으로 식별하고 추출 및 정량화를 연구하는 기술이다. 오피니언 마이닝은 특정 주제나 상황에 따라 감성에 따라 반영해야 하기 때문에 사용자가 어떤 의도를 갖고 있는지 분석하여 처리하는 것이 중요하다[7].

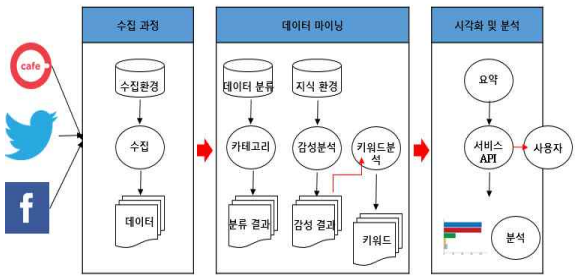


그림 3. 오피니언 마이닝 절차

오피니언 마이닝 과정은 3단계로 분류하여 진행된다. 첫 번째로 웹 크롤링을 통한 데이터 수집이다. 두 번째로 수집된 데이터 내에 저장되어 있는 텍스트 내에 평가요소와 긍정, 부정을 가리키는 오피니언 관계가 있는 문장을 인식하고 불필요한 특수문자와 감성과 관계없는 품사를 제외한다. 마지막으로 두 과정을 통해 긍정, 부정을 의미하는 단어들이 담긴 특정 주제에 관한 텍스트 데이터를 요약을 통해 분석과 평가한다. 오피니언 마이닝 과정을 통해서 소비자들이 제품에 대해 평가한 오피니언 정보를 판단할 수 있다. 또한, 오피니언 마이닝을 통해 분석된 객관적인 자료로 인하여 제품의 신뢰성을 향상시키고 호감도를 판단해 제품의 구매 판단 여부를 확인할 수 있다[8, 9, 10].

본 논문에서는 신조어가 단순히 불용어 처리되는 것이 아니라 신조어에 대한 의미와 감성이 적용되어 알고리즘 성능 저하를 막고 특정 주제에 관련하여 구체적인 감성 판별 연구가 진행될 수 있도록 하는 연구를 진행하였다.

## Ⅲ. 신조어 감성 판별 기법

### 1. 감성 판별 시스템 구성

본 절에서는 대량의 텍스트 데이터와 언어의 폭이 넓어짐에 따라서 다량의 소셜 미디어 신조어가 발생하는 댓글이나 리뷰를 수집한 뒤, 신조어의 긍정, 부정 감성 판별을 진행한다. 그림 4는 텍스트 데이터에서 신조어 추출 후 신조어의 긍정, 부정 감성 판별을 진행하는 시스템 구성도이다. 본 논문에서 제안하는 시스템 구성도는 국내 포털사이트에서 제공하는 카페나 블로그 데이터를 위주로 수집하고, 수집된 데이터들은 신조어 추출을 위해 '한나눔' 한국어 형태소 분석기를 이용하여 진행한다[11- 12]. 기존의 긍정, 부정 감성 분석에 대한 연구에서는 리뷰나 소셜 미디어 댓글의 어휘에 한정하여 통계적 수치나 자연어 처리 기법을 사용하였으나 문맥에 따라 감성어의 의미가 다르게 분류되는 경우가 발생하거나 신조어에 의해 문장의 의미가 퇴색되고 정확한 결과 값을 찾아내기 어렵다는 문제가 발생하였다[13-14]. 본 연구에서는 이러한 문제점들을 보안하기 위해 신조어의 긍정, 부정 분류를 토대로 실험을 진행하였다.

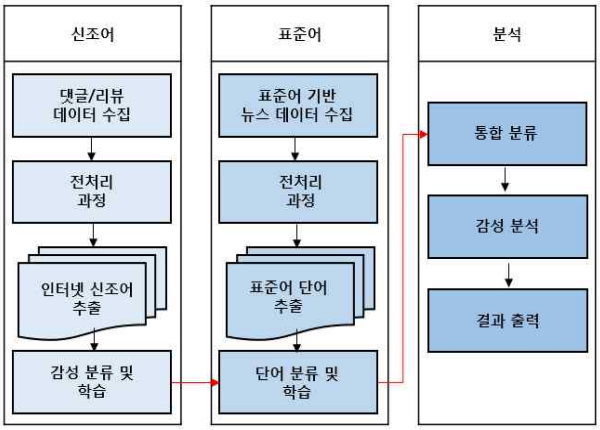


그림 4. 신조어 감성 판별 시스템 구성도

### 2. 신조어 긍정 부정 감성 분석

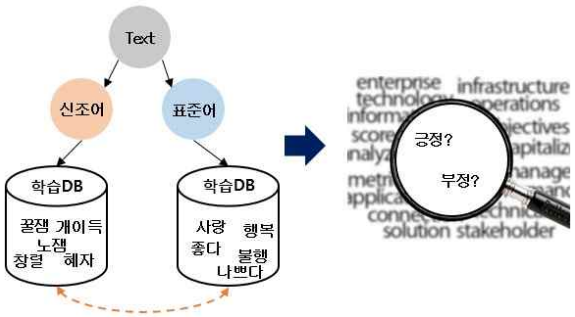


그림 5. 신조어와 표준어 긍정, 부정 분류 판별 과정

신조어를 추출하기에 앞서서 예외적인 부분은 소셜 미디어 신조어 중 대부분 접두사 '개'를 붙여 사용하는 경우가 많기 때문에 그러한 단어들은 욕설이 아닌 신조어로 판별하고 분류하였으나 지역비하, 비속어, 인신공격 등 오해의 소지를 불러일으킬 수 있는 단어들은 추출 후 제거하였다. 소셜 미디어 상의 댓글, 리뷰를 통해 수집된 데이터를 토픽별로 문서에 저장하여 분석을 한다. 신조어는 품사로 분류 시 의미가 없는 단어이므로 형태가 분명치 않아지기 때문에 문장 내 공백 혹은 키워드를 기준으로 나누기 위해 토크나이징(Tokenizing) 기법을 이용하고, 품사 태깅(POS tagging)을 진행하지 않는다. 표 1은 소셜 미디어 데이터의 토크나이징 적용 전후 데이터를 나타낸다.

표 1. 소셜 미디어 데이터의 토크나이징

| | 댓글/리뷰 원문 |
|---|---|
| **원본 데이터** | 1. 레알 창렬 오브 창렬 ㅋㅋㅋㅋㅋㅋㅋㅋㅋ. 이게 과연 5만원 상인가 싶어서 웃음만 나오더라고요<br>2. 야 어벤져스 그전에 마블영화 하나도 안보고보면 노잼이야<br>3. 솔직히 이렇게 욕하는거 드물잖아 아무리 개존못이라도ㅋㅋ<br>⋮ |
| **토크 나이징 적용 후 데이터** | 1. 레알, 창렬, 오브, 창렬, 이게, 과연, 5만원, 상인가, 싶어서,웃음만, 나오더라고요,<br>2. 야, 어벤져스, 그, 전에, 마블 영화, 하나도, 안, 보고보면, 노잼, 이야,<br>3. 솔직히, 이렇게, 욕, 하는거, 드물잖아, 아무리, 개존못, 이라도<br>⋮ |

소셜 미디어 신조어를 판별하기 위해서 불용어를 제거하고 인터넷 검색 시 검색 용어로 사용하지 않는 단어로써 관사, 접속사 같은 품사 등은 의미가 없는 단어이지만 포털 사이트의 검색 엔진마다 동일하지 않기 때문에 다르다. 형태소 분석을 통해 한국어 코퍼스에서 고빈도 불용어 리스트를 표 2와 같이 작성하여 빈도수가 많이 존재하는 단어들도 제외하였다.

표 2. 한국어 불용어 리스트

| 형태 | 비율 | 형태 | 비율 |
|---|---|---|---|
| 이 | 0.01828 | 나오 | 0.000725 |
| 있 | 0.011699 | 가지 | 0.00072 |
| 하 | 0.009774 | 씨 | 0.00071 |
| 것 | 0.005723 | 만들 | 0.000704 |
| ⋮ | ⋮ | ⋮ | ⋮ |

표 3과 같이 불용어 제거가 완료된 텍스트 데이터들은 각 문서의 키워드를 활용하여 빈도수 측정을 통한 신조어 키워드를 추출한다. 이후 분석 작업을 통해 추론된 단어들은 텍스트 문서로 결과 값을 저장하고 'WordCount'를 통해 빈도수 상위 20개의 단어들만 추려서 저장한다.

표 3. 토크나이징 데이터의 불용어 제거

| | 댓글/리뷰 |
|---|---|
| **불용어 제거 완료 데이터** | 1. 레알, 창렬, 오브, 창렬,<br>2. 어벤져스, 마블, 노잼,<br>3. 개존못<br>⋮ |

빈도수 계산은 표 4와 같이 상위 단어들 목록을 저장하여 상위 랭크에 저장될수록 사용량이 많은 단어로 추측이 가능하다. 하지만 분석을 통해 저장된 단어이므로 특정 주제에 의해 사용량이 다르다. 실제로 댓글과 리뷰는 사용하는 단어들의 패턴이 다르며 표에 소개된 수치는 크게 의미가 없다.

표 4. 빈도수로 추론된 상위 단어

| 단어 | 빈도수 |
|---|---|
| 꿀잼 | 211 |
| 개꿀잼 | 154 |
| 노잼 | 103 |
| 창렬 | 98 |
| 갑분싸 | 78 |
| 가즈아 | 63 |
| 핵노답 | 61 |
| 노답 | 59 |
| 개이득 | 54 |
| 혜자 | 51 |
| 존못 | 49 |
| 극혐 | 44 |
| ⋮ | ⋮ |

신조어는 표 5와 같이 위키피디아를 참조하여 긍정, 부정을 분류하고 각각의 단어가 지니는 의미를 인식하고 감성 학습을 위해 긍정과 부정 문서별로 분류한 뒤, 파이썬을 사용하여 학습을 진행하였다.

표 5. 위키피디아를 참조한 긍정, 부정 분류와 의미

| | 단어 | 의미 |
|---|---|---|
| **긍정** | 꿀잼 | 너무 재미있다. |
| | 개이득 | 완전 이익/이득 |
| | 혜자 | 가격대비 (품질/맛) 좋다 |
| | 위꼴 | 위가 뒤틀릴 정도로 맛있다(맛있어 보인다) |
| **부정** | 노잼 | 너무 재미없다 |
| | 노답 | (사람 또는 특정주제) 답이 없다 |
| | 창렬 | 가격대비 (품질/맛) 좋지 않다 |
| | 극혐 | 극도로 혐오 |

분류된 감성 값이 긍정인지 부정인지 판별하기 위해 구축된 환경을 사용하여 신조어가 포함된 문장이 학습을 통해 구축된 데이터 안에서 적용되었을 때 어떤 결과를 도출할 수 있는지 실험을 진행하였고 표 6에 따라 표 7과 같이 결과 값을 확인할 수 있다.

표 6. 신조어 긍정, 부정 분류에 대한 알고리즘

```
txt = emoji_pattern.sub(r'',txt)
txt = spliter.pos(txt, norm=True, stem=True)
for n,c in txt:
        return_list.append(n)
for li in return_list:
        if li in positves and len(li) >= 2:
            po.append(li)
for li in return_list:
        if li in negatives and len(li) >= 2:
            ne.append(li)

                po = count(po)
            ne = count(ne)
            outpo = ''
            outne = ''
            for posi in po:
             outpo += posi
            for nega in ne:
             outne += nega

        print(filename)
        wb.save(filename)
```

표 7. 신조어 감성에 따른 긍정, 부정 판별

| 단어 | 감성 | 단어 | 감성 |
|---|---|---|---|
| 꿀잼 | positive | 혜자 | positive |
| 개이득 | positive | 창렬 | negative |
| 노잼 | negative | 극혐 | negative |
| 헬조선 | negative | 노답 | negative |

### 3. 표준단어 긍정, 부정, 감성 분석

소셜 미디어 신조어를 통해 구축된 신조어에 대한 긍정과 부정 값에 대한 단어들만 감성이 분류되기 때문에 표준어에 대한 긍정, 부정 분석기법과 통합하여 실험을 진행한다. 신조어 구축과 마찬가지로 첫 번째 단계로 진행했던 웹 크롤링으로 수집한 댓글, 리뷰에 대하여 문장들을 형태소 분석을 통해 긍정과 부정을 검증하였다. 'Doc2vec'의 'Genism'패키지를 활용하여 긍정과 부정을 분류하였다. 한국어 문서들의 분류는 'KoNLP'를 사용하고 'NLTK'패키지 안의 'Naive Bayes Classifier'라이브러리를 사용한다. 토픽 모델링을 지원하는 'Gensim'은 여러 문장이나 문서에 내재되어 있는 규칙 또는 토픽들을 찾아내는데 용이하다. 표준어 분석은 신조어 감성 학습과 구성은 동일한 형태로 진행하나 과정은 반대로 진행이 된다. 전처리 과정에서 불용어 제거와 품사 태깅 명사 추출 단계를 거쳐 진행한다. 품사 태깅 과정은 신조어와는 달리 표준 단어는 사전적인 의미를 담고 있어 형태소 분석이나 실험을 진행해도 무방하다. 표준 단어의 사용이 잦은 네이버 뉴스의 데이터를 수집하여 수집된 뉴스 데이터에서 명사만을 활용하여 형태소 분석을 다음과 같은 과정을 이용하여 진행한다[15]. 빈번하게 발생하는 불필요한 어휘와 신조어와 다르게 의미를 알 수 없는 단음절 체언 및 용언을 제거하고, 최종적으로 추출된 명사의 감성 점수화를 통하여 데이터의 감성 값을 도출하였다.

$$TF = \frac{\text{문서 내 단어의 개수}}{\text{문서 내 모든 단어의 수}}$$
$$IDF = \log\left(\frac{\text{문서 전체 갯수}}{\text{단어를 포함한 문서의 수}}\right) \quad (1)$$

문서 내에서 등장하는 단어의 빈도를 나타나는데 단어와 문서간의 중요도를 나타내기 위해 수식(1)과 같이 TF 값을 적용한다. 문서 내에서 많이 출현할수록 상대적으로 중요하다. DF란 특정 단어가 문서에 등장한 횟수를 뜻하고 IDF는 단어 수를 해당 단어의 DF로 나눈 뒤 로그를 취한 값이다. 그 값이 클수록 특이한 단어라는 걸 알 수 있다. TF-IDF를 이용하면 불용어를 걸러 낼 수 있으며 단어별 가중치를 알 수 있다. 감성 분류에서는 빈도수보다는 해당 단어가 자체가 있느냐 없느냐가 더 중요할지도 모른다. 해당 단어가 출현을 하면 1로 간주하고 출현 빈도에 가중치 값을 곱한다. 표 8은 도출된 가중치에 나이브 베이즈 분류법을 적용한 결과를 나타낸다.

표 8. 나이브 베이즈 기법에 가중치 적용

| 문서 | 좋다 | 별로다 | 부족하다 | 감성 |
|---|---|---|---|---|
| doc1 | 3.36 | 0 | 0 | pos |
| doc2 | 5.75 | 0 | 1.03 | pos |
| doc3 | 0 | 2.88 | 2.12 | neg |
| doc4 | 0.42 | 0.92 | 3.03 | neg |

## Ⅳ. 실험 평가 및 결과

### 1. 신조어 분석

본 절에서는 신조어 및 표준어 긍정, 부정 분석하여 특정 주제에 관한 문서의 감성 수치를 비교하여 성능을 평가하였다. 실험은 3단계를 거쳐 진행하였다. 첫 번째로 신조어 감성을 분석하여 신조어가 포함된 문장이 긍정인지 부정인지 판단한다. 'Naive Bayes Classification'을 사용하여 문장의 감성 값을 도출 후 다음으로 표준어 감성 분류 수치도 동일한 조건으로 구동

하고 도출하였다. 마지막으로 추출된 감성 값을 토대로 신조어와 표준어가 함께 사용하여 신조어가 포함된 문장이 감성 분류가 잘 도출되는지 확인하였다[16]. 여러 문서 안의 신조어가 포함된 문장들 중에서 조건부 확률 값을 구하기 위해 파이썬을 사용하여 코딩을 진행하였고 도출된 감성 값은 표 9과 같이 확인할 수 있다.

표 9. 추출된 신조어 감성 분류 예

| 문장 | 긍정:부정 | 결과 |
|---|---|---|
| “이 영화 완전 꿀잼 ... 또 보고싶다” | 1.0:2.8 | positive |
| “티셔츠 이값에?... 개이득 룰루” | 1.0:1.8 | positive |
| “소문 듣고 와서 봤는데... 노잼이군 ” | 2.2:1.0 | negative |
| “현지화 다됐네... 역시 헬조선 ” | 1.8:1.0 | negative |

## 2. 표준어 분석

신조어 분석 연구와 결합하여 사용하게 될 표준어 단어의 긍정, 부정 감성 분석은 결과 수치의 정확성을 위하여 표준어 문장의 사용 빈도가 잦은 ‘뉴스’ 데이터와 신조어가 포함된 일반적인 댓글, 리뷰를 사용하여 표 10과 같이 수집을 통해 얻어진 데이터와 비교분석을 진행하였다.

표 10. 뉴스와 댓글, 리뷰 비교를 위한 데이터 수집

| 뉴스 | 댓글/리뷰 |
|---|---|
| 북한의 일부 고위 간부는 남북정상이 합의한 '판문점 선언'에 대해 회의적 반응을 나타내는 것으로 알려졌다. 자유아시아방송(RFA)은 평양의 한 소식통을 인용해 "최근 중앙의 일부 고위 간부 사이에서 남북정상회담의 결과물인 '한반도의 완전한 비핵화'를 둘러싸고 회의적 반응이 나오고 있다"면서 이들은 김정은 국무위원장이 "북한의 목숨과 같은 핵을 완전히 포기할 리 없다고 생각 한다"고 보도했다. 소식통은 북한 매체가 남북정상회담 이후 "한반도 통일의 문이 활짝 열린 것처럼 요란하게 선전하고 있다"면서..... | 여러 제품을 써봤지만 이 제품처럼 만족스러운 구매는 없었던 것 같아요. ^^ 꿀템은 처음이예요~<br>.....<br>(중략)<br>.....<br>애들도 좋아하고 맛도 좋고 특유의 비린맛도 없어서 ~~ 호호 요즘 애들말로 개이득~ 이라고 한다죠 꿀템 ^^ |

실험을 통해 도출된 값은 표 11과 같이 비교분석을 통해 도출된 결과를 확인할 수 있다. 뉴스 기사의 텍스트는 일반적으로 특수한 주제가 아닌 경우에는 표준어를 사용하기 때문에 댓글, 리뷰의 텍스트 데이터와 비교 시에 차이가 분명하게 드러난다. 특정 제품의 경우 30대와 40대가 자주 구매하는 제품의 리뷰를 수집해서 소셜 미디어 신조어의 사용이 일반적인 댓글에 비해 적으나 표준어 기반 감성 분석을 적용했을 시에는 분명한 차이를 보였다.

표 11. 표준어 감성 분류를 적용한 비교분석

| | 뉴스 | 댓글/리뷰 |
|---|---|---|
| **사용한 데이터** | ‘북한’키워드 | 특정 제품 |
| **사용언어** | 표준어 | 표준어 |
| **정확도** | 0.90231 | 0.54822 |

## 3. 평가 및 결과

본 절에서는 제안한 방법을 이용하여 최종적인 결과 도출을 위해 평가를 진행하였다. 신조어와 표준어 긍정, 부정 감성 분석을 통한 연구를 토대로 각각의 감성별로 학습과 실험을 반복하여 결과의 정확성을 높였다. 이러한 과정을 거쳐 문서에 포함된 단어의 긍정, 부정 수치를 확인하고, 학습된 데이터 값은 웹 크롤링을 통해 수집된 텍스트 데이터를 활용하여 댓글, 리뷰 및 뉴스 데이터의 데이터가 어떠한 긍정, 부정 감성 값을 도포하고 있는지 확인하였다. 그리고 실험을 통해 진행된 결과 값은 정확도와 감성 수치에 대한 결과를 도출하고 기존에 연구되었던 표준어 감성 분석 기법과 비교를 통해 수치 비교를 진행하였다. 제안하는 신조어 감성과 표준어 감성에 대하여 결합을 통해 결과 값 표 12를 통해 각 문서 안의 문장에 신조어와 표준어가 함께 쓰인 단어들을 확인하고 각 문서별로 결과 값을 출력하였다.

표 12. 문서별 표준어와 신조어 분석

| 문서 | 문장 |
|---|---|
| doc1 | 이 겜 완전 꿀잼 (...) 정말 잘 샀다능...! 만족 만족! ㅎ |
| doc2 | 헬조선 취업 정말 어렵다 (...) 불합격 통지 꿀잼..(...) |
| doc3 | 어쩔 수 없이 샀는데 가성비 창렬 (...) 이렇게 비싸고.. |
| ⋮ | ⋮ |

확인된 결과는 표 13과 같이 표준어와 신조어가 포함된 문장이 긍정인지 부정인지 최종 결과를 출력하고 도출된 결과 값은 문장 수치를 확률 값으로 전환하고 도출된 결과 값을 확인하였다. ‘doc2’의 경우에 ‘doc1,3’의 결과보다 낮은 결과 값을 확인할 수 있는데 이는 ‘doc2’의 문장에서 부정의 의미가 내포되어 있으나

표 12에 '꿀잼'이라는 긍정의 단어가 포함되어 있어서 부정의 결과를 나타내지만 낮은 결과 값이 도출된 것을 확인할 수 있다.

표 13. 표준어와 신조어 결과 값 도출

| 문서 | 표준어 | 신조어 | 감성 | 결과 |
|---|---|---|---|---|
| doc1 | 만족 | 꿀잼 | Pos. | 0.9362 |
| doc2 | 어렵다 | 헬조선 | Neg. | 0.7893 |
| doc3 | 비싸다 | 창렬 | Neg. | 0.9113 |
| ⋮ | ⋮ | ⋮ | ⋮ | |

기존의 표준어 분석 알고리즘을 통해 추출된 정확률과 본 논문에서 제안하는 결과 값의 정확도를 평균화하여 표 14와 같이 도출하였다.

표 14. 기존 연구와 제시한 연구에 대한 비교 수치

| 텍스트 분류를 통한 연구 | | 신조어/표준어 감성 연구 | |
|---|---|---|---|
| 문서 | 감성 | 문서 | 감성 |
| doc1 | 0.8967 | doc1 | 0.9362 |
| doc2 | 0.6832 | doc2 | 0.7893 |
| doc3 | 0.8799 | doc3 | 0.9113 |
| ⋮ | ⋮ | ⋮ | ⋮ |

실험을 통해 진행된 결과는 표 14를 토대로 그림 6으로 시각화 하였다.

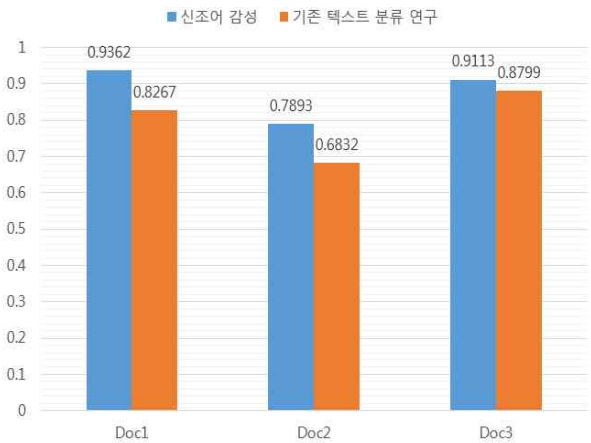


그림 6. 감성 비교분석 및 평가 수치

기존의 알고리즘은 텍스트 분류를 통한 단순 빈도수로 결과 값을 도출하고 불용어 수준의 단어를 정확하게 처리하기가 어렵다는 단점이 존재하여 객관적인 수치를 나타내는데 어려움이 있다. 기존의 연구에서 진행된 표준어 감성에 비해 제안하는 알고리즘이 평균 0.21 정도의 미세한 수치 상승이 나타나 제안하는 방법에 대한 효율성을 입증할 수 있는 결과를 시각화를 통해 그래프로 확연한 차이를 확인할 수 있다. 그림 6에서 나타난 문서 'doc1', 'doc2', 'doc3'은 댓글과 리뷰에 대하여 저장된 문서이며 평가를 진행 후 상대적으로 낮은 수치를 보이는 'doc2'는 기존의 알고리즘이나 본 논문에서 제안하는 방법을 적용하여도 긍정적인 단어가 포함되어 있어 낮은 결과 값이 도출되었다. 실제로 문서를 확인하였으나 컴퓨터가 이해하기 어려운 내용을 갖고 있어 다른 문서에 비해 높은 결과 값을 도출하기 어렵다.

## V. 결론

본 논문에서는 소셜 미디어 상의 신조어 때문에 발생하는 데이터 마이닝의 성능 저하와 데이터 손실을 막기 위한 신조어 긍정, 부정 감성 판별을 진행하였다. 본 논문에서 제안하는 방법은 웹 크롤링을 통한 댓글, 리뷰 데이터들을 수집하여 신조어 추출을 진행해 긍정, 부정 값을 통해 신조어 감성 분석의 기반을 마련하고 종래의 표준어 기반 감성 분석 방법과 결합하여 구체적인 데이터 마이닝 기술에 근접하는 것이다. 신조어 분석과 종래 기술인 표준어 기반 감성 분석을 결합하여 기존에 불용어로 제거되었던 신조어가 감성 의미를 가짐으로써 기능을 다하여 결과 값이 향상되는 점을 확인하였다. 본 연구에서는 신조어의 감성 분석에 대한 연구로 단어에 대한 기능을 할 수 있도록 하였으며 신조어는 더 이상 제거 대상이 아닌 필수 요소로 자리매김해야 할 것이다. 또한, 신조어는 소셜 미디어의 한 문화에서 파생된 단어로 표준어와 같이 반드시 써야 할 단어는 아니지만, 소셜 미디어의 문화가 점점 계속 될수록 단어는 앞으로도 계속 생길 것이며 그에 대한 지속적인 대처가 필요할 것이고 그러한 환경 구축에 대한 연구를 진행할 예정이다.

## REFERENCES

[1] 장경현, "신조어 연어의 형성원리," *인문논총,* 제66권, 269-297쪽, 2011년 12월
[2] 강아름, 이상연, 이건, "매스미디어 상 인터넷 용어 처리를 위한 은닉 마코프 모델기반 신조어 추출," *한국지능시스템학회 학술발표 논문집,* 제25권, 제1호, 119-120쪽, 2015년 4월
[3] 안정은, "Text Mining 기법을 이용한 표준특허기술의 유사도 측정방법," *한국정보과학회 학술발표 논문집,* 제36권, 제1호, 1-5쪽, 2009년 6월

[4] 이한동, 김종배, “복합명사를 포함하는 개선된 키워드 추출 방법,” *예술인문사회융합멀티미디어논문지,* 제7권, 제10호, 857-864쪽, 2017년 10월
[5] 이성직, 김한준, “TF-IDF의 변형을 이용한 전자뉴스에서의 키워드 추출 기법,” *한국전자거래학회지,* 제14권, 제4호, 59-73쪽, 2009년 11월
[6] 대한민국의 인터넷 신조어 목록(2018), https://ko.wikipedia.org/wiki/대한민국의_인터넷_신조어_목록 (accessed Mar., 12, 2020).
[7] 홍택은, 김정인, 신주현, “인스타그램 이미지와 텍스트 분석을 통한 사용자 감정 분류,” *스마트미디어저널,* 제5권, 제1호, 61-68쪽, 2016년 3월
[8] 장경애, 박상현, 김우제, “인터넷 감정기호를 이용한 긍정/부정 말뭉치 구축 및 감정분류 자동화,” *정보과학회논문지,* 제42권, 제4호, 512-521쪽, 2015년 4월
[9] 이종화, 레환수, 이현규, “오피니언 마이닝을 통한 국내와 수입 의류 제품에 대한 고객 평판 연구,” *인터넷전자상거래연구,* 제15권, 제3호, 223-234쪽, 2015년 6월
[10] 김동성, 김종우, “온라인 여론의 감성분석을 위한 감성용어 자동화 추출 방안 연구,” *한국경영정보학회 학술대회논문집,* 제2016권, 제6호, 187-189쪽, 2016년 6월
[11] 한나눔(2018), http://semanticweb.kaist.ac.kr/home/index.php/HanNanum (accessed Mar., 12, 2020).
[12] 조하나, 정연오, 이재동, 이지형, “인터넷 뉴스 댓글의 감성 분석을 통한 오피니언 마이닝,” *한국지능시스템학회 학술발표 논문집,* 제23권, 제1호, 149-150쪽, 2013년 4월
[13] 박승현, 이은지, 김판구, “한글 편집거리 알고리즘을 이용한 한국어 철자오류 교정방법,” *스마트미디어저널,* 제6권, 제1호, 16-21쪽, 2017년 3월
[14] 차준석, 김정인, 김판구, “단어 간 의미적 연관성을 고려한 어휘 체인 기반의 개선된 자동 문서요약 방법,” *스마트미디어저널,* 제6권, 제1호, 22-29쪽, 2017년 3월
[15] 최성자, 손민영, 김영학, “키워드 기반 블로그 마케팅을 위한 연관 키워드 추천 시스템,” *정보과학회 컴퓨팅의 실제 논문지,* 제22권, 제5호, 246-251쪽, 2016년 5월
[16] 안광모, 김윤석, 김영훈, 서영훈, “Levenshtein 거리를 이용한 영화평 감성 분류,” *한국디지털콘텐츠학회 논문지,* 제14권, 제4호, 581-587쪽, 2013년 12월

## 저 자 소 개


**김정인(정회원)**

2011년 조선대학교 컴퓨터공학과 학사 졸업(공학사).
2019년 조선대학교 컴퓨터공학과 박사 졸업(공학박사).
2019년~현재 (BK21+)스마트인터넷기반 융합콘텐츠기술 인력양성사업팀 연구원(Post-Doc).

<주관심분야 : 인공지능, 딥러닝, 빅 데이터 처리, 자연어 처리>

**박상진(정회원)**

2016년 조선대학교 제어계측로봇공학과 학사 졸업.
2018년 조선대학교 소프트웨어융합공학과 석사 졸업.
2019년~현재 ㈜솔로몬테크노서플라이 운영지원팀 재직.

<주관심분야 : 자연어 처리, 텍스트 마이닝, 오피니언 마이닝>

**김형주(정회원)**

1999년 조선대학교 전산통계학과 학사 졸업.
2002년 원광대학교 컴퓨터공학과 석사 졸업.
2018년 조선대학교 교육대학원 정보·컴퓨터교육전공 석사 졸업.
2018년~현재 조선대학교 컴퓨터공학과 박사 과정.

<주관심분야 : 빅데이터 처리, 인공지능, 딥러닝>

**최준호(정회원)**

1997년 호남대학교 컴퓨터공학과 학사 졸업.
2000년 조선대학교 전자계산학과 석사 졸업.
2004년 조선대학교 전자계산학과 박사 졸업.
2014년~현재 조선대학교 자유전공학부 부교수.

<주관심분야 : 지식베이스, 자연어처리, 지능형보안, 정보추출>

김한일(정회원)

1995년 서울대학교 컴퓨터공학과 박사 졸업(공학박사).
1995년~현재 제주대학교 컴퓨터교육과 교수.
2009년~현재 제주대학교 스토리텔링학과 교수.

<주관심분야 : 컴퓨터교육, 문화기술, 스토리텔링>

김판구(정회원)

1988년 조선대학교 컴퓨터공학과 학사 졸업(공학사).
1990년 서울대학교 컴퓨터공학과 석사 졸업(공학석사).
1994년 서울대학교 컴퓨터공학과 박사 졸업(공학박사).
1994년~현재 조선대학교 컴퓨터공학과 교수.

<주관심분야 : 인공지능, 정보 검색, 시맨틱 웹, 자연어 처리>