2018년 8월

석사학위 논문

나이브 베이즈 분류를 적용한 소셜 미디어 상의 신조어 감성 판별 기법

조선대학교 산업기술융합대학원

소프트웨어융합공학과

박 상 진

2018년 5월

석사학위논문

나이브 베이즈 분류를 적용한 소셜 미디어상의 신조어 감성 판별 기법

박 상 진

나이브 베이즈 분류를 적용한 소셜 미디어 상의 신조어 감성 판별 기법

Sensitivity identification method for new words of social media using the naïve bayes classification

2018년 4월

조선대학교 산업기술융합대학원

소프트웨어융합공학과

박 상 진

나이브 베이즈 분류를 적용한 소셜미디어 상의 신조어 감성 판별 기법

지도교수 김 판 구

이 논문을 공학석사학위신청 논문으로 제출함.

2018년 5월

조선대학교 산업기술융합대학원

소프트웨어융합공학과

박 상 진

박상진의 석사학위논문을 인준함

위원장 조선대학교 교수 신 주 현 (인)

위 원 조선대학교 교수 최 준 호 (인)

위 원 조선대학교 교수 김 판 구 (인)

2018년 5월

조선대학교 산업기술융합대학원

목 차

표 목차

그림 목차

ABSTRACT

Sensitivity identification method for new words of social media using the naïve bayes classification

SangJin Park
Advisor : Prof. PanKoo Kim Ph.d
Department of Software Convergence Engineering
Graduate School of Industry Technology Convergence, Chosun University

From PC communication to the development of the internet, a new term has been coined on the internet, and the internet 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 internet-words can have many advantages, including the use of short sentences to solve the problems of various letter-limited messengers and reduce data. However, internet words does not have a dictionary meaning and there are limitations and degradation of algorithms such as data mining.

Recently, with the influence of fourth industrial revolution, data has been formed as valued. 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 Internet 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 newly Internet 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.

I.서론

A. 연구 배경 및 목적

인터넷의 발달과 스마트폰의 보급화로 인한 그에 따른 인터넷 문화가 형성 되면서 사용자들 또한 문화와 시대에 발맞춰 녹아들면서 발맞춰 나아가고 있다. 이러한 문화에서 가장 중요한건 소통이다. 그리고 그 문화에서 빼놓을 수 없는 부분은 바로 신(新) 인터넷 용어 이다. 인터넷 보급화 초기 시절 PC통신 게시판이나 채팅에서는 특수문자를 사용하여 사용자의 감정이나 느낌점 등을 나타내어 소통하는 것이 전부였으나 메신저나 'SNS'의 등장과 그들의 가교역할을 해주는 스마트폰의 보급화로 신 인터넷 용어의 출연이 생기고 빈번하게 사용하고 있는 추세다.

먼저 용어의 등장은 인터넷 주 사용 층인 1·20대 사용자들이 주로 사용하고 있고 사용자들의 이용을 통해서 또 다른 용어의 탄생과 발전이 되어 지고 있다. 새로운 인터넷 용어의 사용은 여러 장단점을 가져올 수 있는데 장점 중 가장 큰 강점은 빠른 의미전달이다. 긴 문장을 짧은 문장으로 글자 수 제한이 있는 'Twitter'의 타임라인이나 기타 메신저의 문제점이 존재하는데 신조어 사용으로 짧은 단어로 함축적인 의미전달이 가능해 최소한의 문제점의 해소가 가능하다. 예를 들어 '이 영화는 정말 재미있다.'라는 문장을 인터넷 용어로 바꾸면 '이 영화 꿀잼.'으로 바꿀 수 있는데 '꿀잼'과 '정말 재미있다'는 비슷한 뜻을 갖고 있다. 기존의 표준 문장에서 인터넷 용어의 전환으로 인해 약 2배의 데이터 축소가 이루어지는데, 한글 텍스트 데이터는 한 글자 당 2Byte로 실제 웹상에서 쓰이고 있는 텍스트 데이터의 크기와 보다 긴 문장을 사용 했을 시에 절감되는 데이터의 양은 실로 많을 것이다.

최근 빅 데이터는 가장 각광 받고 있는 기술 중 하나이다. 스마트 기기의 보급화와 발전으로 사용자가 사용 후 발생 되는 텍스트 나 이미지 등 정형되지 않는 많은 양의 비정형 데이터가 발생한다. 특히, 사용자를 통해 생성되는 데이터들은 특정한 장소나 물건에 대해 다양한 정보와 의견을 표현한다. 빅 데이터는 3V로 정의되는데 데이터의 양(Volume), 속도(Velocity), 다양성(Variety)으로 되는데 최근에는 가치(Value)와 복잡성(Complexity)을 덧붙여서 사용자를 통해 창출되는 데이터가 단순히 텍스트와 이미지로 끝나는 것이 아니라 가치로써 의미가 더해진다.

이러한 관점에서 빅 데이터는 4차 산업혁명에서 혁신과 경쟁력, 생산성 강화를 향한 중요한 원천으로 간주되고 있다. 하지만 주사용 층 이외에 일부의 사용자들은 단어와 의미를 이해하는데 어려움이 있으며 데이터 마이닝(Data Mining)기술 이나 빅 데이터(Big Data)같은 연구에는 사전적인 의미를 갖고 있지 않아 알고리즘의 성능 저하와 연구에 제약사항이 발생한다. 과거 인터넷 용어는 은어로써 사용을 자제 하고 기피하는 현상이 있었으나 현재 인터넷 문화에서는 빼놓을 수 없는 부분으로 자리매김 했다.

이를 바탕으로 본 논문에서는 인터넷 신조어 분석을 통해 인터넷 신조어에 대해 긍·부정 분류 연구를 진행 하고자 한다. 현재 위키피디아(Wikipedia)에 구축 되어 있는 인터넷 신조어 목록의 단어들과 문서에서 신조어와 동시에 사용하는 표준어 단어들을 수집하여 의미를 통한 긍정 값과 부정 값의 분류를 진행 한다. 분류를 통해 추출된 긍정/부정 단어들은 파이썬(Python)을 통해 학습시킨다. 구축된 데이터를 이용하여 데이터 마이닝 기법의 성능을 향상시키는 방법을 제안하고 기존의 연구와 비교 실험을 통해 나은 성능평가를 진행한다.

B. 연구 내용 및 구성

인터넷 웹(Web)상에서는 시간이나 공간을 막론하고 실시간으로 데이터가 발생하고 있다. 사용자는 각각의 취향에 있어서 관심 분야나 불특정 주제에 관하여 텍스트와 영상 이미지 등 거대한 데이터가 생성 된다. 이러한 실시간으로 발생 되는 비정형화 데이터는 4차 산업혁명의 일부분으로써 가치가 형성 되고 효과적인 분석 연구를 요구한다. 이러한 분석을 위해 인터넷에서 파생 되는 단어들에 대해 웹 크롤링(Crawling) 통한 텍스트 데이터를 추출하고 텍스트 마이닝과 오피니언 마이닝을 통해 의미부여 및 단어들에 대한 긍정/부정 분류 통한 문장의 오피니언 파악을 진행하고자 한다.

본 논문의 구성은 다음과 같다.

1장 서론에서는 연구의 배경과 목적에 관련하고, 2장 관련 연구에서는 새로운 인터넷 용어에 대한 형성원리와 Text Mining(텍스트 마이닝)과 Opinion Mining(오피니언 마이닝)에 대해 서술하고, 긍·부정 감성 분석에 관련한 기존의 선행연구와 알고리즘에 관련하여 기술한다.

3장에서는 인터넷 신조어 추출하기 위한 방법과 'Naïve Bayes' 기법을 통해 신조어 감성단어의 수치와 TF-IDF를 사용하여 표준어 감성의 긍정/부정의 수치 도출을 통하여 긍·부정 분류 방법을 제시 한다.

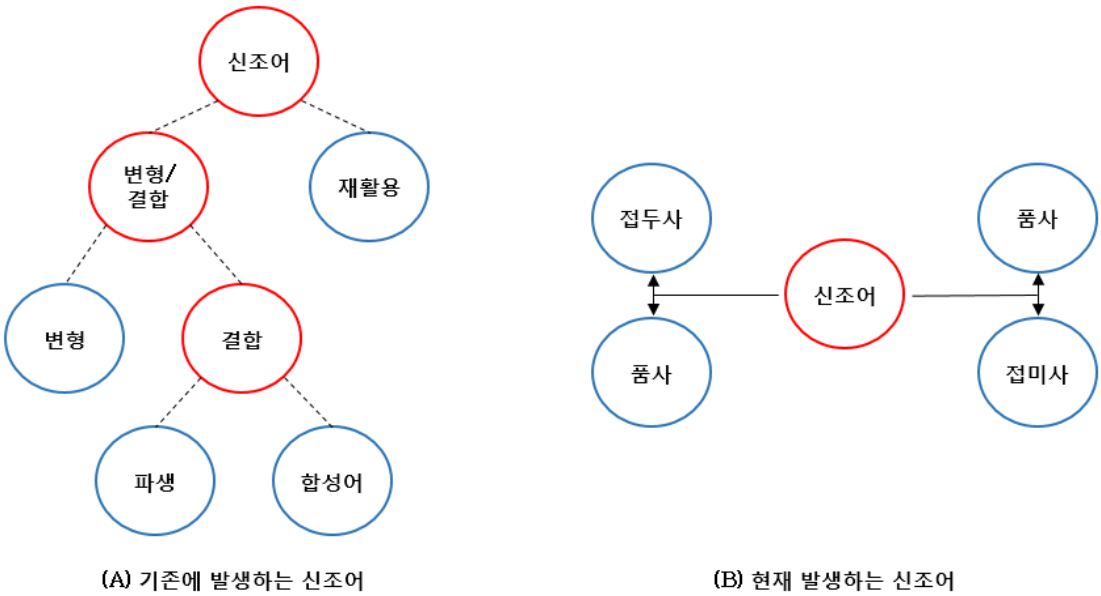
4장에서는 본 논문에서 제안 하는 방법에 대한 실험과 환경 구축 후 기존의 연구의 비교 실험을 진행한다.

마지막 5장에서는 본 논문의 결론과 향후 연구 방향에 대해 제언한다.

Ⅱ.관련 연구

A. 인터넷 신조어 형성 원리

인터넷 문화 중 하나인 새로운 인터넷 용어는 인터넷 내에서는 물론이고 상당수가 현실 세계에서도 사용한다. 인터넷의 용어들은 대부분 유행에 예민하여 빠르게 확산되어 타올랐다가 식어버리는 취약부분이 존재하여 단어의 의미 이외에 구조나 탄생에 대해서는 의의가 필요하지 않다. 하지만 신조어를 형성과 탄생은 흥미로운 부분이고, 단어가 유행하는 기간에는 단어에 대한 평가가 필요 하다.



[그림 2-1] 신조어 형성과정

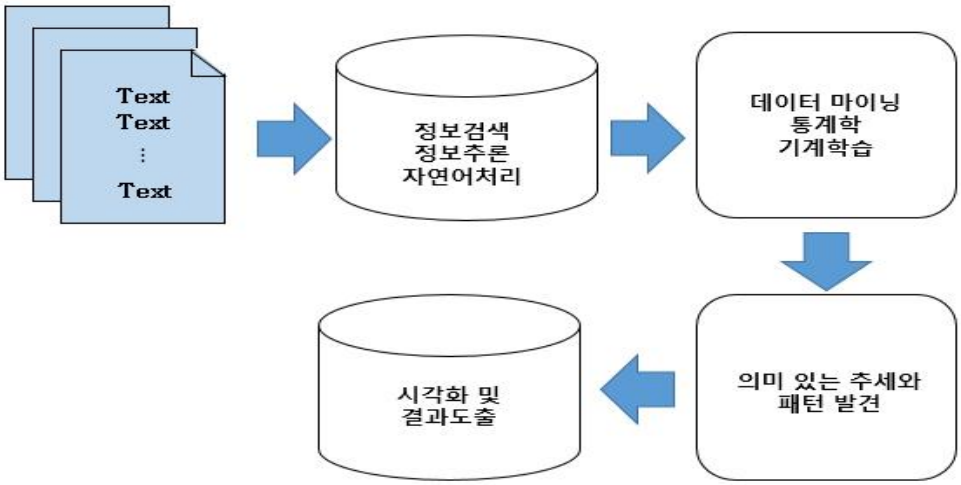
인터넷 신조어는 '동사+형용사', '형용사+명사', '명사+명사'나 반대의 조합과 조합으로 이루어져 있거나 긴 단어를 줄여 줄임말 형태의 단어로 불규칙적인 형태로 구성되어 있다[1, 2]. 또한 해당 시대의 유행하는 어떤 사물이나 사람의 형태를 가져와 탄생하는 단어도 존재한다. 예를 들어 배우 '김혜자'씨와 가수 '김창렬'씨의 이름을 따와 '혜자'와 '창렬'이 있다. '창렬'은 가격에 비해 품질이 좋지 못하는 경우 사용하는 단어로써 가수'김창렬'이 모델로 한 제품의 품질이 가격에 비해 좋지 못하여 생겨난 부정적인 의

미를 지닌 단어 이다. ‘혜자’의 경우 그 반대의 의미로써 긍정의 의미를 지닌 단어이고 몇 해 전부터 관용어처럼 사용 되어 지고 있다. 보통의 경우 유행하는 단어들은 보통 어느 커뮤니티에서 사용하다가 점차 퍼져나가 오랫동안 사용자들에 의해 쓰여 지게 되면 관용어처럼 탄생한다. 이렇게 탄생한 신조어는 진화 하거나 변형시켜 또 다시 재탄생하게 된다.

조합의 형태로 탄생하는 신조어는 한 분야에 열중하는 사람의 뜻을 지닌 일본어 ‘오타쿠’가 진화와 변형의 형태로 ‘오덕후’의 형태가 되고 성소수자가 자신의 성정체성을 대중에게 공개하는 뜻을 지닌 ‘커밍아웃’과 비슷하게 ‘오덕후’가 자신의 취미를 대중에게 공개하는 ‘덕밍아웃’이 탄생하게 되었다. ‘-밍아웃’은 이와 같은 진화와 변형의 형태로 탄생하게 되어 ‘무엇을 알리게 된다.’ 라는 뜻을 쓸 때엔 단어의 ‘-밍아웃’을 단어의 접미사에 붙여 사용한다.

이런 단어는 관심사가 비슷한 커뮤니티에서는 긍정의 의미를 지니지만 객관적으로 봤을 때 부정 적인 시선으로 보는 경우도 존재하여 단어의 긍정과 부정의 판별이 모호하다. 이를 토대로 단어의 뜻과 탄생하게 되는 이유를 들여다봤을 때에는 상당히 흥미로운 부분이다. 우리말의 특성상 한글 언어는 사람에 따라서 다양한 단어와 문장의 형태로 구성 되어 있기 때문에 오랜 기간의 연구와 노력이 필요하다.

B. Text Mining을 이용한 신조어 추출

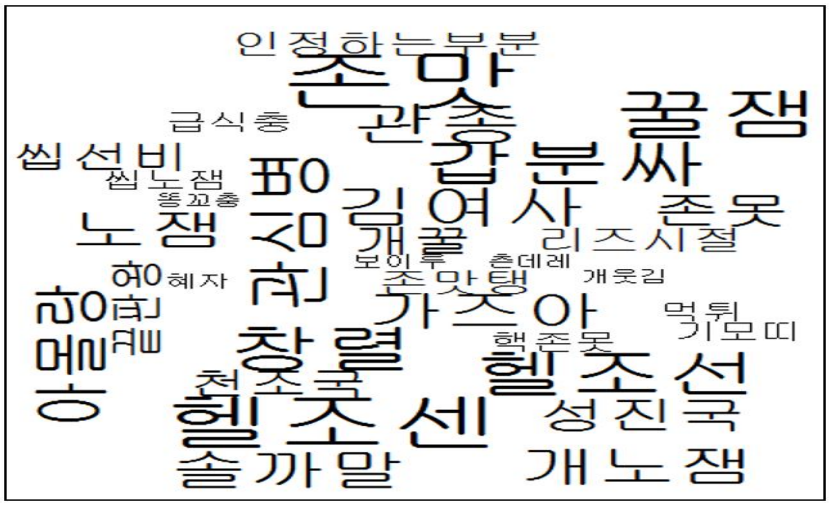


[그림 2-2] Text Mining 과정

데이터 마이닝 기술 중 하나인 Text Mining(텍스트 마이닝)은 자연어 처리와 정보 추출 등의 분야를 연구하는데 유용한 기술 중 하나이다.

'SNS'나 인터넷 카페 커뮤니티에서 흔히 찾을 수 있는 데이터들은 구조가 완전하지 않는 형태로 구성되어 있고 가공 되지 않은 데이터로 그 안에서 불분명한 형태 안에 필요한 키워드 추출하는 작업은 중요하다[3].

대규모 문서 나 유저를 통해 실시간으로 생성되거나 비정형적 데이터 안에서 텍스트의 특징 정보를 추출 하여 키워드 형태로 표현하고 텍스트간의 유사도를 확인하여 군집화 한다. 정형화된 데이터는 새로운 정보를 생성하고 찾고자 하는 패턴이나 키워드를 찾는데 [그림 2-2]와 같은 형태로 진행 한다[4]. 아울러, 추출된 데이터를 분석해서 의미와 의도 그리고 경향을 파악 하는데 유용하다. 현재 국내 정보 검색 엔진은 주로 키워드 매칭을 이용하여 사용자에 따라 입력된 정보를 통해 검색의 정확도를 높이기 위한 방법으로 쓰인다. 'SNS'나 인터넷 카페 커뮤니티에서 흔히 찾을 수 있는 데이터들은 구조가 완전하지 않는 형태로 구성되어 있고 가공 되지 않은 데이터로 그 안에서 불분명한 형태 안에 필요한 키워드 추출하는 작업은 중요하다[4].



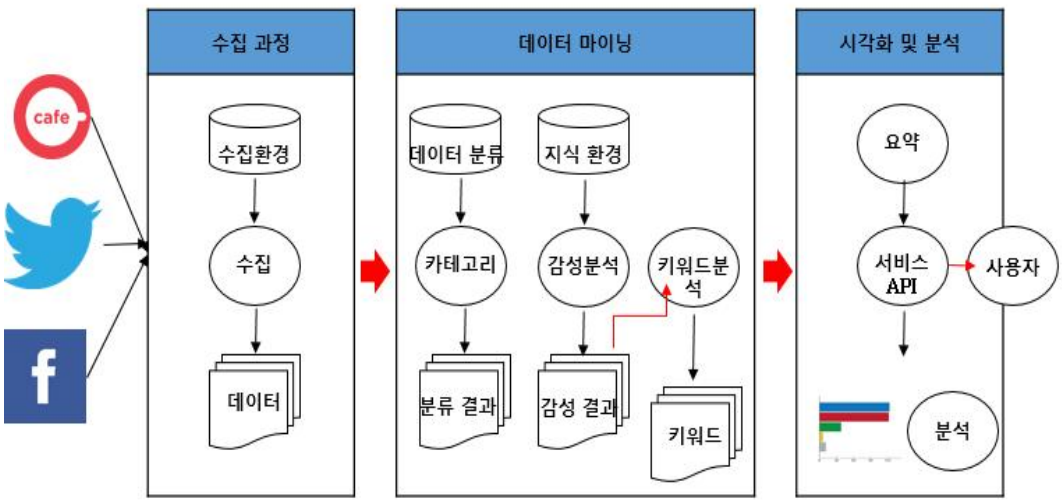
[그림 2-3] 'WordCloud'를 사용하여 추출된 텍스트 데이터 시각화

[그림 2-3]은 인터넷 포털사이트 카페와 Twitter에서 매순간 실시간으로 사용자가 본인의 의견이나 감성을 작성한 텍스트 데이터를 추출하여 'R'프로그래밍에서 'Konlp' 패키지를 활용하여 텍스트 데이터 시각화 한 것이다. 인터넷에서 무수히 많은 사용자들이 인터넷 신조어를 사용하고 있는 것을 그림[2-3]을 통해 알 수 있다.
신조어 추출은 문서내의 단어의 빈도수를 구하는데 유용한 'TF-IDF'를 이용하여 수치를 통해 키워드 추출 및 분석을 진행한다.[5].

추출된 인터넷 신조어는 단순히 의미를 갖고 있지 않아 위피피디아(WikiPedia)에서 제공하는 '대한민국 인터넷 신조어'목록을 통해 임의로 긍정/부정 분류를 진행하고 분류 된 데이터는 오피니언 마이닝을 통해 학습하여 사용한다[6].

C. Opinion Mining(오피니언 마이닝)

오피니언 마이닝 기법과 감성 분석 연구는 연구자에 따라서 동일한 기술로 보거나 그렇지 않은 기술로 보는 경우가 존재 한다. 오피니언은 마이닝은 우리말로 감성 분석이다. 따라서 본 논문은 전자의 경우로 가정 하에 연구를 진행 하였다. 오피니언 마이닝은 자연 언어 처리, 텍스트 분석 등을 사용해 정서적인 상태와 주관적인 정보를 체계적으로 식별 하고 추출 및 정량화를 연구하는 기술이다. 다량의 문서나 인터넷에서 유저들의 텍스트 데이터를 통해 추출된 데이터를 이용하여 사용자의 리뷰와 설문 조사 및 'SNS'에서 마케팅과 고객 관리 및 의견 분석을 통해 반영되는데 유용하기 때문에 여론이나 사람들의 관심이 어떻게 변하고 있는지 특정 기업에서 제품이나 서비스에 대해 좋고 싫음을 분석한다. 오피니언 마이닝은 특정 주제나 상황에 따라 감성에 따라 반영해야하기 때문에 사용자가 어떤 의도를 갖고 있는지 분석하여 처리하는 것이 중요하다.



[그림 2-4] 오피니언 마이닝 절차

오피니언 마이닝은 [그림2-4] 과정을 거쳐 연구를 진행한다. [그림2-4]의 과정을 크게 3가지로 분류 하여 진행 할 수 있다.

먼저 첫 번째로 웹 크롤링을 통한 데이터 수집이다. 사용자의 생각이나 감성이 많이

담겨 있는 특정 제품에 대한 리뷰와 'SNS' 등을 통해 수집을 진행한다. 한국어에 대한 사용자 평점이나 '좋아요.'점수를 통해 나타난 단어를 긍정 의미로 지정하고 평점이 낮거나 '싫어요.'점수를 통해 나타난 단어들은 부정 의미로 표현 될 수 있다.

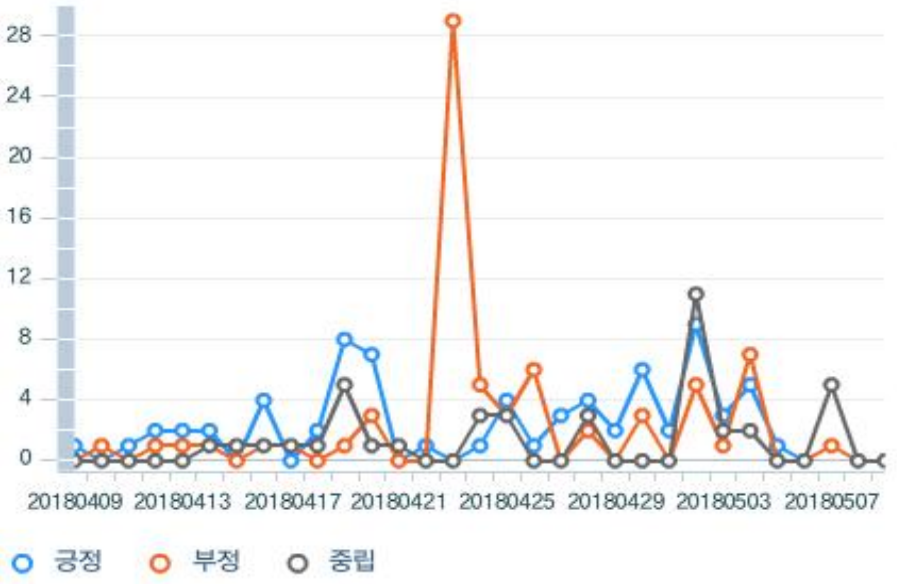
[그림 2-5] 영화 평점 오피니언

하지만, [그림 2-5]처럼 점수와 단순한 문장의 분석을 통한 방법은 한계를 보여준다. 두 영화는 실제로 상반되는 평가를 보여주지만 일명 '평점테러'로 인하여 극단적인 점수의 수치를 보여주기 때문에 정확한 결과 값을 도출하는 방법으로는 적당하지 않다.

두 번째로 수집된 데이터 내에 저장 되어 있는 텍스트 내에 평가요소와 긍정/부정을 가리키는 오피니언 관계가 있는 문장을 인식하고 불필요한 특수문자와 감성과 관계없는 품사를 제외한다. 두 번째 과정에서는 문장이 긍정/부정이 시스템 내에서 인식 하는지를 판별해야 한다. 긍정/부정 오피니언인지 문장이나 문서에서 분류하기 위하여 통계 기반이나 패턴이나 룰에 의한 방법을 동시에 적용 할 수 있다. 보통의 일반적인 표준어의 경우에는 품사 분류나 자동 띄어쓰기기법을 학습 데이터를 구축하여 'Naïve

Bayes'이나 'SVM'등의 알고리즘을 적용 하여 기계 학습을 진행한다.

마지막으로 두 과정을 통해 긍정/부정을 의미하는 단어들이 담긴 특정 주제에 관한 텍스트 데이터를 요약을 통해 분석과 평가한다.



[그림 2-6] '갤럭시 노트7'에 대한 긍정/부정 추이

[표 2-1] '갤럭시 노트7'에 대한 오피니언 마이닝 분석표

제품 특징	긍정(%)	부정(%)	제품 특징	긍정(%)	부정(%)
품질	0.25	0.75	업데이트	0.2	0.8
가격	0.5	0.5	부가 기능	0.7	0.3
성능	0.8	0.2	기타	0.35	0.65

[그림 2-6]은 삼성사의 스마트폰 갤럭시 노트7에 대한 18.04.09. ~ 18.05.07까지의 인터넷 상 여론에 대한 긍정/부정 감성 추이를 나타내는 그래프이고 이를 기반으로 [표

2-1]은 종래에 구축된 데이터를 이용하여 추이에 기반 하여 제품에 대해 오피니언 마이닝 분석을 통해 추출된 긍정/부정 평가이다. 18.04.21~25기간에 실제로 핸드폰 업데이트나 부가 서비스에 대한 불만사항이 여론을 통해 확인되었고 오피니언 분석을 통해 부정적인 내용이 포함되고 있어 사용자들의 여론을 쉽게 파악 할 수 있다. 이러한 오피니언 마이닝 결과 값을 통해 소비자들이 제품에 대한 평가에 대해 좋고 싫음의 정보를 알 수가 있고, 객관적인 자료로 인하여 제품의 신뢰성을 향상 시키고 호감도와 불호감도를 판단해 제품의 구매 판단 여부를 확인 할 수 있다[7, 8, 9].

D. 오피니언 마이닝을 통한 감성 분류에 대한 연구

종래 기술로 연구된 감성 분석 연구의 알고리즘은 오피니언 마이닝을 통해 분류된 긍정/부정 값은 수집을 통해 저장된 텍스트 데이터를 사용하여 문장의 어휘나 품사 등을 분류하여 감성 분류를 진행한다[10].

감성 분류의 대표 구축 모델인 'SentiWordNet(SWN)'은 'WordNet'에 긍정/부정 의미를 일정한 값을 지정하여 구축된 기술로 활용 되고 있다. 기존의 한국어 감성 사전 관련 연구 형태로는 감성을 나타내는 명사, 형용사, 동사에 따르기 때문에 품사의 단어를 간추려 요약 후 감정 분류를 한다[11].

긍정/부정 적인 의미를 지정하기 위해 단어 어휘의 극성 정보를 판단하여 의미를 부여해야 한다. 이 방법은 1990년대 말부터 시도 되었다. 이러한 시도에 대해 추론 된 방법은 확률에 기반 한 'PMI'기법을 적용하여 결과 값을 도출 하였다. 이 기법은 비교적으로 간단하지만 가장 좋은 결과 값이 도출 될 수 있어서 현재까지 많은 연구 기법에 적용 되고 있다[12].

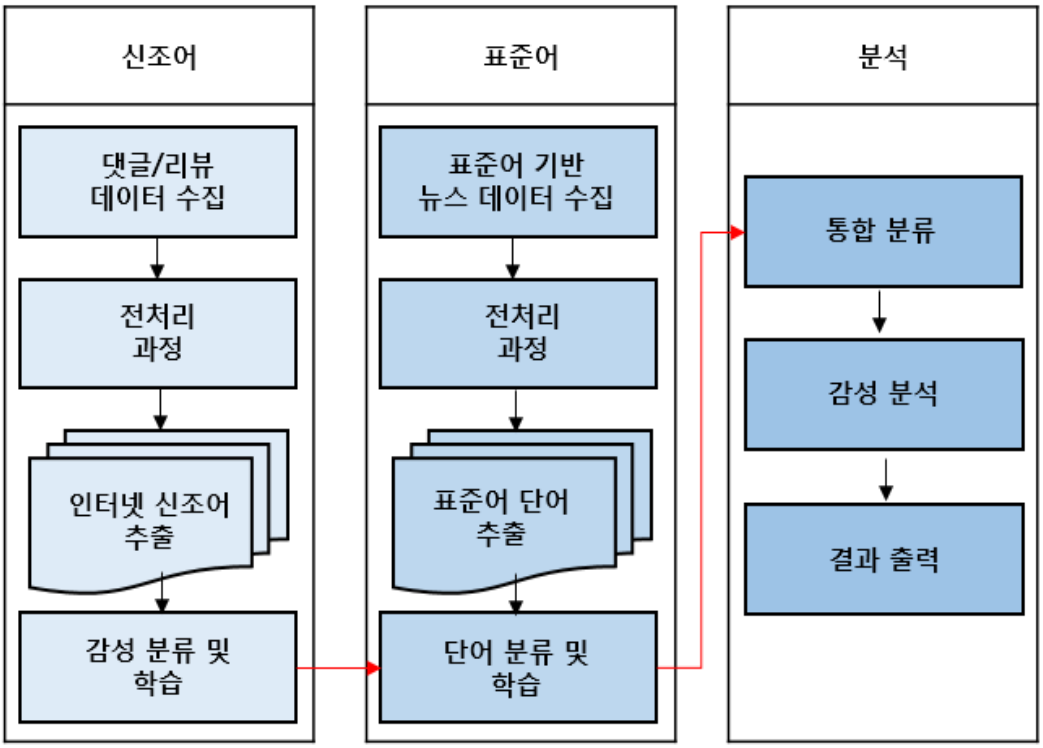
$$PMI(W1, W2) = \log \frac{p(W1, W2)}{p(W1)p(W2)} \quad (1)$$

수식(1)은 'PMI'기법에 관하여 수식의 형태로 나타낸 것이다. w1,과 w2은 각각의 단어의 연관성을 구하고자 하는 것을 나타내며, 같은 문서 안에 두 단어가 나타날 확률 값으로 표현된다. 만약 문서 안에 두 단어가 나타날 확률이 서로 독립적인 형태라면 확률 값은 '0'이 될 것이다. 수식을 통해 도출된 확률 값이 양수이면 비슷한 의미를 갖는다는 것을 뜻하며 반대로 음수이면 다른 의미를 갖는 다는 것을 의미 한다[13].

본 연구를 통해 이러한 기술들을 이용하여 신조어에 대한 분석을 통해 긍정/부정의 오피니언을 적용하여 무분별하게 사용되고 있는 신조어가 단순히 불용어 처리 되어 없어질게 아니라 단어들에 대한 의미와 감성이 적용되어 알고리즘 성능 저하를 막고 특정 주제에 관련하여 구체적인 오피니언 연구가 진행 될 수 있도록 하는 연구를 진행하였다.

Ⅲ. 신조어 긍·부정 감성 판별 기법

A. 긍·부정 판별에 대한 시스템 구성



[그림 3-1] 시스템 구성도

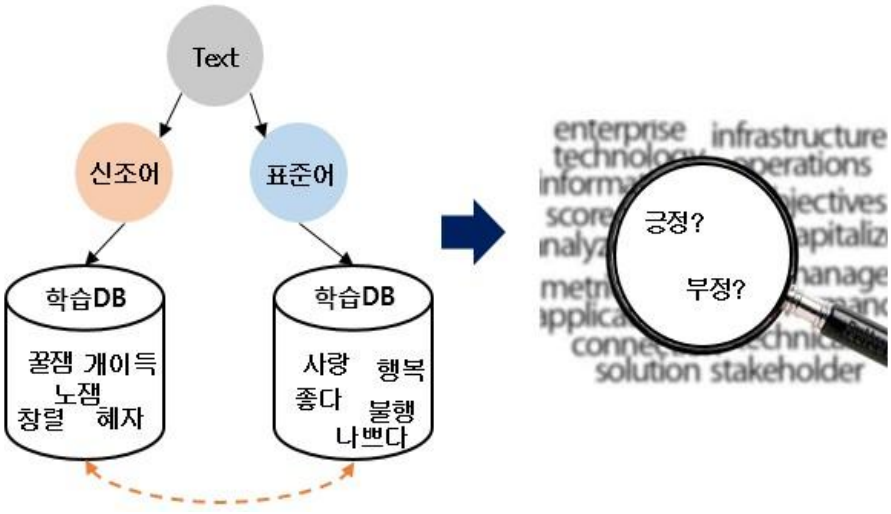
본 장에서는 서두에 소개한 인터넷 신조어에 대한 감성 분류에 따라서 대량의 텍스트 데이터와 언어의 폭이 넓어짐에 따라서 다량의 인터넷 신조어가 발생하는 댓글이나 리뷰를 통해 수집 후 긍정/부정 분류를 통해 기존의 긍·부정 감성 분석 연구보다 좀 더 알고리즘 성능 저하를 막고 사용자의 생각이나 의견을 구체적으로 알아보고자 한다.

[그림 3-1]은 텍스트 데이터에서 신조어 분류 후 긍정/부정의 분석을 통해 기존의 긍·부정 감성 분석 연구와의 조화를 통한 긍·부정 분석을 제안하는 시스템 구성도이다.

본 논문에서 제안하는 시스템 구성도는 국내 포털사이트에서 제공하는 카페 나 블로그 데이터를 수집 하고 수집된 데이터들은 신조어 추출을 위해 '한나눔' 한국어 형태소 분석기를 이용하여 진행한다[14, 15].

기존의 긍정/부정 감성 분석에 대한 연구에서는 리뷰나 인터넷 댓글의 어휘에 한정하여 통계적 수치 나 자연어처리 기법을 사용 하였으나 문맥에 따라 감성어의 의미가 다르게 분류 되는 경우가 발생 하거나 신조어에 의해 문장의 의미가 퇴색 되고 정확한 결과 값을 찾아내기 어렵다는 점의 문제가 발생 하였다. 본 연구에서는 이러한 점들을 보안하기 위해 신조어의 긍정/부정 분류를 토대로 실험을 진행 한다.

B. 신조어 긍·부정 감성 분석



[그림 3-2] 신조어/표준어 긍·부정 분류 판별 과정

신조어를 추출하기에 앞서서 예외 적인 부분은 인터넷 신조어중 대 부분 접두사 '개'를 붙여 사용 하는 경우가 많기 때문에 그러한 단어들은 욕설이 아닌 신조어로 판별하고 분류 하였으나 지역비하, 비속어, 인신공격 등 오해의 소지를 불러 일으킬 수 있는 단어들은 추출 후 제거 하였다.

1. 신조어 추출 (Extract New words)

인터넷상의 댓글, 리뷰를 통해 수집된 데이터를 토픽별로 문서에 저장 하여 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 3-1] 댓글/리뷰 원문 'Tokenizing' 적용

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

[표 3-2] 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
들	0.006869	지금	0.0007
⋮	⋮	⋮	⋮

수집 후 토크나이징을 적용한 텍스트 데이터는 [표 3-1]과 같이 저장 하였다. 인터넷 신조어를 판별하기 위해서 불용어 제거를 진행한다. 인터넷 검색 시 검색 용어로 사용하지 않는 단어로써 관사, 접속사 같은 품사 등은 의미가 없는 단어 이지만 포털 사이트의 검색 엔진 마다 동일하지 않기 때문에 다르다.

의미가 담긴 품사와 뒤에 붙는 (보)조사, 숫자, 특수문자를 제거한다. 형태소 분석을 통해 한국어 코퍼스에서 고빈도 불용어 리스트를 [표 3-2]와 같이 작성하여 빈도수가 많이 존재하는 단어들도 제외 한다.

[표 3-3] '[표 3-1]'의 데이터 불용어 제거

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

[표 3-3]과 같이 불용어 제거가 완료된 텍스트 데이터들은 각 문서의 키워드를 활용하여 빈도수 측정을 통한 신조어 키워드를 추출 한다.
이후 분석 작업을 통해 추론 된 단어들은 텍스트문서로 결과 값을 저장하고 'WordCount'를 통해 빈도수 상위 20개의 단어들만 추려서 저장 한다.

[표 3-4] 빈도수로 추론된 상위 단어

Word.	Count.
꿀잼	211
개꿀잼	154
노잼	103
창렬	98
갑분싸	78
가즈아	63
핵노답	61
노답	59
개이득	54
혜자	51
존못	49
극혐	44
호갱님	39
헬조선	36
명존세	33
천조국	31
위꼴	29
병맛	20
파오후	14
오덕후	10

빈도수 계산으로 [표 3-4]와 같이 상위 단어들 목록을 저장 하였다. 상위 랭크에 저장 될수록 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.

2. 신조어 긍정/부정 학습(Training New words)

신조어 추출로 인한 데이터들은 'Python'을 활용하여 학습을 통해 긍정/부정 값을 확인한다. 추출된 단어는 위키피디아에 있는 목록을 참고하여 단어의 의미를 확인 후

긍정 과 부정 문서에 저장 한다.

[그림 3-3] 대한민국 신조어 목록(위키피디아)

[그림 3-3]와 같이 위키피디아에 대한민국 인터넷 신조어에 대한 의미와 기원에 대하여 가나다순으로 나열해 신조어에 대해 긍정 부정을 확인 할 수 있다.

긍정/부정의 단어를 위키피디아를 참조하여 추출된 단어들은 문서별로 파이썬에 학습을 시켜 진행한다.

[표 3-5] 위키피디아를 참조한 긍정/부정 의미와 분류

Class	Word	Mean
Positive	꿀잼	너무 재미있다.
	개이득	완전 이익/이득
	혜자	가격대비 (품질/맛) 좋다
	위꼴	위가 뒤틀릴 정도로 맛있다(맛있어 보인다)
Negative	노잼	너무 재미없다
	노답	(사람 또는 특정주제) 답이 없다
	창렬	가격대비 (품질/맛) 좋지 않다
	극혐	극도로 혐오

분류된 단어들은 [표 3-5]와 같이 위키피디아를 참조하여 긍정/부정을 분류하고 각각의 단어가 의미에 인식하고 감성 학습을 위해 긍정과 부정 문서별로 구축을 진행 하였다.

분류된 감성 값이 긍정인지 부정인지 1차원적으로 판별을 위해 구축된 환경을 사용하여 신조어가 포함된 문장이 학습을 통해 구축된 데이터 안에서 적용 되었을 때 어떤 결과를 도출 할 수 있는지 실험을 진행 하였고 [표3-7]에 따라 [표 3-6]과 같이 결과 값을 확인 할 수 있다.

[표 3-6] 신조어 감성에 따른 긍정/부정 판별

Word	Sensitivity	Word	Sensitivity
꿀잼	positive	혜자	positive
개이득	positive	창렬	negative
노잼	negative	극혐	negative
헬조선	negative	노답	negative

[표 3-7] 신조어 긍·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)
```

C. 표준단어 긍·부정 감성 분석

인터넷 신조어를 통해 구축된 신조어에 대한 긍정과 부정 값에 대한 단어들만 감성이 분류되기 때문에 표준어에 대한 긍·부정 분석 기법과 통합하여 실험을 진행 한다.

신조어 구축과 마찬가지로 첫 번째 단계로 진행 했던 웹 크롤링으로 수집한 댓글/리뷰에 대하여 문장들을 형태소 분석을 통해 긍정과 부정을 검증을 한다. 'Doc2vec'을 활용하여 'Genism'패키지를 통해 긍정과 부정을 분류 한다. 한국어 문서들의 분류는 'KoNLP'를 사용하고 'NLTK'패키지 안의 'Naïve Bayes Classifier'라이브러리를 사용한다. 토픽 모델링을 지원하는 'Gensim'은 여러 문장이나 문서에 내재 되어 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.

표준단어의 사용이 잦은 네이버 뉴스의 데이터를 수집하여 수집된 뉴스 데이터에서 명사만을 활용하여 형태소 분석을 다음과 같은 과정을 이용하여 진행한다[16]. 빈번하게 발생하는 불필요한 어휘와 신조어와 다르게 의미를 알 수 없는 단음절 체언 및 용언을 제거한다. 최종적으로 추출된 명사의 감성 점수화를 통하여 데이터의 감성 값을 도출한다.

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

문서 내에서 등장하는 단어의 빈도를 나타나는데 단어와 문서간의 중요도를 나타내기 위해 수식(1)과 같이 TF 값을 적용한다. 문서 내에서 많이 출연 할수록 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.

$$TF-IDF(t,d,D) = TF(t,d)IDF(t,D) \quad (2)$$

수식(2)의 TF-IDF는 TF와 IDF를 곱하여 두 지표를 동시에 고려하는 가중치를 산출 하는 방법으로 어떤 단어가 얼마나 많이 쓰였는지 얼마나 특이한지 모두 반영 할 수 있는 수식 이다. 따라서 이 값을 적용하면 불용어를 걸러 낼 수 있으며 단어별 가중치를 알 수 있다.

감성 분류에서는 빈도수보다는 해당 단어가 자체가 있느냐 없느냐가 더 중요 할지도 모른다. 해당 단어가 출현을 하면 1로 간주하고 출현 빈도에 가중치 값을 곱한다. 도출된 가중치는 나이브 베이즈 분류법을 이용해 적용한다.

[표 3-8] Naïve Bayes 기법에 가중치 적용

	좋다	별로다	부족하다	neg/pos
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

본 장에서는 신조어 및 표준어 긍·부정 분석하여 특정 주제에 관한 문서의 감성 수치를 비교하여 성능을 평가한다.

먼저 실험은 3단계를 거쳐 진행한다. 첫 번째로 신조어 감성을 분석하여 신조어가 포함 된 문장이 긍정인지 부정인지 판단한다. 'Naïve bayes classification'을 사용하여 문장의 감성 값을 도출 후 다음으로 표준어 감성 분류 수치도 동일한 조건으로 구동하고 도출한다. 마지막으로 추출된 감성 값을 토대로 신조어와 표준어가 함께 사용하여 신조어가 포함된 문장이 감성 분류가 잘 도출되는지 확인한다[17].

A. 신조어 분석

시스템 구성에서 신조어의 감성만 분류하는 댓글/리뷰에 관한 문장들을 이용하여 긍정/부정을 분류 할 수 있는지에 대한 실험을 진행 한다.

$$p(c|x) = \frac{p(x|c)p(c)}{p(x)} \tag{1}$$

기존의 연구되었던 단순 베이즈 분류나 PMI수치를 통한 실험은 논문에서 제시하는 분류 방법에 비해 평균수치가 낮아지는 현상이 발생하여 본 연구에서는 신조어가 포함된 문장 안의 긍·부정 감성에 대해 조건부 확률을 이용하여 나이브 베이즈 분류에 대한 수식(1)과 같이 진행한다[18]. 신조어가 포함된 문장을 'x'라 하고 'c'는 긍정인지 부정인지 판단하는 'class'라 지칭한다.

$$p(x_1, x_2, x_3, \ldots x_n|c) = p(x_1|c)p(x_2|c)p(x_3|c)\ldots p(x_n|c) \tag{2}$$

수식(1)에 대하여 입력 문장이 여러 단어로 이루어져 있기 때문에 개수를 n으로 놓고 수식(2)와 같이 사용한다.

[표 4-1] 신조어 감성 확률 값 도출을 위한 알고리즘

```
def naive_bayes_classifier(test, train, all_count):
    counter = 0
    list_count = []
    for i in test:
        for j in range(len(train)):
            if i == train[j]:
                counter = counter + 1
        list_count.append(counter)
        counter = 0
    list_naive = []
    for i in range(len(list_count)):

list_naive.append((list_count[i]+1)/float(len(train)+all_count))
    result = 1
    for i in range(len(list_naive)):
        result *= float(round(list_naive[i], 6))
return float(result)*float(1.0/3.0)
```

[표 4-2] 추출된 신조어 감성 분류 예시

문장(Sentence)	pos:neg	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... **노잼**이군 ”	2.2:1.0	negative
“현지화 다됐네... 역시 **헬조선** ”	1.8:1.0	negative

여러 문서안의 신조어가 포함된 문장들 중에서 조건부 확률 값을 구하기 위해 ‘Python’을 사용하여 [표 4-1]과 같이 ‘Python’ 코딩을 진행 하였고 도출 된 값은 [표 4-2]와 같이 감성 값을 확인 할 수 있다.

B. 표준어 분석

신조어 분석 연구와 결합하여 사용 하게 될 표준어 단어의 긍·부정 감성 분석은 결과 수치의 정확성을 위하여 표준어 문장의 사용 빈도가 잦은 '뉴스' 데이터와 신조어가 포함된 일반적인 댓글/리뷰를 사용하여 [표 4-3]과 같이 수집을 통해 얻어진 데이터와 비교 분석을 진행한다.

[표 4-3] 뉴스/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 생각한다"고 보도했다. 소식통은 북한 매체가 남북정상회담 이후 "한반도 통일의 문이 활짝 열린 것처럼 요란하게 선전하고 있다"면서 "그러나 일부 고위 간부는 하급 간부들에게 노골적으로 판문점 선언에 대해	여러 제품을 써봤지만 이 제품처럼 만족스러운 구매는 없었던 것 같아요. ^^ 꿀템은 처음이예요~ ..... (중략) ..... 애들도 좋아하고 맛도 좋고 특유의 비린맛도 없어서 ~~ 호호 요즘 애들 말로 개이득~ 이라고 한다죠 꿀템 ^^

실험을 통해 도출된 값은 [표 4-4]와 같이 비교 분석을 통해 도출된 결과를 확인 한다. 뉴스 기사의 텍스트는 일반적으로 특수한 주제가 아닌 경우에는 표준어를 사용하기 때문에 댓글/리뷰의 텍스트 데이터와 비교 시에 차이가 분명하게 드러난다. 특정 제품의 경우 3·40대가 자주 구매하는 제품의 리뷰를 수집해서 인터넷 신조어의 사용이

일반적인 댓글에 비해 적으나 표준어 기반 감성 분석을 적용 했을 시에는 분명한 차이를 보인다.

[표 4-4] 표준어 감성 분류를 적용한 비교분석

	뉴스	댓글/리뷰
사용한 데이터	'북한'키워드	특정 제품
사용언어	표준어	표준어
정확도	0.90231	0.54822

C. 평가 및 결과

본 논문에서 제안하는 방법을 이용하여 최종적인 결과 도출을 위하여 [그림 4-1]과 같이 생성과정을 통하여 실험을 진행 한다.

신조어와 표준어 긍·부정 감성 분석을 통한 연구를 토대로 각각의 감성별로 학습과 실험을 반복하여 결과의 정확성을 높인다. 이러한 과정을 거쳐 문서에 포함된 단어의 긍·부정 수치를 확인한다.

학습된 데이터 값은 웹 크롤링을 통해 수집된 텍스트 데이터를 활용하여 댓글/리뷰 및 뉴스 데이터의 데이터가 어떠한 긍·부정 감성 값을 도포 하고 있는지 확인한다.
실험을 통해 진행된 결과 값은 정확도와 감성수치에 대한 결과를 도출하고 기존에 연구되었던 표준어 감성 분석 기법과 비교를 통해 수치 비교를 진행한다.

[표 4-5] 문서별 표준어/신조어 분석

Document	Contents
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**..
⋮	⋮

[표 4-6] 표준어/신조어 결과 값 도출

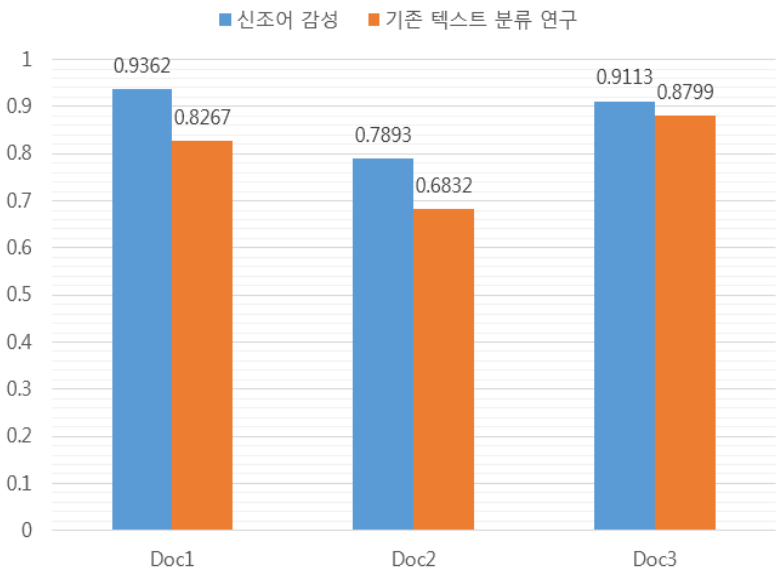
Document	표준어	신조어	Pos/Neg	결과
doc1	만족	꿀잼	Pos.	0.9362
doc2	어렵다	헬조선	Neg.	0.7893
doc3	비싸다	창렬	Neg.	0.9113
⋮	⋮	⋮	⋮	

제안하는 신조어 감성과 표준어 감성에 대하여 결합을 통해 결과 값 [표 4-5]를 통해 각 문서안의 문장에 신조어와 표준어가 함께 쓰인 단어들을 확인하고 각 문서별로 결과 값을 출력 한다. 확인된 결과는 [표 4-6] 같이 표준어와 신조어가 포함된 문장이 긍정인지 부정인지 최종 결과를 출력하고 도출된 결과 값은 문장 수치를 확률 값으로 전환 하고 도출된 결과 값을 확인한다. ‘doc2’의 경우에 ‘doc1,3’의 결과 보다 낮은 결과 값을 확인 할 수 있는데 이는 ‘doc2’의 문장에서 부정의 의미가 내포되어 있으나 [표 4-5]에 ‘꿀잼’이라는 긍정의 단어가 포함되어 있어서 부정의 결과를 나타내지만 낮은 결과 값이 도출 된 것을 확인할 수 있다.

기존의 표준어 분석 알고리즘을 통해 추출된 정확률과 본 논문에서 제안하는 결과 값의 정확도를 평균화 하여 다음과 같이 도출한다.

[표 4-7] 기존 연구와 제시한 연구에 대한 비교 수치

텍스트 분류를 통한 연구		신조어/표준어 감성 연구	
Document	Result	Document	Result
doc1	0.8967	doc1	0.9362
doc2	0.6832	doc2	0.7893
doc3	0.8799	doc3	0.9113
⋮	⋮	⋮	⋮



[그림 4-2] 비교 분석 및 평가 수치

실험을 통해 진행된 결과는 [표 4-7]과 [그림 4-2]를 토대로 시각화 하여 비교 분석 한다. 비교분석을 진행하기 위해 기존에 연구된 표준어 감성 알고리즘과 비교 분석을 진행 한다. 기존의 알고리즘은 텍스트 분류를 통한 단순 빈도수로 결과 값을 도출하고 불용어 수준의 단어를 정확하게 처리하기가 어렵다는 단점이 존재하여 객관적인 수치

를 나타내는데 어려움이 존재한다.

기존의 연구에서 진행된 표준어 감성에 비해 제안하는 알고리즘이 평균 0.21 정도의 미세한 수치상승이 나타나 제안하는 방법에 대한 효율성을 입증할 수 있는 결과를 시각화를 통해 그래프로 확연한 차이를 확인할 수 있다.

[그림4-2]에 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 도출된다. 실제로 문서를 확인 하였으나 컴퓨터가 이해하기 어려운 내용을 갖고 있어 다른 문서에 비해 높은 결과 값을 도출하기 어렵다.

V.결론 및 제언

본 논문에서는 인터넷의 특성상 신조어가 계속 생겨남에 따라 표준어와 달리 매순간마다 계속 생겨나고 있으며 인터넷 발달과 스마트폰의 발달로 인해 실시간으로 수많은 데이터들이 생겨나며 의미를 갖지 않는 신조어로 인하여 데이터 마이닝 기술을 하는데 어려운 문제점이 발생한다. 이러한 문제점으로 알고리즘의 성능 저하와 데이터 손실을 막기 위해 신조어 긍·부정 감성 분석을 진행한다.

본 논문에서 제안하는 방법으로 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 불용어로써 제거되었던 신조어가 감성 의미를 가짐으로써 기능을 다하여 결과 값이 향상 되는 점을 확인 하였다.

본 연구에서는 신조어에 대한 감성 분석에 대한 연구로 인해 단어에 대한 기능을 할 수 있도록 하였으며 신조어는 더 이상 제거 대상이 아닌 필수 요소로 자리 매김해야 할 것이다. 또한, 신조어는 인터넷의 한 문화에서 파생된 단어로 표준어와 같이 반드시 써야 할 단어는 아니지만 인터넷의 문화가 점점 계속 될수록 단어는 앞으로도 계속 생길 것이며 그에 대한 지속적인 대처가 필요 할 것이고 그러한 환경구축에 대한 연구를 진행 할 예정이다.

참고 문헌

[1] 장경현. “신조어 연어의 형성원리”, 인문노총, 제66집, pp.269-297, 2011.

[2] 강아름, 이상연, 이 건. “매스미디어 상 인터넷 용어 처리를 위한 은닉 마코프 모델 기반 신조어 추출”, 한국지능시스템학회 : 한국시스템학회 학술발표 논문집, 제25권 제1호, pp.119-120, 2015.4.

[3] 안정은. “Text Mining 기법을 이용한 표준특허기술의 유사도 측정방법”, 한국정보과학회 학술발표 논문집, 제36권 제1호, pp.1-5, 2009.6.

[4] 이한동, 김종배. “복합명사를 포함하는 개선된 키워드 추출 방법”, 예술 인문 사회 융합 멀티미디어 논문지, 제7권 제10호, pp.857-864, 2017.

[5] 이성직, 김한준. “TF-IDF의 변형을 이용한 전자뉴스에서의 키워드 추출 기법”, 한국 전자 거래학회지, 제14권 제4호, pp.79-73, 2009.11.

[6] https://ko.wikipedia.org/wiki/대한민국의_인터넷_신조어_목록

[7] 장경애, 박상현, 김우제. “인터넷 감정기호를 이용한 긍정/부정 말뭉치 구축 및 감정분류 자동화”, 제42권 제4호, pp.512-521, 2015.04.

[8] 이종화, 레환수, 이현규. “오피니언 마이닝을 통한 국내와 수입 의류 제품에 대한 고객 평판 연구”, 제15권 제3호, pp.223-234, 2015.06.

[9] 김동성, 김종우. “온라인 여론의 감성분석을 위한 감성용어 자동화 추출 방안 연구”, 한국경영정보학회 학술대회논문집, 제2016권 제6호, pp187-189, 2016.

[10] 이영민, 권 필, 유기윤, 김지영. “한국어 장소 리뷰를 이용한 공간 감성어 사전 구축 방법”, 제25권 제2호, pp.3-12, 2017.6.

[11] 박인조, 민경환. “한국어 감정단어의 목록 작성과 차원 탐색”, 한국 심리 학회, 한국 심리 학회지: 사회 및 성격, 제19권 제1호, pp.109-129, 2005.2.

[12] 송상일, 이동주, 이상구. “PMI를 이용한 우리말 어휘의 의미 극성 판단”, 한국정보과학회 학술 발표 논문집, 제37권 제1호, pp.260-265, 2010.6.

[13] Turney and M. Littman. "Measuring praise and criticism: Inference of semantic orientation from association", Proceedings of ACL-02, 40th Annual Meeting of the Association for Computational Linguistics, pp.417-424, 2002.7.

[14] 조하나, 정연오, 이재동, 이지형. “인터넷 뉴스 댓글의 감성 분석을 통한 오피니언 마이닝”, 한국 지능 시스템 학회 학술발표 논문집 제23권 제1호,

pp 149-150, 2013.04.

[15] 한나눔 형태소 분석기.(http://semanticweb.kaist.ac.kr/home/index.php/HanNanum)

[16] 최성자, 손민영, 김영학. “키워드 기반 블로그 마케팅을 위한 연관 키워드 추천 시스템”, 정보과학회 컴퓨팅의 실제 논문지, 제22권 제5호, pp.19-22, 2015.

[17] 안광모, 김윤석, 김영훈, 서영훈. “Levenshtein 거리를 이용한 영화평 감성 분류”, 한국디지털콘텐츠학회논문지, 제14권 제4호 pp.581-587, 2013.12.

[18] 최용화, 이일병. “음절 발생 조건부 확률을 이용한 음절 추천 시스템”, 한국 정보과학회 한국컴퓨터종합학술대회 논문집, 제36권 제1호, pp336-340, 2009.6.