ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2016 September: 22(3): 91~111 http://dx.doi.org/10.13088/jiis.2016.22.3.091

다국어 소셜미디어에 대한 감성분석 방법 개발: 한국어-중국어를 중심으로

최미나
경희대학교 일반대학원 경영학과
(*minna818@gmail.com*)

진윤선
경희대학교 일반대학원 경영학과
(*dudnrha@khu.ac.kr*)

권오병
경희대학교 경영대학
(*obkwon@khu.ac.kr*)

소비자들이 소셜미디어 상에 기록한 글을 통해 기업은 제품 또는 기업 이미지에 대한 감성분석을 수행하는데 이는 소셜미디어 기반 마케팅에서 중요한 활동 중에 하나다. 특히 글로벌 소셜미디어의 경우 국적을 불문하고 다양한 고객이 늘어남에 따라 여러 언어권의 소비자들이 각자의 언어로 다양한 의견을 표명하고 있다. 이처럼 다양한 언어로 작성된 텍스트를 감성분석하기 위해서는 기존 방법과 달리 동일한 언어로 통일시켜야 하는 번역 작업이 필요하다. 하지만 번역을 하게 될 경우, 언어와 관련된 배경이나 문화, 용어사용의 차이 등으로 본래 문서에 있는 모든 단어나 문법을 정확히 표현할 수 없는 문제점이 있다. 따라서 본 연구에서는 다중 언어로 수집되는 텍스트를 번역하지 않고 해당 언어별로 텍스트를 분리한 다음 감성분석을 진행하여 각각의 극성치를 종합하는 방법을 제안하고자 한다. 본 연구에서 제안한 다국어 감성분석 알고리즘을 검증하기 위해 다중언어 문장을 한국어, 중국어로 번역한 감성분석의 극성치 편차인 RMSE 값을 비교하였다. 그 결과, 번역을 통한 다중언어의 감성분석보다 언어별로 분리한 감성값이 실제 감성값에 가장 근접하는 것으로 나타나 본 연구에서 제안한 방법론의 우수성을 입증하였다. 본 연구는 다수의 유사한 연구에서 사용했던 알고리즘을 사용하지 않고 원문 그대로 다중언어 감성분석을 시도했다는 점에서 의의가 있다.

주제어 : 감성분석, 다국어, 소셜미디어 마케팅, 텍스트마이닝, SentiWordNet

논문접수일 : 2016년 6월 24일 논문수정일 : 2016년 8월 10일 게재확정일 : 2016년 8월 11일
원고유형 : 일반논문 교신저자 : 권오병

1. 서론

최근 텍스트마이닝 기법을 이용한 언어 감성분석 알고리즘 관련 연구는 Word-of-mouth (WOM) 마케팅이나 실시간 감성분석 등과 같은 비즈니스 영역에서 많이 활용되고 있다(Boiy and Moens, 2009). 텍스트마이닝에 의한 감성분석을 위해서는 다양하면서도 고품질의 감성어휘 사전을 확보하는 것이 성능에 많은 영향을 준다. 따라서 언어별로 감성어휘를 확보하고 각 언어적 특징을 반영한 감성 분석을 실시하게 된다. 초기 연구는 대부분 영어로 된 리뷰나 댓글을 감성분석 하는데 집중되었다(Pak and Paroubek, 2010; Gamallo and Garcia, 2014; Go and Huang, 2009). 반면 최근에는 아랍어(Abdul-Mageed et al., 2014), 중국어(Zhang, 2009), 네덜란드어(Atteveldt et al., 2009), 프랑스어(Ghorbel and Jacot, 2011), 터키어(Vural et al., 2013) 등 다양한 언어를 대상으로 감성분석을 진행하는 연구가 확산되고 있다. 이런 시도들은 각 단일 언어 체계에서의 감성분석에 도움을 주고 있다.

전 세계적으로 문화가 서로 융합됨에 따라 비

즈니스 측면에서의 고객도 다언어권으로 바뀌어 가고 있으며, 각 언어권의 사용자들이 동일한 콘텐츠를 보고 서로 간의 경험이나 느낌 등을 한 사이트에 공유할 수 있게 되었다. 한국도 한류의 영향으로 동북아 및 동남아 등에 한류 마케팅이 가능하게 되었다. 이에 따라 동일한 콘텐츠에 여러 가지 언어로 쓰인 리뷰나 댓글이 등장하는 현상이 점점 많아지고 있다. 예를 들면, 대표적으로 동영상 공유 사이트인 유튜브나 사진 공유를 위한 인스타그램, 호텔예약 사이트 Expedia 등에서 사람들은 언어에 제한 받지 않고 자신의 의견이나 관점을 댓글로 자유롭게 표현하고 있다. 그 이유는 Facebook이나 Twitter와 같이 지인들 사이의 정보 공유나 의사소통을 돕는 유형이 상호 이해 가능한 언어로 작성되는 것과는 달리, 이 사이트들은 공유된 콘텐츠에 대해 각자의 생각을 담기 때문에 다양한 언어로 구성되기 때문이다. 따라서 단일언어 감성분석 위주로 되어 있는 기존 방법으로는 다국어로 표현된 사이트에 대한 텍스트 분석을 하기가 불편하다(Bal, 2011).

다언어 감성분석을 위한 기존의 접근 방식은 다양한 언어의 댓글 등에 대해 우선 단일 언어로 번역한 후 번역된 텍스트로 분석하는 것이다(Bautin et al., 2008; Martín-Valdivia et al., 2013; Hajmohammadi, et al., 2015). 그러나 번역 알고리즘을 사용하여 한 가지 언어로 통일시킨 후 감성분석을 진행하는 것은 다음과 같은 문제들이 존재한다. 첫 번째는 두 언어에 관련된 배경이나 문화가 차이가 있고, writing style, 사용되는 linguistic term이 다르기 때문에 번역을 한다고 해도 원 문서에 있는 모든 단어나 문법을 정확히 표현할 수 없다(Hajmohammadi, et al., 2014). 두 번째는 번역 알고리즘에서 사용하는 사전에 의해 번역하는 것 자체에 착오가 존재한다는 것이다(Hajmohammadi et al., 2014). 즉, 번역 과정에서 단어의 뜻이 달라지는 경우가 있어 감성의 미묘한 왜곡이 일어나는 단점을 가진다. 예를 들어 중국어 “离谱”는 형용사로 “황당하다”라는 뜻으로 가장 많이 쓰임에도 불구하고 인터넷 네이버 사전(http://dic.naver.com)에서는 “노래를 부르는 것이 악보에 맞지 않다”, 인터넷 구글 사전(translat.google.com)에서는 “난폭한”으로 번역된다. 그러므로 언어의 번역과정에서 일어나는 극성치의 변경 현상을 최소화하는 방법이 필요하다.

따라서 본 연구는 다중 언어로 수집되는 텍스트를 번역하지 않고, 해당 언어별로 텍스트를 분리한 다음 각각 감성분석을 진행하고, 나중에 각각의 극성치를 종합하는 방법을 제안하고자 한다. 구체적으로 유튜브나 인스타그램과 같이 한 콘텐츠에 여러 언어권의 텍스트가 존재하지만 한 단위 글에는 하나의 언어로만 작성된 경우의 다국어 감성분석을 제안하고자 한다. 또한 본 연구는 여러 언어들 중에서도 중국어와 한국어가 결합된 경우의 감성분석에 집중하고자 한다.

본 연구의 구성은 다음과 같다. 먼저 2장에서는 다중언어 감성분석에 대한 기존 연구를 고찰하고, 3장에서는 제안한 방법론을 기술하였다. 본 연구에서 제안한 방법의 활용 가능성을 검증하기 위하여 4장에서는 실험을 진행하였으며, 5장과 6장에는 각각 연구의 의의를 토론하고 결론을 맺었다.

2. 문헌연구

기존의 연구에서는 단일 중국어, 단일 한국어를 대상으로 감성분석을 진행해왔다. 중국어는

Machine Translation Services, Machine Learning Method, N-Grams, Document Frequency Method 등과 같은 기법을 이용하여 감성분석 연구를 진행하고 있다(Wan, 2008; Tan and Zhang, 2008; Zheng et al., 2015). 또한 중국어 감성분석 알고리즘을 기반으로 '기차표 실명제(火车票实名制)'에 관한 블로그 글, '7.23 Wenzhou Train Collision'에 관한 공공 블로그 글, 모바일 폰에 관한 리뷰 등을 감성 분석하는 연구들이 진행되어 왔다(Fu et al, 2013; Shi et al., 2013; Zheng et al., 2015).

한국어 감성분석 기법으로는 Machine Learning, Transformation-based Learning method 등이 있고(Yang and Ko, 2014;), 감성분석을 응용한 연구로 트위터 감성분석을 통해 최근 동향에 대한 여론을 분석하는 연구, 뉴스를 감성 분석하여 주가 시장을 예측하는 연구 등이 있다(Lee and Lee, 2013; Kim et al., 2014).

다중 언어로 구성된 텍스트의 감성분석에 관해서도 많은 연구가 진행되고 있다. 그 중 다중언어 사전과 기계번역 방법을 적용한 기법이 가장 많이 응용되는데, 특히 기계번역기술을 사용하여 영어 외 언어를 영어로 번역한 후, 감성분석을 진행하는 방법이 가장 보편화된 방법이다. Bautin(2008) 등의 연구에서는 Arabic, Chinese, English 등 9가지 언어를 IBM WebSphere Translation Server를 이용하여 영어로 번역한 후 Lydia sentiment analysis system을 통해 감성분석을 진행하였다. Denecke(2008)의 연구에서는 독일어와 영어를 대상으로 감성을 측정하는 연구를 하였으며, 이 연구에서는 독일어를 영어로 번역한 후 SentiWordNet을 토대로 감성분석 실험을 진행하였다. 그 외에 영어 감성분석 시스템을 machine-translated data 및 감성사전 구축을 통하여 스페인어에 적용시키는 것을 성공하면서 다른 유럽권 나라의 언어(독일어, 프랑스어 등)도 적용하여 다중언어 감성분석의 정확도를 평가하는 노력도 있었다(Balahur and Perea-Ortega, 2015).

이와 같이 번역을 통한 감성 분석 외에도 다른 방법을 사용하여 다중 언어 감성분석을 진행하고자 하는 시도가 증가하고 있다. Boiy(2009) 등의 연구에서는 영어, 네덜란드어, 프랑스어로 된 블로그 글 및 리뷰를 대상으로 다중 언어 감성분석 연구를 진행하였다. 이 연구에서는 여러 가지 분류 알고리즘(Support Vector Machine, Multinomial Naive Bayes, Maximum Entropy)과 feature selection method(Unigrams, Stems, Negation, Discourse features)를 응용하였으며, 감성분석 정확도는 영어 83%, 네덜란드어 70%, 프랑스어 68%로 나타났다. 또한 Bal(2011) 등의 연구에서는 영어와 네덜란드어를 대상으로 언어 독립적 방법(language-independent method)을 이용하여 감성분석을 진행하였고, 그 결과 Dutch 79%, 영어 71%라는 정확도를 나타냈다. 그러나 이 연구는 여러 가지 언어가 포함된 상태에서 감성분석을 수행하지 않은 한계점이 있다. 한편 SentiCorr라는 다중 언어 감성분석 툴(Tromp, 2011)은 language identification, part-of-speech tagging, subjectivity detection and polarity detection 등 네 가지 과정을 통하여 다중언어의 감성을 측정하였다. 결과적으로는 긍정적인 감성과 부정적인 감성 측정이 가능하나 중립적인 감성을 따로 구분하여 측정하지 않았다. 따라서 다중언어로 구성된 댓글의 감성점수를 알기 위하여 한 언어에 맞추어 통합하는 비효율적이고 오류가 많은 방식을 탈피한 개선된 감성분석 알고리즘이 필요하다.

〈Table 1〉 Multilingual Sentiment Analysis

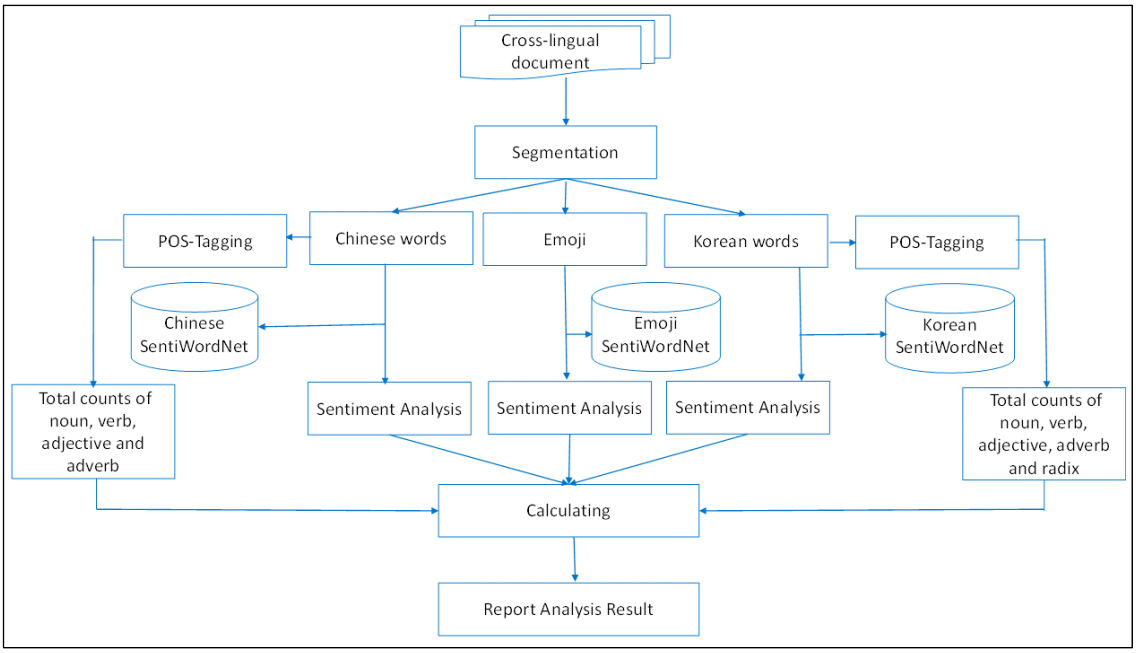
Developers	Language	Sentiment Analysis Method
Bautin et al., 2008	Arabian, Chinese, English and 6 more languages	IBM WebSphere Translation Serve Lydia sentiment analysis system
Denecke, 2008	German, English	Language translation severce Language classification Language standardisation Sentiment classification
Balahur and Perea-Ortega, 2015	English, Spanish	Machine translation Support vector machine Sequential Minimal Optimization
Boiy et al., 2009	English, Dutch, French	Supervised classification algorithm Feature selection method
Bal et al,. 2011	English, Dutch	Language-independent method
Tromp, 2011	English, Dutch	language identification, Part-of-speech tagging, Subjectivity detection Polarity detection

3. 연구 방안

3.1 시스템 프레임워크

본 연구에서 제안하는 다중언어 감성분석 프레임워크는 <Figure 1>과 같다. 우선 입력 받은 언어 별로 문장을 분리하는(segmentation) 과정을 거쳐 중국어 단어와 한국어 단어 즉, 단어를 단위로 나누어 저장한다. 본 연구에서는 한국어 형태소 분석기 Rhino 1.0을 사용하였고, 중국어 Segmentation Tool로는 Jieba를 사용하였다. 또한 중국어 형태소 분석기로는 Stanford POS Tagger (Toutanova et al., 2003) 중 Chinese Stanford Tagger을 사용하였다.

다음으로 중국어 단어를 형태소 분석기에 돌려 입력 받은 중국어 문장에 포함된 명사, 동사, 형용사, 부사의 개수를 센다. 한국어도 마찬가지로 한국어 형태소 분석기를 통해 입력 받은 한국어 중 명사, 동사, 형용사 및 부사의 개수를 센다. 또한 단어 단위로 저장한 중국어를 중국어 감성사전(예: HowNet)을 통해 중국어의 감성점수를 구한다(Dong&Dong, 2000). 한국어도 중국어와 같이 한국어 감성사전(예: 한글 SentiWordNet)을 통해 감성점수를 얻는다(Baccianella et al., 2010). 그 외에 댓글이나 코멘트에 많이 등장하는 이모티콘도 감성을 가지고 있기 때문에 본 연구의 다중언어 감성분석에 추가하였다. 이모티콘 감성사전은 Emoji Sentiment Ranking 1.0(Novak et al., 2015)을 사용하였다. 마지막으로 한국어, 중국어 및 이모티콘의 감성점수 그리고 명사, 동사, 형용사, 부사, 어근(한글만) 및 감성 이모티콘의 개수를 통해 전체 다중언어의 감성점수를 구하게 된다.



〈Figure 1〉 The Proposed Framework

3.2 언어별 감성사전 구축

감성분석을 하기 위해서는 감성어휘사전이 필요하다. 그러나 한글감성사전 중 이미 알려진 Korean SentiWordNet을 사용하기에는 사전에 포함된 긍·부정 단어의 개수가 충분하지 않다. 따라서 본 연구에서는 한글감성사전의 긍정, 부정 단어의 개수를 추가하여 정확도가 향상된 감성분석을 진행하고자 자체로 한글감성사전을 구축하는 작업을 진행하였다. 한글감성사전을 구축하는 절차는 아래와 같다.

단계 1: 영문 워드넷 SentiWordNet 3.0의 단어를 선택한다. 이 때, 형태소와 영어단어를 함께 참조하며 반복되는 것이 있으면 제거한다.

단계 2: 영문 단어와 형태소를 참조하여 구글 번역기를 통해 한글 번역을 진행한다. 영문 단어의 형태소를 참조했기 때문에 형태소에 해당하는 번역결과 중 첫 번째 한글단어를 선택한다.

단계 3: 한글 번역 결과 중 영어문자 혹은 숫자가 들어간 단어나 띄어쓰기가 존재하는 단어(구절)는 제거한다. 도출된 결과는 형태소와 함께 참조하며 구글 번역기를 이용해 영어로 다시 번역한다. 이 단계에서도 단계 2와 같이 형태소에 해당하는 번역결과 중 첫 번째 영어단어를 선택한다.

단계 4: 위 단계를 통해 파악된 영어단어 및 형태소를 참조하여 SentiWordNet 3.0에서 감성점수를 찾는다.

단계 5: 감성점수가 찾아지면 영어단어의 감성점수를 한글단어의 감성점수로 선정한다. 만약 찾아지지 않으면 한글단어의 감성값은 0으로 책정한다.

단계 6: 번역과정 중에서 형태소가 원 번역하고자 하는 단어의 형태소와 일치하지 않을 경우 'unknown_tag'라고 표시하고 추후 한글감성사전 구축 시 제거한다.

단계 7: 얻어진 한글감성단어를 Rhino 1.0 형태소 분석기를 이용하여 각 단어들을 분리한 후 결과값 중 첫 번째로 분리된 값과 그 형태소를 가지고 본 연구에서 사용한 한글감성사전을 구축한다. 예를 들면, "인정하는"라는 단어를 Rhino 1.0을 통해 형태소를 분석하면 "인정/NNG +하/XSV +는/ETM"이 된다. 이 중 본 연구에서는 첫 번째 결과값" 인정/NNG"로 "인정하는"을 대체하여 한글감성사전을 구축하였다. 또한 Rhino 1.0을 통해 분리되지 않는 단어는 그 단어 자체(형태소 포함)로 한글감성사전에 추가하였다. 예를 들어 "출산"이라는 단어의 Rhino 1.0 형태소분석기 결과값은 "출산/NNG"임으로 "출산/NNG" 그대로 한글감성사전에 추가하였다.

단계 8: Rhino 1.0 형태소 분석기를 통해 단어 분리가 끝난 후 일련의 정제작업을 진행한다. 첫 번째는 분리된 결과의 첫 번째 값이 그 단어를 표현할 수 없다고 판단했을 경우 그 단어를 예외 값으로 처리하고 한글감성사전에 추가하지 않는다. "빈틈없는"이라는 단어를 예로 들 때 Rhino 1.0 형태소 분석기 결과가 "비/NNG +ㄴ/ETM + 틈/NNG + 없/VA + 는/ETM"으로 그 첫 번째 결과값 "비/NNG"가 "빈틈없는"이라는 단어를 표현할 수 없다. 따라서 "빈틈없는" 이 단어를 한글감성사전에 추가하지 않았다.

단계 9: 위 과정이 끝난 후, 두 번째 정제작업을 진행한다. Rhino 1.0로 분석한 단어, 즉 영어를 한글로 번역했을 때의 한글단어 및 그 형태소를 세트로 반복되는 단어를 삭제하는 과정을 거친다. 삭제 원칙은 아래와 같다. 한글단어와 그 형태소가 세트로 반복되는 것이 있을 때, 한글을 영어로 번역한 영어단어와 원본 워드넷 영어단어가 같은 세트를 선택하고, 나머지 반복되는 세트는 삭제한다. 만약 번역한 영어단어와 워드넷 영어단어 중 같은 단어가 존재하지 않으면 한글단어를 Rhino 1.0으로 분석한 결과 중 첫 번째 값의 형태소와 같은 세트를 선택한다. 그 중 같은 형태소로 되어 있는 경우가 많으면 임의로 한 세트를 선택한다. Rhino 1.0의 첫 번째 결과의 형태소와 단어의 형태소가 다르더라도 반복되는 단어의 형태소가 모두 같으면 위와 같이 임의로 선택한다. 이렇게 할 수 있는 원인은 영어단어가 다름에도 불구하고 한글로 번역한 결과(단어와 형태소) 및 감성점수가 모두 동일하기 때문이다. 그 외에 한글은 같아도 형태소가 서로 다르면 반복되더라고 삭제하지 않고 그대로 두었으며, 이 부분은 Rhino 1.0 첫 번째 결과값의 반복되는 부분을 정제과정에서 해결한다.

단계 10: 위 과정이 끝나고 Rhino 1.0 첫 번째 결과값 중 반복되는 것을 처리하는 과정

을 거친다. 예하면 “웅대한/adj”의 Rhino 1.0 첫 번째 결과값은 “웅대/XR”이고, “웅대/noun”의 Rhino 1.0 첫 번째 결과값도 “웅대/XR”이다. 따라서 본 연구에서는 “웅대/XR”의 감성점수를 얻기 위하여 두 경우의 평균 감성점수를 구했다. 즉, 첫 번째 형용사일 경우 감성점수는 0.2749, 두 번째 명사일 때는 0.2917로 “웅대/XR”의 감성점수는 0.2833이 된다.

이렇게 총 10단계를 거쳐 구성한 한글감성사전은 한글단어, Rhino 1.0을 통해 단어를 분리한 첫 번째 값과 Rhino 1.0 형태소 분석기를 통해 얻은 각 단어의 형태소 및 각 한글단어의 감성점수로 구성되었다.

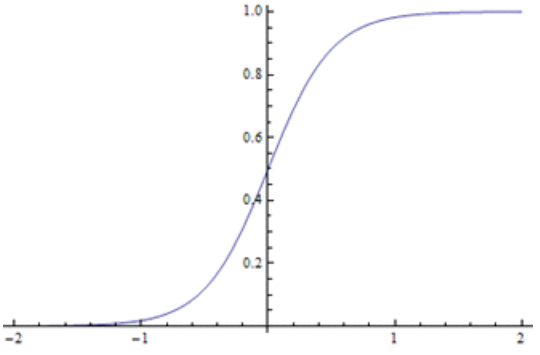
다음은 영문 SentiWordNet 3.0을 기반으로 한글감성사전을 만드는 과정 중 감성단어 개수의 변화를 보여준다(<Table 2> 참조).

〈Table 2〉 Procedure to construct Korean senti-wordnet

Step	Number of words
Original SentiWordNet 3.0	155,287
Korean senti-wordnet (getting rid of alphabets, space, numbers, etc.) (Step 1~3)	45,338
Korean senti-wordnet (getting rid of the words which do not appear in original SentiWordNet) (Step 6)	25,218
Korean senti-wordnet (getting rid of the redundant words, and further refinement) (Step 8~10)	10,719

$$Y = \frac{1}{1+e^{-\alpha x}} \quad (1)$$

구축된 한글감성사전은 중국어 감성사전 HowNet과 마찬가지로 긍정단어의 감성점수는 1, 부정단어의 감성점수는 -1, 중성단어의 감성점수는 0으로 정수화하여 감성분석을 진행한다. 이를 위해서 한글 감성사전 감성값을 Sigmoid function(수식1)에 입력하여 그 결과값이 일정 상한선 수준이면 긍정단어로, 일정 하한선 수준이면 부정단어로, 일정수준 범위 내에 있으면 중립단어로 판정하였다. Sigmoid function 중 α 값은 9로 정하였다.



〈Figure 2〉 Standard logistic sigmoid function

이렇게 총 10 단계를 거쳐 구성한 한글감성사전 Sigmoid function을 통해 10,719개의 한글감성단어로 구성된 감성사전은 긍정 단어는 1,082개, 부정 단어는 1216개, 중립의 성향을 가지고 있는 단어는 8,421개로 구성되었다.

한편 중국어 감성사전은 HowNet을 사용하였다. HowNet의 감성단어에는 감성점수가 따로 부여되지 않고 긍정단어, 부정단어로만 분류되어 있기 때문에 긍정단어의 감성점수는 1, 부정단어

의 감성점수는 -1로 정하였고, 나머지 중성단어들의 감성점수는 0으로 판단하였다.

이모티콘 감성사전은 Emoji Sentiment Ranking 1.0을 사용하였다. 이 이모티콘 감성사전에 따라 긍정 이모티콘은 1, 부정 이모티콘은 -1로 정하고 나머지 중성값을 가진 이모티콘과 여러 가지 감성이 복합되어 구체적인 감성판단이 어려운 이모티콘은 감성값을 0으로 정했다. 예를 들면 Emoji Sentiment Ranking 1.0 에서 "📷"는 부정점수 0.375, 중립점수 0.375, 긍정점수 0.25로 감성값이 -0.125로 제시하고 있다. 다음으로 최종 감성값을 표현하는 sentiment bar에서는 긍정, 부정, 중립 모두 표현되었는데, 특히 부정과 중립은 같은 면적으로 표시되어 있기 때문에 이와 같은 이모티콘은 그 자체만으로 구체적인 감성판단이 어려워 0값으로 산정했다.

3.3 다중언어 감성분석 프로세스

다중언어 감성분석을 위한 기법은 크게 3단계로 구성된다. 우선 감성분석을 하고자 하는 문장을 입력받으면 그 문장을 단어 단위로 쪼갠 후 형태소 분석기를 통해 각 단어의 형태소를 판단하고, 형태소 중 명사, 동사, 형용사, 부사의 개수를 센다. 특히 한글의 경우 단어구성이 중국어와 다르기 때문에 한글은 명사, 동사, 형용사, 부사 외에 어근도 추가하여 센다. 또한 이모티콘의 감성값이 긍정 혹은 부정이면 이모티콘의 개수를 센다. 이와 같이 개수의 합을 한국어는 $\sum W_k$, 중국어는 $\sum W_c$, 이모티콘은 $\sum E_c$라고 표현한다.

두 번째 단계에서는 감성사전을 통한 감성분석을 진행한다. 만약 분리 된 단어가 감성사전에 속해 있는 단어이면, 감성사전 중의 감성점수로 그 단어의 감성값을 반영한다. 이를 수식(1)과 수식(2)로 표현할 수 있다. S_j는 특정 단어의 감성점수, W_j는 다중 언어로 구성된 구절을 Segmentation과정을 거쳐서 얻은 특정단어이며 SDW는 감성사전에 포함되어 있는 감성단어이다. 특히 수식(2)는 특정단어가 감성사전 속 단어에 포함되어 있지 않는 조건을 만족하는 경우로 이럴 때에는 특정단어의 감성점수 S_j값은 0이 된다.

$$S_j = S(W_j)\ (W_j \in \sum SDW) \quad (1)$$

$$S_j = S(W_j)\ (W_j \notin \sum SDW) \quad (2)$$

마지막으로 언어 별로 얻은 총 감성점수를 명사, 동사, 형용사, 부사, 어근(한국어의 경우에만 해당) 및 긍·부정 감성값이 존재하는 이모티콘 개수(이모티콘의 경우에만 해당)의 합으로 나눈 후 나온 결과와 각 언어에서 할당된 가중치를 반영하여 최종 얻고자 하는 다중언어 감성점수를 구하게 된다. 이것을 수식(3)을 통해 계산할 수 있다.

W_c는 중국어가 전체 다중언어에 차지하는 가중치, W_k는 한국어가 차지하는 가중치이며 W_e는 이모티콘이 차지하는 가중치이다. S_c는 감성분석을 하고자 하는 모든 중국어의 감성점수 S_j의 합이고, S_k는 중국어와 마찬가지로 모든 한국어 감성점수 S_j의 합이며, S_e는 모든 이모티콘의 감성점수 S_j의 합이다. S는 마지막 우리가 얻고자 하는 다중언어의 감성점수이다.

$$S = W_c * \frac{S_c}{\sum W_c} + W_k * \frac{S_k}{\sum W_k} + W_k * \frac{S_e}{\sum W_e} \quad (3)$$

4. 실험

4.1 실험절차

본 연구에서 제안한 다중언어 감성분석 알고리즘을 검증하기 위하여 다중언어로 구성된 댓글이나 리뷰를 통해 감성분석을 아래와 같은 방법으로 진행하고자 한다. 비교 대상은 다음과 같다.

방법 1: 다중언어 문장을 한국어로 번역한 후 감성 분석
방법 2: 다중언어 문장을 중국어로 번역한 후 감성 분석
방법 3: 본 연구에서 제안한 방법으로 다중언어 감성분석 기법으로 다중언어를 감성분석

측정치는 정확도(overall accuracy)로 하였다. 테스트용 데이터 셋은 실제 사람들이 작성한 글로 하며, 먼저 각 언어별로 세 사람의 encoder에 의해 내용분석을 실시하고 극성치를 결정하게 한다. 그런 후에 encoder 들의 평가한 결과의 평균치로 극성치를 결정하게 된다. 이렇게 하여 만들어진 극성치는 검증을 위한 데이터에 class로서 추가된다. 그런 후에 위의 세 방법 각각을 통해 검증용 데이터 셋에 적용하여 얻어진 극성값 추정치와 사람들에 의하여 작성된 극성치와의 편차에 대해 RMSE를 구한다.

4.2 데이터셋

본 연구에서는 동영상 공유 사이트인 유튜브에서 한국어와 중국어로 구성된 댓글을 수집하였다. 동영상 콘텐츠 도메인은 요즘 중국에서 인기가 많은 한류스타, 엔터테인먼트 등에 관한 동영상을 위주로, 그 동영상에 달린 댓글 중 중국어와 한국어로 구성된 댓글을 수집하였다. 댓글은 2016년 4월 24일부터 29일까지 수집하였고 수집된 댓글은 총 180개이다.

실험을 시작하기 전, 우선 한국어 댓글 90개, 중국어 댓글 90개를 데이터셋으로 만들고 설문을 통해 encoder들이 그 댓글의 극성치를 구하는 작업을 진행하였다. 한국어 댓글은 한국사람 대상으로 설문을 받아 그 극성치를 구하고, 중국어 댓글은 중국사람 대상으로 극성치를 구했다. 이 데이터셋의 댓글은 KPop, Kbeauty, 한중 예능프로그램에 관련된 댓글이고, 한국어 댓글은 사전실험을 하기 전 비속어를 완화하고 띄어쓰기를 수정하였다. 댓글 번역 시에는 구글 번역기를 사용하였다.

댓글은 구체적으로 아래와 같이 구성되었다. 유튜브에서 KPop(가수, 노래 등)에 관련된 동영상을 선정한 후 그 동영상에 작성된 중국어 댓글, 한국어 댓글 각 10개씩 수집하였다. 이런 기준으로 KPop 동영상 3개 즉 3세트의 댓글을 수집하였다. Kbeauty도 마찬가지로 한개 동영상에 달린 댓글 중 중국어 댓글 10개, 한국어 댓글 10개 씩 총 3개 동영상의 댓글을 수집하였으며, 한중 예능프로그램 관련 동영상도 위 방법과 같이 중국어 댓글 10개, 한국어 댓글 10개씩 3세트를 수집하였다.

4.3 설문조사 방법

댓글 극성치를 구하는 작업은 수집된 댓글을 언어별 각 3명의 encoder에게 설문조사를 하는 방식으로 진행하였다. 설문조사 대상은 20~30대로 유튜브 등 사이트를 많이 이용하며, 중국과 한국에서 유행하는 엔터테인먼트 콘텐츠에 대해

잘 알고 있는 encoder을 대상으로 선정하였다. 설문조사에서는 댓글의 극성치를 판단하기 위하여 7점 척도를 이용하였다. 1점～3점은 부정적인 측면, 4점은 중립, 5점～7점은 긍정적인 측면으로 판단할 수 있게 설문지를 설계하였다. 또한 댓글 설문사항 외에 피설문자의 기본 인적사항(성별, 연령, 학력 등)을 묻는 문항을 추가하였다. 이렇게 구성된 설문지를 통해 얻을 수 있는 댓글의 감성값은 1점～7점 사이의 감성점수이다. 응답한 설문지를 회수한 후 감성점수 비교의 편의성을 위해 응답한 각 댓글의 감성치를 본 연구에서 개발한 다중언어 감성분석 방법론의 결과값에 맞추어 -1점부터 1점 사이로 조정을 진행하였다. 따라서 -1점부터 0점 사이의 극성치가 나오면 부정적 성향을 띤 댓글, 0점은 중립적인 댓글, 0점부터 1점 사이의 극성치는 긍정적 성향을 가진 댓글로 판단할 수 있다.

다음은 3명의 encoder를 통해 각 댓글 극성치를 판단하는 과정을 소개한다. 만약 3명의 encoder가 만장일치로 부정이거나 긍정 혹은 중립의 점수를 선택하면, 그 점수를 합산하여 평균값을 구하는 방식으로 댓글의 극성치를 정했다. 그러나 3명 encoder의 의견이 만장일치 하지 않은 경우, 한명은 긍정 댓글이라고 생각하고 긍정적인 점수를 줬지만 나머지 2명 encoder은 부정 댓글이라고 여서 부정점수를 선택했을 시, 3명의 encoder를 모여 댓글 긍부정성을 다시 결정하는 토론 과정을 거쳤다.

이런 과정을 거쳐 그 중의 encoder이 1차 댓글 긍부정성에 대한 판단 오류를 인정한고, 3명의 결과가 만장일치가 되면 다시 원 방법을 통해 점수를 합산하여 평균치를 구하는 방식으로 해당 댓글의 감성점수를 정했다. 만약 댓글 긍부정성 판단 토의 과정을 거쳤지만 여전히 3명 encoder의 의견이 다른 경우에는 이런 댓글을 따로 분리한 후, encoder 설문 결과 점수의 평균을 구해 극성치를 정했다. 결론적으로는 모두 합산하여 평균치를 구하는 방식이지만, 절차 중 긍부정성 판단 토의 과정을 거침으로 인해 댓글 극성치 판단의 정확성을 높일 수 있었다.

4.4 실험결과

실험결과의 정확성을 평가하기 위하여 원본 댓글을 한국어로 번역하여 얻은 감성값, 중국어로 번역하여 얻은 감성값, 원본 그대로 감성분석을 진행하여 얻은 감성값, 이 세 가지 경우의 값과 encoder를 통해 얻은 극성값과의 편차에 대해 RMSE를 구했다. <Table 3>, <Table 4>, <Table 5>는 각각 KPop, Kbeauty, K엔터테인먼트에 해당되며, 본 연구에서 제안하는 언어별 감성분석을 수행한 결과값과 댓글을 한 가지 언어로 통일하여 번역작업을 거친 후, 얻은 감성분석의 결과값이다. 이 중 어떤 값이 encoder에 의해 판단된 감성점수와 유사한지 RMSE를 통해 검증하였다.

<Table 3>은 KPop 관련 콘텐츠 3개 중 1개의 결과를 표시한 것이다. 댓글을 보면 “裡面的人好漂亮”으로 시작해서 앞 10개는 중국어 댓글, “진짜 예쁘다” 부터 뒤 10개는 한국어 댓글로 구성되었다. 각 댓글의 감성분석 결과를 살펴보면, “裡面的人好漂亮” 원본 댓글을 그대로 본 연구에서 제안하는 감성분석 방법으로 감성분석을 진행했을 때 감성값은 0.5, 댓글을 한국어로 번역했을 때 감성값은 0.25, 3명의 encoder을 통해 얻은 실제 극성값은 0.556이다. 또한 “진짜 예쁘다”라는 댓글의 결과를 보면, 원본의 감성값은 0.5, 번역의 값성값은 0, 실제 극성값은 0.556이다. 댓글의 감성값을 보면, 원본을 대상으로 한

〈Table 3〉 Results of sentiment analysis for KPop-related content #1

Comments/Replies	Original polarity	Translated polarity	Polarity by the encoders
裡面的人好漂亮	0.5	0.25	0.556
少女時代人美歌又好聽	0.167	-0.333	0.778
允兒好漂亮喔 □	0.733	0.667	0.889
非常好	0.5	1	0.667
少女時代好美喔□	0.6	0.5	0.778
我永遠愛少女時代	0.333	0.25	0.889
姐姐們都好美□我開始想夏天了□	0.524	0.375	0.667
少女時代好漂亮 很迷人	0.5	0.333	0.778
真好听，喜欢少女时代	0.75	0.25	0.667
好聽！！！	0.5	0	0.556
진짜 예쁘다	0.5	0	0.667
소녀시대 뮤비가 제일 세련됐다~~	0.333	0.429	0.556
우왕 !!(^o^) 티파니 예쁘다	0.75	0.8	0.889
수영이 키 큰건 알고있었는데 윤아까지 키 커보이게 잘 찍었네요ㅎㅎ	0.167	0.115	0.444
그래도 태연이 제일 예쁘다 ~~!	0.333	0.2	0.667
노래가 너무 상쾌하고 좋다 ㅠㅠ □	0.6	0.583	1
소녀시대 엄청 예뻐요 ♥♥♥	0.667	0.0476	1
아 진짜 예쁘다	0.5	0	0.667
너무 좋다 ♡	0.667	0.667	0.778
배경이 약간 에이핑크 리멤버랑 비슷하다는 생각 든사람?? 배경은 비슷해도 각각의 특성을 잘 나타낸것 같네요^^	0.364	0.2	0.333
Method 1: RMSE = 0.408			
Method 2: RMSE = 0.402			
Method 3 (proposed): RMSE = 0.271			

감성값이 번역한 값보다 실제 극성값에 더 접근해 있는 것을 알 수 있다. 따라서 KPop 관련 콘텐츠 1의 RMSE 결과는 한국어 번역본 0.408, 중국어 번역본 0.402, 원본 댓글은 0.271로 원본댓글의 RMSE가 가장 작게 나온 것을 알 수 있다.

<Table 4>는 Kbeauty 관련 콘텐츠 3개 중 1개의 결과를 나타낸다. 댓글을 보면 "第一眼看起来很像林志玲这个妆容"으로 시작으로 10개의 중국어 댓글, "언니 너무 매력있어요" 등 10개는 한국어 댓글로 구성되었다. 각 댓글의 감성분석 결과에 대해 보면, "第一眼看起来很像林志玲这个妆容" 원본 댓글을 그대로 본 연구에서 제안하는 감성분석 방법으로 감성분석을 진행했을 때 감성값은 0, 댓글을 한국어로 번역했을 때 감성값은 0.25로 3명의 encoder을 통해 얻은 실제 극성값은 0.111이다. 또한 "언니 너무 매력있어요"라는 댓글의 결과를 보면, 원본의 감성값은 0.333, 번역의 값성값은 0, 실제 극성값은 0.556이다. 댓글의 감성값을 보면, 원본을 대상으로 한 감성값이 번역한 값보다 실제 극성값에 접근해 있는 것

〈Table 4〉 Results of sentiment analysis for Kbeauty-related content #1

Comments/Replies	Original polarity	Translated polarity	Polarity by the encoders
第一眼看起来很像林志玲这个妆容	0	0.25	0.111
真的非常喜欢这一P的妆容！太美了！	0.222	0.208	0.667
整天在脸上画画有意思吗	0.25	0.4	-0.667
真的非常喜欢这一P的妆容！太美了！	0.222	0.208	0.667
非常美丽	0.5	1	0.778
最后的灯光打的真强大啊！效果好的不要不要的	0.333	0	0.556
素颜超像范范的!!	0	0	0.444
很有实力的化妆师 ， 喜欢！	0.25	0	0.667
皮肤化妆看起来很自然哦， 喜欢(❤ ω ❤)	0.481	0.533	0.556
喜欢她的化妆风格，解释的也很详细，如果把化妆道具名称加进去就更完美了	0.231	0.143	0.556
언니 너무 매력있어요	0.333	0	0.556
보조개 있으시면 이하늬 똑 닮으실것 같다 □ 진짜 여신 □	0.2	0.133	0.778
피부 관리 어떻게 하세요? 피부가 너무너무 좋으세요 ～!!!!!♡♡♡	0.533	0.25	0.889
아이라인 그리는거 예술이네요 ㅠㅠㅠ 전 아무리해도 안되던데	0.2	0.175	0.556
너무 예쁘게 메이크업 하신다 ㅋ	0.333	0	0.556
렌즈 정보 좀 주세요 진짜 너무 예쁘세요..❤□	0.333	0.36	0.667
시간 도둑 포니언니 ! 이번 영상은 붉은 화장도 없는데 장미가 어울리는 여자 같아요 아래 속눈썹 예쁘구 립스틱 볼에 발라서 신기했는데 다 펴발르고 나니 여리여리한 볼이 되네요 예뻐요 ㅎㅎ□	0.199	0.09	1
포니님 진짜 이 메이크업 어마어마하게 예뻐요!!	0.2	-0.25	0.667
□□□□□□□□□ 그저 감탄합니다... 포니갓	0.909	0.897	0.778
원래 메이크업동영상들 볼 때 좀 데일리하고 연한 메이크업은 잘 안보는데... 이번 메이크업 정말 예쁘네요...! 그리고 설명을 차분히 세심하게 얘기해주셔서 좋았습니다 :) 꼼꼼이 얘기해주셔서 또한번 배워가네요ㅎㅎ 요새 일이 많이 바쁘고 힘드시죠ㅠㅠ?! 전보다 낯빛이 어두우신것같아요ㅠㅠ 너무 무리하시지 마시고 쉬엄쉬엄하세요~:) 항상 예쁜 메이크업들 감사합니다♡	0.205	0.192	0.889
Method 1: RMSE = 0.486			
Method 2: RMSE = 0.530			
Method 3 (proposed): RMSE = 0.443			

을 알 수 있다. 따라서 Kbeauty 관련 콘텐츠 1의 한국어 번역본은 0.486, 중국어 번역본은 0.580, 원본 댓글은 0.443으로 원본 댓글의 RMSE가 가장 작게 나온 것을 알 수 있다.

마찬가지로 <Table 5>는 한중 예능프로그램 관련 콘텐츠 3개 중 1개의 결과이다. 댓글을 보면 “这期好点了啊, 如果剪辑更注意节奏就好了。请继续加油!” 등 10개의 중국어 댓글, “장위안 중국에도 출연했군요. 멋져요.” 등 10개는 한국어 댓글로 구성되었다. 각 댓글의 감성분석 결과에 대해 보면, “这期好点了啊, 如果剪辑更注意节奏就好了。请继续加油!” 원본 댓글을 그대로

〈Table 5〉 Results of sentiment analysis for Kentertainment-related content #1

Comments/Replies	Original polarity	Translated polarity	Polarity by the encoders
这个节目好轻松，好好笑！好看！好评！为什么大家说泰国小哥不帅呢，估计是开玩笑的，我觉得他很可爱啊，张玉安可能是之前参加韩国节目时候因为代表中国，所以比较谨慎，感觉有点儿无趣，估计应该也没有那么大男子主义，都让女生在外面可以很强势了，又怎么能说他大男子主义呢，他有点儿无聊是真的，太正经了，希望以后可以放轻松就好了。	0.222	0.0185	0.556
节目非常好，但我想提个小意见。在投完票以后，希望可以列出一个表格。上面要起头，问题是什么《》，有哪些国家的代表同意，哪些国家代表不同意，这用比较清晰明了。	0.333	0	0.444
居然比国内网站还清晰....我觉得第二期比第一期进步很多！希望以后越来越好~	0.308	0.059	0.556
但是节目还是很喜欢滴～ 祝越办越好	0.25	0	0.556
太短了，没看够	0	0.25	0.111
越来越好	0.5	0	0.778
这期好点了啊，如果剪辑更注意节奏就好了。请继续加油！	0.182	0.083	0.556
这一期感觉进步了，但是配音有点奇怪（btw张玉安和韩冰都好可爱）	0.375	0.183333	0.333
很可爱啊大家	0.5	0	0.556
很好！	0.5	1	0.667
중국어 공부하고 있지만 매 번 느끼는건 청취가 참 어렵다는 것이다	-0.333	0.111	-0.111
한국거랑 완전 똑같네 헐	-0.33333	0.4	-0.333
장위안 잘하네	0	0	0.333
짱위안~~! ㅎㅎ	0	0	0.444444
한국의 비정상회담의 제작진이 얼마나 훌륭한지 이제야 알았다... 한국인으로써 한동수씨가 말하는 분량이 적어 아쉽기도 하고...	0.111	0.174	-0.111
장위안 중국에도 출연했군요 .멋져요.	0.5	0	0.333
아 장위안 중국말 잘하네.	0	0	0.333
난 장위안이 좋다. 그러니 자막 좀 만들어줘ㅠㅠ	0.4	0.167	0.333
남의 방송에 감 놔라 배 놔라 할 자격은 없지만 비정상회담 아끼는 시청자로써 몇 마디 하는데 장위안 나온다하여 엄청 기대하고 봤더니 이 프로가 안건에 대해 본인들의 나라는 어떤지 소개하고 토론하는 프로가 맞는지 궁금하다...멤버들 분량도 엉망이고 엠씨들이 말이 더 많음ㅋㅋ 새삼 비정상회담 제작진이 편집과 연출을 참 잘했구나, 엠씨들이 진행 잘하고 재밌고 본인들 말 줄여서 배려 많이 해줬구나 느낌. 이 자리를 빌려서 감사해요	0	0.041	-0.222
아직 2회 밖에 안되었잖아요 처음에 어색하지만 첫회보다 많이 늘었다고같아요. 짱위안 오빠 너무 잘했지만 역시 한국 비정상회담과 더 잘 맞는것같아요	0.077	-0.108	0
Method 1: RMSE = 0.383			
Method 2: RMSE = 0.304			
Mehtod 3: RMSE = 0.239			

본 연구에서 제안하는 감성분석 방법으로 감성분석을 진행했을 때 감성값은 0.182, 댓글을 한국어로 번역했을 때 감성값은 0.083, 3명의 encoder을 통해 얻은 실제 극성값은 0.556이다. 또한 “장위안 중국에도 출연했군요 .멋져요.”라는 댓글의 결과를 보면, 원본의 감성값은 0.5, 번역의 값성값은 0, 실제 극성값은 0.333이다. 댓글의 감성값을 보면, 원본을 대상으로 한 감성값이 번역한 값보다 실제 극성값에 접근해 있는 것을 알 수 있다. 따라서 예능프로그램 관련 콘텐츠 1의 RMSE 결과는 한국어 번역본은 0.383, 중국어 번역본은 0.304, 원본 댓글은 0.239로 원본댓글의 RMSE가 가장 작게 나온 것을 알 수 있다.

모든 데이터셋의 RMSE 분석 결과는 <Table 6>과 같다. <Table 6>은 본 연구의 각 데이터셋 감성분석 결과에 대한 RMSE 결과이다. 첫 번째로 KPop관련 콘텐츠 1의 한국어 번역본 RMSE 값은 0.408, 중국어 번역본 RMSE 값은 0.402, 원본 댓글의 RMSE 값은 0.271이다. KPop 관련 콘텐츠2의 한국어 번역본 RMSE 값은 0.446, 중국어 번역본 RMSE 값은 0.481, 원본 댓글의 RMSE 값은 0.361이다. KPop관련 콘텐츠 3의 한국어 번역본 RMSE 값은 0.366, 중국어 번역본 RMSE 값은 0.286, 원본 댓글의 RMSE 값은 0.236이다. KPop 관련 콘텐츠의 결과는 모두 방법 3인 원본 댓글을 그대로 감성분석을 할 때, 번역에 의한 방법 1과 방법 2보다 실제 감성값과 가장 유사하게 나온다는 것을 알 수 있다.

두 번째로 Kbeauty 관련 콘텐츠 1의 한국어 번역본 RMSE 값은 0.486, 중국어 번역본 RMSE 값은 0.530, 원본 댓글의 RMSE 값은 0.443이다. Kbeauty 관련 콘텐츠 2의 한국어 번역본 RMSE 값은 0.256, 중국어 번역본 RMSE 값은 0.333, 원본 댓글의 RMSE 값은 0.198이다. Kbeauty 관련 콘텐츠 3의 한국어 번역본 RMSE 값은 0.547, 중국어 번역본 RMSE 값은 0.425, 원본 댓글의 RMSE 값은 0.360이다. Kbeauty 관련 콘텐츠의 결과도 모두 방법 3인 원본 댓글을 그대로 본 연구에서 제안한 방법론에 따라 감성분석을 할 때 번역을 통한 방법 1과 방법 2에 비해 실제 감성값과 가장 접근하여 나온다는 것을 알 수 있다.

마지막으로 예능프로그램 관련 콘텐츠 1의 한국어 번역본 RMSE 값은 0.383, 중국어 번역본 RMSE 값은 0.304, 원본 댓글의 RMSE 값은 0.239이다. 예능프로그램 관련 콘텐츠 2의 한국어 번역본 RMSE 값은 0.440, 중국어 번역본

〈Table 6〉 Performance evaluation with 9 data sets (RMSE)

Comments/Replies	Original polarity	Translated polarity	Polarity by the encoders
KPop 관련 콘텐츠1	0.408	0.402	0.271
KPop 관련 콘텐츠2	0.446	0.481	0.362
KPop 관련 콘텐츠3	0.366	0.286	0.236
Kbeauty 관련 콘텐츠1	0.486	0.530	0.443
Kbeauty 관련 콘텐츠2	0.256	0.333	0.198
Kbeauty 관련 콘텐츠3	0.547	0.425	0.360
예능프로그램 관련 콘텐츠1	0.383	0.304	0.239
예능프로그램 관련 콘텐츠2	0.440	0.476	0.374
예능프로그램 관련 콘텐츠3	0.500	0.411	0.315

RMSE 값은 0.476, 원본 댓글의 RMSE 값은 0.374이다. 예능프로그램 관련 콘텐츠 3의 한국어 번역본 RMSE 값은 0.500, 중국어 번역본 RMS E값은 0.411, 원본 댓글의 RMSE 값은 0.315이다. 예능프로그램 관련 콘텐츠의 결과도 모두 방법 3인 원본 댓글을 그대로 본 연구에서 제안한 방법론에 따라 감성분석을 할 때 번역을 통한 방법 1과 방법2 에 비해 실제 감성값과 가장 유사하게 나온다는 것을 알 수 있다.

따라서 번역을 통해 다중언어를 한 가지 언어로 통일시킨 후 진행하는 감성분석보다는 본 연구에서 제안한 방법인 언어 별로 분리하여 감성분석을 한 후 감성값을 얻는 방법이 결과적으로 볼 때 실제 감성값에 더 접근하는 것을 알 수 있었다. 즉, 이는 번역을 통한 다중언어 감성분석보다 더 정확한 결과를 얻을 수 있다는 것을 의미한다.

5. 토의 및 결론

5.1 학술적 의의

본 연구는 다중 언어 감성분석을 연구함에 있어서 다중 언어로 수집된 텍스트를 번역하지 않고, 해당 언어별로 텍스트를 분리한 다음 각각 감성분석을 진행하고, 나중에 각각의 극성치를 종합하는 방법을 제안하였다. 유튜브나 인스타그램과 같이 한 페이지에 여러 언어권의 텍스트가 존재하지만 한 단위 글에는 하나의 언어로만 작성된 경우의 다국어 감성분석을 제안하였고, 많은 언어들 중에서 중국어와 한국어가 결합된 경우의 감성분석에 집중하였다.

본 연구의 학술적 의의를 살펴보면 첫 번째로 다수의 유사한 연구에서 사용했던 번역 알고리즘을 사용하지 않고, 원문 그대로 다중언어 감성분석을 시도한 점이다. 기존 다중언어 감성분석 연구에서는 번역한 다중언어를 한 가지 언어로 통일 한 후, 그 언어를 대상으로 감성분석을 진행하는 방법을 시도하였다. 이는 다중언어 감성분석이 가능할지라도 번역과정에서 오는 뜻의 왜곡, 문법의 변형 및 번역 알고리즘 자체 오류 등을 고려하지 않았기 때문에 원문자체를 감성분석하는 것에 비해 정확도가 낮다.

따라서 본 연구에서는 이런 점을 보안하고자 다중언어를 언어 별로 분리한 다음 감성분석을 진행하고, 나중에 각각의 극성치를 종합하는 방법을 제안하였다. 이 방법은 번역 알고리즘에서 오는 일련의 단점을 극복하고, 원문자체로 감성분석을 진행했기 때문에 번역 후 감성분석을 진행했을 때 보다 훨씬 더 사람이 직접 판단한 감성값에 접근할 수 있었다.

둘째, 다중언어 감성분석을 진행하면서 언어 감성분석 뿐만 아니라 이모티콘의 감성분석도 추가하여 최종 감성분석 결과를 도출한 것이다. 많은 사람들은 댓글의 감성 즉, 긍부정적을 판단할 때 원문내용과 원문에 달린 이모티콘에 영향을 많이 받는다. 따라서 댓글 감성분석을 진행하려면 이모티콘의 감성분석을 추가하는 것이 결과의 정확도를 높이는데 큰 영향을 준다. 이런 점들을 고려하여 본 연구에서는 EmojiSentimentRanking1.0을 이용하여 이모티콘이 대중적으로 표현하는 긍부정성을 다중언어 감성분석에 추가함으로써 기존보다 사람이 판단하는 댓글의 극성값에 더 접근하는 결과값을 얻은 것이다.

셋째, 많은 연구에서 진행하지 않은 아시아권 나라의 두 가지 언어를 대상으로 다중언어 감성

분석을 시도한 것이다. 대부분 감성분석 관련 연구는 주로, 영어권이나 유럽권 언어를 대상으로 진행되었고, 아시아권의 언어는 많이 진행되지 않은 편이다. 더 나아가서 다중언어 즉, 두 가지 이상의 언어를 대상으로 언어가 모두 아시아권 언어 경우의 연구는 극히 드물다. 따라서 감성분석 측면에서 볼 때 많이 진행되지 않은 연구를 시도함으로써 이 후 더 많은 아시아권 언어 감성분석연구에 도움을 줄 수 있다.

5.2 실무적 의의

소비자들은 제품이나 브랜드를 평가하기 위해 소셜미디어를 이용하여 댓글을 작성하는 형식으로 자신의 의견을 표현하고, 기업에서는 이러한 소비자의 댓글을 모아서 분석하는 것은 소셜미디어를 이용한 중요한 마케팅활동으로 되고 있다. 실제로 많은 기업에서는 소비자들의 댓글을 분석하고자 실시간으로 댓글정보를 모아 인력으로 댓글의 긍부정성을 파악하는 작업을 하고 있다. 만약 전 세계적으로 제품을 수출하는 기업일 경우, 그 나라의 소비자뿐만 아니라, 해외시장에서의 소비자 반응도 중요하다. 이러한 시점에서 본 연구는 크게 아래와 같은 두가지 실무적 의의를 가진다.

첫째, 감성분석 연구를 통해 인력으로 댓글의 긍부정성을 판단하는 것을 기계가 하기 때문에 더 적은 인력자원이 투입되어 원하는 결과를 얻을 수 있다. 기존에는 사람이 댓글을 하나하나 읽으면서 댓글의 긍부정성을 판단했다면, 본 연구에서 개발된 알고리즘을 이용한다면 기계를 통해 댓글의 감성정도를 실시간 파악할 수 있다. 또한 사람이 읽는 것보다 기계가 분석하는 것이 더 빠르기 때문에, 인력자원뿐만 아니라 시간도 절약할 수 있다. 인력자원과 시간의 절약을 바로 비용절감의 효과를 가져 올 수 있어 전반 회사의 이익에 직접적으로 영향을 미친다.

둘째, 중국어 및 한국어를 대상으로 감성분석을 진행하기 때문에 두 나라간의 다양한 무역활동에도 도움을 줄 수 있다. 한류 열풍으로 많은 나라에서는 한국이라는 나라에 관심을 보이고 있다. 특히 이웃나라인 중국에서 한국의 KPop, Kbeauty 및 엔터테인먼트 콘텐츠를 따라 배우거나, 협력하여 프로그램을 만들고자 하는 노력을 많이 하고 있다. 이러한 시점에서, 한국에서 본국의 문화 사업을 중국시장을 넓혀 추진하고자 한다면 한국의 장점인 엔터테인먼트 콘텐츠에 대한 한국 대중의 태도 및 중국에서의 반응이 구체적으로 어떤지, 또한 중국 측에서도 이런 문화사업이 과연 한국과 중국 대중의 인기를 얼마나 끌고 있는지 등을 분석해야 한다. 따라서 중국어와 한국어를 동시에 감성분석을 할 수 있는 알고리즘을 개발하는 것은 현 시점에서 양 나라의 무역활동과 문화사업 발전에 큰 의미가 있다고 생각한다.

5.3 한계점 및 추후 연구방향

본 다중언어 감성분석방법론 개발연구는 기존의 번역과정을 통한 감성분석에서 벗어나 다중언어를 원문자체로 감성 분석하는 시도를 했고, 또 현재까지 많은 연구가 진행되지 않은 아시아권 언어를 대상으로 연구를 진행하였다. 따라서 전반적인 연구는 아직 초기단계 수준에 머물러 있기 때문에 아래와 같은 한계점을 가지고 있다.

첫째, 오픈된 체계적인 한국어 감성사전이 없음으로 자체적으로 구축한 한국어 감성사전을 사용한 것이다. 감성분석의 정확도는 감성사전

의 품질에 많이 의존된다고 할 수 있다. 때문에 우수한 감성사전을 가지고 있는 것은 감성분석의 결과 정확도 향상에 중요한 역할을 한다. 그러나 본 연구에서는 자체적으로 한글감성사전을 영문 SentiWordNet 3.0기반으로 구축했기 때문에 많은 한계점들이 존재한다. 따라서 추후 연구에서는 한국의 국문학전문가들이 직접 개발한 한글감성사전으로 감성분석 연구를 진행하는 것이 더 정확한 결과를 얻을 수 있는데 큰 역할을 할 것으로 본다.

둘째, 본 연구의 알고리즘을 검증함에 있어서 데이터셋의 부족함이다. 더 많은 데이터셋을 이용하여 검증을 해야 했지만, 중국어, 한국어가 동시에 존재하는 동영상을 선정해야 하는 조건 때문에 다른 감성분석 연구에 비해 적은 데이터셋을 사용하였다. 따라서 향후 연구에서는 더 많은 데이터셋을 확보하여 다중언어 감성분석 알고리즘 검증을 진행해야 한다.

또한 한국어 댓글 중 비속어 처리 문제도 고려해야 한다. 한국어 댓글은 다양한 비속어와 신조어의 사용 그리고 철자 오타가 발생하는 경우가 많다. 또한 많은 네티즌들은 표준어 사용보다 인터넷에서 통용되는 언어를 본능적으로 사용하는 현상이 비일비재하다. 따라서 이런 경우의 감성분석은 어떻게 진행할 지에 대해 추후 연구로 진행해야 한다.

5.4 결론

본 연구에서는 번역 알고리즘을 사용하지 않고 다중언어 감성분석을 진행하는 연구를 제안하였다. 즉 언어별로 텍스트를 분리한 다음 각각 감성분석을 진행하고 나중에 감성값을 통합하는 방법론이며, 많은 언어 중에서 중국어와 한국어를 대상으로 연구를 진행하였다. 본 연구에서 제안하는 방법론을 검증하고자, 유튜브 동영상사이트에서 KPop, Kbeauty, 예능프로그램 등 요즘 열풍 도는 한류 관련 콘텐츠를 대상으로, 9개 동영상을 선정하고, 각 동영상에 달린 중국어, 한국어 댓글을 수집하여 검증 데이터셋을 구성하였다.

본 연구의 실험은 전체 댓글을 한국어로 번역하는 방법, 전체 댓글을 중국어로 번역하는 방법, 본 연구에서 제안하는 방법 이 세 가지를 실제 측정값과 비교하여 RMSE를 구하는 방법으로 정확도를 측정하였다. 실제 측정값은 3명의 encoder를 설문조사를 진행하는 방법으로 댓글에 대한 긍부정성 점수를 얻어 평균을 구했다. 검증 결과, 한국어, 중국어로 번역하는 방법 1, 2에 비해, 본 연구에서 제안하는 각 언어별 분리하여 감성분석을 하고 나중에 감성값을 통합하는 알고리즘의 RMSE값이 9개 데이터셋 모두 작게 나왔다. 이것은 본 연구에서 제안하는 방법론의 결과가 실제 사람이 측정한 긍부정 점수에 가장 근접하게 나타나는 것으로 알 수 있으며, 번역을 통해 나타난 감성분석 방법보다 더 우수한 정확도를 가지고 있다.

본 연구의 정확도를 더 높이기 위해서는 한국 국문학전문가를 통한 한국어 감성사전의 구축이 필요하여 또 한국어 댓글에서 많이 나타나는 비속어, 신조어의 사용, 철자의 오류 등 문제점들을 어떻게 해결할 지에 관해 연구를 더 진행하여야 한다. 또한 추후에 한국어, 중국어가 동시에 출현하는 콘텐츠를 더 많이 확보하고 관련 댓글을 수집한 후 본 연구의 알고리즘에 대한 2차 검증을 진행하여 더 좋은 정확도를 위한 다중언어 감성분석 알고리즘 개선을 진행할 필요성이 있다.

참고문헌(References)

Abdul-Mageed M., M. Diab and S. Kübler, "SAMAR: Subjectivity and sentiment analysis for Arabic social media," *Computer Speech & Language*, Vol.28, No.1(2014), 20~37.

Baccianella S., A. Esuli and F. Sebastiani, "SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining," *LREC*, Vol.10(2010), 2200~2204.

Balahur A. and J. M. Perea-Ortega, "Sentiment analysis system adaptation for multilingual processing: The case of tweets," *Information Processing & Management*, Vol.51, No.4 (2015), 547~556.

Bal, D., M. Bal, A. Van Bunningen, A. Hogenboom, F. Hogenboom and F. Frasincar, "Sentiment analysis with a multilingual pipeline," *In International Conference on Web Information Systems Engineering*, Springer Berlin Heidelberg, (2011), 129~142.

Bautin M., L. Vijayarenu and S. Skiena, "International Sentiment Analysis for News and Blogs," In *ICWSM*, (2008).

Boiy E. and M. F. Moens, "A machine learning approach to sentiment analysis in multilingual Web texts," *Information retrieval*, Vol.12, No.5(2009), 526~558.

Denecke K., "Using sentiwordnet for multilingual sentiment analysis," In *Data Engineering Workshop, 2008. ICDEW 2008. IEEE 24th International Conference on IEEE*, (2008), 507~512.

Dong Zhengdong and Dong Qiang, "Introduction to Hownet," In HowNet, http://www.keenage.com, 2000.

Gamallo P. and M. Garcia, "Citius: A Naive-Bayes Strategy for Sentiment Analysis on English Tweets," *Proceedings of SemEval*, (2014), 171~175.

Ghorbel H. and D. Jacot, "Sentiment analysis of French movie reviews," In Advances in Distributed Agent-Based Retrieval Tools, Springer Berlin Heidelberg, (2011), 97~108.

Go A., R. Bhayani and L. Huang, "Twitter sentiment classification using distant supervision," CS224N Project Report, *Stanford,* (2009).

Hajmohammadi M. S., R. Ibrahim and A. Selamat, "Bi-view semi-supervised active learning for cross-lingual sentiment classification," *Information Processing & Management*, Vol.50, No.5(2014), 718~732.

Hajmohammadi M. S., R. Ibrahim and A. Selamat, "Cross-lingual sentiment classification using multiple source languages in multi-view semi-supervised learning," *Engineering Applications of Artificial Intelligence*, Vol.36(2014), 195~203.

Hajmohammadi M. S., R. Ibrahim, A. Selamat and H. Fujita, "Combination of active learning and self-training for cross-lingual sentiment classification with density analysis of unlabelled samples," *Information Sciences: 317*, (2015), 67~77.

Kim Y., S. R. Jeong and I. Ghani, "Text opinion mining to analyze news for stock market prediction," *Int. J. Advance. Soft Comput. Appl,* Vol.6, No.1(2014).

Lee G. H. and K. J. Lee, "Twitter Sentiment Analysis for the Recent Trend Extracted from the Newspaper Article," *KIPS Transactions on Software and Data Engineering*, Vol.2,

No.10(2013), 731~738.

Martín-Valdivia M. T., E. Martínez-Cámara, J. M. Perea-Ortega and L. A. Ureña-López, "Sentiment polarity detection in Spanish reviews combining supervised and unsupervised approaches," *Expert Systems with Applications*, Vol.40, No.10(2013), 3934~3942.

Novak P. K., J. Smailović, B. Sluban and I. Mozetič, "Sentiment of emojis," *PloS one*, Vol.10, No.12(2015), e0144296.

Pak A. and P. Paroubek, "Twitter as a Corpus for Sentiment Analysis and Opinion Mining," In *LREC*, Vol.10(2010), 1320~1326.

Shi W., H. Wang and S. He, "Sentiment analysis of Chinese microblogging based on sentiment ontology: a case study of '7.23 Wenzhou Train Collision'," *Connection Science*, Vol.25 No.4(2013), 161~178.

Tan S. and J. Zhang, "An empirical study of sentiment analysis for chinese documents," *Expert Systems with Applications*, Vol.34, No.4(2008), 2622~2629.

Toutanova K., D. Klein, C. D. Manning and Y. Singer, "Feature-rich part-of-speech tagging with a cyclic dependency network," In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, Association for Computational Linguistics, (2003), 173~180.

Tromp E. and M. Pechenizkiy, "Senticorr: Multilingual sentiment analysis of personal correspondence," In *2011 IEEE 11th International Conference on Data Mining Workshops(ICDMW),* (2011), 1247~1250.

Van Atteveldt, W., J. Kleinnijenhuis, N. Ruigrok and S. Schlobach, "Good news or bad news? Conducting sentiment analysis on Dutch text to distinguish between positive and negative relations," *Journal of Information Technology & Politics*, Vol.5, No.1(2008), 73~94.

Vural, A. G., B. B. Cambazoglu, P. Senkul and Z. O. Tokgoz, "A framework for sentiment analysis in turkish: Application to polarity detection of movie reviews in turkish," *In Computer and Information Sciences III,* Springer London, (2013), 437~445.

Wan, X., "Using bilingual knowledge and ensemble techniques for unsupervised Chinese sentiment analysis," In *Proceedings of the conference on empirical methods in natural language processing*, Association for Computational Linguistics, (2008), 553~561.

Xianghua F., L. Guo, G. Yanyan and W. Zhiqiang, "Multi-aspect sentiment analysis for Chinese online social reviews based on topic modeling and HowNet lexicon," *Knowledge-Based Systems,* Vol. 37(2013), 186~195.

Yang S. and Y. Ko, "Classifying Korean comparative sentences for comparison analysis," *Natural Language Engineering*, Vol.20, No.4(2014), 557~581.

Zhang C., D. Zeng, J. Li, F. Y. Wang and W. Zuo, "Sentiment analysis of Chinese documents: From sentence to document level," *Journal of the American Society for Information Science and Technology*, Vol.60, No.12(2009), 2474~2487.

Zheng L., H. Wang and S. Gao, "Sentimental feature selection for sentiment analysis of Chinese online reviews," *International Journal of Machine Learning and Cybernetics*, (2015), 1~10.

Abstract

A Method of Analyzing Sentiment Polarity of Multilingual Social Media: A Case of Korean-Chinese Languages

Meina Cui*·Yoonsun Jin*·Ohbyung Kwon**

It is crucial for the social media based marketing practices to perform sentiment analyze the unstructured data written by the potential consumers of their products and services. In particular, when it comes to the companies which are interested in global business, the companies must collect and analyze the data from the social media of multinational settings (e.g. Youtube, Instagram, etc.). In this case, since the texts are multilingual, they usually translate the sentences into a certain target language before conducting sentiment analysis. However, due to the lack of cultural differences and highly qualified data dictionary, translated sentences suffer from misunderstanding the true meaning. These result in decreasing the quality of sentiment analysis. Hence, this study aims to propose a method to perform a multilingual sentiment analysis, focusing on Korean-Chinese cases, while avoiding language translations. To show the feasibility of the idea proposed in this paper, we compare the performance of the proposed method with those of the legacy methods which adopt language translators. The results suggest that our method outperforms in terms of RMSE, and can be applied by the global business institutions.

Key Words : Sentiment Analysis, Multilingual Data Analysis, Social Media Marketing; Text Mining, SentiWordNet

Received : June 24, 2016 Revised : August 10, 2016 Accepted : August 11, 2016
Publication Type : Regular Paper Corresponding Author : Ohbyung Kwon

* School of Management, Kyung Hee University
** Corresponding author: Ohbyung Kwon
School of Management, Kyung Hee University
26 Kyungheedae-ro, Dongdaemun-gu, Seoul 130-701, Korea
Tel: +82-2-961-2148, Fax: +82-2-961-0515, E-mail: obkwon@khu.ac.kr

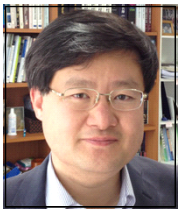
저 자 소 개

최 미 나

중국 연변과학기술대학교에서 경영학과 학사학위를 취득하였으며, 경희대학교에서 일반대학원 경영학과 빅데이터경영전공 석사학위를 취득하였다. 인간 친화적 로봇 서비스 환경에서 판단 적합성 90%이상인 복합지식 기반 판단 및 의미기반 로봇 표현 기술 개발 프로젝트에 참여하였으며, 연구 관심분야는 텍스트 마이닝, 감성분석, 빅데이터 분석 등이다.

진 윤 선

숙명여자대학교에서 e비즈니스학으로 석사학위를 취득하고, 현재 경희대학교에서 빅데이터경영 전공으로 박사과정에 재학 중이다. 2012년~2015년까지 웹발전연구소에서 선임연구원으로 재직한 바 있으며, 현재 개인 행복 증진을 위한 큐레이션 커머스용 연구를 주로 수행하고 있다. 주요 관심분야는 빅데이터, 공공데이터, 데이터마아닝, 텍스트마이닝 등이다.



권 오 병

현재 경희대학교 경영학과 교수로 재직 중이다. 1988년 서울대학교 경영학과(경영학사), 1990년 한국과학기술원 경영과학과(공학석사), 1995년 한국과학기술원 경영과학과 공학박사)를 졸업하였다. 2001년~2002년에는 카네기멜론대학 전산학부에서 방문과학자로 근무한 바 있으며 2009년~2011년에는 샌디에고주립대학 경영정보학과의 겸직교수로 재직한 바 있다. 관심분야는 빅데이터분석, 사물인터넷, 의사결정지원시스템 등이다.