

소셜 데이터의 감성 분석을 위한 신조어 및 이모티콘 감성 사전 구축

양진솔*, 윤경일**, 조영훈***, 정광식****,교신저자
*,**,***,****한국방송통신대학교 정보과학과

e-mail : sanbbang@naver.com, pofour@knou.ac.kr, yabukijoe@knou.ac.kr, kchung0825@knou.ac.kr

Building a Newly-coined Words and Emoticon Emotional Dictionary for Emotional Analysis of Social Data

Jin-Sol Yang*, Kyoung-Il Yoon**, Yeong-Hoon Jo***, Kwang Sik Chung****
Dept of Computer Science, Korea National Open University

요약

SNS 의 발전으로 기업이나 공공단체는 소셜 데이터가 가지고 있는 감성이나 의견, 여론 등을 분석해서 신흥 가치를 창출하려 한다. 소셜 데이터를 기반으로 하는 감성 분석은 사람들의 소비 측면 및 제품 평가 파악은 물론 기업 매출 및 정책 수립 등에서 도움이 된다. 하지만 소셜 데이터는 각종 신조어 및 이모티콘이 다수 포함되어 있어 기존 감성 분석 방법으로는 정확한 분석을 하기 어렵다. 이러한 문제를 해결하기 위해 본 논문에서는 신조어 및 이모티콘 감성 사전을 구축하고, 분석 과정에서 기존 감성 사전과 본 논문에서 구축된 신조어 및 이모티콘 감성 사전을 사용하여 감성 분석 정확도를 비교한다.

1. 서론

소셜 데이터는 소셜 미디어상에서 생산되는 막대한 양의 비정형 데이터로써 그 양은 기하급수적으로 증가하고 있으며 빠르게 확산하고 있다. 이러한 소셜 데이터가 가지고 있는 감성이나 의견, 여론 등을 분석해서 기업의 마케팅 및 전략 수립에 활용할 경우 기업들은 온라인 시장의 신흥 가치를 창출할 수 있다. 일반적으로 언론 매체 등에서 제공하는 인터넷 뉴스는 표준어를 기본으로 사용한다. 하지만 개인이 작성하는 소셜 데이터는 비표준어인 신조어 및 이모티콘 등을 다수 포함한다. 신조어 및 이모티콘은 SNS에서 활용되고 있으며 직접적인 감성적 의미가 있다. 이처럼 신조어 및 이모티콘은 감성 분석의 중요한 부분임에도 불구하고 감성 사전 및 분석에서 제외되어 감성 분석의 정확도를 저하하는 중요 요인으로 작용하고 있다. 신조어 및 이모티콘을 별도의 감성사전으로 구축하여 기존 감성 사전과 함께 사용한다면 감성 분석의 정확도를 높일 수 있다. 이러한 문제를 해결하기 위해 본 논문에서는 소셜 데이터에서 신조어 및 이모티콘 추출하고, 감성 사전 구축을 위해 웹 기반으로 이루어진 별도의 전용 도구를 개발하여 감성사전으로써 활용이 가능한 신조어 및 이모티콘을 분류하였다. 이후 분석 과정에서 기존 감성 사전과 본 논문에서 구축된 신조어 및 이모티콘 감성 사전을 사용하여 감성 분석 정확도를 비교하였다.

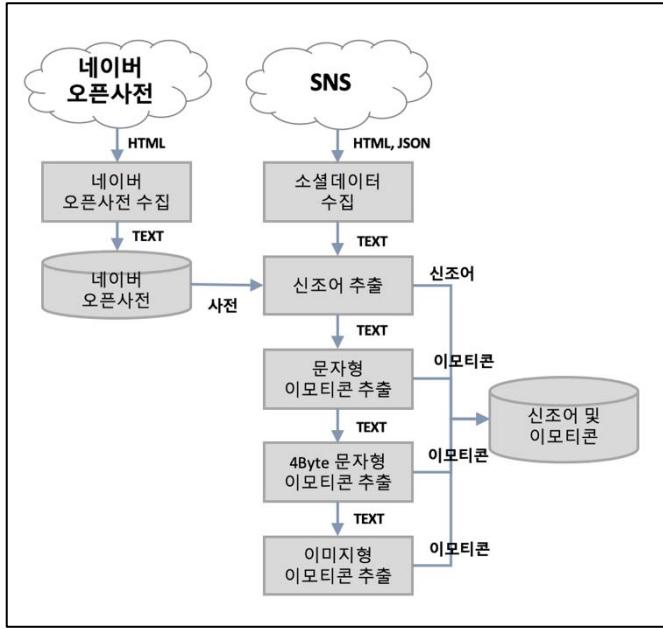
2. 관련연구

송은지(2015)는 온라인상의 문장은 철자와 띄어쓰기 오류가 많고 문장의 길이가 짧아 정확한 의미를 파악할 수 없는 경우가 많으므로 기존의 형태소 분석 기로는 정확한 분석을 할 수가 없다고 하였다. 이러한 문제점을 해결하기 위해 어절 패턴 및 초/중성 사전을 이용해서 보정하고 문장 내 품사의 우선순위를 이용한 의미 선택 방법을 사용하였다. 송은지는 기계 학습을 통한 감성 분석 기법인 SVM 알고리즘을 사용하였으며, 감성 사전을 기반으로 한 본 논문과는 차이가 있다.[1]

한국감정 분석 코퍼스(KOSAC)는 감성 분석 및 의견 분석 연구에 필수적인 한국 감정 말뭉치를 만드는 방법을 제안했다. KOSAC는 감성 분석에 필수적인 한국어 감정 코퍼스 구축을 위해 332 개의 신문기사와 7,744 개의 문장을 주석 대상으로 삼아 총 17,582 개의 감정 표현이 주석되어 있다. 본 논문에서는 KOSAC를 기본 감성사전으로 사용하고 신조어 및 이모티콘 감성 사전을 이용한 감성 분석 방법을 제안한다.[2][3]

3. 신조어 및 이모티콘 추출

본 논문에서 신조어 및 이모티콘 감성 사전 구성 위해 소셜 데이터를 수집한다. 수집된 소셜 데이터에서 신조어, 문자형 이모티콘, 4byte 문자형 이모티콘, 이미지형 이모티콘을 추출한다. (그림 1)은 신조어 및 이모티콘의 추출 프로세스의 전체 과정을 나타낸다.



(그림 1) 신조어 및 이모티콘 추출 프로세스

신조어 및 이모티콘 추출을 위한 소셜 데이터로 트위터와 네이버 블로그를 수집한다. 트위터는 검색 API인 Twitter4j 라이브러리를 이용하여 실시간 트위터의 트윗을 무작위로 수집한다. Twitter4j 라이브러리는 트윗이 작성된 디바이스 운영체제에서 사용하는 언어를 제공한다. 본 논문에서는 트윗이 작성된 디바이스 운영체제에서 사용 중인 언어가 한국어인 트윗을 수집한다. 네이버 블로그의 경우 자체 검색 API를 제공하지 않기 때문에 네이버 블로그를 수집하는 크롤러를 개발하여 사용한다. 본 논문에서 개발한 웹 크롤러는 "엔터테인먼트/예술" 카테고리에 포함된 네이버 블로그의 내용을 주기적으로 수집한다. 이후 무작위로 수집된 트윗에서 URL, 사용자 아이디, 해시태그를 제거한다.[4][5]

소셜 데이터에서 신조어 추출을 위해 네이버 오픈 사전을 기반 데이터로 사용한다. 이용자 참여형 오픈 사전인 네이버 오픈 사전에는 국어, 영어, 중국어, 일본어 등 32 개 언어의 신조어가 웹사이트 (opendict.naver.com)에 등록되어 있다. 본 연구에서는 네이버 오픈 사전을 추출하기 위한 전용 웹 크롤러를 개발하였다. 본 논문에서 개발한 웹 크롤러는 네이버 오픈 사전 웹 사이트에서 검색 조건이 한국어이고, '좋아요'가 10 개 이상 달린 '실시간 단어' 페이지를 수집한다. 이후 신조어 추출 단계에서 소셜 데이터에 오픈 사전의 단어가 포함된 경우 신조어로 판단한다.

소셜 데이터에서 이모티콘 추출은 문자형 이모티콘, 4Byte 문자형 이모티콘, 이미지형 이모티콘의 3 가지 형태로 추출한다. 문자형 이모티콘은 한글의 초성, 중성, 종성으로 이루어져 있으며 단일 요소로는 문자를 표현할 수 없다. 하지만 문자형 이모티콘은 특수문자 및 한글 초성만으로 이루어진 경우가 대다수이다. 이러한 특성을 이용하여 문장의 어절에서 초성으로만 이루어진 단어를 이모티콘으로 판단한다. 예를 들어, "좋다^^" 는 "^"만 남게 되고, "치인건디TT" 는

"TT"만 남게 된다. 4Byte 문자형 이모티콘의 경우 기존 2Byte 유니코드 문자에서 4Byte로 확장되면서 그림 형태의 문자 표현이 가능해졌다. 이로 인해 트위터나 페이스북 등의 SNS 업체들은 4Byte 문자형 이모티콘을 지원하기 시작했다. 4Byte 문자형 이모티콘 추출 과정은 소셜 데이터의 문자를 검사해 4Byte 유니코드로 인코딩된 문자를 추출한다. 이미지형 이모티콘은 일반적으로 블로그나 카페 등 웹사이트 기반의 HTML 형태로 구성된다. 이미지형 이모티콘 추출 과정은 소셜 데이터에서 태그 추출 후, SNS 업체들만의 고유한 패턴을 찾아낸다. 이미지 형태의 이모티콘 태그에서 alt 속성에 "스티커 이미지"라는 값이 있으면 태그를 추출한다. 다만 이미지형 이모티콘의 경우 SNS 업체들만의 고유한 이모티콘이기 때문에 광범위한 감성 사전으로 사용될 수 없다. 그러므로 이미지형 이모티콘은 특정 SNS에서만 한정적으로 사용된다. 예를 들어 네이버 블로그에서 추출된 이미지형 이모티콘 감성사전은 네이버 블로그를 감성 분석할 경우만 사용된다.

4. 신조어 및 이모티콘 감성 사전 구축

신조어 및 이모티콘 감성 사전 구축을 위해 등록자는 추출된 신조어 및 이모티콘에서 유효한 신조어 및 이모티콘을 감성사전으로 등록한다. 유효한 신조어 및 이모티콘은 감성사전으로 사용될 가치가 있는 신조어 및 이모티콘을 의미한다. 본 논문에서는 유효한 신조어 및 이모티콘 분류의 수작업 문제를 해결하고, 분류를 명확하게 하도록 (그림 2)와 같은 전용 도구를 개발하였다.

The screenshot shows a user interface for manual classification. At the top, there's a header with a close button (X) and a title '분류'. Below it is a search bar with placeholder text '소셜건수' and '총 8 / 85,943건'. A button '소셜 데이터 보기' is also present. The main area has sections for '설명' (Description), '감성' (Sentiment), '가중치' (Weight), '사전수정' (Dictionary Modification), '추가 키워드' (Additional Keywords), '제외 키워드' (Excluded Keywords), and '공백주기' (Blank Period). The '설명' section contains the text: '막내온탑' and '막내들이 막내임에도 불구하고 자신보다 나이많은 형동의 따위를 별명으로 슬파시친것.' The '감성' section has a radio button for '긍정' (Positive) which is selected. The '가중치' section has radio buttons for 1, 2, 3, 4, and 5, with 3 selected. The '사전수정' section contains the text '막내온탑'. The '추가 키워드' and '제외 키워드' sections are empty. The '공백주기' section has checkboxes for '앞 공백주기' and '뒤 공백주기', both of which are unchecked. At the bottom right are buttons for '저장' (Save) and '취소' (Cancel).

(그림 2) 전용 도구를 이용한 신조어 및 이모티콘 감성 사전 등록

전용 도구는 신조어 및 이모티콘의 상세 검색이 가능하며, 감성 사전의 기본 요소인 극성과 가중치를 지정할 수 있다. 상세 검색은 신조어, 문자형 이모티콘, 4Byte 문자형 이모티콘, 이미지형 이모티콘의 선택 및 텍스트 검색이 가능하다. 극성의 값은 긍정, 부정, 혼합으로 이루어진다. '인싸','^^'와 같은 긍정적인 성향의 어절은 극성이 긍정으로 분류되고, '찐따','여성혐오'

와 같은 부정적인 어절은 극성이 부정으로 분류된다. 눈물을 의미하는 문자형 이모티콘인 'TT'는 기쁨의 눈물을 의미할 수 있으며, 슬픔의 눈물을 의미할 수 있다. 이처럼 긍정적인 의미로도 사용되고, 부정적인 의미로도 사용되는 어절은 극성이 혼합으로 분류된다. 가중치는 어절의 긍정적 성향이나 부정적 성향의 강도에 따라 1~5 가지의 분류가 가능하다. 발음은 같지만 의미가 다른 동음이의어는 감성분석에서 분석 정확도를 떨어뜨린다. 예를 들어 '한남'이란 어절은 한국 남자를 비하하는 신조어를 의미하며, 표준 국어사전에서는 한강(漢江) 남쪽 유역의 땅을 의미한다. 긍정 성향의 트윗인 "오늘따라 한남대교 정체가 없어서 너무 좋다!!!!"는 표준 국어사전의 '한남'을 의미하지만, 감성 분석 시 신조어 '한남'으로 분류되어 잘못된 결과가 나타날 수 있다. 이러한 문제를 해결하기 위해 유효한 신조어 및 이모티콘을 감성사전으로 등록 시 필수 항목인 강도 및 가중치 이외에 선택항목으로 키워드 추가, 키워드 제외, 사전 앞뒤 공백을 설정할 수 있다. <표 1>은 본 논문에서 제시한 신조어 및 이모티콘 감성 사전을 나타낸다. [7][8][9]

<표 1> 신조어 및 이모티콘 감성사전 예시

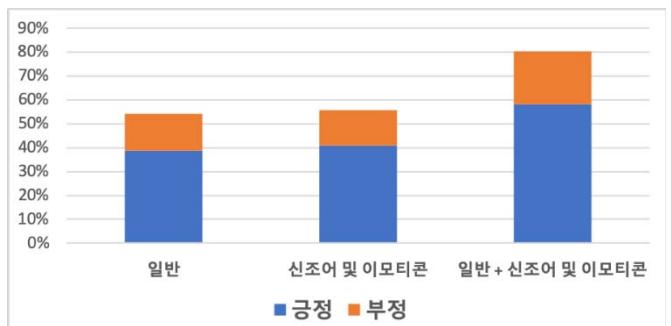
No	사전	극성	가중치	종류
1	ㅋㅋ	positive	3	문자형
2	^^;	positive	2	문자형
3	귀요미	positive	4	문자형
4	😍	positive	5	4Byte형
5	💔	negative	4	4Byte형
6	😡	negative	5	이미지형
7	😊	positive	3	이미지형

5. 실험

본 논문에서 제시한 신조어 및 이모티콘 추출 프로세스를 이용하여 감성 사전을 구축하였다. 신조어 및 이모티콘 추출 프로세스는 리눅스 CentOS 6.5 OS 와 JAVA JDK 1.6 플랫폼 환경에서 구동되며, 소셜 데이터를 저장하기 위한 데이터베이스로 mysql 5.5 가 사용된다. 감성 사전 구축에 사용될 기반 데이터로 트위터의 트윗이 사용되었다. 2018년 3월 7일부터 2018년 5월 7일까지 트윗이 작성된 디바이스 운영체제에서 사용 중인 언어가 한국어인 트윗을 무작위로 수집

하였다. 수집된 총건수는 4,210,744 건이다. 수집된 트윗은 전처리 과정을 통해 URL, 사용자 아이디, 해시태그가 제거되었다. 전처리 과정이 완료된 트윗에서 신조어 추출을 위해 JAVA 언어로 개발된 크롤러를 이용하여 90,373 개의 네이버 오픈 사전을 수집하였다. 수집된 네이버 오픈 사전을 이용하여 트윗에서 55,996 건의 신조어가 추출되었다. 또한 본 논문에서 제시한 이모티콘 추출 알고리즘을 통해 문자형 이모티콘 31,340 건, 4Byte 문자형 이모티콘 1,171 건이 추출되었다. 이미지형 이모티콘의 경우 트위터에서 제공하지 않으므로 제외되었다. 트윗에서 추출된 신조어 및 이모티콘은 감성사전으로 사용하기 위해 별도의 분류 과정이 필요하다. 본 논문에서는 분류 과정의 시간을 최소화하기 위해 감성 분석 시 표본 데이터로 사용될 트윗에 포함된 신조어 및 이모티콘만을 분류하였다. 이처럼 분류된 감성 사전은 신조어 410 건, 이모티콘 201 건이다.

분류된 신조어 및 이모티콘 감성사전으로 감성 분석에 적용할 경우 분석 성능이 얼마나 향상되었는지 실험하였다. 감성 분석 실험에서 사용할 기반 데이터로 트위터에서 2018년 8월 6부터 2018년 9월 7일까지 '아이돌'이란 키워드가 포함된 긍정적 성향의 트윗 926 건, 부정적 성향의 트윗 327 건을 수집하였다. 본 실험의 평가 항목으로 일반 감성 사전을 이용한 감성 분석 정확도, 신조어 및 이모티콘 감성 사전을 이용한 감성 분석 정확도, 일반 감성 사전과 신조어 및 이모티콘 감성 사전을 결합한 감성 분석 정확도를 측정하였다. 일반 감성 사전은 한국어 감정 분석 코퍼스(KOSAC)가 사용된다. 측정 결과 (그림 3)과 같이 일반 감성 사전의 분석 정확도는 54.11%, 신조어 및 이모티콘 감성 사전의 분석 정확도는 55.62%, 일반 감성 사전과 신조어 및 이모티콘 감성 사전을 결합한 분석 정확도는 80.16%로 나타났다. 일반 감성 사전과 신조어 및 이모티콘 감성 사전을 결합하여 분석할 경우 일반 감성사전으로 분석할 때 보다 분석 정확도가 약 26% 향상되었다.



(그림 3) 감성 분석 정확도 결과

6. 결론

모바일 장치가 개발되고 대중화됨에 따라 일상생활에서 스마트폰이 필수품이 되었으며 스마트폰을 사용하여 사회적 문제를 교환하고 개인적인 의견을 교환하는 것이 활성화되었다. 그 결과 SNS에서의 감성

분석의 중요성이 커지고 있다. 특히, 기업과 정부는 소셜 데이터에 대한 감성 분석의 결과에 관심을 보인다. 그러나 신조어 및 이모티콘과 관련한 감성 분석 연구는 제한적이다. 본 논문에서는 신조어 및 이모티콘을 추출하고 감성 사전으로 구축하는 방법을 제안하였다. 그리고 신조어 및 이모티콘 감성 사전의 감성 분석 효과를 평가하기 위해 기존 분석 방법과 신조어 및 이모티콘 감성 사전을 이용한 분석 방법을 비교하였다. 그 결과 신조어 및 이모티콘 감성 사전을 이용한 감성분석방법은 기존 감성분석방법보다 분석 정확도가 향상되었다.

참고문헌

- [1] 송은지, 소셜 미디어 상 고객피드백을 위한 감성 분석, 한국정보통신학회논문지, 2015
- [2] 김문형, 장하연, 조유미, 신효필, KOSAC (Korean Sentiment Analysis Corpus): 한국어 감정 및 의견 분석 코퍼스, 한국정보과학회 학술발표논문집, 2013
- [3] 신효필, 김문형, 박수지. (2016). 한국어 감정분석 코퍼스를 활용한 양상정보 기반의 감정분석 연구. 언어학, (74), 93-114.
- [4] 안정국, 김희웅, 집단지성을 이용한 한글 감성어 사전 구축, 지능정보연구, 2015
- [5] 윤한중, 한국어 트위터 데이터의 감성 분석 알고리즘 구현, 서울과학기술대학교, 2015
- [6] 김준교, 싸이월드와 네이버블로그 브랜드 커뮤니케이션 활용에 관한 연구, 한국일러스트레이션학회, 2005
- [7] Bollen, J., Mao, H., & Zeng, X. Twitter mood predicts the stock market. Journal of computational science, 2(1), 2011, pp.1-8.
- [8] Yadollahi, A., Shahraki, A. G., & Zaiane, O. R.. Current State of Text Sentiment Analysis from Opinion to Emotion Mining. ACM Computing Surveys (CSUR), 50(2), 25. 2017
- [9] DiGrazia, J., McKelvey, K., Bollen, J., & Rojas, F. More tweets, more votes: Social media as a quantitative indicator of political behavior. PloS one, 8(11), 2013, e79449.