SNS 텍스트 데이터를 이용한 영화평 분석*

차소윤, 이봉기, 이호, 위석철, 이수원
숭실대학교 컴퓨터학부
e-mail:seamonstersc@gmail.com

Opinion Mining on Movie Reviews using SNS Text Data

Soyun Cha, Bong Gi Lee, Seokcheol Wi, Ho Lee, Soowon Lee
School of Computer Science and Engineering, Soongsil University

요 약

오늘날 스마트폰의 보급으로 SNS는 급속도로 성장하였고, 매일 엄청난 분량의 텍스트 데이터가 생성되고 있다. 본 연구에서는 다른 매체에 비해 개인의 의견이 좀 더 거침없이 올라오는 SNS의 특징에 주목해 SNS의 텍스트 데이터를 대상으로 하는 평판 분석 기법을 제안한다. 제안 방법은 분석하고자 하는 대상에 대한 SNS 데이터를 수집하여 DB에 저장한 다음, 광고 제거 과정과 자동 띄어쓰기 과정 및 형태소 분석을 거친 후 감성 포함 여부 확인 과정과 극성 분류 과정으로 구성된다. 평판 분석을 위해 본 연구에서는 감성 단어 사전의 쾌-불쾌 수치와 활성화 수치를 사용한다. 분석 결과 모든 문서에 대한 극성 분류 정확도는 55%였고, 감성 포함 여부 확인 과정이 올바르게 수행된 문서에 대한 극성 분류 정확도는 82%였다.

1. 서론

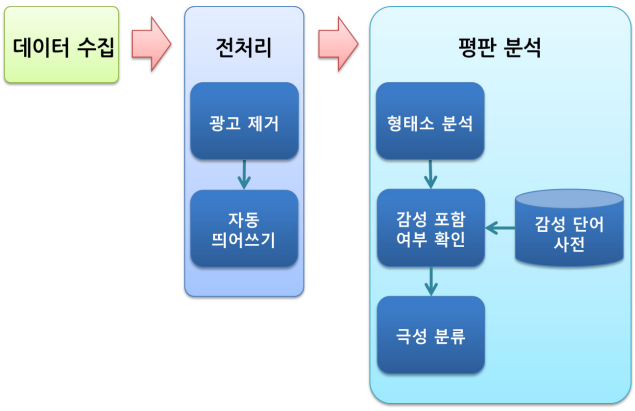
오늘날 SNS(Social Network Service)가 급속도로 성장하게 되었고, 그에 따라 매일 엄청난 분량의 텍스트 데이터가 생성되고 있다. 이렇게 많은 양의 텍스트 데이터 내에는 다양한 내용과 의견들이 포함되어 있다.

SNS에 올라오는 문서는 블로그 등에 올라오는 문서에 비해 좀 더 거침없는 의견이 포함되어 있는 경향이 있다. 본 연구에서는 아직 다른 매체에 비해 지능적인 광고성 문서가 적고 다른 매체에 비해 개인의 의견이 여과없이 표현되는 SNS의 특징에 주목해 SNS의 텍스트 데이터를 대상으로 하는 평판 분석 기법을 제시한다. 기존 연구 [1]에서는 상품평 분석을 위해 반자동으로 구축되는 감성 단어 사전을 사용하였으며, 기존 연구 [2]에서는 영어권 감성 단어 사전인 SentiWordNet을 번역하여 감성 단어 사전을 구축하여 평판 분석에 사용하였다. 기존 연구 [3]에서는 [4]에 첨부된 감성 단어 사전 중 쾌-불쾌 수치를 기반으로 감성 분석을 하였다. 본 연구에서는 [4]에 첨부된 감성 단어 사전의 쾌-불쾌 수치와 활성화 수치를 함께 사용하고 SNS에서 자주 사용되지만 감성 단어 사전에 없는 단어들에 대하여 매핑(mapping) 테이블을 이용하여 SNS 문서를 분석하는 방법을 제시한다. 본 연구에서는 SNS 텍스트 중 영화에 대한 선호도만을 평가하는 것으로 목표를 한정한다.

* 이 논문은 2012년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (No. 2012-047687)

2. 분석 시스템

본 연구에서 제시하는 시스템의 구조는 그림 1과 같다. 데이터 수집 과정에서는 분석하고자 하는 대상에 대한 SNS문서를 검색하여 DB에 저장하고, 전처리 과정에서는 텍스트 데이터를 분석 가능한 형태로 가공하며, 평판 분석 과정에서는 분석하고자 하는 대상에 대한 각 문서의 선호도를 계산한다.



(그림 1) 전체 시스템 구조

전처리 과정은 광고 제거 과정과 자동 띄어쓰기 과정으로 나뉜다. SNS에는 개인적인 문서뿐만 아니라 수많은 광고 문서도 올라오기 때문에 광고 제거 과정이 필요하다. 그리고 일반적으로 사람들은 SNS에 문서를 올릴 때 띄어쓰기를 올바르게 하지 않는 경향이 있으며, 분석 과정에서

정확히 형태소를 분석하기 위해서는 띄어쓰기가 올바르게 되어있어야 하기 때문에 자동 띄어쓰기 과정이 필요하다.

광고 제거 과정에서는 여러 번 올라온 동일한 내용의 문서들을 제거한다. SNS에 여러 번 올라온 동일한 내용의 문서들을 살펴보면 높은 확률로 광고 문서라는 것을 알 수 있다. 이러한 문서들을 제거하기 위해서 새로운 문서가 추가될 때마다 해시(hash) 값이 같은 문서를 찾아 내용이 완전히 같은 경우 광고 문서로 판단해 광고 문서 테이블에 저장하고 입력 데이터에서 삭제한다.

광고 제거 과정이 끝나면 남은 텍스트 데이터를 대상으로 자동 띄어쓰기를 한다. 띄어쓰기를 하기 위해 [5]에서 개발한 형태소 분석기의 자동 띄어쓰기 기능을 이용한다.

분석 과정은 형태소 분석 과정과 감성 포함 여부 확인 과정, 그리고 극성 분류 과정으로 나뉜다. 형태소 분석 과정에서는 해당 문서를 형태소 단위로 구분하고, 감성 포함 여부 확인 과정에서는 각 문서가 감정을 가지고 있는 문서인지 확인하며, 극성 분류 과정에서는 각 문서를 긍정적인 문서와 부정적인 문서로 분류한다.

형태소 분석은 [6]에서 개발한 한국어 형태소 분석/품사 태깅 모듈을 이용한다. 해당 문서가 형태소 분석 과정을 거치게 되면 모든 단어가 형태소 단위로 분리되고 각 형태소의 품사가 태깅된다.

감성 포함 여부 확인 과정에서는 해당 문서에 미리 구축한 감성 단어 사전의 단어가 포함되어 있는지 확인한다. 감성 단어 사전에 단어를 어미가 제거된 형태로 저장해둔 다음, 감성 단어 사전의 모든 단어와 형태소 분석 과정을 거친 문서의 각 형태소의 앞부분이 일치하는지 여부를 확인한다. 예를 들어 문서에 '떨떠름하네염'이라는 단어가 포함되어있는 경우 형태소 분석기가 완벽히 분석하지 못하고 '떨떠름하네염'이 하나의 형태소라고 분석할 수 있는데 이런 경우 사전의 '떨떠름'이란 단어와 앞부분이 일치하므로 '떨떠름하네염'은 '떨떠름'의 의미라고 생각할 수 있다. 만약 감성 단어 사전에 n개의 단어가 있고 문서에 포함된 형태소가 m개면 문서 내의 모든 감성 단어를 찾는 데는 $n \times m$의 시간이 소요된다.

감성 단어 사전 구축에는 [4]의 부록 1에 수록된 단어들의 쾌-불쾌 수치와 활성화 수치를 사용한다. 감성 단어 사전 구축에 사용된 단어의 예는 표 1과 같다. 쾌-불쾌 수치는 단어가 가진 긍정 또는 부정적인 느낌의 정도를 나타내고 활성화 수치는 단어가 가진 느낌이 얼마나 강렬한지를 나타낸다. 쾌-불쾌 수치는 부정적인 느낌이 강한 단어일수록 0에 가깝고 긍정적인 느낌이 강한 단어일수록 7에 가깝다. 본 연구에서는 부정적인 느낌은 음수로 표현하고 긍정적인 느낌은 양수로 표현하기 위해, 쾌-불쾌 수치가 3.81보다 낮은 단어는 대부분 부정적인 느낌의 단어들이고 3.81보다 높은 단어는 대부분 긍정적인 느낌의 단어들이라고 판단해 쾌-불쾌 수치에서 3.81을 뺀 값을 사용했다. 활성화 수치는 단어가 가진 느낌이 강렬하지 않을수록 0에 가깝고 강렬할수록 7에 가깝다.

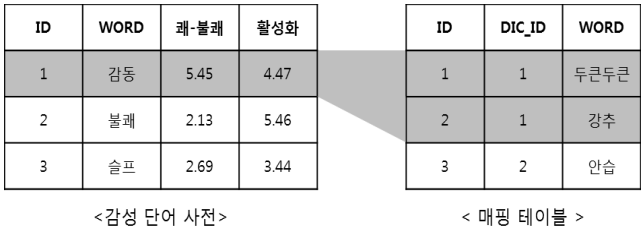
<표 1> 감성 단어 구축에 사용된 단어 예

단어	쾌-불쾌	활성화
노엽	2.48	5.32
매혹	4.84	4.59
뿌듯	5.75	4.63
스산	2.85	3.72
애석	2.67	3.78
우수	4.28	3.38
쾌감	5.79	5.91
처참	1.64	4.5

그리고 SNS에서 사용되는 단어들 중 [4]의 부록 1에 없는 단어들을 조사해 정리한 다음, 설문조사를 통해 수치를 결정해 감성 단어 사전에 추가하여 사용하였다. 설문조사는 [4]의 부록 1의 단어들 중 샘플을 뽑아 추가할 단어들이 샘플로 뽑힌 단어들 중 어느 단어와 비슷한 느낌을 가졌는지 선택하게 하는 방식으로 수행되었다. 추가되는 단어들은 샘플로 뽑힌 단어들에 매핑시켜 해당 수치를 사용하였다. 매핑된 단어의 예는 표 2와 같다. 매핑 테이블과 감성 단어 사전의 매핑 예시는 그림 2와 같다. 매핑 테이블의 WORD 속성은 새로 추가된 감성 단어를 의미하고, DIC_ID 속성은 감성 단어 사전 내의 매핑될 WORD의 ID 값을 가진다.

<표 2> 매핑된 단어의 예

단어	감성 단어 사전의 단어
빡	가엾
안습	가엾
상당	부럽
유쾌	공감
감사	공감
엄청	행복
인상	공감
갠차	부럽



ID	WORD	쾌-불쾌	활성화
1	감동	5.45	4.47
2	불쾌	2.13	5.46
3	슬프	2.69	3.44

<감성 단어 사전>

ID	DIC_ID	WORD
1	1	두큰두큰
2	1	강추
3	2	안습

< 매핑 테이블 >

(그림 2) 감성 단어 사전 및 매핑 테이블

본 연구에서는 극성의 분류와 강도를 구하기 위해 수식 1을 제안한다. 집합 $W(d)$는 문서 d에서 추출된 형태소의 집합 $A(d)$와 감성 단어 집합 E의 교집합이다. 그리고 $word_{d,k}$는 집합 $W(d)$의 원소이며, $|W(d)|$ 는 문서 d에 포함된 감성 단어의 수, $posneg()$는 단어의 쾌-불쾌 수치를 나타내는 함수, $active()$는 단어의 활성화 수치를 나타내는 함수, $pref(d)$는 문서 d의 선호도를 나타내는 함수이다.

$$pref(d) = \sum_{k=1}^{|W(d)|} ((posneg(word_{d,k}) - \alpha) \times active(word_{d,k}))$$

$$word_{d,k} \in W(d) = A(d) \cap E$$

(수식 1) 문서의 극성 분류 및 극성의 강도 계산식 ($\alpha = 3.81$)

3. 실험

3.1 광고 제거 효율

광고 제거 과정의 성능을 평가하기 위해 광고 제거 전과 후의 텍스트 데이터에서 문서를 100개씩 뽑아 광고 문서의 비율을 확인해 보았다. 광고 문서인지 아닌지에 대한 판단 기준은 지능적인 광고 문서인지 아닌지는 알 수 없다는 가정 하에 문서를 직접 읽고 광고를 하기 위한 의도가 분명하면 광고 문서로 판단하였다. 확인 결과 광고 제거 과정 이전 15%였던 광고 문서의 비율이 12%로 감소하였다.

3.2 감성 포함 여부 확인 정확도

감성 포함 여부 확인 과정의 성능을 평가하기 위해 감성 포함 여부 확인 과정을 거친 텍스트 데이터에서 문서를 100개 뽑아 직접 읽고 감정을 가지고 있는지 아닌지를 확인한 다음 그 비율을 평가해 보았다. 평가 결과 시스템이 감정을 포함하고 있다고 분류한 문서 중 67%가 실제로 감정을 포함하고 있었다. 실제 분류 예는 표 3과 같다.

<표 3> 감성 포함 여부 확인 예

문서	시스템 분류	실제 분류
@RunJungjin 피에타 진짜 재미있게봤어요 또 보고싶네용~.~	포함	포함
[영화] 피에타, 사랑. http://t.co/1farNMuJ 피에타를 보는 새로운 관점.	비포함	비포함
늑대아이 흑흑...짱	포함	포함
늑대아이 좋나요?	비포함	비포함

3.3 극성 분류 정확도

극성 분류 과정의 성능을 평가하기 위해 극성 분류 과정을 거친 텍스트 데이터에서 문서를 100개 뽑아 직접 읽고 극성이 올바르게 분류되었는지 확인한 다음 올바르게 분류된 비율을 평가해 보았다. 문서가 긍정적인 의견과 부정적인 의견을 모두 포함하는 경우 일단 올바르게 분류한 것으로 간주하였다. 감성 포함 여부 확인 과정을 거친 데이터에 극성 분류 과정을 거치게 한 결과 55%의 문서가 올바른 극성으로 분류되었다. 감성 포함 여부 확인 과정을 거친 데이터 중 올바르게 분류된 67%의 문서만을 대상으로 극성 분류 정확도를 계산하면 $\frac{55}{67} \times 100 \fallingdotseq 82\%$의 문서가 올바른 극성으로 분류되었다. 실제 분류 예는 표 4와 같다. 표 4에서 각 문서의 감성 단어로 분류된 단어 옆에는 각 단어에서 3.81을 뺀 쾌-불쾌 수치와 활성화 수치를 소수점 아래 세 번째 자리에서 반올림해 표기 하였다.

<표 4> 극성 분류 예

문서	선호도	실제 분류
@RunJungjin 피에타 진짜 재미(쾌-불쾌 수치: 1.91, 활성화 수치: 4.79)있게봤어요 또 보고싶네용~.~	9.1489	긍정
@mamami0921 근데 전 본레거시 재밌(쾌-불쾌 수치: 1.26, 활성화 수치: 3.74)더라구요...ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ첨에도 그랬고 세번째 보고서도 재밌딱ㅋㅋㅋㅋ(쾌-불쾌 수치: 1.26, 활성화 수치: 3.74)하고나와서ㅋㅋㅋㅋㅋㅋ	9.4248	긍정
늑대아이 흑흑...짱(쾌-불쾌 수치: 0.19, 활성화 수치: 4.37)	0.8303	긍정
공모자들 뒤편좀봤는데 아무리 역할이지만 최다니엘 패고싶더라..미친(쾌-불쾌 수치: -0.81, 활성화 수치: 3.33)..납치같은거해서 장기매매하는 인간들은 죽어서 평생(쾌-불쾌 수치: -0.81, 활성화 수치: 3.33) 불구덩이에서 살아야돼.. 자식을 기다리는 부모도있고 아직 창창...	-5.3946	부정

4. 결론

본 연구에서는 SNS 텍스트 데이터를 대상으로 [4]에 첨부된 부록 1의 쾌-불쾌 수치와 활성화 수치를 사용하고, 사전에 없는 단어는 매핑 테이블을 이용하도록 하여 평판 분석을 하는 방법을 제안하였다. 구축된 시스템의 평판 분석 정확도는 55%였고, 감성 포함 여부 확인 정확도가 100%라 가정하였을 때의 정확도는 82%였다.

감성 포함 여부 확인 과정이 정확도에 큰 영향을 미치는 것으로 판단되므로 좀 더 정확한 분석을 위해서는 감성 단어 사전을 확장하거나 다른 감성 포함 여부 확인 방법을 도입하는 것이 필요하다. 감성 단어 사전의 경우 영화 평에 자주 등장하는 단어 위주로 구축되었는데, 단어가 일반적으로 쓰일 때와 영화 평에서 쓰일 때의 느낌이 다르므로 각 단어의 쾌-불쾌 수치를 조절하고 영화 평에서 자주 사용되는 단어를 좀 더 많이 추가해야 할 필요가 있다. 그리고 문맥에 따라 단어의 의미가 달라지는 경우도 있으므로 문맥을 인지해 단어의 수치를 다르게 사용하는 방법에 대한 연구가 요구된다. 광고 제거 효율도 높일 필요성이 있는데, 문장이 완전히 같지 않지만 비슷한 문구가 반복되는 문서들을 제거할 방법을 찾을 필요성이 있다. 그리고 극성 분류 과정도 단순히 단어의 앞부분만 비교하는 방법 뿐 아니라 단어가 변화하는 여러 가지 형태를 고려하는 방법을 찾을 필요가 있다.

참고문헌

[1] 명재석, 이동주, 이상구, "반자동으로 구축된 의미 사전을 이용한 한국어 상품평 분석 시스템", 2007, 정보과학회논문지, 제35권, 제6호 (pp.392~403)

[2] 김정호, 김명규, 차명훈, 인주호, 채수환, "한국어 특성을 고려한 감성 분류", 2010, 한국감성과학회지, 제13권, 제3호 (pp.449~458)

[3] 김명규, 김정호, 차명훈, 채수환, "텍스트 문서 기반의 감성 인식 시스템", 2009, 한국감성과학회지, 제12권, 제4호 (pp.433~442)
[4] 박인조, 민경환, "한국어 감정단어의 목록 작성과 차원 탐색", 2005, Korean Journal of Social and Personality Psychology, Vol. 19, No. 1 (pp.109~129)
[5] http://nlp.kookmin.ac.kr
[6] http://air.changwon.ac.kr