효과적인 감정 자질을 이용한 한국어 문서 감정 분류 시스템

황재원,고영중°
동아대학교 컴퓨터공학과
sftcap@gmail.com, yjko@dau.ac.kr

A Korean Sentiment Classification System Using Effective Emotion Features

Jaewon Hwang and Youngjoong Ko°
Dept. of computer Engineering, Dong-A University

1. 서 론

텍스트로부터 추출할 수 있는 유용한 정보 중에 하나가 작자가 해당 문서의 주제에 대해 표현한 감정 혹은 의견(sentiment or opinion)이다[1]. 근래에 들어 인터넷을 통해 상품에 대한 평가(review)를 온라인으로 손쉽게 수집할 수 있게 됨에 따라, 텍스트 문서들에서 자동으로 감정과 의견을 추출할 수 있다면, 저비용으로, 그리고 자동으로 의견 조사가 가능할 것이다. 최근 외국에서는 이러한 작자의 의견이 담겨 있는 문서로부터 작자의 감정을 자동으로 판별하는 연구가 활발히 진행되고 있다. 전통적인 문서 분류가 문서의 주제(topic)에 초점을 맞추었다면, 감정 분류(sentiment classification)는 저자의 주제에 대한 긍정 감정과 부정 감정에 초점을 맞춘 연구 분야로서, 고객 평가의 요약, 공공 의견 조사, 고객 성향 분석 등의 응용 영역을 가지고 있다.

본 논문에서는 문서의 감정을 분류하기 위해 효과적으로 자질을 추출하기 위한 방법을 제안하고, 이를 통해 문서를 표현하여 기계 학습 기법 중 하나인 지지 벡터 기계(SVM)를 사용하여 성능을 평가하여 추출된 감정 자질의 유용성을 평가한다.

2. 감정 자질 추출과 문서 감정 분류 실험 및 평가

효과적인 문서 감정 분류를 위해서 가장 중심이 되어야 하는 부분이 자질의 선정 방법이다. 본 논문에서는 문서의 감정 분류를 위해서 사용될 충분한 양의 감정 자질을 추출하기 위해서, 영어 단어 시소러스의 유의어 정보[2]를 이용하여 단어를 확장하고, 이를 한영사전을 통해 번역하여 감정 자질을 추출한다. 생성된 감정 자질을 이용하고 기존의 문서 분류 기법을 적용하여, 문서에 대한 감정을 분류한다.

최종적으로 본 논문에서는 다음의 3가지의 감정 자질을 생성하고 비교, 평가한다.

[표 1] 자질 단어의 구성

자질 구분	내용
내용어(자질1)	형태소 분석으로 추출한 내용어(명사, 동사, 형용사, 부사)
감정 자질 (자질2)	감정 대표 어휘의 유의어 확장 단어 집합 {긍정:861개/부정:1,834개}
균형 감정 자질 (자질3)	부정 자질의 수를 줄여서 균형을 맞춘 단어 집합 {긍정:861개 / 부정:844개}

선정된 감정 자질의 가중치는 TF-IDF 가중치 기법으로 가중치를 책정하고, 문서 분류기는 지지 벡터 기계(Support Vector Machine)를 사용한다.

실험 데이터는 총 2,479개의 문서이며, 3개의 분야(신문 기사, 제품 리뷰, 영화 리뷰)를 나누어 수집하였다. 성능 평가 방법으로는 10-fold cross validation 방법을 사용하고, 정보 검색 분야에서 일반적으로 사용되는 정확률(precision)과 재현율(recall)을 사용한다. 최종 성능은 정확율과 재현율을 하나의 값으로 표현해주기 위해서 F_1-Measure를 사용한다.

최종 실험 결과는 다음 [표 2]와 같다.

[표 2] 최종 실험 결과 (SVM, TF-IDF 기법 사용)

구분	자질	긍정	부정	평균	비고
BaseLine	자질1	77.18	75.75	76.47	-
Proposed Method	자질2	84.15	82.00	83.07	+6.6
Proposed Method	***자질3***	***85.20***	***83.39***	***84.30***	***+7.83***

[표 2]에서 알 수 있듯이, 자질2가 자질1에 비해 6.6%의 성능 향상을 얻었고, 자질3이 자질2에 비해 1.23%의 성능 향상을 얻을 수 있었다. 최종적으로, 일반적인 문서 분류에서 사용하는 내용어 자질(자질 1)에 비해 본 논문에서 추출한 감정 자질(자질 3)을 사용했을 경우에 7.83%의 높은 성능 향상을 얻을 수 있었다.

3. 결 론

본 논문에서는 한국어 감정 분류 시스템을 위한 효과적인 자질 추출 방법을 제안하고, 그 유용성을 평가 하였다. 한국어 문서 감정 분류를 위해서는 일반적인 정보 검색에서 사용하는 형태소 분석을 통해 추출된 내용어 자질보다는 의미에 기반한 새로운 감정 자질의 생성이 매우 중요하다는 것을 알 수 있었다. 본 논문에서 제안한 방법으로 추출된 감정 자질은 한국어 문서 감정 분류에 적용했을 때, 84.3%의 높은 성능을 보였다.

참고 문헌

[1] M. Rimon, "Sentiment Classification: Linguistic and Non-Linguistic Issues," Hebrew University.
[2] http://eedic.naver.com/list_thesaurus.naver 네이버 영어 단어 시소러스.
[3] P.D. Turney and M.L. Littman, "Measuring Praise and Criticism: Inference of Semantic Orientation from Association," In *Proceedings of the ACM Transactions on Information Systems*, pp.315-346, 2003.
[4] A. Esuli and F. Sebastiani, "Determining the Semantic Orientation of Terms through Gloss Classification," In *Proceedings of the CIKM*, pp.617-624, 2005.