http://dx.doi.org/10.5392/JKCA.2012.12.12.471

트윗 감정 분류를 위한 다양한 기계학습 자질에 대한 비교 연구

Comparative Study of Various Machine-learning Features for Tweets Sentiment Classification

홍초희, 김학수
강원대학교 컴퓨터정보통신공학전공

Cho-Hee Hong(nlpchhong@kangwon.ac.kr), Hark-Soo Kim(nlpdrkim@kangwon.ac.kr)

요약

문서를 대상으로 한 다양한 감정 분류 연구가 진행되어 왔으며, 최근에는 트윗 감정 분류에 그대로 적용되고 있다. 그러나 이러한 연구들은 트윗의 구조, 이모티콘, 철자 오류 그리고 신조어와 같은 트윗의 특징을 고려하지 않아 좋은 성능을 보이지 못하고 있다. 본 논문에서는 기계학습을 기반으로 다양한 자질을(이모티콘 극성, 리트윗 극성, 사용자 극성, 대체 어휘)사용하여 실험하여 트윗 감정 분류 성능의 영향을 확인하였다. 기계 학습기 SVM(Support Vector Machine) 기반의 감정 분류 실험으로 이모티콘 극성 자질과 사용자 극성 자질이 트윗 감정 분류 모델의 성능 향상에 기여를 하는 것을 알 수 있었다. 이와 비교하여 리트윗 극성과 대체 어휘 자질은 트윗 감정 분류 모델에 큰 영향이 없는 것을 알 수 있었다.

■ 중심어 : | 감정분류 | 트위터 | 트윗 | 기계학습 자질 |

Abstract

Various studies on sentiment classification of documents have been performed. Recently, they have been applied to twitter sentiment classification. However, they did not show good performances because they did not consider the characteristics of tweets such as tweet structure, emoticons, spelling errors, and newly-coined words. In this paper, we perform experiments on various input features (emoticon polarity, retweet polarity, author polarity, and replacement words) which affect twitter sentiment classification model based on machine-learning techniques. In the experiments with a sentiment classification model based on a support vector machine, we found that the emoticon polarity features and the author polarity features can contribute to improve the performance of a twitter sentiment classification model. Then, we found that the retweet polarity features and the replacement words features do not affect the performance of a twitter sentiment classification model contrary to our expectations.

■ keyword : | Sentiment Classification | Twitter | Tweets | Machine-learning Feature |

I. 서 론

트위터(twitter)는 소셜 네트워크(social network)와 메시지가 결합된 형태의 커뮤니케이션 플랫폼이다. 트위터 사용자들은 트윗(tweet)이라는 짧은 문구를 통하여 자신의 소식을 업데이트한다. 또한 사회 현상 혹은

* 본 연구는 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (NO.2010- 0009875)
접수번호 : #120926-008
접수일자 : 2012년 09월 26일
심사완료일 : 2012년 11월 02일
교신저자 : 김학수, e-mail : nlpdrkim@kangwon.ac.kr

특정 상품에 대한 감정을 표현하여 다른 사용자들과 정보를 공유한다[1]. 최근 스마트폰 사용의 증가로 트위터에 대한 접근이 용이해져 2010년 기준 한글 트위터 사용자 수는 약 90만 명이었으나 2012년 기준 한국 사용자는 600만 명을 넘어 사용자가 폭발적으로 증가하였다[2]. 사용자의 폭발적 증가로 인하여 트위터가 사회적 영향력을 얻게 되었고, 이러한 영향력 때문에 기업 마케팅, 광고, 캠페인 등 다양한 분야에 활용되고 있다[3]. 따라서 이를 분석하기 위한 감정 분류 연구가 활발하다. 그러나 트윗은 길이의 제약 때문에 메시지 길이가 짧고, 맞춤법의 오류와 신조어 및 축약어를 포함하고 있어 감정 분류에 어려움이 많다.

본 논문에서는 트위터 데이터의 다양한 내부자원을 활용하여 감정 분류 시스템 구현하고, 성능 비교 실험으로 트위터 데이터 감정 분류를 위한 적합한 내부 자질 활용을 소개 한다.

II. 관련 연구

1. 감정 분류를 위한 자질 추출 연구

감정 분류를 위한 연구는 외부 자원을 활용한 방법과, 내부 자원을 활용한 방법으로 구분 할 수 있다. 먼저 감정 사전이나 정보검색 스니핏(snippet) 같은 외부 자원을 이용하는 방법이 있다. 외부 자원을 활용하는 방법은 수집된 말뭉치의 신뢰도를 높일 수 있는 장점이 있다. 그러나 외부 자원 활용은 많은 비용을 필요로 한다는 단점이 있다. 또한 외부 자원에 의해 감정 분류 시스템의 성능이 결정될 수 있다[4][5]. 내부 자원을 활용한 방법은 음절 *n*-그램(*n*-gram)이나 슬라이딩 윈도우를 기반으로 하는 것이 있다. 이 방법들은 외부 자원 활용 방법보다 비용이 적게 들지만 말뭉치에 의존적이라는 단점이 있다[6].

외부 자원 및 내부 자원을 이용하는 연구는 대부분 형태소 분석 단계를 거치게 된다. 그러나 트윗은 자의적인 철자 오류와 신조어를 포함하고 있어서 형태소 오분석 확률이 높다. 따라서 형태소 오분석을 보완하는 연구나 형태소 분석기를 사용하지 않고 자질을 추출하여 사용하는 연구가 필요하다. 본 논문에서는 형태소 분석기의 오분석을 보완할 수 있는 자질 적용 방법을 소개한다.

2. 트위터 데이터 기반 연구

최근 트위터에 대한 정의, 트위터 현상 분석에 대한 관련 연구가 많아졌다. 그러나 트위터 데이터를 활용하기 위한 연구는 그리 많지 않으며, 대부분 기존에 일반 문서에 적용하는 방식을 그대로 적용하는 경우가 많다. 그러나 일반 문서와 비교하여 트윗은 추출 할 수 있는 자질이 적고, 노이즈(noise)가 많아서 일반적인 문서를 대상으로 한 시스템보다 낮은 성능을 보이고 있다. 트윗 대상의 감정 분류 연구는 감정 자질 추출이 힘들어 감정 문자 사전을 이용하거나 감정 분류 대상을 한정하여 감정 자질 추출을 하는 연구가 있으며[7][8], 트위터 대상의 차이는 있으나 정밀도 83%의 성능을 보이고 있다[9]. 그러나 다양한 감정 표현 추출의 어려움과 대부분 외국어로 작성된 트위터 데이터 기반 연구이기 때문에 한국어로 적용을 했을 때 좋은 성능을 기대 하기 어렵다. 한국어 트위터를 대상으로 감정 유무를 판별하는 연구는 정밀도 86%, 감정 종류(기쁨, 걱정, 슬픔, 분노)를 판단 한 결과는 정밀도 75%를 보였다[10]. 또한 긍정, 부정, 중립의 감정 분류 성능은 정밀도 69.50% 성능을 보였다[11].

III. 감정 분류 시스템

1. 트윗의 특징 및 데이터 정제

트위터는 사용자가 게시한 트윗에 다른 사용자가 답을 할 수 있으며 이를 응답(reply)이라 한다. 또한 관심있는 다른 사용자의 트윗을 계속적으로 볼 수 있도록 등록하는 것을 팔로우(follow)라하며 팔로우 한 사람들은 팔로워(follower)라고 한다. 트위터는 특정 사용자의 트윗을 자신의 팔로워들이 볼 수 있도록 재전송하는 기능을 갖고 있으며 이를 리트윗(RT; retweet)이라 한다[12]. 트윗은 한글과 영문을 1글자로 동일하게 취급하며, 140 글자로 길이의 제약이 있다. 그리고 대부분의

트윗이 입력 공간이 협소한 스마트 폰으로 작성되기 때문에 문법적 오류가 많으며, 인터넷에서 사용되는 신조어나 이모티콘 등을 포함하는 경우가 많다. 또한 트윗은 사용자 아이디(identifier), 링크(link), 이메일(email) 등의 특수 문자열을 포함하는 경우가 많으며, 특수 문자열들이 다양한 형태로 출현하기 때문에 기계 학습 기반의 감정 분류 시에 형태 불일치에 따른 정보량의 분산을 초래한다. 이러한 정보량 분산 문제를 줄이기 위해서 본 논문에서는 아이디를 “USERID”로, 링크를 “URL”로, 이메일 주소를 “EMAIL”과 같은 형태로 변환하는 정제 과정을 거친다.

2. 자질 추출

본 논문에서는 95% 정도의 형태소 단위 정확률을 보이는 형태소 분석기를 이용하여 추출한 어휘 자질을 기본 자질로 사용한다[13][14]. 기본 어휘 자질은 “명사”, “동사”, “형용사”, “부사”, “보조용언”, “미등록어”, “외국어”로 총 8가지의 품사를 사용하였다. [표 1]은 형태소 분석 자질 추출 예를 나타낸다.

표 1. 형태소 분석 자질 추출 예

원 글	더 이상의 피해자가 없기를 바라며 RT 부탁드려요. 올레 KT의 스마트폰 해외로밍 무제한? 거짓말 하지마
형태소 분석결과	더/MAG 이상/NNG 의/JKG 피해자/NNG 가/JKS 없/VA 기/ETN 를/JKO 바라/VV 며/EC RT/SL 부탁/NNG 드리/VV 어요/EF ./SF 올레/IC KT/SL 의/JKG 스마트/NNG 폰/NNG 해외로밍/NA 와이/NNG 파/NNG 이/JKS 무/XPN 제한/NNG ?/SF 거짓말NNG 하/VV 지/EC 말/VX 아/EC
기본 자질 추출	더/MAG 이상/NNG 피해자/NNG 없/VA 바라/VV RT/SL 부탁/NNG 드리/VV KT/SL 스마트/NNG 폰/NNG 해외로밍/NA 와이/NNG 파/NNG 제한/NNG 거짓말/NNG 하/VV 말/VX

그리고 기본 자질에 트윗 컨텐츠(tweet contents)로부터 추출할 수 있는 4가지 자질을 추가하면서 감정 분류 시스템의 성능을 비교 실험한다.

첫 번째 실험 자질은 이모티콘(emoticon)의 극성(polarity)이다. 트윗 사용자들은 자신의 감정을 축약하여 표현하기 위해서 이모티콘을 자주 사용한다. 그러나 이모티콘은 특별한 형식이 없이 다양한 형태로 나타나기 때문에 감정 분류 시스템의 자질로 사용하는데 어려움이 많다. 이러한 문제를 해결하기 위해서 본 논문에서는 이모티콘에 많이 사용되는 몇 가지의 문자와 영어권에서 많이 사용되는 이모티콘을 대상으로 긍정과 부정으로 나누어 극성을 부여한다[7]. [표 2]는 이모티콘의 감정 극성을 판별하기 위한 긍정 문자와 부정 문자를 나타낸다.

표 2. 이모티콘 감정 문자표

긍정 문자		부정 문자
ㅋ	〈	ㅜ
ㅎ	〉	ㅠ
^	♥	-
*	:)	_
+	=)	;
b	:-)	:(
=D	:D	=(

[표 3]는 실제 이모티콘의 감정 극성의 판별 과정을 보여준다.

표 3. 이모티콘의 감정 극성 판별 과정의 예

입력 문자열	단계1	단계2	단계3	단계4	단계5	단계6
^^	+1^	+1+1	+2	긍정		
ㅜㅜ	-1ㅜ	-1-1	-2	부정		
^^;;;	+1^;;;	1+1;;;	2-1;;;	+1-1;	-1	부정

각 입력 문자열은 감정 극성 값을 ‘0’으로 초기화하고 입력 문자가 긍정인 경우 ‘+1’, 부정인 경우 ‘-1’값을 부여 한다. [표 3]의 첫 번째 입력 문자열 “^^”은 긍정 문자 ‘^’가 2번 입력되어 +2의 값을 가지며 긍정으로 이모티콘 극성이 판별된다. 두 번째 입력 문자열 “ㅜㅜ”는 부정 문자 ‘ㅜ’가 2번 입력되어 -2의 값을 가지며 부정으로 이모티콘 극성이 판별 된다. 마지막 입력 문자열 “^^;;;”은 앞의 “^^”는 +2, 뒤의 “;;;”은 -3의 값을 가지므로 최종적으로 부정 이모티콘으로 판별된다. [표 4]는 실제 트윗 내에서 이모티콘의 극성을 판별하고 자질 형태로 대체한 예를 보여준다.

표 4. 이모티콘 대체 실제 예

입력문장	ㅡㅡ케이티 쓰리지 진짜..병맛 ^^ 와이파이도 ㅠ.ㅠ 안터질때는대박임ㅋㅋ
결과문장	EMNEG 케이티 쓰리지 진짜 병맛 EMPOS 와이파이도 EMNEG 안터질때는대박임 EMPOS

[표 4]에서 보는 것과 같이 입력된 이모티콘에 대하여 감정을 판별한 후에 긍정은 “EMPOS”, 부정은 “EMNEG” 형태로 자질을 표현한다. 한정된 이모티콘 감정 문자로 다양한 형태의 이모티콘을 모두 판별할 수 없기 때문에 만약 이모티콘의 감정 판별이 어려운 특수문자열이 있으면 노이즈로 판단하고 제거하였다.

두 번째 실험 자질은 리트윗 극성이다. 리트윗 극성 자질은 트윗 사용자가 다른 사용자의 글을 리트윗 할 때 리트윗 된 트윗의 감정과 리트윗한 사용자의 감정이 연관성을 가진다는 가정 하에 추가한 자질이다. 우선 사용자에 의하여 처음으로 작성된 트윗을 Seed 트윗으로 정의하고, 기본 어휘 자질을 추출한 Seed 트윗과 감정 태그를 쌍으로 Seed 코퍼스를 작성한다. 이 때, Seed 트윗의 감정 태그는 학습데이터에서 찾으며 없는 경우는 중립을 태그를 부여하였다. [그림 1]은 리트윗을 포함한 트윗에서 리트윗 자질을 추가하는 방법의 예이다. 학습 시 리트윗을 포함한 트윗의 경우 Seed 코퍼스에서 Seed 트윗을 검색하여 완전히 일치하는 경우 감정 태그를 찾아 감정 태그를 추가한다. 태그가 긍정 경우 “RTPOS”를, 부정인 경우 “RTNEG”로 변환하여 추가한다.

세 번째 실험 자질은 사용자의 극성이다. 사용자 극성 자질은 한 사용자가 어떤 이슈에 대해 글을 작성할 때 일관적인 감정을 갖고 있다는 가정 하에 적용한 자질이다. 앞서 이모티콘 극성 판별 때와 마찬가지로 트윗을 작성한 사용자를 기준으로 처음 값을 ‘0’으로 설정한 후, 해당 사용자가 게시한 긍정 트윗의 수와 부정 트윗의 수를 계산하여 최종 극성을 부여한다. 최종 값이 0보다 크면 긍정의 극성을, 0보다 작으면 부정의 극성을, 0이면 중립의 극성을 가진다고 판단한다. 이러한 계산 결과를 바탕으로 사용자의 극성이 긍정이면 “AUPOS”를, 부정이면 “AUNEG”를, 중립이면 “AUNEU”를 자질로 추가한다.

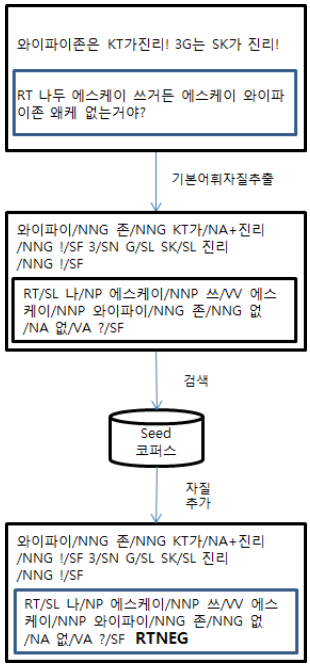

그림 1. 리트윗 자질 적용 예

네 번째 내부 자질은 대체 어휘이다. 트윗의 경우에 형태소 분석기에 등록되지 않은 미등록어가 많이 발생하기 때문에 학습 자질에 없는 어휘가 입력 자질로 추출되는 경우가 빈번하다. 이를 보완하기 위해서 본 논문에서는 입력된 어휘 자질이 학습 자질에 포함되지 않는 경우에 문자간 유사도를 계산하여 가장 큰 학습 자질을 대체 어휘로 사용한다. 어휘 간 문자 유사도는 수식 (1)에서 보는 것과 같이 편집거리(edit distance)와 최장 공통 부분수열(LCS; longest common subsequence)을 사용하여 계산한다[15].

$$RFM(t',t) = \arg\max_{t'} \frac{LCSRatio(t',t)}{editDistance(t',t)} \quad (1)$$

수식 (1)에서 t' 은 학습 어휘 자질이며, t는 입력 어휘 자질이다. $LCSRatio(t',t)$는 t'과 t 사이의 가장 긴 공통 부분문자열의 비율이다. $editDistance(t',t)$는 t'과 t 사이의 편집거리를 의미한다. 입력 어휘 자질과 학습 어휘 자질에 대하여 유사도를 계산하고 가장 큰 값을 기준으로 정규화를 수행한 후, 임계값(실험치 0.4) 이상의

최대값을 갖는 학습 어휘 자질을 대체 어휘로 선택한다. [그림 2]은 대체 어휘 자질의 적용 예를 보여준다.

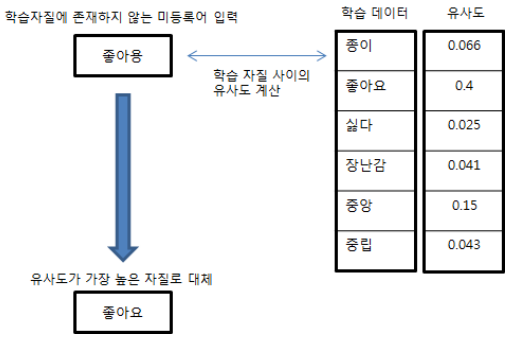

그림 2. 대체 자질 예

[그림 2]의 단계 1과 같이 입력된 자질 “좋아용”이 학습 어휘 자질에 없으면 단계 2와 같이 학습 어휘 자질과 입력 어휘 자질 사이의 유사도를 구한다. 그리고 단계 3에서 유사도가 가장 큰 자질인 “좋아요”를 대체 어휘 자질로 선택하여 “좋아용” 대신에 “좋아요”를 자질로 추출한다.

Ⅳ.실험 데이터 및 실험 결과

1. 실험 데이터 및 실험 환경

본 논문에서 사용된 트윗 데이터는 “KT 무선 네트워크 3G 품질”을 주제로 수집한 것으로 1명의 자연어 처리 전공 석사과정 학생이 긍정과 부정에 해당하는 감정 태그를 부착하였다. 중립에 해당하는 트윗은 실험의 대상으로 삼지 않았다. [표 5]는 각 범주에 대한 트윗 개수와 총 데이터 개수를 나타낸다.

표 5. 범주 별 트윗 개수

범주	개수
긍정	4,977
부정	6,245
전체	11,222

본 논문의 목적은 기계학습 모델의 구현이 아니라 트윗 감정 분류에 영향을 미치는 여러 자질들을 비교 평가하는 것이기 때문에 문서 분류에 좋은 성능을 보이는 SVM(Support Vector Machine)을 기준 모델로 선택하여 트윗 감정분류 시스템을 구축하였다[16]. 실험 결과를 평가가기 위해서는 10배 교차 검증(10-fold cross validation)으로 진행하였으며, 재현률(recall), 정확률(precision), 정밀도(accuracy), F1-척도(F1-measure)를 이용하여 성능을 비교하였다.

2. 실험 결과

본 논문에서는 앞서 설명한 4가지의 자질에 적용에 따라 총 10가지의 실험을 진행 하였다. [표 6]은 실험 데이터 평균과 각 자질의 비율을 나타낸다.

표 6. 실험 데이터 평균 개수와 자질 비율

전체 실험 데이터 평균 개수	1122.2
이모티콘 극성 자질 개수 및 비율	1099.56(97.98%)
리트윗 극성 자질 개수 및 비율	145.86(13%)
사용자 극성 자질 개수 및 비율	593(52.84%)
대체 어휘 자질 개수 및 비율	138.6(12.35%)

[표 6]에서 전체 실험 데이터에서 이모티콘 자질은 약 98%, 리트윗 자질은 약 13%, 사용자 자질은 약 52%의 비율을 갖고 있는 것을 확인 할 수 있다. 그 중 대체 자질은 각 실험 데이터 평균 자질 수는 21625.7이며, 약 0.007%를 차지한다. [표 7]은 자질 적용에 따른 감정 분류 실험 결과이다. [표 7]에서 A는 형태소 분석을 통한 어휘 자질을 사용했을 때의 감정 분류 시스템이며 F1-척도 83.26%의 성능을 보였다. B, C, D ,E는 A를 기준으로 각각 이모티콘 극성 자질, 리트윗 극성 자질, 사용자 극성 자질, 대체 어휘 자질을 적용했을 때의 시스템 성능을 나타낸다. A와 비교하여 이모티콘 극성 자질을 적용한 B는 F1-척도 84.96%로 약 1.7%의 성능 향상을 볼 수 있었다. 리트윗 극성 자질을 사용한 C와, 사용자 극성 자질을 사용한 D는 A와 비교하여 F1-척도 결과가 83.27%, 83.29%로 각 각 0.01%, 0.03%로 큰 성능 향상을 얻을 수 없었다. 대체 어휘를 적용한 E는 A와 비교하여 F1-척도 결과가 83.21%로 약 0.05% 낮은 성능 결과를 보였다. B, C, D, E의 실험으로 이모티콘 극성

표 7. 자질 적용에 따른 감정분류 실험 결과

	내부 자질	재현률	정확률	정밀도	F1-척도
A	기본 어휘 자질	83.27	83.25	83.48	83.26
B	A + 이모티콘 극성	84.97	84.95	85.16	84.96
C	A + 리트윗 극성	83.28	83.26	83.50	83.27
D	A + 사용자 극성	83.30	83.28	83.51	83.29
E	A + 대체 어휘	83.22	83.21	83.44	83.21
F	A + 이모티콘 극성 + 리트윗 극성 + 사용자 극성 + 대체 어휘	84.83	84.83	85.03	84.82
G	F - 이모티콘 극성	81.49	81.68	81.74	81.58
H	F - 리트윗 극성	84.82	84.83	85.02	84.82
I	F - 사용자 극성	83.86	83.97	84.1	83.92
J	F - 대체 어휘	84.76	84.77	84.97	84.77

자질은 사용자의 감정을 담고 있어서 감정 분류 시스템의 성능에 영향을 주는 것을 알 수 있다. 그러나 리트윗 극성 자질, 사용자 극성 자질, 대체 어휘 자질은 감정 분류 시스템에 영향력이 적다는 것을 알 수 있다. 이에 원인으로 리트윗 극성 자질의 경우 리트윗 내용이 매우 짧아서 추출 가능한 자질이 적어지기 때문에 리트윗 대상 트윗과 다른 감정을 표현하는 경우 정확한 감정 판별이 어렵게 되는 것을 뽑을 수 있다. 또한 사용자 극성 자질을 추가한 실험 D의 경우는 분류 대상의 트윗의 사용자가 학습 데이터 내에 존재하지 않는 경우가 발생되기 때문이다. 따라서 사용자 극성 정보가 없는 경우는 자질이 추가 되지 않아 성능 향상을 저해하는 요소가 될 수 있다. [표 8]은 사용자 극성이 학습 데이터에 없을 경우와 있을 경우 각각 정밀도를 나타낸다.

표 8. 사용자 유무에 따른 실험 결과

사용자 극성 유무	정밀도
사용자 극성이 없는 경우	82.89
사용자 극성이 있는 경우	84.20

사용자 극성이 있는 경우는 84.20%로 사용자 극성이 없는 경우와 비교하여 약 1.31% 성능이 높은 것을 볼 수 있으며, 실험 A와 정밀도 결과를 비교하여 0.72% 성능이 높은 것을 볼 수 있다.

실험 F는 기본 자질에 이모티콘 극성 자질, 리트윗 극성 자질, 사용자 극성 자질, 대체 어휘 자질을 적용한 결과이다. 그리고 G, H, I, J는 F를 기준으로 각 각 이모티콘 극성 자질, 리트윗 극성 자질, 사용자 극성 자질, 대체 어휘 자질을 제외한 감정 분류 시스템 결과이다. 실험 H, I, J는 실험 A와 비교하여 F1-척도가 각각 1.56% ,0.66%, 1.51%로 약간의 성능 향상을 보였다. 그러나 실험 F와 비교하여 실험 G는 F1-척도 결과가 81.58%로 약 3.24% 낮아지는 것을 볼 수 있으며, 이를 근거로 이모티콘 극성 자질이 감정 분류 시스템 성능 향상에 가장 큰 영향을 주는 것을 알 수 있었다. 또한 이모티콘 극성 자질의 성능 영향을 검증하기 위하여 실험 A와 B 사이의 T-검정을 사용하였다. 그 결과 신뢰수준 95%에서 실험 A와 실험 B 간에 유의한 차이가 있는 것을 확인하였다(p-value=0.00004).

[표 9]는 이모티콘 자질이 적용되어 긍정적인 효과를 나타낸 경우의 예이다.

표 9. 이모티콘 자질 적용 예

	트윗	감정
A	@olleh 굉장히 빠른 답변 감사합니다. 답변시간은 상관없이 24시간운영하느라 수고많으십니다. 핸드폰 랙걸린건지 3G가 약했던건지 모르겠지만 해결됐습니다. 월급 많이 받으십시요 ^^	긍정
B	@olleh 검색해도 olleh WiFi로밍안나오네요ㅜㅜ	부정
C	우리학교좀 보소 —— 나있을땐 올레와이파이 안달아주더니 나 휴학하니까 전 건물에 다달았네?	부정

[표 9]에서 A, B는 이모티콘 판별로 각 각 긍정과 부정으로 기본 어휘 자질만 사용한 결과와 다른 결과를 얻어 긍정적인 효과를 보였다. 그리고 C와 같은 경우는

글의 내용만으로는 감정을 판별하기 어려운 경우로 볼 수 있으며, 이모티콘 자질을 추가하여 부정으로 바르게 분류 한 경우이다.

V. 결론 및 향후 연구

트윗 데이터의 감정 분류 연구는 높은 필요성에도 불구하고 트윗 데이터의 특징 때문에 높은 성능의 감정 분류 시스템 구현이 어렵다. 본 논문에서는 이와 같은 문제를 위하여 트윗 데이터의 특징을 활용하여 다양한 내부 자질 적용하고 감정 분류 모델 구현하였다. 그리고 각 모델 성능 비교 실험으로 적용된 자질이 감정 분류 모델에 끼치는 영향을 확인하였다. 우선 이모티콘 감정 극성 자질의 적용으로 특수 문자열에 의한 노이즈 발생을 최소화 할 뿐만 아니라 감정 분류 시스템 성능 향상을 볼 수 있었다. 다음으로 리트윗 극성 자질은 사용자와 다른 사용자를 연결하는 감정 네트워크를 형성하기 때문에 감정 분류 시스템에 적합한 자질로 예상되었지만, 리트윗 내용이 매우 짧아서 추출 가능한 자질이 적고, 리트윗 된 트윗과 다른 감정을 표현하는 경우가 많아 큰 성능 향상을 볼 수 없었다. 그러나 리트윗은 사용자 사이의 관계에 대한 정보가 있기 때문에 향후 사용자들 사이에 관계 추출 연구에 중요한 역할을 할 수 있을 것으로 기대된다. 다음 사용자 극성 자질은 특정 현상에 대해서 각 사용자가 작성한 트윗 전체에 대하여 감정을 판별하고 자질로 사용함으로써 감정 분류 시스템 성능 향상이 있었으나, 학습 데이터에 입력된 사용자 극성이 없는 경우는 사용자 극성 자질 적용이 어렵기 때문에 이를 보완하기 위한 연구가 필요하다. 마지막으로 대체 어휘 자질에 대한 감정 분류 실험은 자의적인 문법 오류 때문에 많은 미등록어가 발생되어 추출 할 수 있는 자질이 부족하게 되는 문제를 보완하기 위한 목적이었다. 그러나 문자 사이의 유사도만을 사용하여 선택된 대체 어휘 자질은 감정 분류 성능에 큰 영향이 없었다. 따라서 대체 어휘 자질을 선택하기 위한 다양한 방법을 고안하여 적합한 대체 어휘 추출한다면 감정 분류 시스템 성능 향상을 기대 할 수 있을 것이다.

참 고 문 헌

[1] L. Barbosa and J. Feng, "Robust sentiment detection on Twitter from biased and noisy data," In Proceedings of the 23rd International Confere-nce on Computational Linguistics, pp.36-44, 2010.

[2] http://www.bloter.net/archives/74190

[3] 홍초희, 김학수, "트윗 분류를 위한 효과적인 자질 추출", 한국정보과학회 학술발표논문집, 제38권, 제1호, pp.229-232, 2011.

[4] 신준수, 김학수, "강건한 한국어 상품평의 감정 분류를 위한 패턴 기반 자질 추출 방법", 정보과학회논문지 소프트웨어 및 응용, 제37권, 제12호, pp.946-950, 2010.

[5] 황재원, 고영중, "감정 분류를 위한 한국어 감정 자질 추출 기법과 감정 자질의 유용성 평가", 인지과학, 제19권, 제4호, pp.499-517, 2008.

[6] H. Cui, V. Mittal, and M. Datar, "Comparative Expe- riments on Sentiment Classification for Online Product Reviews," In Proceedings of the 21st National Conference on Artificial Intelligence, Vol.2, pp.1265-1270, 2006.

[7] L. Jiang, M. Yu, M. Zhou, X. Liu, T. Zhao, "Target-dependent Twitter Sentiment Classi-fication," In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, pp.151-160, 2011.

[8] W. Wu, B. Zhang, M. Ostendorf, "Automatic Generation of Personalized Annotation Tags for Twitter Users," In Proceedings of Human Language Technologies 2010, pp.689-692, 2010.

[9] A. Go, R. Bhayani, L. Huang, Twitter Sentim-ent Classification using Distant Supervision, CS224N Project Report Stanford, 2011.

[10] 김동균, 허지용, 조지훈, 박수영, 김용혁, "기계학습 기반의 감정 트위터 봇", 한국정보과학회 학술발표 논문집, 제38(2B)권, pp.379-382, 2011

[11] 홍초희, 김학수, “신뢰도 높은 트윗 감정 분류를 위한 하이브리드 자질 추출 기법”, 강원대학교 정보통신논문지, 제16권, pp.38-41, 2012

[12] 박수영, 하용호, 김용혁, “트윗터 정보 검색 분야의 최근 연구들”, 한국정보과학회 2010 한국컴퓨터종합학술대회 논문집, 제37권, 제2(C)호, pp.25-29, 2010.

[13] 최맹식, 김학수, “기계학습에 기반한 한국어 미등록 형태소 인식 및 품사 태깅”, 정보처리학회논문지, 제18-B권, 제1호, pp.45-50, 2011.

[14] 심광섭, 양재형, “인접 조건 검사에 의한 초고속 한국어 형태소 분석”, 한국정보과학회논문지 소프트웨어 및 응용, 제31권, 제1호, pp.89-99, 2004.

[15] Z. Xue, D. Yin, and B. D. Davison, “Mormalizaing MicroText,” In Proceedings of AAAI-11 workshop on Analyzing Microtext, pp.74-79, 2011.

[16] B. Pang, L. Lee, and S. Vaithyanathan, “Thumbs up? Sentiment Classification Using Machine Learning Techniques,” In Proceedings of the EMNLP, pp.79-86, 2002.

저 자 소 개

홍 초 희(Cho-Hee Hong) 준회원

- 2011년 2월 : 강원대학교 컴퓨터정보통신공학(공학사)
- 2011년 3월 ~ 현재 : 강원대학교 컴퓨터정보통신공학(공학석사)

<관심분야> : 정보검색, 감정분류, 키워드 추출, 문서분류

김 학 수(Hark-Soo Kim) 정회원

- 1996년 2월 : 건국대학교 전자계산학과(공학사)
- 1998년 2월 : 서강대학교 컴퓨터학과(공학석사)
- 2003년 2월 : 서강대학교 컴퓨터학과(공학박사)
- 2004년 ~ 2005년 : CIIR in UMass, Amherst (박사후연구원)
- 2005년 ~ 2006년 : 한국전자통신연구원(선임연구원)
- 2006년 ~ 현재 : 강원대학교 컴퓨터정보통신공학전공 교수

<관심분야> : 자연어처리, 대화시스템. 정보검색, 질의응답시스템