http://dx.doi.org/10.3745/KIPSTB.2011.18B.5.325

형식적 및 비형식적 어휘 정보를 반영한 문장 감정 분류

조 상 현[†] · 강 행 봉[††]

요 약

최근 트위터, 페이스북과 같은 소셜 네트워크 서비스(Social Network Service : SNS)가 활성화됨에 따라 서비스 사용자들에 의해 작성된 막대한 텍스트들로부터 의미 있는 정보를 찾기 위한 연구가 많은 주목을 받고 있다. 특히 문장에 담겨 있는 감정은 활용 범위가 매우 넓은 정보로서 문장에 대한 감정을 분류하는 일은 매우 유용한 일이라고 할 수 있다. 본 논문에서는 문장의 감정을 분류하기 위해 문장에 포함되어 있는 형식적 어휘 정보와 이모티콘이나 인터넷 용어와 같은 온라인상에서 많이 이용되는 다양한 형태의 비형식적 어휘 정보를 이용한 새로운 문장 감정 분류 방법을 제안한다. 기존에는 문장의 감정을 분류하기 위해 사전을 기반으로 한 형식적 어휘 정보를 이용했지만, 최근 인터넷 사용자들은 인터넷 용어나 이모티콘과 같은 비형식적 어휘를 많이 사용해 기존의 형식적 어휘 정보만으로는 정확한 감정 분류가 어렵다. 제안한 방법은 형식적 어휘 정보와 비형식적 어휘 정보를 이용해 다양한 형태의 어휘를 포함하는 인터넷 상의 문장들에 대해 보다 정확한 감정 분류 결과를 보여준다. 또한, 같은 어휘라도 도메인별로 다른 감정을 나타내는 경우가 많으므로 제안한 방법에서는 도메인별로 다른 감정 어휘정보를 이용했다. 각 감정 어휘 정보를 통해 특징벡터로 표현된 문장은 Support Vector Machine(SVM) 분류 방법을 통해 감정을 분류하고 그 성능을 평가했다.

키워드 : 감정 분류, 감정 특징, 오피니언 마이닝, SVM

A Sentence Sentiment Classification reflecting Formal and Informal Vocabulary Information

Sang-Hyun Cho[†] · Hang-Bong Kang[††]

ABSTRACT

Social Network Services(SNS) such as Twitter, Facebook and Myspace have gained popularity worldwide. Especially, sentiment analysis of SNS users' sentence is very important since it is very useful in the opinion mining. In this paper, we propose a new sentiment classification method of sentences which contains formal and informal vocabulary such as emoticons, and newly coined words. Previous methods used only formal vocabulary to classify sentiments of sentences. However, these methods are not quite effective because internet users use sentences that contain informal vocabulary. In addition, we construct suggest to construct domain sentiment vocabulary because the same word may represent different sentiments in different domains. Feature vectors are extracted from the sentiment vocabulary information and classified by Support Vector Machine(SVM). Our proposed method shows good performance in classification accuracy.

Keywords : Sentiment Classification, Sentiment Feature, Opinion Mining, SVM

1. 서 론

최근 스마트폰의 대중적인 보급으로 페이스북과 같은 소셜 네트워크 서비스(Social Network Service, SNS)가 활성화되고 있다. SNS는 기존의 인맥을 강화하고 새로운 인맥을 형성하여 폭넓은 인적 네트워크를 형성할 수 있도록 해주는 서비스로서 많은 사람들은 이러한 서비스를 통해 서로에게 댓글을 달아주는 형태로 막대한 양의 텍스트 정보를 생성하고 있다. 최근에는 바이럴 마케팅(viral marketing), 즉 입소문을 통한 마케팅 방법이 많이 이용되고 있어, 이러한 SNS에서 생성된 텍스트를 이용하여 의미 있는 정보를 추출하기 위한 다양한 분석이 시도되고 있다. 특히 문장에 포함되어 있는 감정은 활용 범위가 매우 넓은 정보로서 문장의 감정을 분류하는 일은 최근 많은 주목을 받고 있다.

※ 본 과제는 2011년도 가톨릭대학교 교비연구비의 지원 및 문화체육관광부 및 한국콘텐츠진흥원의 2011년 문화콘텐츠산업기술지원사업의 지원 및 지경부산하 NIPA에서 추진한 "2010년 대기업 연계 IT/SW 창의연구과정" 사업의 지원으로 수행되었음.
† 준 회 원 : 가톨릭대학교 컴퓨터공학과 박사과정
†† 종신회원 : 가톨릭대학교 디지털미디어학부 교수(교신저자)
논문접수 : 2011년 6월 20일
수 정 일 : 1차 2011년 8월 4일, 2차 2011년 8월 17일
심사완료 : 2011년 9월 9일

문장의 감정은 기본적으로 감정을 포함하고 있는 어휘들에 의해 결정된다. 이러한 감정이 포함된 어휘인 감정 특징은 문장의 감정을 분류하는 데 매우 중요한 역할을 한다. 하지만, 기존의 감정 특징은 사전 정보에 기반을 두어 구성되기 때문에 일반적으로 비표준어 표현을 많이 사용하는 인터넷 댓글이나 SNS에서 사용되는 텍스트의 감정을 분류하는 데는 적합하지 않다. 또한 최근 인터넷 사용자들은 이모티콘을 통해 자신의 감정을 표현하는 경우가 많지만 사용자 취향이나 기타 여러 가지 요인으로 인해 매우 불규칙적으로 표현되는 경우가 많아 이를 이용하는 것은 쉽지 않다.

본 논문에서는 문장의 감정을 분류하기 위해 문장에 포함되어 있는 사전 기반의 형식적 어휘 정보뿐만 아니라 인터넷 용어, 이모티콘 정보와 같은 비형식적 어휘 정보를 이용한 새로운 문장 감정 분류 방법을 제안한다. 문장의 감정을 분류하기 위해 제안한 방법에서는 감정사전을 이용해 문장으로 특징벡터로 표현한다. 감정 사전은 품사별 어휘와 감정 가중치로 구성된다. 이 감정사전에는 이모티콘도 포함된다. 하지만 앞서 언급했듯이, 이모티콘은 불규칙적으로 사용되는 경우가 많아 제안한 방법에서는 베이지안 프레임워크를 이용해 불규칙적으로 사용된 이모티콘을 감정 사전에 포함된 참조 이모티콘으로 변환하여 사용한다. 감정 사전을 이용해 특징 벡터로 표현된 문장은 SVM을 통해 감정이 분류된다.

본 논문의 구성은 다음과 같다. 2장에서는 본 연구의 관련 연구에 대해 살펴보고, 3장에서는 본 논문에서 제안하는 문장 감정 분류 시스템에 대해서 설명한다. 그리고 4장에서는 실험 결과를 기술하고, 마지막으로 5장에서는 결론을 기술한다.

2. 관련 연구

문서의 감정 분류는 문서 분류의 한 분야로서 여러 기계학습 방법들이 문서 감정 분류에 이용되었다. 영화 평론, 상품 평가 등의 특정 영역에서의 문서를 다양한 기계 학습 기법을 이용해 긍정과 부정의 범주로 분류하는 연구가 수행되었다[1-4].

일반적으로 문서 분류에 관한 연구는 크게 문서 특징 추출에 관한 연구와 문서 분류 모델에 관한 연구로 나눌 수 있다.

문서 특징의 추출은 문서 분류를 위한 중요한 문제로서 다양한 연구가 진행되었다. 문서의 특징 추출 방법은 학습 데이터로부터 형태소 분석기를 이용해 해당 내용어(content word)를 추출하고 추출된 단어에 대한 가중치를 계산하는 것이 일반적이다. 따라서 감정 분류시스템에서 감정을 분류하기 위한 감정 특징을 추출하기 위해 영어권 선행 연구에서는 WordNet과 같은 어휘 의미망을 이용한 감정 분류에 적합한 특징을 추출하는 연구가 수행되었으며[5-6] 영어권 어휘 자원을 이용해 감정 특징의 가중치를 정하는 연구도 진행되었다[7]. 이러한 특징의 가중치를 계산하는 방법으로 흔히 Term-frequency(TF)와 TF-IDF(Inverse Document Frequency)가 많이 이용된다. TF가 어휘의 빈도수만을 이용하는데 비해 TF-IDF는 TF와 역문서 빈도수(Inverse Document Frequency)의 곱으로 표현된다. TF-ISF(Inverse Sentence Frequency) 가중치 기법도 많이 이용되는데 이는 문서의 빈도수 대신 문장의 빈도수를 사용하는 것이다[8-10].

또한 최근에는 문서에 포함된 감정을 정확하게 분류하기 위해 기존의 uni-gram 기반의 부정어 처리 방법대신 uni-gram 및 bi-gram에 걸친 다양한 특징을 이용한 부정어 처리에 관한 연구가 진행되었다[11].

그리고 문서를 정확하게 분류하기 위해 문서빈도(document frequency), 상호정보(mutual information), 카이

〈표 1〉 감정 특징 사전 구축을 위한 문장 DB 구성

출처	감정	총계	일반	통신	인물	여행	음식	영화
Me2day	긍정	2,981	36	795	247	704	1,199	·
	부정	1,284	24	608	178	334	140	·
	중립	956	59	549	55	76	217	·
Twitter	긍정	2,800	1,180	925	110	·	585	·
	부정	2,268	799	887	281	·	301	·
	중립	2,525	1,902	368	32	·	223	·
Naver 영화	긍정	1,249	·	·	·	·	·	1,249
	부정	245	·	·	·	·	·	245
	중립	96	·	·	·	·	·	96
합계		14,404	4,000	4,132	903	1,114	2,665	1,590

제곱 통계량(χ^2 statistic), 정보 획득량(information gain)등의 특징에 관한 연구도 진행되었다. 특징을 통해 문서를 표현하는 방법으로 보통 벡터 공간 모델(vector space model)이 이용된다. 이 방법은 문서 전체의 특징을 분석하여 문서를 하나의 벡터로 표현하는 방법으로 특징의 위치 정보와 같은 문서의 구조적인 정보를 반영하지 못하는 단점을 가지고 있다. 이러한 단점을 극복하기 위해 문서의 구조적 정보를 이용한 방법이 연구되었으나 신문기사와 같은 형식적인 문서에만 적용할 수 있는 단점을 가지고 있다[12]. 이를 해결하기 위해 제목과 문장 간의 유사도를 이용해 특징의 가중치에 적용하는 연구가 수행되었다[13].

문서 분류 모델 중 신경망(neural network), Naive Bayes, Maximum Entropy 등 다양한 기계 학습 방법을 이용해 문서를 분류하는 방법들이 연구되었다. 최근에는 학습에 대한 빠른 처리 및 고차원 데이터 처리 성능이 높은 SVM을 이용한 방법이 많이 사용되고 있다[8][9][10][14].

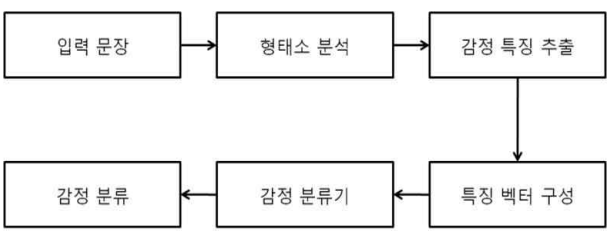
3. 감정 특징을 이용한 문장 감정 분류

3.1 개요

본 시스템의 개요는 (그림 1)과 같다. 먼저 형태소 분석기를 이용해 입력된 문장에 포함된 어휘나 단어들을 품사별로 분리한다. 본 논문에서는 기존의 공개 한글 형태소 분석기인 Korea Language Technology(KLT)를 이용했다[15]. 품사별로 구분된 단어들은 각 품사별 감정 사전을 이용해 문장에 대한 감정 특징을 추출한다. 품사별 감정 사전은 각 품사별로 각 감정에 따른 단어와 그 감정세기를 포함하고 있다. 감정 세기는 -3점부터 3점까지로 매겨져 있으며, 이 감정 세기는 온라인상에서 획득한 각 감정별 샘플 문장을 분석하여 획득하였다. 본 논문에서는 명사, 동사, 형용사, 부사, 이모티콘 등 총 5개 특징을 이용했으며, 특징 벡터는 문장이 포함하고 있는 부사를 제외한 4개 특징에 대한 감정 가중치이다. 이렇게 구성된 특징벡터는 미리 훈련된 SVM 분류기에 의해 3개 감정(긍정, 부정, 중립) 중 하나로 분류한다.

3.2 감정 특징 사전 구축

감정 특징은 감정을 느끼게 하는 단어의 어휘 자원으로 다양한 출처로부터 획득된 문서를 분석하여 기본 감정 특징을 추출했다. 하지만 이렇게 획득된 감정 특징은 그 수가 제한적이므로 이를 확장할 필요가 있다.



(그림 1) 제안한 시스템 개요

이러한 제한된 감정 특징을 확장하기 위해 기존의 여러 연구에서는 시소러스를 활용하여 단어를 확장하였다. 즉, 이미 확보된 감정 특징에 대한 시소러스 정보를 이용해 감정 특징을 확장하는 것이다. 본 논문에서는 네이버(Naver) 국어사전에 있는 관련 어휘 정보(비슷한 말, 반대말)를 활용해 품사별로 획득된 기본 감정 특징을 확장했다.

위와 같이 사전에 기반을 두어서 확장된 형식적 감정 특징들은 대부분 표준어 기반이다. 하지만 온라인상에서 사용되는 문장들은 인터넷 용어나 이모티콘과 같이 비표준어 어휘들을 많이 이용되어 있으므로 기존의 형식적 감정 특징만으로는 온라인상에서의 비형식적 문장의 감정을 분석하는 데에는 여러 가지 제약이 있다.

또한 특정 도메인에 따라 같은 어휘가 다른 감정을 나타내는 경우가 있다. 예를 들어 '가볍다'라는 어휘는 '인물' 도메인에서는 부정적인 의미를 나타내지만, '통신' 도메인에서는 긍정적 의미를 나타낸다. 이와 같이 같은 어휘가 특정 도메인에 따라 감정이 달라지므로 보다 높은 정확한 감정 분류를 위해 도메인별 감정 특징을 따로 구축하는 것이 필요하다.

따라서 본 논문에서는 이러한 기존의 형식적 감정 특징 사전 외에 <표 1>과 같이 온라인상에 등록된 사용자 등록글을 수집 및 분석하여 비형식적 감정 특징 사전, 그리고 도메인별 감정 사전을 구축했다. 표 2는 최종적으로 구축된 감정 사전의 구성을 나타낸 것이다.

<표 2>를 보면 다른 품사들에 비해 감정 사전에 포함된 이모티콘의 수가 현저히 적은 것을 확인할 수 있다. 그 이유는 다음과 같다.

〈표 2〉 구축된 감정 사전 구성

도메인	감정 특징	명사	동사	형용사	부사	이모티콘
일반	긍정	227	86	72	42	19
	부정	314	209	73		18
통신	긍정	35	14	7	·	·
	부정	45	21	11		·
인물	긍정	10	6	3	·	·
	부정	7	5	8		·
여행	긍정	31	18	13	·	·
	부정	28	11	10		·
음식	긍정	41	23	10	·	·
	부정	30	17	7		·
영화	긍정	30	10	5	·	·
	부정	24	15	5		·
합계	긍정	374	157	110	42	19
	부정	448	278	114		18

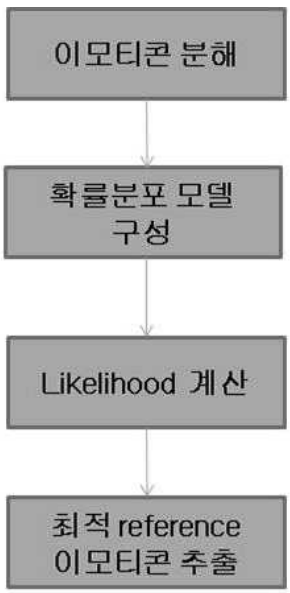
첫째, 인터넷 댓글이나 SNS 문장들을 분석한 결과, 인터넷 사용자들이 주로 사용하는 이모티콘들은 대부분 '^^'와 같은 간단한 것들이다. 이는 대부분의 SNS 사용자들이 컴퓨터 키보드나 스마트폰의 가상 키보드를 통해 글을 작성한다. 이는 키보드로는 복잡한 이모티콘을 사용하는 것이 매우 불편하고, 자신의 감정을 나타내는데 충분한 간단한 이모티콘이 있는데 굳이 복잡한 이모티콘을 사용할 필요가 없기 때문이다.

둘째, 같은 의미를 가지지만 불규칙적으로 사용되는 이모티콘들은 배제했다. 실제로 인터넷 사용자들은 이모티콘을 자주 사용하지만 많은 경우 같은 의미의 이모티콘들을 개인의 취향이나 습관, 오타 또는 기타 다양한 원인으로 인해 매우 불규칙적으로 사용한다. 이러한 불규칙성을 다루기 위해 본 논문에서는 베이지안 프레임워크를 이용한다. 이에 대한 자세한 내용은 다음 절에서 다룬다.

3.3 이모티콘 처리

원래 웃는 얼굴이 주류를 이루었기 때문에 스마일리(Smiley)로 불리던 이모티콘은 SNS 서비스에서 뿐만 아니라 이메일, 메신저와 같은 디지털 커뮤니케이션에서 표현되기 힘든 감정이나 신체의 상태나 동작을 문자, 기호, 숫자들의 조합으로 나타내는 것으로 텍스트에 의한 실시간 커뮤니케이션에서 빠르고 간결하게 감정을 내포한 의사전달을 가능하게 한다.

따라서 이모티콘은 문장의 감정을 분류하는데 매우 중요한 요소이다. 하지만 이모티콘은 사용자의 취향이나 오타, 그리고 기타 여러 가지 요인으로 인해 같은 의미를 가짐에도 매우 불규칙하게 쓰여 그 자체를 특징으로 사용하기가 매우 어렵다. 간단한 예를 들면 '^_^'과 '^__________^'은 같은 의미이지만 개인에 따라 다르게 '_'의 개수를 다르게 사용된다. 이러한 불규칙 이모티콘들을 그대로 사용하기는 어려우므로 본 논문에서는 이렇게 불규칙적으로 자주 사용되는 이모티콘 중 제일 간단한 이모티콘 형태를 참조 이모티콘이라고 하고 불규칙 이모티콘을 참조 이모티콘으로 변환하여 감정 특징으로 사용하는 방법을 이용한다. 이를 위해 불규칙 이모티콘 구성 요소들을 벡터화하고 이를 기반으로 베이지안 프레임워크를 이용한다. (그림 2)는 본 논문에서 제안한 불규칙 이모티콘 처리과정을 도식화한 것이다.



(그림 2) 제안한 시스템 개요

본 논문에서 이모티콘을 벡터화 하기 위해 사용한 <표 3>과 같은 특수문자들을 이모티콘의 구성요소로 이용한다. 각 이모티콘은 <표 3>과 같은 문자들을 bin으로 하는 히스토그램을 이용해 정규화 과정을 거쳐 다음과 같은 확률 분포 형태로 표현할 수 있다.

〈표 3〉 이모티콘 구성 요소

^	=	_	-	.	;
+	o	@	~	3	O
▽	♡	♥	★	,	$
(	)	/	>	<	。
ㅡ	ご	#	ㅋ	ㅎ	:
!	$	%	&	*	?

$$q = \{q^{(u)}\}_{u=1,\cdots,m} \tag{1}$$

여기서, $q^{(u)} = C_p \sum_{i=1}^{n} \delta[b(x_i) - u]$, δ는 kronecker delta 함수, $b(\mathbf{x_i})$는 $\mathbf{x_i}$의 히스토그램 bin(이모티콘 구성요소) 색인을 대응하는 함수, C_q는 정규화 상수이다.

q를 불규칙 이모티콘, q_{ref}를 참조 이모티콘이라 하면, 베이즈 정리에 의해 다음이 성립한다.

$$p(q_{ref} \mid q) \propto p(q \mid q_{ref})\, p(q_{ref}) \tag{2}$$

베이지안 프레임워크에서 불규칙적으로 표현된 이모티콘을 사전에 포함되어 있는 적절한 참조 이모티콘으로 변환하기 위해 본 논문에서는 maximum likelihood 방법을 이용한다.

$$\arg\max_{q_{ref}} p(q_{ref} \mid q) \tag{3}$$

식 (1)에서 불규칙 이모티콘과 참조 이모티콘의 likelihood를 계산하기 위해 본 논문에서는 분포간 유사도를 구하는데 식 (4)와 같이 battacharyya coefficent 를 이용한다. battacharyya coefficent는 두 개의 확률분포 사이의 겹쳐진 정도를 나타내는 측정치로서 두 확률분포의 유사도를 계산하는데 많이 이용된다.

$$p(q \mid q_{ref}) = \sum_{u=1}^{m} \sqrt{q^{(u)} \cdot q_{ref}^{(u)}} \quad (4)$$

3.4 특징 벡터 구성

입력 문장을 형태소 분석 후 구축된 감정 사전을 통해 각 품사별 감정 가중치를 도출하고 이를 기반으로 문장의 특징 벡터를 생성한다. 본 논문에서 사용한 특징벡터는 다음과 같다.

<line> .=.
<target><feature>:<value><feature>:<value>
… <feature>:<value>
<target> .=. +1 | 0 | -1 (+1:긍정, 0:중립, -1: 부정)
<feature> .=. <integer> (1:명사, 2:동사, 3: 형용사, 4:부사, 5:이모티콘)
<value> .=. <float> (감정 세기)

(그림 3) 특징 벡터의 구성

3.5 감정 분류기

문서나 문장을 분류하기 위해 다양한 기계 학습 방법이 이용되어 왔지만 최근에는 고차원 데이터 처리에서 좋은 성능을 가지는 SVM이 많이 이용되고 있다. SVM(Support Vector Machine)이란 Vapnik에 의해 제안된 supervised learning 방법으로 경험적 분류 오류(empirical classification error)를 최소화하면서, 기하학적 마진(geometrical margin)을 최대로 하는 방법으로 최대 마진 분류기(maximal margin classification)이라고도 한다.

SVM은 다층 퍼셉트론 분류기(Multi-Layer Perceptron classifier)의 대안적인 학습방법으로 주어진 패턴을 고차원 특징 공간(high-dimensional feature space)으로 사상할 수 있고, 대역적으로 최적의 식별이 가능하다는 특징을 가지고 있다. 또한 신경망을 포함하여 통계적 패턴 인식 방법 등 전통적인 대부분의 패턴 인식 기법들이 학습 데이터의 수행도를 최적화하기 위해 경험적인 위험 최소화(Empirical Risk Minimization)방법에 기초하고 있는 반해 SVM은 고정되어 있지만 알려지지 않은 확률 분포를 가지는 데이터에 대해 잘못 분류하는 확률을 최소로 하는 구조적인 위험 최소화(Structure Risk Minimization)방법에 기초하고 있다. 본 논문에서는 SVM 공개 라이브러리인 LIBSVM을 이용해 감정 분류기의 훈련 및 분류를 수행했다.

4. 실험 및 결과

4.1 실험데이터

실험은 트위터, 페이스북, 미투데이와 같은 SNS 서비스에서 사용자가 작성한 글들을 무작위로 수집하여 수행되었다. 수집된 문장의 구성은 <표 4>와 다음과 같다.

〈표 4〉 테스트 데이터 구성

도메인	이모티콘 사용	무	유	무	유	계
	인터넷 용어	무	무	유	유	
일반	긍 정	304	276	263	153	996
	부 정	249	241	210	201	901
	중 립	1,429	119	511	44	2,103
통신	긍 정	180	31	6	46	263
	부 정	110	9	0	2	121
	중 립	517	37	17	45	616
인물	긍 정	78	56	36	4	203
	부 정	26	6	3	1	43
	중 립	34	15	16	3	54
여행	긍 정	145	76	45	21	287
	부 정	33	8	32	9	82
	중 립	71	18	6	2	97
음식	긍 정	86	56	36	4	182
	부 정	13	6	3	1	23
	중 립	268	15	16	3	302
영화	긍 정	316	127	100	40	583
	부 정	68	13	8	3	92
	중 립	87	2	9	1	99
	합 계	4,014	1,111	1,317	583	7,047

4.2 실험 결과

본 논문에서 사용된 SVM 분류기는 Radial Basis Function(RBF) 커널을 이용했으며 이모티콘 및 인터넷 용어 사용 유무에 따라 해당 데이터를 10개의 부분집합으로 나눈 뒤 10-fold cross validation 방식으로 실험했다.

문장의 감정 분류의 성능을 평가하기 위해 다음과 같이 정확률(precision)과 재현율(recall)을 이용한 F1-measure을 사용하였다.

$$\text{정확율}(p) = \frac{\text{해당 감정으로 분류된 실제 해당 감정 문장 수}}{\text{해당 감정으로 분류된 문장 수}} \quad (5)$$

$$\text{재현율}(r) = \frac{\text{해당 감정으로 분류된 실제 해당 감정 문장 수}}{\text{해당 감정 전체 문장 수}} \quad (6)$$

$$F_1(r,p) = \frac{2rp}{r+p} \quad (7)$$

〈표 5〉 도메인별 성능 비교

도메인	구분	감정	Precision	recall	F1-measure
일반	일반 감정 사전 사용	긍정	0.6287	0.5592	0.5919
		부정	0.6377	0.4560	0.5318
		중립	0.6841	0.7993	0.7372
	일반 감정 사전 + 이모티콘 처리	긍정	0.6986	0.7891	0.7411
		부정	0.7111	0.7514	0.7307
		중립	0.8393	0.7675	0.8018
통신	일반 감정 사전 사용	긍정	0.7169	0.6302	0.6707
		부정	0.7922	0.5495	0.6488
		중립	0.2101	0.5238	0.2999
	일반 감정 사전 + 이모티콘 처리	긍정	0.7335	0.672	0.7014
		부정	0.7633	0.6103	0.6782
		중립	0.243	0.4841	0.3235
	도메인 감정 사전 사용	긍정	0.7336	0.6235	0.6740
		부정	0.809	0.6123	0.6970
		중립	0.235	0.5419	0.3278
	도메인 감정 사전 + 이모티콘 처리	긍정	0.7564	0.6743	0.7129
		부정	0.7789	0.6788	0.7254
		중립	0.282	0.5038	0.3615
인물	일반 감정 사전 사용	긍정	0.9207	0.7438	0.8228
		부정	0.6585	0.6279	0.6428
		중립	0.4842	0.8518	0.6174
	일반 감정 사전 + 이모티콘 처리	긍정	0.9454	0.7684	0.8477
		부정	0.5669	0.6976	0.6254
		중립	0.518	0.7962	0.6276
	도메인 감정 사전 사용	긍정	0.9147	0.7931	0.8495
		부정	0.7714	0.6279	0.6922
		중립	0.5056	0.8333	0.6293
	도메인 감정 사전 + 이모티콘 처리	긍정	0.9325	0.8177	0.8713
		부정	0.6521	0.6976	0.6740
		중립	0.5394	0.7592	0.6306
여행	일반 감정 사전 사용	긍정	0.8995	0.655	0.7580
		부정	0.6987	0.7073	0.7029
		중립	0.3965	0.7113	0.5091
	일반 감정 사전 + 이모티콘 처리	긍정	0.9099	0.7038	0.7936
		부정	0.566	0.7317	0.6382
		중립	0.4492	0.6391	0.5275
	도메인 감정 사전 사용	긍정	0.917	0.655	0.7641
		부정	0.6904	0.7073	0.6987
		중립	0.4067	0.7422	0.5254
	도메인 감정 사전 + 이모티콘 처리	긍정	0.9324	0.7212	0.8133
		부정	0.5607	0.7317	0.6348
		중립	0.4598	0.6494	0.5383
음식	일반 감정 사전 사용	긍정	0.9252	0.5439	0.6850
		부정	0.3714	0.5652	0.4482
		중립	0.7616	0.9205	0.8335
	일반 감정 사전 + 이모티콘 처리	긍정	0.9393	0.6813	0.7897
		부정	0.3	0.6521	0.4109
		중립	0.8246	0.8874	0.8548
	도메인 감정 사전 사용	긍정	0.9035	0.5659	0.6959
		부정	0.423	0.4782	0.4489
		중립	0.7629	0.9271	0.8370
	도메인 감정 사전 + 이모티콘 처리	긍정	0.9295	0.7252	0.8147
		부정	0.3333	0.5652	0.4193
		중립	0.8282	0.894	0.8598
영화	일반 감정 사전 사용	긍정	0.9048	0.6363	0.7471
		부정	0.4628	0.6086	0.5257
		중립	0.3045	0.7474	0.4327
	일반 감정 사전 + 이모티콘 처리	긍정	0.9178	0.6706	0.7749
		부정	0.4765	0.663	0.5544
		중립	0.3409	0.7575	0.4701
	도메인 감정 사전 사용	긍정	0.9099	0.6758	0.7755
		부정	0.5888	0.576	0.5823
		중립	0.2988	0.7575	0.4285
	도메인 감정 사전 + 이모티콘 처리	긍정	0.9213	0.7032	0.7976
		부정	0.5686	0.6304	0.5979
		중립	0.3348	0.7676	0.4662

테스트 문장에 대한 문장 감정 분류 결과는 <표 5>와 같다. <표 5>에서 볼 수 있듯이 기존의 형식적 어휘만을 이용하는 경우와 비교했을 때 형식적 감정 어휘 정보와 비형식적 어휘 정보를 함께 이용한 제안한 방법이 보다 높은 F1-measure를 가짐을 확인할 수 있다. 특히 긍정 감정과 부정 감정의 경우에는 사전의 종류나 이모티콘의 고려 여부에 따라 F1-measure의 차이가 나타남을 확인할 수 있다. 이는 SNS 댓글이나 인터넷 댓글 같은 문장의 감정을 분류하는데 형식적 어휘뿐만 아니라 비형식적 어휘도 매우 중요한 정보임을 나타낸다.

5. 결 론

본 논문에서는 사전 정보에 기반을 둔 형식적 감정 어휘와 이모티콘과 같은 비형식적 감정 어휘를 이용한 문장 감정 분류 시스템을 제안했다. 제안한 방법은 기존의 형식적 감정 특징을 이용해 문장의 감정을 분류하는 방법에 비해 인터넷 상의 댓글과 같은 다양한 문장의 감정을 분류하는데 우수한 성능을 보여주었다.

하지만 실험 결과를 보면 통신과 영화 분야에서의 결과가 다른 분야에 비해 분류 결과가 좋지 않은데 이는 통신과 영화 분야의 댓글들이 우회적으로 표현한 것이 다른 분야에 비해 상대적으로 많았기 때문이다.

향후 연구 과제로써 이러한 우회적 표현에 대해 보다 정확한 감정 분류 방법을 연구할 예정이다. 또한 현재는 정해진 감정 특징 사전을 이용하기 때문에 새로운 단어나 이모티콘을 신속하게 반영하기 어렵다. 따라서 새로운 형태의 감정 특징을 신속하게 반영하기 위해 on-line 학습방법을 이용해 새로운 감정 특징을 반영하는 시스템을 연구할 예정이다.

참 고 문 헌

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

[2] J. Yi, T. Nasukawa, R. Bunescu, and W. Niblack, “Sentimental Analyzer : Extracting Sentiments about a Given Topic using Natural Language Processing Techniques,” In Proceedings of International Conferenceon Data Mining, pp.427-434, 2003.

[3] N. Hiroshima, S. Yamada, O. Furuse and R. Kataoka, “Searching for Sentences Expressing Opinions by Using Declaratively Subjective Clues,” In Proceedings of the Workshop on Sentiment and Subjectivity in Text, pp.39-46, 2006.

[4] 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.

[5] S.M. Kim and E. Hovy, “Determining the Sentiment of Opinions,” In Proceedings of the COLING conference, pp.1367-1373, 2004.

[6] A. Esuli and F. Sebastiani, “Determining the Semantic Orientation of Terms through Gloss Classification,” In Proceedings of the CIKM, pp.617-624, 2005.

[7] A. Esuli and F. Sebastiani, “PageRanking WordNet Synsets: An Application to Opinoin Mining,” In Proceedings of the ACL, pp.424-431, 2007.

[8] 김묘실, 강승식. “SVM을 이용한 악성 댓글 판별 시스템의 설계 및 구현”, 한글 및 한국어 정보처리 학술대회, pp.285-289, 2006.

[9] 황재원, 고영중. “감정 자질을 이용한 한국어 문장 및 문서 감정 분류 시스템”, 정보과학회논문지 : 컴퓨팅의 실제 및 레터, 14(3): pp.336-340, 2008.

[10] 황재원, 고영중. “문장 감정 강도를 반영한 개선된 자질 가중치 기법 기반의 문서 감정 분류 시스템”, 정보과학회논문지 : 소프트웨어 및 응용, 36(6): pp.491-497, 2009.

[11] 정유철, 최윤정, 맹성현, “감정 기반 블로그 문서 분류를 위한 부정어 처리 및 단어 가중치 적용 기법의 효과에 관한 연구”, 인지과학, 19(4): pp.477-497, 2008.

[12] M. Murata, Q. Ma, K. Uchimoto, H. Ozaku, H. Isahara, and M. Utiyama, “Japanese Information Retrieval Using Location and Category Information,” Journal of the Association for Natural Language Processing, Vol.7, No.2, pp.81-88, 2000.

[13] Y. Ko, J. Park, and J, Seo, “Automatic Text Categorization using the Importance of Sentences,” In Proceedings of the 19th International Conference on COLING, pp.474-480, 2002.

[14] Joachims, T. “Text Categorization with Support Vector Machines: Learning with Many Relevant Features”. Machine Learning ECML98, Tenth European Conference on Machine Learning, pp.137-142, 1998.

[15] http://nlp.kookmin.ac.kr/HAM/kor/download.html

조 상 현

e-mail : cshgreat@catholic.ac.kr

2003년 가톨릭대학교 수학과

2005년 가톨릭대학교 컴퓨터공학과(석사)

2005년~현 재 가톨릭대학교 컴퓨터공학과 박사과정

관심분야 : 컴퓨터 비전, 패턴인식, 컴퓨터 그래픽스

강 행 봉

e-mail : hbkang@catholic.ac.kr

1980년 한양대학교 전자공학과

1986년 한양대학교 전자공학과(석사)

1989년 Ohio State Univ. 컴퓨터공학(석사)

1993년 Rensselaer Polytechnic Institute 컴퓨터공학(박사)

1994년~1997년 삼성종합기술원 수석연구원

1997년~현 재 가톨릭대학교 디지털미디어학부 교수

2005년 UC Santa Barbara Visiting Professor

관심분야 : 컴퓨터비젼, 컴퓨터그래픽스, HCI, 인공지능, 기계학습