자질 가중치의 재조정을 통한 감정 분류

서형원○, 김형철, 김재훈, 이공주
한국해양대학교 컴퓨터공학과, 충남대학교 전기정보통신공학부
wonn24@gmail.com, yhdosu@nate.com, jhoon@hhu.ac.kr, kjoolee@cnu.ac.kr

Sentiment Classification Using Feature Reweighting

Hyung-Won Seo, Hyung-Chul Kim, Jae-Hoon Kim, Kong-Joo Lee
Department of Computer Engineering, Korea Maritime University
Division of Electrical and Computer Engineering, Chungnam National University

요 약

이 논문은 한글 뉴스 기사의 댓글에 대한 감정 분류 방법을 제안한다. 제안된 방법은 기계학습을 이용하는데 본 논문에서는 자질의 가중치를 재조정하는 좀 색다른 방법을 제안한다. 일반적으로 댓글은 독자들이 특정 기사에 대해서 어떠한 감정을 가지고 있는지를 파악하는 중요한 단서가 된다. 그런데 독자들의 감정은 가사에 어떤 분야에 속하느냐에 영향을 받는다. 예를 들면 정치 기사는 부정적인 댓글은 많이 포함하고 있으며 인물 기사는 긍정적인 기사를 많이 포함한다. 이 논문은 이와 같은 댓글의 속성을 이용해서 기사의 원문과 기사의 분야 정보를 이용하여 가중치를 조정한다. 제안된 시스템의 성능을 평가하기 위해 신문 기사와 댓글을 수집하여 감정 말뭉치를 구축하였으며 감정 자질을 추출하기 위해 감정 사전을 구축하였다. 제안된 시스템의 F_1 척도는 92.2%였으며 원문의 감정 단어와 분야 정보가 댓글의 감정을 분류하는데 중요한 자질임을 알 수 있었다.

주제어: 감정 분석, 감정 마이닝, 자질 추출

1. 서론

최근 인터넷의 급속한 발달로 웹에서의 커뮤니케이션이 매우 활발해지고 있다. 이로 인해 누구나 쉽게 자신의 의사를 표현하거나 다른 이들의 의견을 수집하는 것이 가능해졌다. 예를 들어, 인터넷에서 뉴스 기사를 읽고 댓글을 달거나 어떤 상품에 대한 평을 자유롭게 작성하는 등, 웹에서의 개인적인 의사 표현은 다른 사람에게 큰 영향력을 미치는 요인 중에 하나가 될 수 있다. 가령 어떤 물건을 구입하려는 구매자에게 그 상품에 대한 사용 후기가 굉장히 큰 영향을 미칠 수 있다. 이런 이유들 때문에 최근 많은 사람들은 주관적인 감정이나 의견, 판단, 추천 등을 자동적으로 수집하고 분석하는 것에 많은 관심을 가지게 되었다[1]. “이 제품에 대해 사람들은 어떻게 생각하고 있을까?” “이 제품에 대해 사람들은 무엇이 불만이고 그 이유는 무엇일까?”와 같은 질문은 왜 주관적인 의견이나 감정을 다루는 ‘감정 마이닝’ 혹은 ‘감정 분석’이 중요한지를 알게 해주는 좋은 이유가 될 수 있다. 예를 들어, “이 영화가 끝날 때까지 지루하지 않게 잘 참고 있을 사람은 아무도 없을 것이다.”라는 문장은 부정적인 의미를 표현한다. 이렇게 주관적인 의견이 들어간 표현과는 다르게, “이 제품은 NB2LH 리튬 이온 밧데리가 기본으로 포함되어 있는 패키지이다.”라는 문장은 중립적인 의미를 표현한다. 이런 표현들은 신문 기사, 개인 블로그, 리뷰 사이트 등에서 쉽게 발견할 수 있다. 대부분의 이런 문서들은 객관적이거나 주관적인 형태로 나타날 수 있는데, 이 말은 즉 실제의 객관적인 사실에 기인한 글이거나 주관적인 견해나 의견을 나타낸 것인지를 의미한다. 하지만 불행히도 대부분의 경우에는 이 두 가지가 모두 혼합되어 있는 문서가 많기 때문에 이런 문서들을 분류하는 것은 좀 더 많은 언어적 처리가 필요하다. 그렇지만 인터넷에 있는 문서 중 어떤 주관적인 의견을 수집하고 분석하는 일은 기업이나 개인에게도 매우 유용한 자료가 될 수 있기 때문에, 감정 분류는 반드시 연구되어야 한다. 게다가 스팸 메일 필터링에 이용할 수 있고, 어떤 이모티콘을 사용했는지에 따라 수신 메일을 좋은 쪽 혹은 나쁜 쪽으로 분류할 수도 있다[2].

많은 연구자들이 이미 다양한 방면에 걸쳐서 이 분야에 대해 연구해오고 있다. 어떤 학자들은 문장 단위로 감정 분석을 했으며[3,4] 어떤 학자들은 문서 단위로 감정 분석을 하였다[5,6]. 몇몇 학자들은 형용사, 동사 혹은 감정 표현과 관련된 n-gram을 자동으로 인식하고 분석하는 것을 연구했다[6-8]. [9]는 bootstrapping 패턴 학습 시스템을 이용하여 감정 표현을 추출하였다. 게다가, 어떤 학자들은 주관적인 표현들을 패턴 형식으로 추출하였다[10]. 대부분의 이런 학자들은 문장이나 문서가 감정적인 표현을 가지고 있다는 전제 하에 단지 긍정인지 아닌지를 다루고 있지만, 이 논문은 원래의 문서(즉, 기사 본문)와 그에 달린 댓글을 같이 분석하여 두 문서간의 어떤 관계가 있는지를 파악하는 것에 중점을 두었다. 그리고 중립적인 표현은 무시하고 긍정 혹은 부정의 표현만을 대상으로 하였다.

이 논문의 구성은 다음과 같다. 2장에서 몇몇 관련된 연구에 대해 기술하고 3장에서 가중치 조정 방법에 대해 기술한다. 4장에서는 전체적인 감정 분류 시스템의 구성에 대해 기술하고 5장에서는 실험, 그리고 나머지 결론과 향후 과제에 대해 기술할 것이다.

2. 관련 연구

2.1 벡터 공간 모델에서의 가중치 조정 방법

벡터 공간 모델(vector space model)은 문서를 순위화하는 대표적인 정보검색 모델이다. 벡터 공간 모델이란 간단히 말해 각 문서를 하나의 벡터로 간주하여 표현되며 각각의 차원(dimension)은 단어(term)에 대응된다. 만약 문서 안에 어떤 특정 단어가 있다면 그에 대한 차원은 0이 아닌 어떤 값을 가지게 된다. 이런 단어 가중치(term weight) 값을 계산하on위해 여i된다식이 이용되는데, 그 중에 몇몇은 확률 이론가지깔초한 방법들가지비해 더단어 가결과를 내기을 한다. 가장 널리 알려진 방법 중 하나인 TF-IDF(Term Frequency-Inverse Document Frequency)[11]는 단어 빈도수(TF)를 역문헌 빈도수(IDF)와 같이 적용하여 어떤 문서를 대표하는 단어들을 효율적으로 찾아주는 알고리즘이며 수식 (1)과 같이 정의된다.

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (1)$$

식 (1)에서 $n_{i,j}$ 는 어떤 특정한 단어 w_i가 특정 문서d_j 안에 출현한 횟수이다. 그리고 $\sum_k n_{k,j}$는 같은 문서 안에서 그 단어들이 출현한 총 횟수이다. 즉, 더 많이 나온 단어가 더 중요하다는 의미에서 나온 것이다. 그러나 어떤 단어가 특정 말뭉치 안에 있는 다른 문서에도 여러 번 나온다면 그것은 중요한 것이 아닐 수도 있다. 어떻게 보면 여기도 나오고 저기도 나오는 그런 평범한 단어일 수 있기 때문이다. 이런 단어를 제외시키기 위해 역문헌 빈도수를 사용한다. DF(Document Frequency)는 어떤 말뭉치 안에 특정 단어 w_i를 포함하는 문서 수이다. 역문헌 빈도수는 모든 문서 수를 문서 빈도수로 나눈 것에 로그를 취함으로써 구할 수 있고 수식 (2)와 같이 정의된다.

$$idf_i = \log \frac{|D|}{|\{d : w_i \in d\}|} \quad (2)$$

식 (2)에서 $|D|$는 말뭉치 안에 있는 문서의 총 수, 그리고 $|d : w_i \in d|$는 단어 w_i를 포함하는 문서 수이다. 하지만 만약 말뭉치 안에서 w_i가 포함되어 있지 않다면 0으로 나누는 문제(division-by-zero)를 야기할 수 있기 때문에 보통 $1+|d : w_i \in d|$을 이용한다. 최종적으로 TF-IDF는 수식 (3)과 같이 정의된다.

$$tf-idf_{i,j} = tf_{i,j} \bullet idf_i \quad (3)$$

TF-IDF 가중치는 문서의 TF가 높고 모든 문서에 대한 DF가 낮을수록 높아진다. 그러므로 말뭉치 안에 있는 모든 문서에 걸쳐 나오는 단어는 제거될 것이다. 이런 이유 때문에 TF-IDF는 벡터 공간 모델 안에 있는 두 문서 사이의 유사도를 구하기 위해 코사인 유사도와 함께 자주 쓰인다.

2.2 자질 선택 및 추출

자질 선택은 패턴인식, 통계, 데이터 마이닝 분야에서 자주 이용되어 왔다[12]. 주된 아이디어는 대량의 문서로부터 자질을 조금씩 축소해 나가면서 효과적인 자질만을 선택하는 것이다. 그래서 이 과정 중 올바른 자질 집합을 찾는 것은 매우 중요하다. 오늘날 문서 분류 문제에 있어서 더 높은 정확도를 가지는 자질을 선택하기 위한 많은 연구가 진행되어 오고 있다[13,14]. 반면에, 자질 추출은 입력 데이터로부터 의미 있는 자질들의 집합을 추출하는 것을 의미한다. 이것은 주로 입력 데이터가 너무 큰 것에 비해 그 안에 정말로 필요한 정보량이 충분하지 않을 때 사용되곤 한다. 추출된 각각의 단어는 가중치 벡터를 만드는데, 더 높은 가중치를 가지는 것이 그 문서의 특징을 더 잘 표현한다고 볼 수 있다. 정보 검색 분야에서 문서로부터 추출된 내용어는 주로 명사와 동사로 이루어지는 경우가 많다. 그러나 감정 분류 문제에서는 형용사나 부사가 자질을 추출하는데 있어서 중요한 역할을 한다.

2.3 감정 분류

넓은 의미에서 감정 마이닝은 웹 문서 안에 있는 어떤 대상에 대해 저자의 개인적인 의견이나 감정을 나타내는 것을 찾아내는 것을 의미한다[15]. 이 대상은 상품, 영화, 서비스, 어떤 주제 등이 될 수 있다.

감정 분류는 조금 다른 의미에서, 감정 마이닝을 문서 분류의 문제로서 접근한다[16]. 그 말은 즉, 분류 문제를 문서 단위에서 접근하고 어떤 문서가 감정적인 표현을 가지는 지에 상관없이 긍정적인지 부정적인 지에 따라 문서를 분류하는 것을 의미한다. 일반적인 데이터 마이닝에서의 문서 분류는 정의된 클래스에 기반을 둔 내용어를 이용하고 이런 내용어는 기본적으로 명사로 되어 있다. 이에 반해 감정 분류는 문제를 '훌륭하다', '좋다', '느리다' 처럼 형용사 혹은 부사로 된 감정 단어로 다룬다[5]. 그리고 감정 분류에서의 클래스는 주로 긍정, 부정 혹은 중립에 포커스를 맞춘다.

본 논문은 문서 단위에서 분류를 하는 것에 초점을 맞춘다. 그리고 본문과 그에 대한 댓글들 사이에 어떤 관계가 있는지 살펴본다.

3. 감정 분류에서의 자질 조정 방법

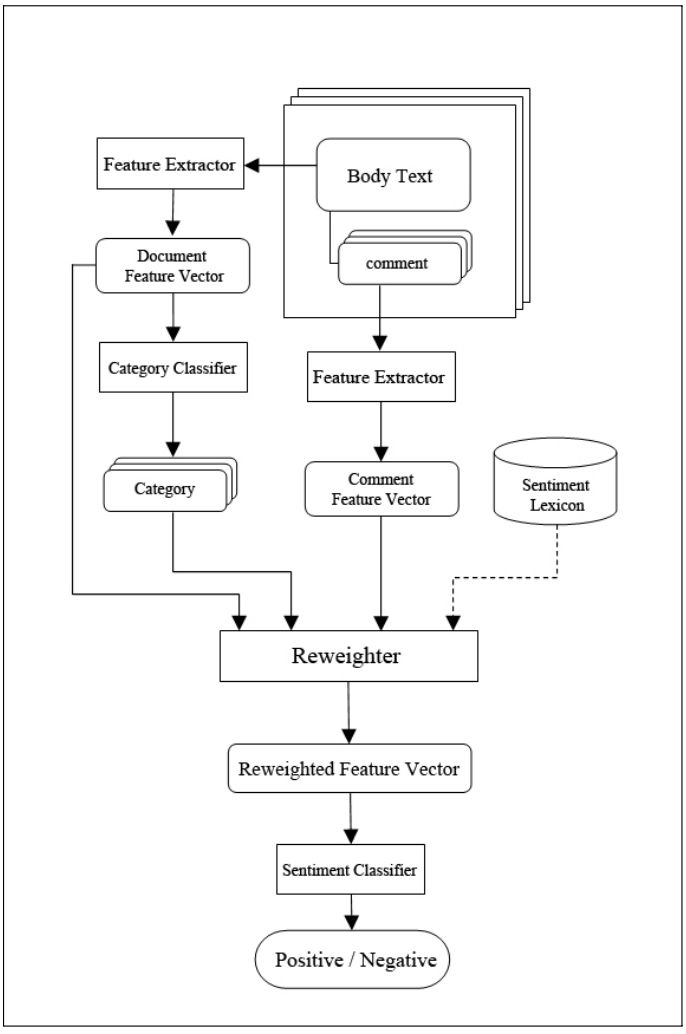
이 장에서는 가중치를 조정한 자질 벡터를 이용한 감정 분류 시스템의 구성에 대해 기술한다.

3.1 감정 분류 시스템

(그림 1)은 이 논문에서 제안한 감정 분류 시스템의 구성도이다. 시스템은 문서를 입력으로 받아서 출력 결과로 긍정인지 부정인지를 나타낸다. 이 논문에서 다루는 문서는 다른 감정 분류 모델들[5,17]과는 달리, 본문과 그것의 댓글로 구성되어 있다. 각 문서는 정치, 경

제와 같은 카테고리를 구성하고 있고, 각 댓글은 기본적으로 긍정, 부정, 중립적인 표현을 포함하고 있다. 그러나 우리는 문서의 모호하고 불규칙적인 언어 특성 때문에 긍정과 부정, 이 2가지 형태의 댓글에 대해서만 수집하고 다룬다.

이 논문에서 제안한 시스템은 2개의 자질 추출기, 카테고리 분류기, 가중치 재조정기, 감정 분류기로 구성되어 있다. 비록 2개의 자질 추출기를 각각 다른 곳에서 사용하지만 그들의 기능은 똑같다. 자질 추출기는 각각 본문이나 댓글을 입력으로 받아서 불용어 목록에 있는 단어들은 제거를 한 후 2음절 혹은 3음절로 나눈다(3.2절에 자세히 설명할 것이다). 카테고리 분류기는 벡터로 표현된 본문을 입력으로 받아서 그 본문이 어떤 카테고리에 가장 적합한지를 기계 학습[18]을 통해 알아낸다.



(그림 1) 감정 인식 시스템 구성

이 논문에서 사용된 분야는 8개이고, 이것에 대해서는 나중에 좀 더 자세히 기술할 것이다. 기본적으로 자질의 가중치를 조정하는 것은 벡터에 표현된 댓글들을 입력으로 받고, 어떤 특정 자질들(3.3절에서 더 다룰 것임)의 가중치를 2차적으로 높이기 위해 추가적으로 본문, 분야 정보와 감정 사전에 대한 자질 벡터를 받는다. 이렇게 최종적으로 모아진 자질 벡터를 가지고 감정 분류기는 댓글이 본문에 대해 긍정인지 부정인지를 분야 분류기처럼 기계 학습을 이용해 결정한다.

3.2 자질 추출

최근 외국의 연구들은 감정을 추출하는 것에 있어서 한 단어 혹은 두 단어에 초점을 맞추고 있다. [5]는 그들의 연구에서 한 단어가 가장 좋은 성능을 보였다고 보고했다. 그러나 우리는 한국어의 특성 때문에 이 연구의 결과를 그대로 사용할 수 없었다. 기본적으로 한국어에는 너무나 다양한 규칙 혹은 불규칙 문법이 존재하고, 게다가 정치 분야처럼 대부분 부정적인 댓글을 야기 시키는 기사가 많은 경우에는 비속어나 채팅용어처럼 이해하기 어려운 단어가 많이 사용되기도 하기 때문이다. 따라서 한 단어를 그대로 사용하는 것 대신에 본 논문은 두 음절 혹은 세 음절을 사용한다. 다음과 같은 이유 때문에 형태소 분석과 같은 특별한 색인기를 사용하지 않아도 그다지 큰 문제가 되지 않는다고 생각한다.

- 한국어에서 단어의 대략 80%는 2음절 혹은 3음절이다[19].
- 한국어에서 2음절 혹은 3음절은 정보 검색 분야에서 충분히 좋은 자질로 사용될 수 있다[20,21].
- 상당수의 댓글은 은어, 속어, 두문어, 심지어 이모티콘처럼 형태학적 분석에서 오류를 범할 수 있는 여러 변칙적인 단어들을 많이 포함한다.

이 논문은 세 가지 형태의 텍스트(즉, 본문, 댓글, 감정 사전)를 두 음절 혹은 세 음절로 나누어 사용하였다. 그리고 변형시킨 후, 그것들이 긍정을 나타내는지 부정을 나타내는지 판단하기 어려울 수 있다. 이런 음절은 중요한 자질이 아니면서 높은 빈도수를 가질 수 있기 때문에 이 논문에서는 불용어를 이용하여 문제를 다루기로 한다.

3.3 가중치 재조정 방법

이 논문에서는 기본적으로 TF-IDF를 가중치 조정 방법으로 이용한다. 이 절에서는 한글 신문에 대한 댓글들을 대상으로 하는 감정 분류 문제에서 자질을 조정하는 조금 색다른 방법을 제시한다. 이 방법은 수식 (3)에서 보는 것과 같이 특정 상태에 따라서 용어 빈도수를 조정한다.

$$tf'_{ij} = tf_{ij} + \alpha, \quad \alpha = \begin{cases} 2 & \text{if } t_{ij} \in (B_{jh(j)} \cap S \cap M_{h(j)}) \\ 1 & \text{if } t_{ij} \in (B_{jh(j)} \cap S) \\ 0 & otherwise \end{cases} \qquad (3)$$

여기서 t_{ij}, $B_{jh(j)}$, S, $M_{h(j)}$은 각각 j번째 댓글에 달린 i번째 단어, 어떤 기사 원문에 달린 j번 째 댓글, 감정 단어, 기사의 원문을 의미한다. $h(j)$는 j번째 댓글에 대한 기사 원문을 찾는 함수이다. 즉, 어떤 댓글이 감정 단어라면 TF를 1만큼 증가시키고, 그것이 원문에도 속하는 단어라면 2만큼 증가시킨다. 또한 식 (4)와 같이 자질이 기사 원문 분야와 일치한 경우에는 원문 기사에

포함된 감정 단어의 수에 비례한 만큼 TF를 증가시킨다(식 (4)). 식 (4)에서 $C_{h(j)}$은 그 원문에 대한 카테고리 정보를 의미하고 이것은 그림 (1)에 설명했듯이 카테고리 분류기를 통하여 결정하였다.

$$tf''_{ij} = tf'_{ij} + \beta, \qquad \beta = \begin{cases} \frac{|B_{jh(j)} \cap S \cap M_{h(j)}|}{2} & \text{if } t_{ij} = C_{h(j)} \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

4. 말뭉치 및 감정 사전 구축

이 논문은 제시된 감정 분류 방법을 평가하기 위해 두 가지의 데이터를 사용한다. 첫 번째 데이터는 제시된 방법을 평가하기 위한 말뭉치이고 두 번째는 자질 재조정 방법을 위한 감정 사전이다. 두 데이터들은 공개된 것이 없기 때문에 직접 만들어서 사용하였다. 어떻게 만들었는지는 다음 절에서 기술할 것이다.

4.1 말뭉치

이 논문에서는 제시한 시스템을 평가하기 위해 한글 인터넷 뉴스 기사[1]로부터 1,377개의 한글 뉴스 기사를 수집하였다(<표 1>). 수집된 신문 기사들은 문서로써 이전에 언급된 고유의 카테고리를 가지고 있고, 하나 이상의 댓글을 포함하고 있다. 댓글을 포함하지 않는 본문은 미리 제거하였고 그 결과, 최종적으로 8,332개의 댓글을 수집하였다. <표 2>는 평가하기 위한 말뭉치로써 수집된 문서의 통계적 정보를 나타낸다.

<표 1> 기사 원문의 분야 정보

분야	문서 수	분야	문서 수
정치	462	IT	13
경제	293	컬럼	44
국제	107	인물	13
사회	321	문화	124
총	1,377		

<표 2> 수집된 말뭉치의 통계적 정보

	본문	댓글
총 수	1377	8,322
평균 문서 수	-	6
단어 총 수	863,379	274,626
평균 단어 수	627	33

<표 2>에서 보는 바와 같이 본문 하나에 대해 평균 댓글 수는 6이고, 댓글에 대해 평균 단어의 수는 33이고, 이는 평균적으로 두세 문장으로 구성되어있음을 뜻한다. 그러므로 이 논문에서는 댓글 자체를 하나의 작은 문서라고 가정한다. 각 댓글은 수동적으로 긍정인지 부정인지에 대한 정보를 부착하였다. 그 결과로 7,513개의 부정, 809개의 긍정 댓글을 수집하였다.

1) www.donga.com

4.2 감정 사전

이 논문은 인터넷[2]을 통해 영어로 된 감정 사전을 미리 수집하였다. 그 감정 사전에는 4,138개의 부정, 2,297개의 긍정 단어로 구성되었다. 이 논문에서는 영어 감정 사전에 있는 단어를 영한사전을 이용하여 반자동적으로 번역하여 한글 감정 단어를 수집하였다. 이에 유의어와 반의어를 이용하여 확장시켰다. 이런 작업을 걸쳐 최종적으로 총 4,046개의 부정 단어와 3,044개의 긍정 단어로 구성된 한글 감정 사전을 구축하였다. 이 사전은 명사, 형용사, 부사로 구성되어 있다(<표 3>).

<표 3> 감정 단어 총 수

	영어	한글
부정	4,138	4,046
긍정	2,297	3,044

5. 예비 실험

이 논문에서는 아래와 같은 질문의 답을 찾기 위해서 예비 실험을 수행하였다. 첫째, 어떤 분류기가 한글 댓글을 감정 분류를 하는데 있어서 가장 적합한가? 둘째, 기사의 본문이 과연 그 댓글에 감정적인 표현들을 인식하는데 있어서 도움을 주는가? 셋째, n-gram을 이용하는 것이 한글 댓글을 감정 분류하는데 있어 충분한가? 다음 절에서는 이런 질문들에 답하기에 앞서 실험을 위한 환경을 기술한다.

5.1 실험 환경

5.1.1 학습 말뭉치와 테스트 말뭉치

이 논문에서는 제안된 시스템을 평가하기 위해 4장에서 기술한 말뭉치를 사용한다. 총 8,322개의 댓글로 구성된 말뭉치는 감정 분류를 하는데 있어서 충분치 않다고 판단하였기 때문에 교차 검증(cross-validation) 방법[22]을 이용한다. 이 논문에서 사용된 모든 평가는 전체 데이터에 대해서 4차 교차 검증(4-fold cross validation)을 이용한다.

5.1.2 정확도 측정

이 논문에서 정확도 측정을 위해 사용된 F_1 척도(F-score 혹은 F-measure)는 정확도와 재현율을 혼합한 보편적인 측정 방법이라고 할 수 있다. F_1 척도는 정확도와 재현율 사이에 적절히 평균적인 값을 갖게 하는 것이라고 해석될 수 있다.

5.1.3 기계학습 도구

일반적인 영역에서, Weka[3]와 AI::Categorizer[4]처럼 다양한 기계학습 도구가 있다. 이 논문에서는 후자를 이용하는데, 이것은 자동적으로 텍스트를 분류하기 위한 framework이고 모듈들로 구성되어 있다.

5.2 감정 분류를 위한 분류기

이 절에서는 앞서 제시된 질문에 답을 소개할 것이

2) www.cs.pitt.edu/mpqa
3) www.cs.waikato.ac.nz/ml/weka/
4) search.cpan.org/~kwilliams/AI-Categorizer-0.09/

다. 즉, "어떤 분류기가 한글 댓글을 감정 분류를 하는데 있어서 가장 적합한가?" 인데 일반적으로 분류기의 성능에 영향을 미치는 수많은 파라미터들이 있다. 가장 적합한 분류기를 선택하기 위해 다음과 같이 3가지 파라미터를 고정시킨다.

- 자질의 수: 800
- 자질 선택 방법: Chi Square
- 자질들: 2음절

이 논문에서는 여러 분류기가 있지만 그 중 KNN, Naïve Bayes, SVM분류기를 후보로 정하였다. <표 4>는 각 분류기에 대한 거시 평균(macro-average)을 나타낸다.

<표 4> 각 분류기의 성능

분류기	재현율	정확률	F_1
Naïve	0.935	0.871	0.902
KNN	0.867	0.936	0.903
SVM	**0.929**	**0.929**	**0.929**

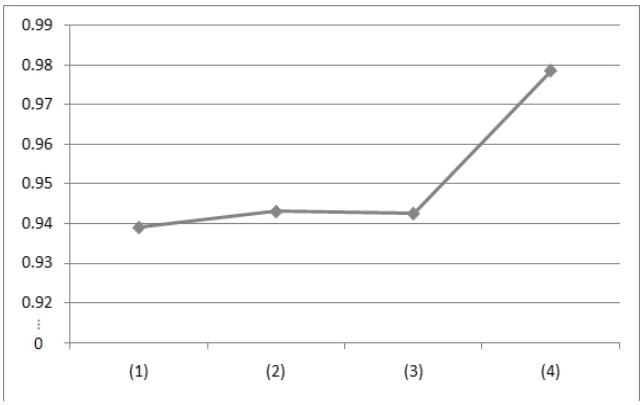
일반적으로 텍스트 분류하는 것에 있어서 SVM이 가장 좋은 성능을 보여 왔다[23]. 이 논문에서 실험한 결과 역시 비슷한 결과를 얻을 수 있었다. <표 4>에서 보이는 바와 같이, SVM 이 가장 좋은 성능을 보였기 때문에 이 논문은 SVM을 기본 분류기로 간주한다.

5.3 실험 결과

이 장에서는 3장과 4장에서 기술한 4가지 타입의 자질을 평가한다.

(1) B (댓글 안의 단어)
(2) B + S (감정 사전 안의 단어)
(3) B + S + M (본문 안의 단어)
(4) B + S + M + C (본문의 카테고리 정보)

이 논문은 3.2절에서 언급된 것처럼 자질 타입에 따라 적절한 자질 재조정 방법을 이용한다. 자질 타입 (1)의 경우, 어떤 재조정 방법이라도 사용되지 않았고 이것을 우리의 기본 모델로 정의한다. 자질 타입 (2)와 (3)의 경우, 각각 방법에 수식 (3-2)과 수식 (3-1)이 적용되었다. 자질 타입 (4)의 경우, 모든 자질과 더불어 카테고리 정보까지 이용되었다.



(그림 2) 감정 분류 시스템의 성능 평가

(그림 2)는 모든 가중치 재조정 방법에 대한 우리의 실험 결과를 나타낸다. 결과에 따르면, 앞서 5장에서 언급된 두 번째와 세 번째 질문에 대한 답을 찾을 수 있다. 즉, 본문과 n-gram을 이용하는 것이 한글 신문 기사의 댓글을 감정 분류하는데 있어서 도움이 된다는 것을 알 수 있다.

6. 결론 및 향후 연구

이 논문은 한글 문서에 대한 감정 분류를 위한 가중치 조정 방법의 조금 색다른 방법을 제시한다. 이 방법은 자질로써 본문과 그에 대한 댓글, 감정 사전, 그리고 분야 정보를 이용한다. 게다가 제시된 방법은 감정 단어나 문서에 관련된 어떤 자질들의 가중치를 높이는 역할을 한다. 비록 이것은 예비 실험이긴 하지만, 앞서 제시된 결과들은 이 방법이 분류의 효과적인 측면에서 거시 평균 F_1 척도를 향상시키는데 도움이 된다는 것을 보였다. 향후 과제로 다양한 자질 선택 방법을 이용하는 것이 측정 방법에 따라 어떤 특정 관계를 파악할 수 있기 때문에 정보 이득(information gain)과 문헌 빈도(Document Frequency)처럼 몇몇 다른 자질 선택 방법을 이용할 것이다. 예를 들면, 어떤 방법은 클래스에 독립적이고 어떤 방법은 클래스에 의존적이기 때문이다. 게다가 각각의 분류기에 대한 최적의 자질 수를 정하기 위해 다양한 양의 자질을 테스트할 것이다.

감사의 글

본 연구의 일부는 2009년도 한국전자통신연구원의 위탁연구로 수행되었습니다.

참조문헌

[1] Y. Cao, J. Xu, T.Y. Liu, H. Li, Y. Huang, and H.W. Hon, Adapting ranking SVM to document retrieval", Proceedings of SIGIR-06, Seattle, USA, 2006.

[2] E. Spertus, "Somkey: Automatic recognition of hostile messages", Proceedings of the 5th International Conference on Intelligent User Interfaces, Providence, RI, U.S.A., July 27-31, pp. 1058-1065, 1997.

[3] J. Wiebe, R. Bruce, and T. O'Hara, "Development and use of a gold standard data set for subjectivity classifications", Proceedings of the ACL-99, 1999.

[4] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis", Proceedings of the HLT, pp. 347-354, 2005.

[5] B. Pang, L. Lee, and S. Vaithyanathan, "Thums up? Sentiment classification using machine learning techniques", Proceedings of EMNLP-02, pp. 79-86, 2002.

[6] P. Turney, "Thumbs Up or thumbs down? Semantic orientation applied to unsupervised classification of reviews", Proceedings of ACL, pp. 417-424, 2002

[7] V. Hatzivassiloglou, and K. McKeown. "Predicting the semantic orientation of adjectives", Proceedings of ACL-EACL, 1997.
[8] J. Wiebe, R. Bruce, and M. Bell, "Identifying collocations for recognizing opinions", Proceedings of the ACL-01 Workshop on Collocation: Computational Extraction, Analysis, and Exploitation, 2001.
[9] E. Riloff, J. Wiebe, and T. Wilson, "Learning subjective nouns using extraction pattern bootstrapping", Proceedings of the CoNLL-03 conference, 2003.
[10] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts", Proceedings of the ACL-04, Main Volume, pp. 240, 2004.
[11] G. Salton, A. Wong, and C.S. Yang, "A vector space model for automatic indexing", Proceedings of the ACM, vol. 18, nr. 11, pp. 613-620, 1975.
[12] M.E. ElAlamia, "A filter model for feature subset selection based on genetic algorithm", Department of Computer Science, Mansoura University, Mansoura 35111, Egypt, 2008.
[13] K. Fukumizu, F.R. Bach, and M.I. Jordan, "Dimensionality reduction for supervised learning with reproducing kernel hilbert spaces. JMLR, vol. 5, pp. 73 - 99, 2004.
[14] L. Song, A. Smola, A. Gretton, K. Borgwardt, and J. Bedo, "Supervised feature selection via dependence estimation", Proceedings of the 24th international conference on Machine learning, pp. 82 -830, 2007.
[15] B. Pang, and L. Lee, "Opinion mining and sentiment analysis", Foundations and Trends in Information Retrieval, vol. 2, no. 1-2, pp. 1-135, 2008.
[16] K. Nigam, and M. Hurst, "Towards a robust metric of opinion", Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications, 2004
[17] A. Kennedy, and D. Inkpen, "Sentiment classification of movie reviews using contextual valence shifters", Computational Intelligence, vol. 22, pp. 110, 2006.
[18] T. Mitchell, "Machine Learning", McGraw Hill. 1997.
[19] C.-S. Kim, and Y.-B. Kim, "Statistical Information of Korean dictionary to construct an enormous electronic dictionary", Journal of Korean Contents Society, vol. 7, no. 6, pp. 60-68, 2007,
[20] J.-H. Lee, H.-R. Park, H.-J. Park, J.-A. Ahn, and M.-H. Kim, "An effective indexing methods for hangul texts", Proceegings of the Korean Society for Information Management Conference, pp. 11-14, 1995.
[21] C.-Y. Jung, "An indexing method based on the mixed n-gram for Korean information retrieval", Master Thesis in Department of Computer Engineering, Korea Maritime University, 2004.
[22] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection", Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence, vol. 2, no. 12, pp. 1137 - 1143, 2004.
[23] C. D. Manning, P. Raghavan, and H. Schűtze, "Introduction to information retrieval", Cambridge University Press, 2008.