한국어 문서 감정 분류를 위한 감정 자질 추출 및 가중치 강화 기법에 관한 연구

지도교수 고 영 중

이 논문을 공학 석사학위
청구논문으로 제출함

2008년 12월

동아대학교 대학원

컴 퓨 터 공 학 과

황 재 원

황재원의 공학 석사학위 청구논문을 인준함

2008년 12월

위 원 장 박 경 환 인

부위원장 고 영 중 인

위 원 권 기 항 인

국문초록

한국어 문서 감정 분류를 위한 감정 자질 추출 및 가중치 강화 기법에 관한 연구

A Study on Sentiment Features Extraction and Their Weight Boosting Method for Korean Document Sentiment Classification

컴퓨터공학과 황 재 원
지 도 교 수 고 영 중

본 논문에서는 한국어 감정 분류에 기반이 되는 감정 자질의 효과적인 추출 방법과 추출된 감정 자질의 가중치 강화를 통해 한국어 문서 감정 분류의 성능 향상을 얻을 수 있는 기법을 제안한다. 한국어 감정 자질 추출은 감정을 지닌 대표적인 어휘로부터 시작하여 확장할 수 있으며, 이와 같이 추출된 감정 자질들은 문서의 감정을 분류하는데 중요한 역할을 한다. 문서 감정 분류에 핵심이 되는 감정 자질의 추출을 위해서 영어 단어 시소러스 유의어 정보를 이용하여 자질들을 확장하고, 영한사전을 이용하여 확장된 자질들을 번역하여 감정 자질들을 추출하였다. 그 후, 추출된 감정 자질을 추가 확장하고 학습 데이터를 이용하여 얻을 수 있는 감정 자질의 카이 제곱 통계량($\chi 2$ statics)값을 이용하여 각 문장의 감정 강도를 구한다. 이렇게 구한 문장의 감정 강도의 값을 TF-IDF 가중치 기법에 접목하여 감정 자질의 가중치를 강화시킨다. 마지막으로 긍정 문서에서는 긍정 감정 자질만 강화하고 부정 문서에서는 부정 감정 자질만 강화하여 학습하였다. 본 논문에서는 문서 분류에 뛰어난 성능을 보여주는 지지 벡터 기계(Support Vector Machine)를 사용하여 제안한 방법의 성능을 평가한다. 평가 결과, 일반적인 정보 검색에서 사용하는 내용어(Content Word) 기반의 자질을 사용한 경우

보다 약 2.0%의 성능 향상을 보였다.

주요어 : 감정 분류, 한국어 감정 자질, 자질 확장, 가중치 책정,
지지 벡터 기계

목 차

표 목 차

그림목차

Ⅰ. 서 론

최근에는 매일 신문이나 라디오 같은 미디어로부터 인터넷과 같은 전자 매체까지 다양한 경로에서 정보를 습득할 수 있게 되었다. 특히, 인터넷의 확산을 통해 이러한 여러 형태의 정보를 하나로 통합하여 사용자에게 제공함으로써 보다 쉽고 편리하게 정보를 얻고 활용하는 단계에 이르렀다. 이와 같이 인터넷이 폭 넓게 보급되어 온라인(on-line)상에서 얻을 수 있는 텍스트(text) 정보의 양이 급증함에 따라 이러한 거대한 텍스트 집합으로부터 의미 있는 지식을 찾아내는 작업은 많은 분야에서 매우 다양하게 요구되고 있다.

텍스트로부터 추출할 수 있는 유용한 정보 중에 하나가 작자가 해당 문서의 주제에 대해 표현한 감정 혹은 의견(sentiment or opinion)[1]이다. 예를 들어, 기업은 자신들의 상품에 대한 소비자들의 평판을 아는 것이 상품개발과 마케팅을 위한 유용한 정보로 사용될 수 있으며, 또한 영화배급사는 영화에 대한 관객들의 평판을 파악하여 개봉관의 수를 적절하게 조절할 수 있을 것이다. 전통적으로 이러한 평판은 비싼 비용을 지불하고 조사(survey)되어 왔으나, 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 담겨있는 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.

일반적인 문서 분류는 사람이 문서에 나타난 자질을 보고 인식하여 정해진 범주로 분류하는 과정을 수학적으로 모델링하여 기계가 동일한 과정으로 학습하여 문서를 분류하도록 하는 것[2]이다. 효과적인 문서 분류를 위해서

가장 중심이 되어야 하는 부분이 자질의 선정 방법[3]과 자질의 가중치 책정 방법이다. 문서 감정 분류를 위한 효과적인 감정 자질의 선정을 위해 고려해야 할 사항은 감정 분류는 문서에 나타나는 단어의 형태뿐만 아니라 단어의 의미에도 기반 해야 한다는 점이다. 감정 분류는 긍정과 부정의 감정에 초점을 두기 때문에 먼저 이를 가장 잘 표현하는 기본적인 단어인 감정 자질의 생성이 중요하다. 인간이 사용하는 말들 중엔 긍정과 감정을 나타내는 표현들이 있다. 그리고 긍정인 문서에선 긍정적인 표현이 많이 나오고, 부정인 문서에서는 부정적인 표현이 많이 나온다. 이러한 단어들을 잘 판단할 수 있다면 감정 분류를 하는데 도움이 될 것이다. 하지만, 인간이 사용하는 모든 긍정적, 부정적 표현을 다 찾아내는 일은 쉬운 일이 아니다. 그러므로, 영어권 선행 연구[4,5]를 바탕으로 대표적인 긍정, 부정을 나타내는 단어를 통해서 그 단어의 유의어를 모은다면, 각 감정을 나타내는 단어들을 모을 수 있을 것이라 판단하고 감정 자질을 추출하였다. 자질로부터 문서의 감정 분류를 위해서 사용될 충분한 양의 감정 자질을 추출하기 위해서 사전상의 유의어 및 반의어의 의미적 정보를 활용하여 단어의 의미 확장을 시도하였으나, 한국어 사전의 유의어 및 반의어의 정보가 빈약하여 충분한 양의 감정 자질을 얻을 수가 없었다. 대안으로 영어 단어 시소러스 유의어 정보를 이용하여 단어를 확장하고 이를 한영사전을 통해 번역하여 감정 자질을 추출하였다.

전통적인 자동 문서 범주화(automatic text categorization)는 미리 정의된 범주(category)에 문서를 자동으로 할당하는 기법과 관련된 연구 분야로서, 대량의 문서의 효율적인 관리 및 검색을 가능하게 하는 동시에 방대한 양의 수작업을 감소시키는 데 그 목적이 있다.

자동 문서 범주화 과정은 문서를 어떤 자질을 통해 표현할 것인가를 다루는 자질 추출(feature extraction) 과정과 추출된 자질로 표현된 문서를 어느 범주로 할당할 것인가를 결정하는 문서 분류(text classification) 과정으로 구성된다. 감정 분류 역시 이러한 자동 문서 범주화 영역에 포함되는 영역이다.

자질 추출 과정에는 추출된 자질로 어떻게 문서를 표현할 것인가에 대한 색인(indexing)과정이 포함되며, 가장 일반적인 색인 방법은 벡터 공간 모델(vector space model)이다. 이 모델은 문장의 구분 없이 전체 문서에 출현한 각 자질의 빈도수(TF)를 가지고 표현하는 방법이다. 그러나 문서 내에 나타나는 문장들 중에는 해당 문서의 감정을 잘 나타내는 문장과 그렇지 못한 문장들이 있으며, 이러한 문장 감정 강도의 차이는 각 문장에 나타나는 감정 자질(sentiment feature)의 중요도에도 영향을 미친다. 그러므로, 본 논문에서는 자질 선택(feature selection) 기법 중 하나인 카이 제곱 통계량 ($\chi 2$ statics)을 이용하여 감정 자질의 중요도를 얻고, 얻어진 카이 제곱 통계량 값을 이용하여 문장이 지닌 감정의 강도를 결정한다. 최종적으로 감정 자질이 어느 정도의 감정 강도를 지닌 문장으로부터 출현했는지를 색인 과정에 적용하고, 긍정 문서에서는 긍정 감정 자질만, 부정 문서에서는 부정 감정 자질만을 강화하여 기계 학습과정에서 감정의 긍정과 부정에 대한 특징을 더 명확하게 학습하는 이점을 얻는다.

본 논문에서는 한국어 문서의 감정을 분류하기 위한 효과적인 감정 자질 추출방법과 학습 데이터를 이용하여 얻을 수 있는 감정 자질의 카이 제곱 통계량 값을 이용하여 각 문장의 감정 강도를 구하는 방법을 제안한다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 앞서 연구된 관련 연구에 대해 살펴보고, 3장에서는 본 논문에서 제안하는 문서 감정 분류를 위한 한국어 감정 자질 추출 방법과 감정 자질들의 가중치 책정 방법에 대해 논의한다. 4장에서는 본 논문에서 제안하는 방법의 유용성을 평가하고 마지막 장에서는 결론 및 향후 과제에 대해서 기술한다.

Ⅱ. 관련 연구

문서 감정 분류는 문서 분류의 특화된 분야이기 때문에 문서 분류에서 사용되어 온 여러 가지 기계 학습 기법들이 문서 감정 분류에도 적용되어 왔다. 영화 평론과 상품 평가와 같은 특정 영역에서 나타나는 감정적 표현을 Naive Bayes, Maximum Entropy, Support Vector Machine 등의 기계 학습을 통해 문서를 긍정과 부정의 범주로 분류하는 연구가 진행되어 왔다[3,4,6,7]. 또한 분류의 대상이 문서뿐만 아니라, 문장[8,9], 구(phrase)[10,11], 토론의 연설가,[12] 그리고 문장의 감정 패턴 분석을 통해 문장의 여러 감정적 표현을 인식하고 분류하는 연구도 수행되었다[13,14].

그리고, 국내에서는 최근 많은 네티즌들이 사용하고 있는 메신저 프로그램내의 대화 내용의 감정을 파악하여 자동으로 그림말을 붙여주는 시스템과 최근 많은 문제가 되고 있는 인터넷의 악성 댓글을 판별하는 시스템[15]에 관한 연구도 진행되고 있다. 하지만, 지금까지 국내의 연구결과가 아직 기초적인 연구에 머무르고 있기 때문에 영어권 선행 연구 결과[4,5]를 바탕으로 한국어에 적용하여 감정 자질을 추출하였다.

자질 추출 과정 중 자질 선택 단계는 문서에 나타난 여러 단어들 중 범주화에 유용하게 사용될 만한 단어들을 선택하는 과정으로 문서 빈도(document frequency), 상호 정보(mutual information), 카이 제곱 통계량(χ2 static), 정보 획득량(information gain) 등의 기법이 있다.

자질 추출 과정 중 색인 단계는 선택된 자질을 통해 문서를 표현하는 단계로서, 일반적으로 벡터 공간 모델이 사용된다. 이 방법은 문서 전체에 나타난 자질들을 이용하여 문서를 하나의 벡터로 표현하는 방법으로 보통 자질의 빈도수와 역 문헌 빈도수(IDF)를 사용하여 문서를 표현한다. 그러나 이러한 기존의 방법은 문서가 가진 자질의 위치 정보나 문장 간 구분 등의 구조적 정보는 고려되지 못한다는 단점을 가진다.

이러한 한계를 극복하기 위하여 다양한 연구가 진행되었는데, 먼저 문서

의 구조적 정보를 이용하기 위해 단어의 위치나 출현한 문장의 위치에 따라 가중치를 차등 적용한 방법이 연구되었으나, 모든 문서를 두괄식 또는 미괄식으로 가정하였기 때문에 신문 기사(article) 등 형식적인 문서를 제외하곤 그 적용이 힘들다[16]. 이런 약점을 보완하기 위해 제목과 문장 간의 유사도를 이용하여 중요한 문장을 결정하여 자질의 가중치에 적용하는 연구가 수행되었다[17].

Ⅲ. 한국어 문서 감정 분류 시스템

일반적인 문서 분류 시스템에서의 자질 선정 방법은 학습 문서에서 형태소 분석을 통해 내용어(content word)를 추출하고 추출된 대상 자질에 대해 가중치를 부여하는 것이 일반적이다. 하지만 아래의 그림 1.처럼 감정 분류 시스템에서는 의미적 문서 분류를 위해서 먼저 긍정과 부정을 나타내는 어휘 즉, 감정 자질(sentiment feature)들을 따로 추출하여야 한다.

1. 한국어 감정 자질 추출

감정 자질들과 일반적인 정보검색에서 사용되는 어휘들과의 가장 큰 차이점은 정보 검색에서 사용되는 어휘들의 품사는 명사, 동사가 중요하게 사용되는 반면 감정 분류에서는 형용사, 부사 등도 중요하게 사용된다는 점이다. 이러한 감정 어휘 집합을 추출하기 위해서는 여러 가지 어휘자원들이 필요한데 외국의 연구에서는 WordNet과 같은 어휘 의미망이 많이 사용되고 있다. 한국어 감정 자질들의 확장을 위하여 한국어 사전을 파싱한 후 DB(data base)를 구축하여 동의어, 반의어 정보를 획득하고자 하였으나, 부정(13개), 긍정(12개)의 감정 자질만 획득하여, 원하는 결과를 얻을 수가 없었다. 한국어 사전에서는 어휘의 동의어와 반의어의 비중이 낮다고 판단하고 영어단어 시소러스 유의어 정보[18]를 이용하였다.

한국어 감정 자질 추출을 위하여 본 논문에서는 한국어에서 긍정과 부정을 나타내는 대표 어휘를 영어권 선행 연구 결과를 바탕으로 표 1.과 같이 대표 어휘를 선정하고 이들 단어들의 종자 어휘로 사용하여 한국어 감정 자질을 추출하고 확장한다.

표 1. 긍정/부정 영어 단어 대표 어휘

긍정	good, correct, positive, excellent, nice, fortunate, superior
부정	bad, nasty, negative, poor, unfortunate, wrong, inferior

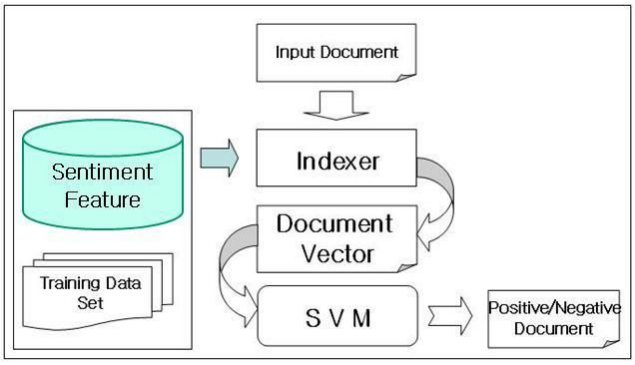

그림 1. 한국어 감정 분류 시스템 구성도

본 논문에서 생성된 자질들은 아래와 같다.

1) 감정 자질(Sentiment Feature)

선정된 대표 어휘를 대상으로 영어 단어 유의어 시소러스 정보를 이용하여 형용사, 부사를 포함한 어휘의 의미를 그림 2.와 같이 대표 유의어(예:harmful)를 순차적으로 확장하였다. 이러한 확장 방법을 각 감정 자질 목록이 더 이상 추가되는 단어가 없을 때까지 수행하여 감정 자질들을 생성하였다. 그 후, 사람이 직접 영한사전을 이용하여 적절한 한국어 감정 자질만 선정하여 긍정 감정 자질(781개)과 부정 감정 자질(1,834개)들을 생성하였다.

2) 균형 감정 자질(Balanced Sentiment Feature)

1)에서 부정 감정 자질(1834개)이 긍정 감정 자질(781개)에 비해 약 2배 많이 생성되었기 때문에, 그림 3.과 같이 영한 번역 작업 시에 출현한 한국어 자질의 횟수가 많은 자질들을 우선하여 감정 자질들을 선정하였다. 2번 이상 출현한 자질들을 선정했을 때 부정 자질의 수가 802개로 긍정 자질(781개)의 수와 거의 균형을 이루었기 때문에 2번 이상 출현한 자질들을 균형 감정 자질로 선택하였다.

최종적으로 생성된 한국어 감정 자질은 표 2.와 같다.

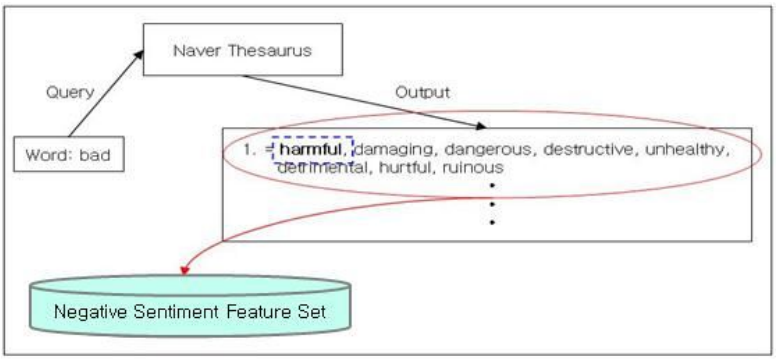

그림 2. 부정 단어(bad)의 확장 예

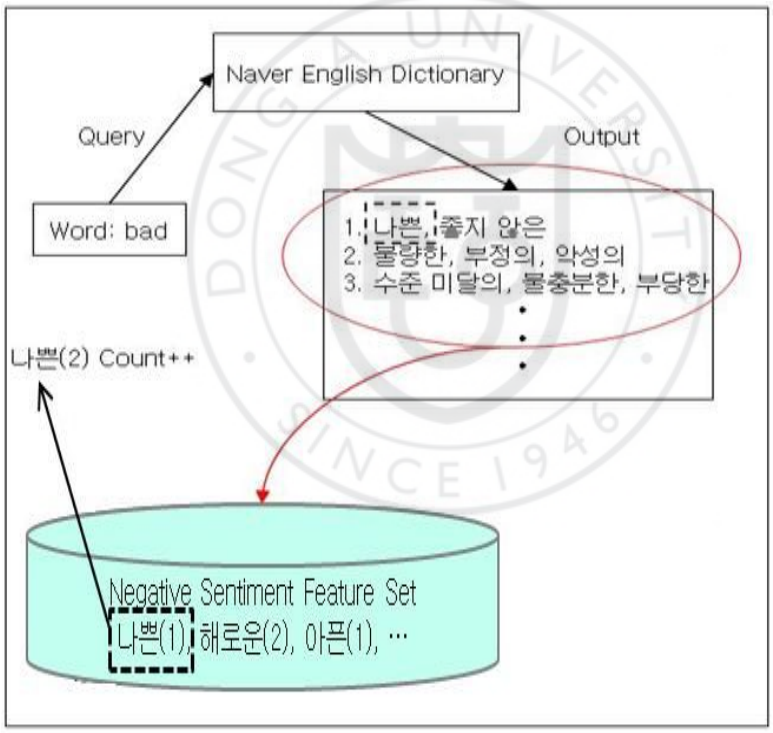

그림 3. 영한 번역시에 한국어 감정 자질 Count의 예

표 2. 생성된 한국어 감정 자질

자질 구분	내용
감정 자질	대표 어휘의 유의어 단어 집합 {긍정: 781개 / 부정: 1834개}
균형 감정 자질	감정 자질의 부정 자질을 줄인 단어 {긍정: 781개 / 부정: 802개}

2. 한국어 감정 자질 확장

앞에서 추출한 한국어 감정 자질이 부족하다고 판단하여 확장하는 작업을 수행하였다. 그 방법은 아래의 단계를 따른다.

1. 학습 문서의 형태소 분석결과 중 추출된 감정 자질 단어를 제외한 명사, 형용사, 부사, 동사의 단어를 추출
2. 한국어에 능숙한 2명의 주석자(annotator)가 각 단어의 긍정, 부정, 중립 여부를 태깅
3. 긍정 또는 부정이라고 태깅된 단어를 대상으로 우리말 국어 대사전 DB(data base)에서 반의어와 동의어 단어를 추출

이렇게 확장된 감정 자질의 수는 아래의 표와 같다.

표 3. 확장된 감정 자질

구분	Senti	Senti DB	Annotation	Annotation DB	총계
긍정	781	137	1123	134	2175
부정	802	291	2155	350	3598

표 3.에서 Senti는 3.1절에서 추출한 감정 자질이며, Senti DB는 이 추출된 감정 자질로 우리말 국어 대사전 DB에서 추출한 반의어, 동의어이다.

Annotation은 학습 문서에서 2명의 주석자가 추출한 단어이며, Annotation DB는 주석자가 추출한 단어를 질의(query)로 우리말 국어 대사전 DB에서 추출한 단어이다.

이렇게 확장한 감정 자질은 긍정이 2175개, 부정이 3598개로 부정의 감정 자질이 더 많았기 때문에 정규화를 수행하였다. 그 과정은 아래와 같다.

1. 학습 문서 내 DF가 1인 부정 Annotation 단어를 삭제
2. 삭제되지 않은 부정 Annotation 단어로 우리말 국어 대사전 DB에서 반의어와 동의어 단어를 추출

표 4. 정규화 된 부정 감정 자질

구분	N-Annotation	N-Annotation DB	총계
부정	1042	223	1265

N-Annotation은 정규화 된 부정 Annotation 단어이며, N-Annotation DB는 정규화 된 부정 Annotation DB 단어이다. 최종적으로, 긍정 감정 자질은 2175개, 부정 감정 자질은 2358개를 추출하였다. 본 논문에서 사용된 감정 자질은 이 정규화 된 감정 자질이다.

3. 한국어 감정 자질 가중치 책정

가. 문장의 감정 강도 계산

문장의 감정 강도는 직관적인 방법에 의해 계산된다. 감정을 지닌 단어는 감정 자질을 통해 쉽게 알 수 있기 때문에 감정 자질이 많이 포함된 문장일수록 감정의 강도가 강하다고 생각할 수 있다.

문장 감정 강도는 식 (1)에 의해서 구한다.

$$Strength(S_i) = 1 + \frac{S_{i_{cnt}}}{D_{\max}}, \quad (1)$$

Strength(Si)는 문장 감정 강도이며 Dmax는 문서 D내의 문장이 가장 감정 자질을 많이 가질 때 그 감정 자질의 수이다. Sicnt는 현재 문장의 감정 자질 수이다.

하지만 약한 감정을 가지는 단어를 많이 포함하고 있다고 해서 감정이 강한 문장이라고 보기 어렵기 때문에 감정 자질의 카이 제곱 통계량을 이용하여 문장 감정 강도를 구하는 방법도 사용하였다. 이 방법은 출현 횟수가 아닌 감정 자질의 카이 제곱 통계량의 수치를 합하여 문장의 감정 강도를 구하는 방법이다.

본 논문에서는 카이 제곱 통계량을 이용하여 문장 감정 강도를 구한다.

나. 카이 제곱 통계량

감정 자질의 중요도를 카이 제곱 통계량으로 결정하였다. 카이 제곱 통계량을 구하기 위한 식은 다음과 같다.

$$\chi^2(t,c) = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)}, \quad (2)$$

여기에서 A는 범주 c에 속해 있는 문서 중 용어 t를 포함하고 있는 문서의 수, B는 범주 c에 속하지 않은 문서 중 용어 t를 포함하고 있는 문서의 수, C는 범주 c에 속해 있는 문서 중 용어 t를 포함하지 않은 문서의 수, 그리고 D는 범주 c에 속하지 않은 문서 중 용어 t를 포함하지 않은 문서의 수이다.

각 범주 별로 얻어진 카이 제곱 통계량 값은 다음과 같은 식 (3)에 의해 가장 큰 값이 해당 용어의 자질 값이 되며 이 값을 감정 자질의 고유 가중치로 사용하며 문장의 감정 강도를 구하기 위해 사용된다.

$$\chi^2\max(t) = \max_{i=1}^{m}\chi^2(t,c_i), \quad (3)$$

다. 감정 자질 가중치 강화

문장의 감정 강도 계산에서 얻어진 문장의 감정 강도는 문서를 하나의 벡터로 표현할 때 감정 자질의 빈도수에 가중치를 강화하기 위해 사용된다. 문서에 출현한 자질의 빈도수는 각 문장에 출현한 자질의 빈도수의 합으로 구해진다. 이때 출현한 문장의 감정 강도에 따라 더해지는 빈도수의 수치가 달라지는데 이를 나타내는 식은 다음과 같다.

$$N(t|d) = \sum_{S_i \in d} tf(S_i,t) \times Strength(S_i), \quad (4)$$

위 식에서 tf(Si,t)는 문장 Si에서 출현한 감정 자질 t의 빈도수이며, N(t|d)은 문서 d에 출현한 문장 감정 강도에 의해 가중치가 강화된 감정 자질 t의 빈도수이다.

위 식에 따르면 각 감정 자질은 출현한 문장의 감정 강도(Strength)만큼의 가중치를 받게 되므로, 감정 강도가 강한 문장에서 나온 감정 자질은 실제로 출현한 빈도수보다 높은 값을 가지게 된다. 하지만 실제로는 모든 감정 자질의 가중치를 강화하는 것이 아닌, 각 카테고리에 해당하는 감정 자질만 강화하게 된다.

4. 문서 표현 및 지지 벡터 기계(SVM)

입력 문서를 형태소 분석[19] 후, 앞 단계에서 추출된 감정 자질을 기준으로 아래식의 TF-IDF 가중치 기법을 사용하여 가중치를 계산한다.

TF-IDF 가중치 기법은 식 (6)과 같이 문서에 어휘 t 가 나타난 어휘 빈도수(tf:term frequency) tf_t 와 역 문서 빈도수(idf:inverse document frequency, 식(5))의 곱으로 나타낸다.

$$idf_t = \log_2 \frac{N}{df_t} \ , \tag{5}$$

여기서 N은 전체 문서의 수이며, df_t는 어휘 t가 출현한 문서의 수이다.

$$weight_t = tf_t \bullet idf_t \ , \tag{6}$$

다음으로, 문서 분류기는 지지 벡터 기계를 사용하였다.

지지 벡터 기계는 두 개의 범주를 구분하는 문제를 해결하기 위해 1995년에 Vapnik에 의해 소개된 학습 기법으로 두 개의 클래스의 구성 데이터들을 가장 잘 분리해 낼 수 있는 초평면(optimal hyperplane)을 찾는 모델이다[20]. 지지 벡터 기계에서의 초평면은 식 (7)과 같이 나타낼 수 있다.

$$\vec{w} \bullet \vec{x} - b = 0 \ , \tag{7}$$

여기서 $\vec{x}$는 분류하고자 하는 문서의 벡터이며 $\vec{w}$와 b는 학습 데이터로부터 학습되어 나온 결과이다. 학습 문서 집합을 $D = \{(y_i, \vec{x_i})\}$과 같이 나타냈을 때, 각각의 학습 문서 벡터($\vec{x_i}$)가 임의의 범주에 속한 문서이면 y_i의 값에 +1을 할당하고, 범주에 속하지 않은 문서에는 -1을 할당한다. 결국 지지 벡터 기계는 식 (8)과 (9)를 만족시키는 $\vec{w}$와 b를 찾는 문제이다.

$$\vec{w} \bullet \vec{x_i} - b \geq +1 \text{ for } y_i = +1 \tag{8}$$

$$\vec{w} \cdot \vec{x_i} - b \geq -1 \text{ for } y_i = -1 \ , \qquad (9)$$

위의 수식들에 따르면 두 개의 클래스를 구분하는 초평면은 무수히 많이 존재하는데, 이들 초평면들 중에서 최적의 초평면은 두 클래스를 구분하는 거리(margin)가 최대가 되는 초평면을 정의할 수 있다. 그림 4.는 벡터를 2차원으로 표현한 한 예로서, 각 x축, y축은 자질들을 나타낸다. 실선은 두 개의 클래스를 구분하는 초평면이고, 점선은 이들 초평면들 중에서 최적의 초평면으로 두 클래스를 구분하는 거리(margin)가 최대가 되는 초평면을 나타낸다. 두 클래스를 나누는 초평면 중에서 초평면들 사이의 거리(d)가 최대인 초평면을 보여주고 있다.

지지 벡터 기계는 직선으로 나눌 수 있는 문제(linearly separable problem)에 사용되는 알고리즘이지만, 다차원의 부드러운 곡선을 이용하여 초평면을 설정하거나, 실제 데이터 벡터를 새로운 자질을 포함한 새로운 벡터 공간에 매핑하는 방법을 통해서 직선으로 나눌 수 없는 문제도 해결 할 수 있다. 지지 벡터 기계 모델을 문서 범주화에 적용되어 좋은 성능을 보여왔다[2].

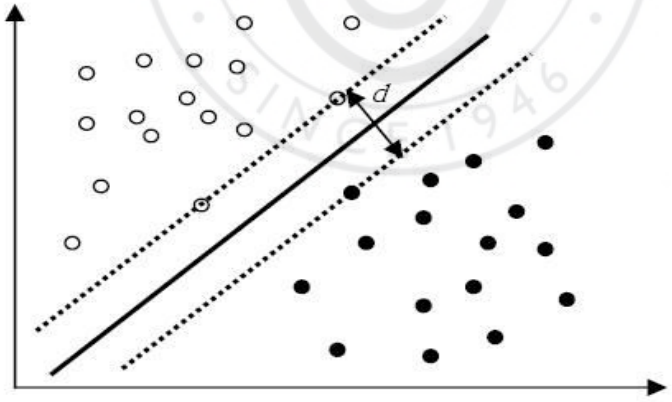

그림 4. 초평면의 거리(Margin)

Ⅳ. 실험 및 결과

1. 실험 데이터

실험에 사용된 문서 데이터는 총 2,480개의 문서이며, 3개의 분야를 나누어 수집하여 신문기사 729개, 영화리뷰 1,356개, 상품리뷰 395개의 문서로 실험하였다. 모든 문서를 사람이 직접 읽고 감정 여부를 판단하여 테스트 말뭉치를 구축하였다.

표 5.실험에 사용한 테스트 말뭉치

분야	긍정	부정	총합
신문기사	417	312	729
영화리뷰	703	653	1356
상품리뷰	205	190	395
총합	1325	1155	2480

2. 성능평가 방법

본 논문에서는 5-fold cross validation 방법으로 실험을 하였으며, 인터넷 사이트상에서 수집된 문서 집합의 평가 방법으로는 정보 검색 분야에서 일반적으로 사용되는 정확률(precision)과 재현율(recall)을 사용하였다.

정확률은 다음 식 (10)과 같이 표현된다.

$$\text{정확률} = \frac{\text{시스템에 의해 판단된 적합 문서수}}{\text{시스템이 적합하다고 판단한 문서수}} \ , \qquad (10)$$

재현율은 다음 식 (11)과 같이 표현된다.

$$\text{재현율} = \frac{\text{시스템에 의해 판단된 적합 문서수}}{\text{적합 문서수}} \ , \qquad (11)$$

정확률과 재현율을 하나의 값으로 표현해주기 위해서 다음 식 (12)와 같이 $F_1 - Measure$를 사용하였다.

$$F_1(r,p) = \frac{2 \cdot r \cdot p}{r+p} \ , \qquad (12)$$

식 (12)에서 r은 재현율에 해당하고 p는 정확률에 해당한다. 본 논문에서는 F1-Measure값으로 실험결과를 표기한다.

3. 실험 결과

실험은 실험 데이터 카테고리의 구분 없이 실험하였다. 먼저, 내용어를 사용한 실험에서는 75.31%의 결과를 얻었다. 본 논문에서는 이 결과를 기본 시스템으로 한다.

가. 확장된 감정 자질 적합성 실험

확장된 감정 자질이 적합한 자질인지를 확인하기 위하여 감정 자질이 포함된 문장만 대상으로 하여 감정 분류 실험을 하였다.

표 6. 확장된 감정 자질 적합성 실험 결과

대상 문장	모든 문장	감정자질포함문장
F1-Measure	75.31	75.36

감정 자질이 포함된 문장만을 대상으로 문서 분류 실험을 한 결과 미세하지만 나은 성능을 보였다.

나. 문장 감정 강도 계산법 비교 실험

3.3절에서 설명한 문장 감정 강도 계산 방법 중 어느 방법이 더 적합한지를 알기 위하여 비교 실험을 수행하였다.

표 7. 문장 감정 강도 계산법 비교 실험 결과

구분	출현 횟수	Sum-chi
F1-Measure	75.68	76.62

출현 횟수는 단순히 감정 자질이 출현한 횟수를 더한 방법이고, Sum-chi는 카이 제곱 통계량 값을 이용하여 문서 감정 강도를 구한 방법이다. 실험 결과 단순히 감정 자질이 많이 나왔다고 해서 높은 중요도를 부여하는 방법보다는 감정의 강도가 강한 자질이 많이 나온 문장에 더 높은 중요도를 부여하는 방법이 더 낫다는 결과를 얻었다.

다. 최종 실험 결과

제안한 방법의 실험은 아래의 단계로 수행되었다.

방법1) 모든 문장을 대상으로 감정 자질 가중치 강화
방법2) 감정 자질이 포함된 문장만을 남겨 감정 자질 가중치 강화
방법3) 방법2의 결과에 해당 카테고리별 감정 자질 강화

방법1의 실험결과는 표 8.과 같다.

표 8. 감정 자질 가중치 강화 실험 결과

구분	기본 시스템	제안한 방법	비고
F1-Measure	75.31	76.62	+1.31

감정 자질의 가중치를 문장의 감정 강도를 고려하여 강화한 방법이 내용어를 사용한 기본 시스템보다 1.31% 나은 성능을 보였다.

방법2의 실험결과는 표 9.과 같다.

표 9. 감정 자질 포함 문장만 대상으로 실험 결과

구분	기본 시스템	제안한 방법	비고
F1-Measure	75.31	76.86	+1.55

감정 자질이 포함된 문장만 대상으로 감정 자질 가중치 강화를 수행한 결과 기본 시스템보다 1.55% 성능 향상을 보였다.

방법3의 실험결과는 표 10.과 같다.

표 10. 카테고리별 감정 자질 강화 실험 결과

구분	기본 시스템	제안한 방법	비고
F1-Measure	75.31	77.23	+1.92

카테고리에 해당하는 감정 자질만을 강화한 결과 기본 시스템보다 1.92% 향상된 성능을 보였다.

최종 성능 비교는 그림 5.와 같다.

감정 자질이 포함된 문장만을 대상으로 문장 감정 강도를 고려하여 감정

자질의 가중치를 강화한 후, 해당 카테고리별 감정 자질만을 강화하여 학습한 방법3이 최대 1.92%의 성능 향상을 보였다.

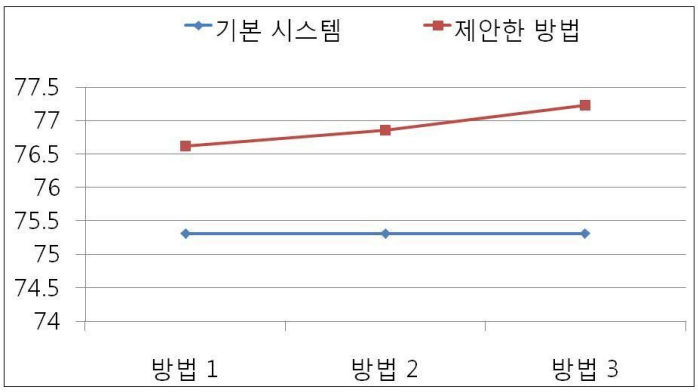

그림 5. 최종 성능 비교

V. 결론 및 향후 연구

본 논문에서는 한국어 감정 분류 시스템을 위한 효과적인 자질 추출 방법과, 문장의 감정 강도를 고려한 자질 가중치 책정 방법을 제안하였다. 한국어 문서 감정 분류를 위해서는 일반적인 정보 검색에서 사용하는 명사, 동사의 품사를 가진 내용어 뿐만 아니라, 형용사, 부사의 품사 역시 중요하며, 단지 형태소 분석을 통한 단어의 형태보다는 그 의미에 기반한 감정 자질의 생성 또한 중요하다. 그리고 모든 문장이 감정을 가진다고 볼 수 없기에 구축된 어휘 자원인 감정 자질을 이용하여 문장의 감정 포함 여부를 판단하고 감정 강도를 구하였다. 문장의 감정 강도를 효과적으로 구하기 위하여 학습 문서를 내의 감정 자질의 카이 제곱 통계량을 이용하였다. 이렇게 구해진 문서내의 문장 감정 강도의 값을 색인 과정에서 각 감정 자질의 빈도에 차등 적용하였다. 그리고 카테고리에 해당하는 감정 자질만을 강화하여 기계학습 과정에서 각 범주의 특징을 명확하게 학습하는 이점을 얻었다.

제안한 방법을 사용했을 경우, 단순히 문서 전체에 출현한 단어의 빈도수를 이용하여 문서를 표현했을 때 보다 약 1.92%의 성능 향상을 얻을 수 있었다.

향후 과제로는 감정 표현의 이중 부정에 관한 패턴을 파악하여 파악된 패턴을 적용할 수 있는 방법에 관한 연구와 문서가 아닌 문장의 감정 분류에 관한 연구도 수행할 것이다. 즉, 문서를 이루는 문장의 분류를 우선적으로 수행하여 해당 범주에 속하는 문장만을 대상으로 문서로 확장하는 방법에 관한 연구를 수행할 것이다.

참고 문헌

1) M. Rimon, "Sentiment Classification: Linguistic and Non-Linguistic Issues," Hebrew University.

2) T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many relevant Features," In Proceedings of the ECML, pp.137-142, 1998.

3) 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 Conference on Data Mining, pp.427-434, 2003.

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) A. Esuli and F. Sebastiani, "Determining the Semantic Orientation of Terms through Gloss Classification," In Proceedings of the CIKM, pp.617-624, 2005.

6) B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up? Sentiment Classification Using Machine Learning Techniques," In Proceedings of the EMNLP, pp.79-86, 2002.

7) 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.

8) B. Pang and L. Lee, "A Sentimental Education: Sentiment Analysis Using Subjectivity Summarization Based on Minimum Cuts," In Proceedings of the ACL, pp.271-278, 2004.

9) Y. Mao and G. Lebanon, "Isotonic Conditional Random Fields and

Local Sentiment Flow," In Proceedings of the NIPS, 2007.

10) P. Turney, "Thumbs up or thumbs down? Sentiment orientation applied to unsupervised classification of reviews," In Proceedings of the ACL, pp.417-424, 2002.

11) Y. Choi, C. Cardie, E. Riloff, and S. Patwardhan, "Identifying sources of opinions with conditional random fields and extraction patterns," In Proceedings of the HLT/EMNLP, pp.355-362, 2005.

12) M. Thomas, B. Pang, and L. Lee, "Get out the vote: Determining support or opposition from congressional floor-debate transcripts," In Proceedings of the EMNLP, pp.327-335, 2006.

13) A. Esuli and F. Sebastiani, "Determining the Semantic Orientation of Terms through Gloss Classification," In Proceedings of the CIKM, pp.617-624, 2005.

14) E. Riloff and J. Wiebe, “Learning extraction patterns for subjective expressions," In Proceedings of the EMNLP, pp.105-112, 2003.

15) 김묘실, 강승식, "SVM을 이용한 악성 댓글 판별 시스템의 설계 및 구현," 한글 및 한국어 정보처리, pp.285-289, 2006.

16) M. Murata, Q. Ma, K. Uchimoto, H. Ozaku, H. Isahara, and M. Utiyama, "Information Retrieval Using Location and Category Information", Journal of the Association for Natural Language Processing, Vol. 7, No. 2, 2000.

17) 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.

18) http://eedic.naver.com/list_thesaurus.naver 네이버 영어단어 유의어 시소러스

19) 강승식, 한국어 형태소 분석 및 정보 검색, 홍릉과학출판사, 2002.

20) V. Vapnik, The Nature of Statistical Learning Theory, Springer, 1995.

Abstract

Study on Sentiment Features Extraction and Their Weight Boosting Method for Korean Document Sentiment Classification

by
Jae Won Hwang

Dept. of Computer Engineering
Graduated School of Dong-A University
Busan, Korea

In this paper, we propose an effective sentiment features extraction method and boosting method of the sentiment features to improve performance in the Korean document sentiment classification. The Korean sentiment features are expanded from several representative sentiment words and they play an important role in building in an effective sentiment classification system. Firstly, the synonym information of the English word thesaurus is used to extract an effective sentiment features and then the extracted English sentiment features are translated into Korean. Secondly, the extracted sentiment features are additional extended. Thirdly, we are able to calculate chi-square statistics of the sentiment features using the training data. The sentiment intensity of each sentence from the document can be number using the obtained chi-square statistic, and this values apply to the TF-IDF weight method for the weight boosting of the sentiment features. Finally, we train that boost the positive sentiment features in the positive document and the negative sentiment features in the negative

document. In this paper, we evaluate our proposed method using support vector machine. Our experimental results show that our proposed method performs about 2.0% better than the baseline using content word based features without consider sentiment intensity of the sentence.