한국마린엔지니어링학회지 제34권 제6호, pp. 871~879, 2010. 9 (ISSN 1226-9549)
Journal of the Korean Society of Marine Engineering

뉴스 댓글의 감정 분류를 위한 자질 가중치 설정

이공주[1] · 김재훈† · 서형원[2] · 류길수[2]

(원고접수일 : 2010년 7월 1일, 원고수정일 : 2010년 8월 16일, 심사완료일 : 2010년 8월 31일)

Feature Weighting for Opinion Classification of Comments on News Articles

Kong-Joo Lee[1] · Jae-Hoon Kim† · Hyung-Won Seo[2] · Keel-Soo Rhyu[2]

요 약 : 본 논문은 뉴스 기사의 댓글에 대한 사용자의 감정을 분류하는 시스템을 제안한다. 제안된 시스템은 댓글의 문서 분류 시스템으로 기계학습에 기반을 두고 있다. 댓글은 일반적인 문서와 달리 본문을 가지고 있으며 본문의 내용이 독자의 감정에 영향을 줄 수 있다. 본 논문에서는 이와 같은 댓글의 특성과 여러 가지 자원을 이용하여 감정 분류를 위한 자질을 제안하고 이들의 가중치 설정 방법을 제안한다. 실험을 통해 이러한 가중치 설정 방법이 한글 뉴스의 댓글에 대한 감정을 분류하는데 효과적임을 알 수 있었다. 또한 댓글과 같이 많은 오류를 포함하는 문서에 대해서 문자 단위의 2음절과 3음절 자질도 충분히 이용 가치가 있음을 확인할 수 있었다. 향후에 뉴스 기사의 댓글뿐 아니라 상품 댓글 등 일반적인 감정 분석에 적용할 계획이다.

주제어 : 감정 분류, 뉴스 댓글, 감정 사전, 자질 가중치 설정

Abstract: In this paper, we present a system that classifies comments on a news article into a user opinion called a polarity (positive or negative). The system is a kind of document classification system for comments and is based on machine learning techniques like support vector machine. Unlike normal documents, comments have their body that can influence classifying their opinions as polarities. In this paper, we propose a feature weighting scheme using such characteristics of comments and several resources for opinion classification. Through our experiments, the weighting scheme have turned out to be useful for opinion classification in comments on Korean news articles. Also Korean character n-grams (bigram or trigram) have been revealed to be helpful for opinion classification in comments including lots of Internet words or typos. In the future, we will apply this scheme to opinion analysis of comments of product reviews as well as news articles.

Key words: Opinion classification, News comments, Sentiment lexicon, Feature weighting

1. 서 론

본 논문은 신문 기사에 대한 댓글의 감정을 파악하는 데 초점을 둔다. 일반적으로 댓글에는 작성자의 감정이 많이 포함되어 있다. 이 때문에 신문 기사의 댓글을 이용해서 해당 기사에 대한 시민들의 여론을 파악할 수 있다[1]. 그러나 신문 기사는 그 종류와 독자들의 관심 대상에 따라 댓글의 유무 혹은 그 양이 아주 다르므로 이들을 수집해서 분석하

† 교신저자(한국해양대학교 컴퓨터공학과, E-mail:jhoon@hhu.ac.kr, Tel: 051-410-4574)
1 충남대학교 정보통신공학과
2 한국해양대학교 컴퓨터공학과

는 일이 참으로 인력과 많은 노력이 필요한 일이다. 이를 돕기 위해 자동으로 댓글을 수집하지만 다음과 같은 댓글의 특성 때문에 분석이 쉽지 않다.

댓글의 길이가 매우 짧다. 대부분 한 두 문장으로 구성되어 있다.

댓글에는 문법에 어긋난 표현 혹은 오탈자가 많이 포함되어 있다. 특히 댓글에는 은어 표현(인터넷 언어, 채팅 용어 등)이 많이 사용된다.

댓글에는 객관적 사실보다 주관적인 내용이 많이 포함되어 있다.

정치나 스포츠 그리고 연예 등과 관련된 분야의 기사에 많은 댓글이 포함되어 있다(**표 1** 참조).

본 논문은 이러한 댓글의 특성을 고려하여 자동으로 웹 사이트에 있는 신문기사의 댓글을 수집하여 독자들의 감정(2가지의 극성인 긍정 혹은 부정)을 분류하는 시스템을 제안한다. 제안된 시스템은 자질 추출 단계, 자질 가중치 계산 단계, 감정 분류 단계로 나뉜다. 자질 추출 단계에서는 신문 기사의 댓글로부터 자질을 추출하는 단계이며 기본적인 자질은 문자 단위의 n-gram이다. 자질 가중치 계산 단계에서는 각 자질의 가중치를 계산하며 기본적인 가중치로 TF-IDF(Term Frequency - Inverse Document Frequency)[2]를 사용한다. 본 논문에서는 시스템의 성능을 개선하기 위해 감정 등에 관련된 자질에 대해서 기본 가중치를 재조정하는 방법을 제안한다(3장 참조). 마지막으로 감정 분류 단계에서는 이전 두 단계에서 구해진 자질 벡터를 이용하여 감정을 긍정 혹은 부정으로 분류한다.

본 논문의 구성은 다음과 같다. 2장에서는 신문기사와 관련한 감정 분류 시스템에 대한 사전 연구에 대해서 기술하고, 3장에서는 제안된 시스템을 자세히 설명한다. 4장에서 제안된 자질의 가중치 재조정 방법이 감정 분류 시스템에 미치는 영향을 살펴본다. 마지막으로 5장에서 결론을 맺고 앞으로의 연구과제에 대해서 기술한다.

2. 관련 연구

최근 인터넷 문화가 발전하고 컴퓨터의 기술이 발전하면서 다양한 형식의 문서를 통해서 개인의 의견이나 감정을 표현한다. 이와 같이 문서에 표현된 다양한 형태의 감정을 찾아내는 방법을 감정 분류이라고 한다[3]. 사실 감정 인식은 매우 다양한 분야에 응용될 수 있다. 예를 들면 여론 분석[4], 쓰레기 편지 여과[5], 논평(상품평, 영화 평론 등) 분석[6] 등이 있다. 본 논문은 댓글을 하나의 문서로 간주하고 주어진 댓글(문서)에 대한 감정을 분류하여 신문 기사의 댓글에 대한 독자들의 의견 혹은 감정을 파악하는 것이므로 이에 관련된 사전 연구들에 대해서 살펴보고자 한다. 먼저 [1]에서는 신문 기사나 블로그(blog)에서 최근 사건에 대해서 어떤 감정이 있는지를 분석하고 다른 사건과 비교하여 이 사건의 상대적인 감정 점수를 구하는 시스템을 제안하였다. 또한 다양한 계층에 따른 감정의 차이를 분석하여 각 계층의 감정 변화를 분석할 수 있었다. [7]에서는 독자의 감정들에 따라 신문 기사를 분류하였다. 이 연구가 본 논문의 목적과 매우 유사하다. 그러나 [7]은 신문 기사 자체를 분류하였으나 본 연구는 신문 기사의 댓글을 분류하고 있다. 또한 [7]은 중국 신문 기사에 대해서 8개의 감정[1]으로 분류하였다. 이와 같이 자세한 감정을 분류할 수만 있다면 이렇게 하는 것이 바람직하다. 그러나 이와 같은 자세한 감정을 분류하는 작업은 전문가도 쉽지 않은 일이다. [8]은 어휘를 기반으로 금융 뉴스 문서에 대해 감정을 분류하였다. 이 연구도 본 논문과 유사하다고 할 수 있으나 특정한 분야에 집중되었으며 사용된 자질은 어휘와 그 어휘들의 결합 정도를 이용하고 있다. [9]는 정치기사에 대한 감정을 분석하고 평가하는 프레임워크를 만들어 미국 대통령 G. W. Bush와 이란 대통령 M. Ahmadinejad에 관련된 기사에 적용하였다.

3. 가중치 재조정 기반 감정 분류 시스템

그림 1은 한글 뉴스 기사의 댓글에 대한 감정 분류 시스템의 구성도이며 전체 시스템의 구성은

1) 'happy', 'angry', 'sad', 'surprised', 'heartwarming', 'awesome', 'bored', 'useful'

[10]과 같다. 본 논문은 [10]의 결과를 토대로 확장하고 개선한 것이다. 제안된 시스템은 웹 뉴스 기사(Web news document)를 입력으로 받아 그 기사에 포함된 각각의 댓글(comment)에 대한 감정을 출력한다. 입력에 해당하는 웹 뉴스 기사는 하나의 본문(news article)과 여러 개의 댓글(news comment)로 구성된다. 출력에 해당하는 감정(sentiment polarity)은 긍정과 부정으로만 간주한다. 전체 시스템은 크게 세 단계로 구성된다. 첫 번째 단계는 자질 추출(feature extraction) 단계이고 두 번째 단계는 자질 가중치 계산(feature weighting) 단계이며, 마지막 단계는 감정 분류(opinion classification) 단계이다. 이하의 절에서는 각 단계에 대해서 자세히 설명할 것이다.

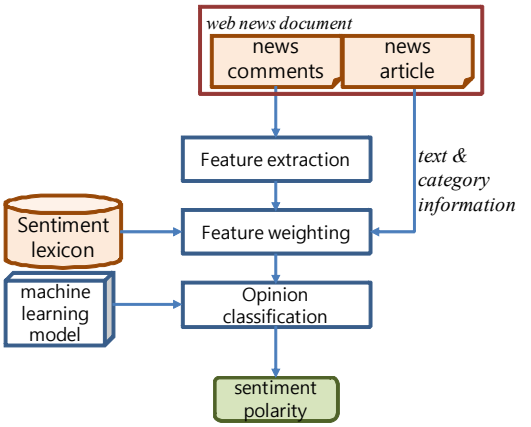


그림 1: 신문 기사의 댓글에 대한 감정 분류 시스템의 구성도

3.1 자질 추출

영어의 경우에는 단어가 문서나 감정 분류의 좋은 자질이 될 수 있다[11-12]. 그러나 한국어에서는 어형의 변화가 심한 교착어이므로 단어 그대로를 자질로 사용할 수 없다. 더구나 댓글과 같은 인터넷 문서는 문법에 맞지 않거나 유행어, 비속어, 은어 등과 같은 말들이 매우 자주 사용된다. 따라서 한국어 형태소 분석기를 비롯한 대부분의 한국어 언어처리기들이 댓글과 같은 문서에 대해서는 올바르게 분석할 수 없다. 이런 이유 때문에 본 논문에서는 단어 대신에 문자 단위의 2-그램(bigram)이나 3-그램(trigram)을 이용한다. 한국어에 대해서 이처럼 n-그램으로 문서를 분류할 수 있는 이유를 살펴보면 다음과 같다.

한국어 단어의 약 80%가 2음절 혹은 3음절로 구성되었다[13].

2음절이나 3음절이 정보검색 분야에서 좋은 자질들로 사용되어 왔다[14-15].

대부분의 댓글은 형태학적으로 분석하기 어려운 유행어, 비속어, 은어 등과 같은 단어가 매우 빈번히 사용된다.

따라서 본 논문에서는 자질로 2-그램 혹은 3-그램을 사용하고 입력 문서인 뉴스 기사와 모든 댓글을 2-그램과 3-그램으로 분리함으로써 자질을 추출한다. 분리된 자질들 중에서 빈도가 높은 자질을 불용어로 간주하여 이를 제거한다.

3.2 자질 가중치 계산

벡터 공간 모델은 정보검색뿐 아니라 감정 분류 분야에서도 널리 사용되어 왔다[3]. 일반적으로 가중치를 계산하는 방법으로 가장 널리 알려진 것은 TF-IDF이며 식 (1)과 같이 정의된다[2].

$$w_{ij} = tf_{ij} \times idf_i \quad (1)$$

여기서 tf_{ij}는 용어빈도(term frequency)로서 자질 t_i가 문서 d_j에 미치는 정도를 나타내고, idf_i는 역문헌빈도(inverse document frequency)로서 자질 t_i가 전체 문서에 미치는 정도이다[2]. 본 논문에서는 식(1)과 같은 가중치를 기반으로 감정과 관련된 자질에 대해 가중치 w_{ij}를 강화하는 새로운 가중치 계산법을 제안한다.

어떤 단어들이 감정을 전달하는가? 사실 이는 간단하게 말할 수 있는 문제는 아니나 많은 연구[3,16]에서는 '좋다(good)', '나쁘다(bad)', '아름답다(beautiful)' 등과 같은 단어가 감정과 관련되

2) $idf_i = \log(\frac{N}{df_i})$이며 N은 총 문서 수이다. tf_{ij}와 idf_i를 구하는 구체적인 방법은 [2]를 참조하기 바란다.

었다고 한다. 본 논문에서는 이와 같은 단어들을 선별하여 감정 사전(sentiment lexicon)을 구축하여 식 (2)과 같은 방법으로 가중치를 조정하며 식 (1)에서 idf_i는 그대로 사용하고 t_{ij}만 조정한다.

$$tf_{ij}^{(1)} = \begin{cases} tf_{ij}^{(0)} + 1 & \text{if } t_i \in S \cap C_j \\ tf_{ij}^{(0)} & \text{otherwise} \end{cases} \quad (2)$$

여기서 $tf_{ij}^{(0)}$는 식 (1)의 tf_{ij}이며, S는 감정 사전이고 C_j는 j번째 댓글이다. 즉 자질 t_i가 감정 사전에 포함되면 그 자질의 빈도에 1을 더한다.

어떤 댓글은 원기사(original article)와 전혀 무관한 스팸 댓글(spam comment)일 경우가 종종 있다. 본 논문에서는 원기사와 전혀 관련이 없는 댓글에 대한 가중치가 낮아지도록 식 (3)과 같이 가중치를 재조정한다.

$$tf_{ij}^{(2)} = \begin{cases} tf_{ij}^{(1)} + 1 & \text{if } t_i \in T \cap C_j \\ tf_{ij}^{(1)} & \text{otherwise} \end{cases} \quad (3)$$

여기서 T는 원기사의 자질 집합을 의미한다. 식 (3)은 감정사전 S과 원기사 T 그리고 댓글 C_j에 모두 포함되어 있는 자질에 대해서 가중치를 재강화한다.

표 1은 원기사의 분야에 따른 댓글의 감정 분포를 보이고 있다. **표 1**에서 보는 바와 같이 정치 분야의 댓글은 대부분 부정적이나 인물 분야의 댓글은 반반 정도이다. 이처럼 댓글은 원기사의 분야에 따라 매우 다른 분포를 가지므로 감정 분류에는 원기사의 분야 정보가 중요한 단서가 될 수 있다. 웹 뉴스 기사에는 대부분 분야 정보를 메타 데이터(metadata)로 포함하고 있다. 따라서 웹 뉴스 기사로부터 분야 정보를 추출하는 작업은 어려지 않다. 본 논문에서는 각 분야에 따른 단어 집합을 정의하여[17], 각 분야 단어에 속하는 자질에 대해서 식 (4)와 같이 가중치를 강화한다. 예를 들어 정치 분야 단어에는 '정당', '국회' 등이 포함되고 경제 분야 단어에는 '환율', '경제', '은행' 등이 포함된다.

$$tf_{ij}^{(3)} = \begin{cases} tf_{ij}^{(2)} + \alpha & \text{if } t_i \in F_c \\ tf_{ij}^{(2)} & \text{otherwise} \end{cases} \quad (4)$$

표 1: 원기사의 분야에 따른 댓글의 감정 분포

분야	긍정 댓글 수	부정 댓글 수
정치	306	4,108
사회	234	1,851
국제	37	262
경제	118	867
문화	54	867
IT	7	20
사설	32	85
인물	17	16
합계	**805**	**8,076**

여기서 α는 매개변수이나 4장에서 기술할 실험에서 $|S \cap C_j|/2$로 설정한다. 즉 j번째 댓글 C_j에 속하는 감정 단어 개수의 절반에 해당하는 값이다. 그리고 F_c는 원기사의 분야 단어 집합이다. 최종적으로 본 논문에서 식 (1)의 t_{ij}대신에 $tf_{ij}^{(3)}$을 사용한다.

3.2.1 자질 가중치 계산의 예

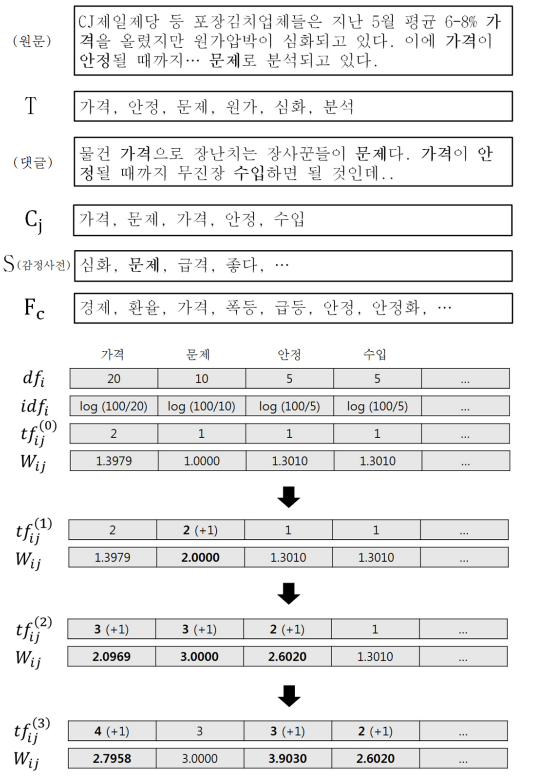


그림 2: 자질 가중치 계산의 구체적인 예

그림 2는 독자들의 이해를 돕기 위해서 자질 가중치의 계산 과정을 보이고 있다. 이 예제는 실험 말뭉치(4장 참조)에서 추출한 원문과 댓글을 사용하고 있으나 지면관계로 일부 생략하였으나 빈도수는 이해를 돕기 위해 가정하였으며 총 댓글 수 N은 100으로 가정하였다. 그림에서 T는 원기사에서 추출한 자질집합이고 C_j는 j번째 댓글에서 추출한 자질집합이다. $tf_{ij}^{(1)}$에서는 자질 '문제'가 감정사전 S에도 속하므로 해당 자질의 가중치가 강화되었다. $tf_{ij}^{(2)}$에서는 자질 '가격', '문제', '안정'의 가중치가 원기사에 속하므로 해당 자질의 가중치가 강화되었다. $tf_{ij}^{(3)}$에서는 자질 '가격', '안정', '수입'에 F_c(원기사가 경제 분야임)에 속하므로 해당 자질의 가중치가 강화되었다. 여기서 편의상 식(4)의 α는 1로 간주하였다. 결과적으로 자질 '가격', '문제', '안정'은 그 가중치가 2번 강화되었고 자질 '수입'의 가중치는 1번 강화되었다.

3.3 감정 분류

본 논문에서는 SVM(Support Vector Machine) 등과 같은 기계학습 방법을 이용해서 감정을 분류한다. 본 논문에서 제안된 감정 분류 단계의 입력은 일반적인 기계학습 기반 분류 시스템과 같은 자질 집합(feature set)이고 출력은 긍정과 부정 이진값으로 표현된다. 자질 집합은 3.1절과 3.2절을 통해서 얻어진 자질 집합을 이용할 수 있으나 일반적으로 이와 같은 자질 집합이 너무 크므로 감정 분류에 불필요한 자질을 제거한다[18]. 이처럼 자질 선정 과정을 통해서 최종적으로 기계학습 알고리즘에 사용될 자질 벡터를 선정한다.

4. 실험 및 토의

이 장에서는 제안된 시스템을 평가하고 그 결과를 논의한다. 이를 위해 먼저 실험 자료(실험 말뭉치, 감정 사전) 및 실험 환경에 대해서 자세히 기술한다.

4.1 실험 자료

제안된 시스템을 시험하기 위해서는 실험 말뭉치로서 감정 말뭉치(sentiment corpus)와 감정 사전(sentiment dictionary)이 필요하다. 이 두 자료 모두 한국어 영역에는 공개된 자료가 없으므로 본 연구진에 의해서 직접 구축되었다.

먼저 한국어 감정 말뭉치 구축 방법에 대해서 살펴보자. 한국어 감정 말뭉치를 구축하기 위해 특정 신문사 웹 사이트[3]로부터 1,377개의 신문 기사와 그에 포함된 댓글들과 분야 정보(정치, 스포츠, IT 등)를 수집하여 모든 댓글에 대하여 수동으로 긍정 혹은 부정 댓글인지를 표시했다. 그 결과는 **표 2**와 같다. 전체 댓글은 8,320개이며, 뉴스 기사 당 평균 6개의 댓글이 포함되어 있다. 각 댓글은 평균 33개의 자질(단어)로 구성되었다. 한국어 문장의 평균 길이가 11임[19]을 감안하면 댓글은 평균 3문장으로 구성된 짧은 문서임을 알 수 있다. **표 1**에서 볼 수 있듯이 본 논문에서 사용된 분야는 정치, 사회, 국제, 경제, 문화, IT, 사설, 인물 분야이다.

표 5: 구축된 한국어 감정 말뭉치의 통계치

	신문 기사	댓글
개수	1,377	8,320
평균 댓글 수	-	6
자질 수	863,379	274,626
평균 자질 수	627	33

다음은 한국어 감정 사전 구축 방법에 대해서 살펴보자. 한국어 감정 사전도 공개되지 않았으므로 본 연구진이 직접 구축하였다. 이를 위해 한국어 사전으로부터 감정 단어를 추출하는 것은 많은 시간이 필요한 작업이므로 본 논문에서는 이미 공개된 영어 감정사전[20]을 사용하여 한국어 감정 사전을 구축하였다. 영어 감정사전에는 2,297개의 긍정 단어와 4,138개의 부정 단어로 구성되었다. 본 연구에서는 이들 단어에 대해서 영한 사전(English-Korean Dictionary)을 이용해서 각 단어를 번역하고 번역된 한국어 단어에 대해서 동의어와 반의어를 이용해 반자동적으로 확장하였다. 최종적으로 중복되거나 감정 단어로 부적합 단어를 제거함으로써 한국어 감정 사전을 구축하였다. 구

축된 한국어 감정 사전은 3,044개의 긍정 단어와 4,046개의 부정 단어로 구성되었고 명사, 형용사, 부사를 포함한다.

4.2 실험 환경

4.1절에서 구축된 감정 말뭉치가 실험을 위해서 충분하지 않으므로 n-교차 검증법(cross validation)[2]과 거시 평균(macro-average)[2]을 사용한다. 본 논문에서는 n은 4이다. 즉 학습 말뭉치(training corpus)의 크기는 6,240이고 검증 말뭉치(test corpus)의 크기는 2,080이다(**표 3** 참조). 평가 방법은 F_1 측도[2]를 사용한다. 기계학습 모델은 SVM(Support Vector Machine), NB(Naïve Bayesian), kNN(k-Nearest Neighbor)를 사용하며 이를 위한 기계학습 도구로서 Perl 모듈인AI::Categorizer[4]를 사용한다. 자질 선정 방법으로는 χ^2(Chi-square)[2]을 사용한다.

4.3 성능 평가

본 논문의 실험 목적은 세 가지이다. 첫째는 한국어 감정 분류에 적합한 기계학습 모델이 무엇인가를 실험으로 살펴보는 것이고, 둘째는 제안된 자질(감정 사전, 원기사, 분야 정보)과 그 가중치 조정 방법이 댓글 감정 분류에 얼마나 영향을 미치는가를 살피는 것이고, 마지막으로 문자 단위의 2음절 및 3음절 자질이 댓글 감정 분류에 미치는 정도를 확인하는 것이다.

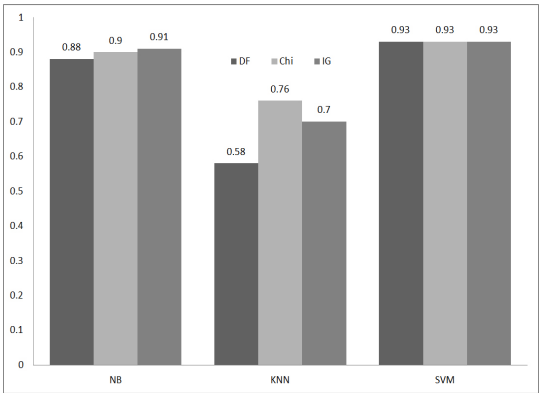


그림 3: 기계학습 방법에 따른 성능 평가

그림 3은 한국어 감정 분류에 적합한 기계학습 방법들(NB, kNN, SVM)을 결정하기 위해 가중치 재조정 없이 기본적인 댓글 정보만으로 각 기계학습 방법의 성능을 비교해보았다. **그림 3**에서 볼 수 있듯이 NB와 SVM은 거의 비슷한 결과를 보였으나, SVM에 약간 우세하므로 이하의 실험에서는 모두 SVM을 사용한다. 또한 자질 선택 방법은 문헌 빈도(document frequency, DF), χ^2(Chi-square, Chi), 정보 이득(information gain, IG)에 따라서 비교해 보았으나 대부분의 경우 거의 비슷한 성능을 보였으며 kNN의 경우 Chi가 우세하였다. 따라서 이하의 모든 실험에서 Chi를 사용할 것이다.

그림 4는 가중치 재조정에 따른 한국어 감정 분류 시스템의 성능을 분석한 그래프이다. X축은 가중치 재조정 방법($tf_{ij}^{(0)}$: 댓글의 n-그램 자질의 TF-IDF, $tf_{ij}^{(1)}$: 감정 사전을 이용한 가중치 조정 방법, $tf_{ij}^{(2)}$: 원기사를 이용한 가중치 재조정 방법, $tf_{ij}^{(3)}$: 원기사의 분야 정보를 이용한 가중치 재조정 방법)이고, Y축은 F_1 측도를 나타낸다. 댓글 정보만을 이용했을 때, 시스템 성능은 2음절과 3음절에 따라 각각 91.8%와 92.7%의 F_1 점수를 보였고 제안된 모든 자질을 이용해서 가중치를 재조정했을 때, 시스템 성능은 각각 96.1%(2음절)와 95.7% (3음절)를 보였다. 본 논문에서 제안한 가중치 조절 방법이 댓글의 감정 인식 영역에서 유용함을 알 수 있었다.

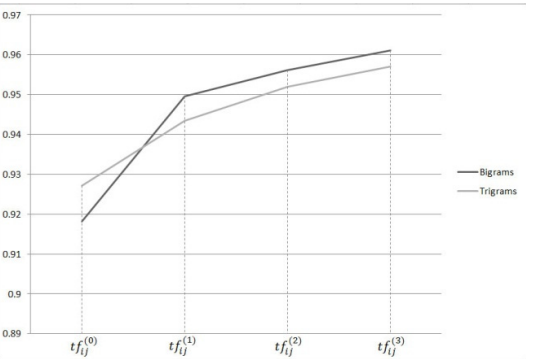


그림 4: 제안된 자질의 가중치 조정에 따른 성능 분석(자질 선택 방법: χ^2, 기계학습 모델: SVM, 자질 벡터의 크기: 300, 평가 측도: F_1 측도)

또한 **그림 4**의 결과는 한글 2음절이나 3음절 자질이 감정 인식 영역에 유용하게 사용될 수 있음을 동시에 알 수 있었다. 이 결과는 실험 환경(언어, 학습 말뭉치, 기계학습 모델, 자질 선택 방법 등)이 달라서 정확하게 비교할 수는 없지만 [20]의 성능보다는 좋은 결과를 보였다.

5. 결론 및 향후 과제

본 논문에서는 한국어 신문 기사의 댓글에 대한 감정 분류 시스템을 제안하였다. 특히 본 논문에서는 댓글에 대한 감정 인식에 적합한 자질로 감정 사전, 뉴스 본문, 그리고 본문의 범주 정보를 추가하였으며 또한 이들의 가중치 조정 방법을 제안하였다. 제안된 가중치 조절 방법은 실험을 통해서 감정 인식 영역에서 유용함을 알 수 있었다. 또한 댓글 등에 자주 등장하는 인터넷 언어 문제를 해결하기 위해 사용된 문자 단위 2음절과 3음절 자질도 감정 인식에 적절함을 알 수 있었다.

향후에 뉴스 기사의 댓글뿐 아니라 상품 댓글 등 일반적인 감정 분석에 적용할 계획이며 또한 한국어 감정 사전과 한국어 감정 말뭉치에 대한 좀더 체계적인 연구가 필요할 것이다.

감사의 글

이 연구는 2009년도 충남대학교 학술연구비에 의해 지원되었음.

참고문헌

[1] N. Godbole, M. Srinivasaiah, and S. Skiena, "Large-Scale Sentiment Analysis for News and Blogs", Proceedings of International Conference on Weblogs and Social Media, 2007.

[2] C. D. Manning, P. Raghavan and H. Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008.

[3] 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.

[4] R. Tong, "An operational system for detecting and tracking opinions in on-line discussions", Working Notes of the SIGIR Workshop on Operational Text Classification, pp. 1 - 6, 2001.

[5] E. Spertus, "Somkey: Automatic recognition of hostile messages", Proceedings of the 5th International Conference on Intelligent User Interfaces, pp. 1058-1065, 1997.

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

[7] K. Lin, C. Yang and H.-H. Chen, "What emotions do news articles trigger in their readers?", Proceedings of SIGIR, pp. 733-734, 2007.

[8] A. Devitt and K. Ahmad, "Sentiment polarity identification in financial news: A cohesion-based approach", Proceedings of the Annual Meeting of the Association of Computational Linguistics, pp. 984-991, 2007.

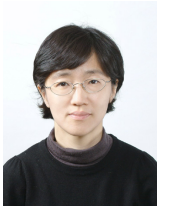
[9] A. Nourbakhsh, C. Khoo, and J.-C. Na, "A framework for sentiment analysis of political news articles", Proceedings of the International Communication Association Conference, 2008.

[10] J.-H. Kim, K. J. Lee, H.-W. Seo, and H.-C. Kim, "Opinion mining for comments on news articles on the Web", Proceedings of the International Conference on Internet, pp. 63-68, 2009.

[11] C. M. Tan and C. D. Lee, "The use of

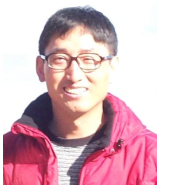
bigram to enhance text categorization", International Journal of Information Processing & Management, pp. 529-546, 2001.
[12] T. Bekkerman and J. Allan, Using Bigrams in Text Categorization, CIIT Technical Report IR-408, 2004.
[13] C. Kim and Y. Kim, "Statistical information of Korean dictionary to construct an enormous electronic dictionary", The Journal of Korean Contents Society, vol. 7, no. 6, pp. 60-68, 2007.
[14] J. Lee, H. Park, J. Ahn and M. Kim, "An effective indexing methods for Korean text", Proceedings of the Korean Society for Information Management Conference, pp. 11-14, 1995.
[15] C. Jung, An Indexing Method Based on the Mixed n-gram for Korean Information Retrieval, Master Thesis, Department of Computer Engineering, Korea Maritime University, 2004.
[16] D. Rao and D. Ravichandran, "Semi-supervised polarity lexicon induction", Proceedings of the 12th Conference of the European Chapter of the ACL, pp. 675 - 682, 2009.
[17] A. Zheng and R Srihari, "Optimally combining positive and negative features for text categorization", Proceedings of the ICML Workshop on Learning from Imbalanced Datasets, 2003.
[18] Z. Zheng, X. Wu, and R. Srihari, "Feature selection for text categorization on imbalanced data", ACM SIGKDD Explorations Newsletter, vol. 6, no. 1, pp. 80-89, 2004.
[19] 홍진표, 차정원, "TextRank 알고리즘을 이용한 한국어 중요 문장 추출", 한국정보과학회 2009 한국컴퓨터종합학술대회 발표논문집, 제36권, 제1호(C), pp. 311-314, 2009.
[20] T. Wilson, J. Wiebe and P. Hoffmann, "Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis", Computational Linguistics, vol. 35, no. 3, pp. 399-433, 2009.

저 자 소 개



이공주(李公主)

1969년생, 1992년 서강대학교 전자계산학과(학사), 1994년 한국과학기술원 전산학과(공학석사), 1998년 한국과학기술원 전산학과(공학박사), 1998년 - 2003년 한국마이크로소프트(유) 연구원, 2003년 이화여자대학교 컴퓨터학과 대우전임강사, 2004년 경인여자대학 전산정보과 전임강사, 2005년 - 현재 충남대학교 전기정보통신공학부 부교수, 관심분야: 자연언어처리, 자연어인터페이스, 기계번역, 정보검색



김재훈(金載熏)

1964년생, 1986년 계명대학교 전자계산학과학사, 1988년 한국과학기술원 전산학과 공학석사, 1996년 한국과학기술원 전산학과 공학박사, 1988년~ 1997년 한국전자통신연구원 선임연구원, 2001년~2002년 Information Sciences Institute in University of Southern California 방문연구원, 2007년~2008년 Beckman Institute in University of Illinois at Urbana-Champaign 방문연구원, 현재 한국해양대학교 컴퓨터공학과 교수, 한국마린엔지니어링학회 편집이사, 한국정보과학회 편집위원, 관심분야: 자연언어처리, 한국어정보처리, 정보검색, 정보추출.

서형원(徐炯源)

1982년생, 2006년 부산외국어대학교 전자컴퓨터공학부학사, 2010년 한국해양대학교 컴퓨터공학과 공학석사, 현재 한국해양대학교 컴퓨터공학과 공학박사 과정, 관심분야: 자연언어처리, 한국어정보처리, 정보검색, 감정인식, 소셜네트웍분석.

류길수 (柳吉洙)

1953년5월생, 1976년 한국해양대학교 기관학과 졸업, 1979년 동대학교 대학원 졸업(석사), 1986년 일본동경공업대학 대학원 졸업(석사), 1989년 동대학원 졸업(박사), 1976-1982년 기관사 승선근무, 1982년-현재 한국해양대학교 IT공학부 교수, 당학회 종신회원.