韓國컴퓨터情報學會 論文誌
第14卷 第5號, 2009. 5.

2009-14-5-1-2

채팅 데이터의 기분 분류 시스템

윤 영 미*, 이 영 호**

Emotion Classification System for Chatting Data

Youngmi Yoon*, Young-Ho Lee**

요 약

온라인 커뮤니케이션 중 인터넷 메신저를 이용한 대화의 비중이 점점 증가하는 추세이나, 이러한 메신저 대화 정보를 효율적으로 재사용할 수 있는 어플리케이션은 많지 않다. 메신저 대화 정보는 사용자의 언어 습관이 반영된다는 특성을 가진다. 이러한 언어 습관은 자주 쓰이는 단어나 이모티콘으로 나타나며, 이들로써 사용자의 기분을 잘 파악할 수 있다는 특성이 있다. 그러므로 본 연구에서는 자주 쓰이는 단어들이나, 기호 등을 이용해서 효과적으로 대화 내용 작성자의 기분 등을 분류할 수 있는 기법을 제안한다. 이러한 기법은 충분한 반복 실험을 통해서 95% 이상의 높은 정확성으로 기분을 분류할 수 있음을 보여주고 있다.

Abstract

It's a trend that the proportion of using an internet messenger among on-line communication methods is getting increased. However, there are not many applications which efficiently utilize these messenger communication data. Messenger communication data have specific characteristics that reflect the user's linguistic habits. The linguistic habits are revealed through frequently used words and emoticons, and user's emotions can be grasped by these. This paper proposes the method that efficiently classifies the emotions of a messenger user using frequently used words or symbols. The emotion classifier from repeated experiments achieves high accuracy of more than 95%.

▸ Keyword : 분류 분석(classification), 기분 분류(emotion classification), 기분 분석(emotion analysis, sentiment analysis), 채팅 데이터 분석(chatting data analysis)

• 제1저자 : 윤영미 교신저자 : 이영호
• 투고일 : 2009. 04. 22, 심사일 : 2009. 05. 14, 게재확정일 : 2009. 05. 19.
* 가천의과학대학교 정보공학부 교수 ** 가천의과학대학교 정보공학부 교수

Ⅰ. 서 론

인터넷의 발달과 함께, 많은 오프라인 커뮤니케이션이 온라인 커뮤니케이션으로 대체되고 있다. 그리고 이러한 온라인 커뮤니케이션에서는 채팅 서버를 통한 다대일 커뮤니케이션보다 인터넷 메신저를 통한 일대일 커뮤니케이션의 비중이 높아지고 있다. 인터넷 메신저가 인기가 있는 이유로 그 편리성 및 모바일 기기의 증가 등을 뽑을 수 있으나, 메신저가 유용한 이유는 단지 편리하거나, 어디서나 사용할 수 있기 때문만은 아니다. 메신저의 유용성은, 첫째로 메신저를 통해서 교류되는 정보가 모두 디지털화 될 수 있으며, 둘째로 대화 상대가 오프라인 관계에 기반 하는 경우가 많고, 대화의 내용이 서버나 기타 경유지를 거치지 않으므로 채팅 서버를 이용한 대화보다 사생활 보호의 정도가 상대적으로 높다는 것이다.

그러므로 메신저를 통해서 교류되는 정보는 인터넷 및 컴퓨터의 보급률과 비례하여 그 양이 점점 증가하고, 사생활이 어느 정도 보호된다는 것을 아는 상태로 대화가 이루어지며, 손쉽게 디지털화되어 저장될 수 있다는 특성을 가진다. 이러한 특성은 업무적인 정보뿐만 아니라 일상생활에서 쉽게 놓칠 수 있는 수많은 정보를 사생활 침해 우려 없이 손쉽게 저장할 수 있게 해준다는 이점을 제공한다. 하지만, 이렇게 축적된 유용한 대화 정보를 이용하는 면에서는 아직 어플리케이션이 부족한 실정이다. 그 주된 이유는 대화 정보가 자연어로 되어 있고, 이러한 자연어의 처리에는 많은 난점이 있기 때문이다.

정보가 자연어로 구성될 경우, 이러한 정보에 대한 검색 작업이나 주제 분류 작업은 확실히 그 정확성과 효율이 떨어진다. 하지만 인터넷 메신저 대화 정보는 위에서 언급한 세 가지 특성 외에도 사용자의 언어 습관(자주 쓰이는 단어, 이모티콘 등)이 반영된다는 특성을 가진다는 점을 주목할 필요가 있다. 특히 이러한 사용자의 언어 습관은 사용자의 기분 상태를 많이 반영한다. 그러므로 인터넷 메신저 대화 정보로부터 대화 내용의 의미를 분석할 필요 없이, 자주 쓰이는 단어나 이모티콘 등의 각 대화 요소들로부터 사용자가 대화를 하고 있을 당시의 기분을 쉽게 알아낼 수 있다.

본 연구에서는 이와 같은 메신저 대화 정보의 특성을 이용해서 마이크로소프트 인터넷 메신저를 통한 대화 정보의 기분을 알아내는 기법을 제안한다. 이 기법은 1) 대화 정보들로부터 유의미한 단어를 추출하여 단어 사전을 구성하고, 2) 대화 정보들을 세션 별로 분리한 후, 대화가 작성될 때 사용자의 기분을 레이블링(labeling)하며, 3) 분리된 대화 세션을 단어 사전을 이용해서 스코어링(scoring)하고, 4) 스코어링된 대화 세션들과, C4.5 decision tree, Bayesian network, Naïve Bayesian, k-nearest neighbor, Support Vector Machine 등의 분류 알고리즘을 이용해서 분류기를 만들어 대화 정보의 기분을 분류함으로써 대화가 이루어질 때 사용자의 기분을 알아맞히는 프로세스로 이루어져 있다. 이러한 프로세스의 세부 과정은 3장에 자세히 서술한다.

본 연구에서 제시된 기분 분류 기법의 결과, 위의 5개 알고리즘 모두 평균 95%를 넘는 정확도로 대화의 기분을 분류하는 결과를 보여주었다. 실험의 방법 및 자세한 결과는 4장에서 서술한다.

Ⅱ. 관련 연구

2.1. 채팅 대화 내용으로 기분을 분류하는 기법

[1]은 모두 1201개의 마이크로소프트 메신저 데이터를 이용했으며, 분노(angry), 슬픔(sad), 걱정(afraid), 역겨움(disgusted), 아이러니함(ironic), 기쁨(happy), 놀람(surprise)의 7가지 감정 중 데이터가 충분한 분노, 기쁨, 놀람의 3가지 감정과 중립(neutral) 감정을 이진 분류(binary classification)했다. 각 대화에서 속성 값을 추출하기 위해서 Chi-Squared metric을 이용해서 주요 단어를 골라내고 그 개수(count)를 속성 값으로 이용했다. 또한 분류는 주로 k-Nearest Neighbor를 이용해서 중립 상태에 대한 분노, 기쁨, 놀람의 분류 결과를 제시했으며, 세 가지 감정의 분류 정확도는 평균 83%로 나타났다. 그러나 [1]에서 제시한 기법은 대화 내용을 분류하기 위해서 이진 분류를 감정의 수만큼 수행해야 한다는 단점을 가지고 있다. 또한 트레이닝 데이터와 테스트 데이터를 구분하지 않고 속성값을 추출한 후 10-fold 교차타당법(cross-validation)을 사용함으로써 트레이닝 데이터와 테스트 데이터의 독립이 이루어지지 않고 있음을 알 수 있다.

[2]는 소프트웨어 에이전트(software agent)가 효과적으로 대화 내용을 분류할 수 있게끔 해주는 프레임워크에 대해서 주로 설명하고 있으며, 구체적인 실험결과는 제시하고 있지 않으므로, 본 연구와의 직접적인 비교가 어렵다. 하지만 기분을 '슬픔-중립-행복', '분노-중립-점잖은', '심각함-중립-신남', '증오-중립-사랑' 의 4가지로 분류하고, 한 대화 세션에 대해서 위의 4가지 분류를 수행함을 가정하고 있는데, 이는 참신하고 유용한 분류 기준이라고 생각된다.

2.2. 블로그 데이터 및 뉴스 기사 의 내용으로 기분을 분류하는 기법

[3]은 최대 6400개의 블로그 텍스트를 트레이닝 데이터로, 400개의 텍스트를 테스트 데이터로 사용하고 있으며, 감정을 37개로 세분류하고 있다. [3]은 자주 등장하는 단어를 PMI-IR이라는 메져(measure)를 사용해서 스코어링하고, 이 결과를 SVM을 이용해서 분류하고 있으며, 그 결과는 대부분 50% 정도를 기록하고 있다. 그러나 [3]에서 감정을 37개로 세분류한 것은 그 객관성이 떨어진다. 예를 들어 기쁨(happy)은 흥분된(excited), 사랑(loved), 희망(hopeful)의 3가지 감정을 포괄할 수 있는 개념이며, 서로 동등하게 분류되기 어려울 수도 있다. 또한 결과로 제시된 50%정도의 정확성은 이 방법이 실용적으로 쓰이기 어려울 수 있다고 생각된다.

[4]는 108,892개의 블로그 텍스트 및 정제된 10,479개의 블로그 텍스트를 바탕으로, 행복(happy), 두려움(fear), 화남(angry), 슬픔(sad)의 4가지로 감정을 분류하고 있다. SVM과 [4]에서 자체적으로 제시한 방법으로 분류한 결과를 비교하고 있으며, 정제된 데이터의 경우 각각 79.4%와 81.8%의 평균 정확도를 보여주고 있다. [4]는 중립적인(neutral) 감정을 고려하지 않음으로써, 4가지 감정만으로는 분류될 수 없는 블로그에 대한 분류 방법을 고려하지 않았다는 약점을 가진다.

[5]는 Yahoo(http://news.yahoo.com)에서 수집한 100개의 뉴스 기사에서 감정과 관련된 단어를 분리하고 기존에 생성된 감정과 관계된 단어의 사전과 EmotionScore라는 메져를 사용해서 스코어링한 후 긍정적, 부정적, 중립적인 감정으로 분류하는 방법을 취하고 있으며, 그 결과 3가지 감정에 대해서 각각86.7%, 94.4%, 87.8%의 정확도를 보여주고 있다. [5]은 다른 연구와 비교해서 감정들을 단순화 시켰다는 특징을 가진다.

[6]은 뉴스 웹사이트의 뉴스 기사들을 대상으로, 화(anger), 역겨움(disgust), 공포(fear), 기쁨(joy), 슬픔(sadness), 놀람(surprise)의 6가지 감정으로 기사를 분류한다. 이를 위해서 [6]은 지식 기반(knowledge based)과 코푸스 기반(corpus based) 방법을 제시했다. 지식 기반 방법은 LSA(Latent Semantic Analysis)를 이용해서 단어와 단어 집합, 문장 등을 백터화하며, 코푸스기반 방법은 8,761개의 블로그 포스트를 기반으로 코푸스를 생성 한 후 Naive Bayes 분류기를 통해서 학습을 시킴으로서 분류기를 생성한다. 결과적으로 뉴스 기사들을 각 감정별로 87%에서 100% 사이의 정확도로 분류할 수 있었다. 국내에서는 인터넷 상의 악성 댓글 판별 시스템 [7]과 관련한 연구가 진행되고 있다.

III. 분류자 생성 기법

3.1. 데이터 설명 및 파싱

본 연구에서는 저자의 2007년 12월에서 2008년 12월까지의 1년간 마이크로소프트 메신저 대화로그를 데이터로 이용했다. 이 로그들을 먼저 대화 상대 및 대화 세션 별로 나누고, 나누어진 각 세션 별 데이터 중에서도 서로 다른 감정으로 분류될 수 있는 대화를 나누는 작업을 선행했다. 이 중 대화의 길이가 4줄 미만이거나, 대화의 대부분이 단답형으로 짧게 이루어져 있는 경우는 수작업으로 제외를 시켰다. 이렇게 나누어진 대화 데이터는 그림 1의 a)처럼 나타난다.

이렇게 나누어진 각 대화 데이터에서 대화가 이루어진 시간 및 대화 상대를 삭제한 후 [8]에서 제공되는 한글 맞춤법 검사기를 사용해서 맞춤법 검사를 수행한 후의 결과는 그림 1의 b)처럼 나타난다. "있엉?"은 "있어?"로, "내가하라는" 은 "내가 하라는"으로 맞춤법 및 띄어쓰기 검사가 된 것을 확인할 수 있다.

맞춤법 검사가 완료된 대화 텍스트는 [8]에서 제공되는 형태소 분석기를 이용해서 일괄적으로 형태소 분석 과정[9]을 거친다. 형태소 분석의 결과가 그림 1의 c)처럼 나타난다. 본 연구에서는 형태소 중 명사를 나타내는 N과 동사 및 형용사를 나타내는 V, 기타 기호를 나타내는 q를 분석의 대상으로 포함시켰다. 그 이유는 N, V, q가 나머지 경우보다 많은 의미를 담고 있을 수 있기 때문이다. 또한 맞춤법 검사는 사용자의 언어 습관 중 대부분의 경우 - 신조어나 인터넷 용어(~있삼, 아놔 등), 습관적으로 쓰는 말(그림 1의 c) 마지막 줄의 "것..."), 이모티콘(-_-;;, ㅠㅠ 등)을 수정하지 않으며, 형태소 분석기는 이러한 단어들을 대부분의 경우 명사를 나타내는 N으로 처리한다. 따라서 N, V, q를 사용하는 것은 사전의 크기를 줄이면서도 합리적으로 분류에 도움이 될 수 있는 단어를 골라낼 수 있다고 판단된다.

형태소 분석의 결과, 하나의 단어가 k 가지의 가능성을 가지고 분석될 경우, 각각의 경우를 분리하여 단어로 포함하되, 단어의 스코어를 1/k로 한다. 예를 들어, 그림 1의 c)에서 '할' 이 동사 '하'와 명사 '할'로 분석되는 경우, 단어 '하'와 '할'

은 각각 0.5의 스코어를 가진다.

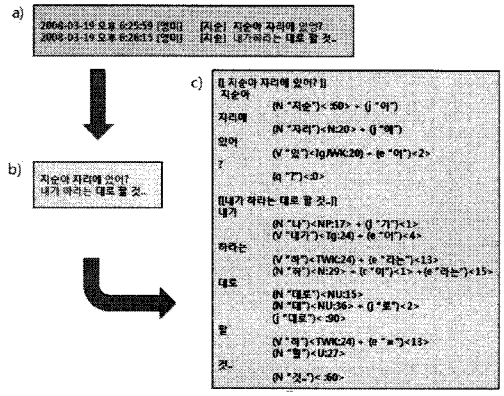

그림 1. 대화 텍스트의 파싱 및 형태소 분석 과정
Fig. 1. Parsing of Talk-text and Morphological Analysis Process

마지막으로 형태소 분석된 각 대화 텍스트에 기분을 레이블한다. 본 연구에서는 기분을 크게 행복(happy), 슬픔(sad), 화남(angry), 중립(neutral)로 나누었고, 나누어진 각 대화 텍스트에 위의 4가지 감정 중 1개를 레이블했다. 각 기분 및 기분에 해당하는 텍스트의 개수는 표 1과 같다.

표 1. 기분 및 기분에 해당하는 대화 텍스트의 개수
Table 1. The Number of Texts for each Emotion Class

Happy	Angry	Sad	Neutral	Total
77	61	46	80	264

3.2. 대화 텍스트의 전 처리

각 대화 데이터는 형태소가 명사(N), 동사 및 형용사(V), 기호(q)인 것만 모인 단어와 그 단어의 빈도수(frequency)의 쌍(pair)의 집합으로 정리된다. 이때 이 쌍(pair)의 *TF* (Term Frequency) 라고 하며, 다음과 같이 표현한다.

$TF\ (w_i,\ doc_i)$ = 문서 doc_i 에서 단어 w_i 의 출현 회수

또한 모든 단어에 대해서 사전을 구성하는 것은 효율성을 떨어뜨리므로 *IDF* (Inverse Document Frequency)[10]의 개념을 도입하여 효율적으로 사전을 구성할 수 있다. *IDF*는 다음과 같이 표현된다.

$$IDF(w_i) = log\left(\frac{n}{DF(w_i)}\right)$$

n : 문서의 총수

$DF\ (w_i)$ = 단어 w_i 가 출현하는 문서의 개수

어떤 단어가 전체 대화 집합에 골고루 분포하면서 자주 나온다면, 그 단어가 어떤 대화의 기분을 반영할 확률은 그만큼 적다고 할 수 있다. *TF* 만을 사용하면 이런 특징을 제대로 표현할 수 없지만, *IDF* 값은 단어가 전체 대화 집합에 자주 나타나는 단어일 경우 *DF* 가 커지면서 *IDF* 의 값은 0에 가까워지므로 효율적으로 이러한 단어를 배제할 수 있다. 따라서 *TF x IDF* 를 사용하면 효율적으로 사전의 크기를 줄이면서 기분을 잘 반영하는 단어를 골라낼 수 있다.

본 연구에서는 *TF x IDF* 의 값이 특정 임계값(threshold) 이상인 단어로만 사전을 구축한다. 결과적으로 전체 데이터는 사전의 모든 단어를 그 속성으로 하는 테이블의 형식으로 구축되며, 각 대화 데이터에 등장하는 각 단어가 구축된 사전에 포함될 경우 그 등장 횟수에 해당 단어의 스코어를 곱한 값을 속성 값으로 하는 하나의 튜플(tuple)로 구성된다. 결과 데이터 테이블은 그림 2와 같이 구성된다. 그림 2에서의 '!!!', '!!!!', '부담', '부산' 등은 사전에 등재된 단어를 뜻하며, 0, 0.6 등의 값은 해당 단어의 스코어와 빈도수를 곱한 값이다.

	!!!	!!!!	…	부담	부산	…	Class
Doc 1	0	0	…	0.6	0	…	Sad
Doc 2	0	0	…	0	0	…	unknown
Doc 3	1	0	…	0	0	…	Angry
…	…	…	…	…	…	…	…
Doc n	0	0	…	0	0.5	…	Happy

그림 2. 데이터 테이블의 예
Fig. 2. Example of Data Table

데이터 테이블을 구성할 때 쓰이는 것은 오직 트레이닝 데이터만 에 해당하는 것으로, 테스트 데이터는 사전의 구축에 관여하지 않는다. 이것은 트레이닝 데이터와 테스트 데이터의 독립성을 보장해준다. 본 연구에서 트레이닝 데이터는 4개의 기분 별 대화 데이터 각각에서 임의로 90%를 선정하며, 테스

트 데이터는 나머지 10%로 이루어진다. 이러한 트레이닝 및 테스트 데이터는 10세트 만들어져서 각각 문서 분류 테스트에 사용한다.

IV. 실험 결과

4.1. 실험 설명 및 실험 환경

본 연구에서는 단어의 허용 *TF x IDF* 값을 뜻하는, 임계값 각각(threshold = 2, 3, 4, 5)에 대해서, 3장 에서 설명한 대로 총 264개의 대화 데이터 중 임의의 90%를 트레이닝 데이터로 선정한 10세트의 트레이닝 데이터에 대한 사전을 구축한다. 따라서 총 40개의 사전이 구축되며 이 사전에 의해서 테이블화 된 전체 데이터셋을, 보편적으로 많이 사용되는 분류 알고리즘인 C4.5 decision tree, Bayesian network, Naïve Bayesian, k-nearest neighbor, Support Vector Machine 의 5가지에 대해서 각각 분류하고 그 결과를 살펴보았다.

실험은 AMD Athlon 64 X2 Dual 2.81GHz의 CPU와 1.93GB 메모리, Windows XP 운영체제가 설치된 PC에서 수행했다. 실험에서 사용된 분류 알고리즘은 Weka[11] v.3.5.8의 구현을 이용했으며, 5가지 알고리즘 모두 Weka tool에서 제공되는 디폴트 파라미터를 사용했다.

4.2. 사전의 크기

먼저, 임계값을 2, 3, 4, 5 로 했을 경우 각각 생성된 10개의 트레이닝 데이터셋을 이용한 사전의 크기의 평균이 그림 3의 그래프에 나와 있다.

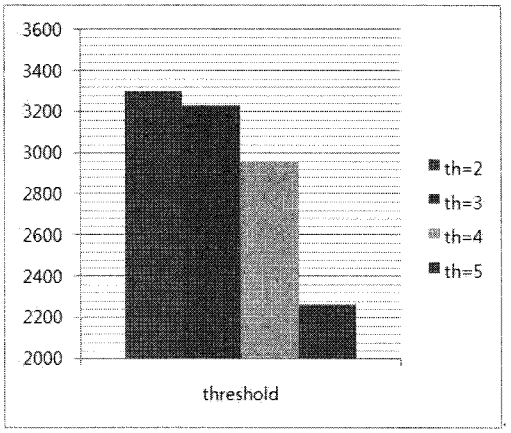

그림 3. Threshold 2, 3, 4, 5에 대한 평균 사전 크기
Fig. 3. Dictionary Size Average for Threshold 2, 3, 4, 5

임계값이 2 이하일 때는 모든 단어가 사전에 등록되었으므로 *TF x IDF* 를 사용하는 의미가 없었다. 또한 임계값이 6 이상일 경우 해당하는 단어가 없었으므로 아예 사전이 만들어지지 않았다.

4.3. 수행 시간 비교

그림 4에 임계값을 달리하며 각 알고리즘을 수행시켰을 때의 평균 수행 시간이 표시되어 있다. 그림4의 결과 k Nearest Neighbor와 Naïve Bayesian이 가장 빠른 수행시간을 보여줌을 알 수 있으며, Bayesian Network의 경우 사전의 크기에 가장 민감하고 가장 느림을 알 수 있다.

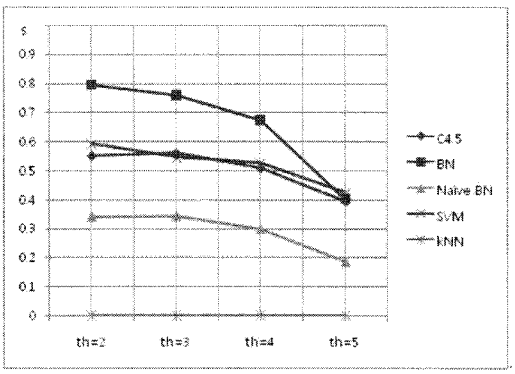

그림 4. 각 분류 알고리즘 별 분류 시간(초)
Fig. 4. Run Time for Each Classification Algorithm(seconds)

이 수행 시간은 분류기를 만들고, 테스트 데이터를 분류하

는 데 소요된 시간이다. 원본데이터의 파싱, 맞춤법 검사, 형태소 분석 및 데이터 테이블 구성 시 소요되는 시간은 포함되지 않았다. 실험 결과에 포함시키지는 않았지만, 이렇게 전처리에 소요되는 시간은 실제 분류 시간보다 훨씬 짧음을 확인했다.

4.4. 정확도 비교

그림 5는 임계값을 달리하며 각 알고리즘을 수행시켰을 때의 평균 정확도를 나타낸다. 그림 5에서 알 수 있듯이, 가장 좋은 성능을 보여주는 것은 두 종류의 Bayesian 분류기이다. 특히 Bayesian network는 임계값이 낮아도 좋은 성능을 보여주는 것을 알 수 있다. 그러나 기본적으로 어느 정도의 편차를 감안할 때, 각 알고리즘의 성능차가 확연히 드러나는 것은 아니므로, 정확도보다는 수행 시간으로 알고리즘을 선택하는 것이 좋음을 알 수 있다.

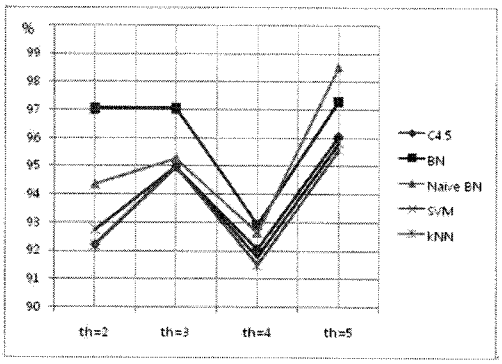

그림 5. 각 분류 알고리즘 별 분류 정확도
Fig. 5. Classification Accuracy for each Classification Algorithm

전체적으로 높은 성능을 보이고 있으며, 특히 사전의 크기가 줄어듦에 따라서 수행 시간뿐만 아니라 정확도까지 높아짐을 확인할 수 있다는 점은 고무적이다. 이런 높은 성능은 전처리 과정에서 형태소 분석의 결과를 효율적으로 이용하고 있으며, *TF x IDF* 메져가 대화 텍스트에서 사용자의 대화 습관을 충분히 반영하고 분류에 결정적인 역할을 하는 단어만을 잘 집어낼 수 있기 때문이라고 추론할 수 있다.

V. 결 론

[1], [2], [3], [4]와의 직접적인 정확도 비교는 어렵지만, 본 연구의 결과는 표면상 가장 높은 정확도를 가진 [1]의 83%와 비교해도 훨씬 우수한 정확도로 채팅 데이터를 분류하는 것을 알 수 있었고, 수행 시간 또한 충분히 실용화될 수 있을 정도로 빠름을 확인할 수 있었다. 저자가 찾아본 바로는 최근 텍스트 분류 기법 중 *TF x IDF* 메져를 사용한 것은 없었으며, 특히 채팅 데이터의 기분 분류에 사용한 연구 또한 처음이다. 본 연구에서는 이 메져가 이러한 채팅 데이터의 기분 분류에 효과적임을 알 수 있었다는 점에서 의미를 찾을 수 있다.

실험에서 나타난 분류 정확도와 분류에 소요된 수행 시간 등을 감안해 볼 때, 본 연구는 많은 분야에서 유용하게 응용되어 사용될 수 있을 것으로 기대된다. 응용될 수 있는 분야로 흥미 있는 것은 라이프 로깅뿐만 아니라, 1) 대화 상대방의 기분을 예측함으로써, 대화 시 실수를 미리 방지할 수 있는 기법, 2) 온라인 게임에서 나의 대화 내용으로 아바타의 표정이나, 행동을 정하는 기법[12], 3) 대화 내용에서 사용자의 기분을 알아냄으로써 온라인 쇼핑몰에서 사용자에게 커스터마이즈된 상품을 제시할 수 있는 서비스 [13, 14] 등을 생각해 볼 수 있다.

하지만 대화 데이터가 쌓여감에 따라서 분류 알고리즘이 처음부터 분류기를 생성한다면 그 속도 상 문제가 될 수 있다. 본 연구가 이러한 응용 분야에 더욱 잘 사용될 수 있게 하기 위해서, 대화가 진행되어 쌓여감에 따라 점층적으로 (incremental) 업데이트되는 분류기를 생성할 수 있는 분류 알고리즘을 개발하는 것을 향후 연구 목표로 하고 있다.

참고문헌

[1] L. Holzman and W. Pottenger, "Classification of emotions in internet chat: An application of machine learning using speech phonemes," Technical Report LU-CSE-03-002, Lehigh University, 2003.

[2] A. Z. Abbasi and Z. A. Shaikh, "An Approach Towards Emotion Estimation During Chat Sessions Using Software Agents,"4th International Conference on Innovations in Information

Technology, pp. 511-515, 2007

[3] G. Mishne, "Experiments with Mood Classification in Blog Posts,"In Workshop on Stylistic Analysis of Text for Information Access, 2005.

[4] Y. Jung, H. Park and S. H. Myaeng, "A Hybrid Mood Classification Approach for Blog Text," In Lecture Notes in Computer Science, Vol. 4099, pp. 1099-1103, 2006.

[5] D. B. Bracewell, J. Minato, F. Ren and S. Kuroiwa, "Determining the Emotion of News Articles," In Proceedings of the 2006 International Conference on Intelligent Computing (ICIC 2006), pp.918-923, 2006.

[6] C. Strapparava and R. Mihalcea, "Learning to Identify Emotions in Text," Proceedings of the 2008 ACM symposium on Applied computing, March 16-20, 2008, Fortaleza, Ceara, Brazil.

[7] 김묘실, 강승식, "SVM"을 이용한 악성 댓글 판별 시스템의 설계 및 구현", 제18회 한글 및 한국어 정보처리 학술대회, 285-289쪽, 2006년 10월

[8] 강승식, "한국어 형태소 분석과 정보 검색", 홍릉과학출판사, 2002년, *http://nlp.kookmin.ac.kr/*

[9] 여상화, "한영 모바일 번역기를 위한 강건하고 경량화된 한국어 형태소 분석기," 한국컴퓨터정보학회, 제14권 제2호, 191-199쪽, 2009년 2월

[10] W. Frakes, "Stemming algorithms," In W. Frakes & R. Baeza-Yates(Eds.), Information retrieval: Data structures and algorithms, Englewood Cliffs, NJ: Prentice Hall, 1992.

[11] I. H. Witten and E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations," Morgan Kaufmann, San Francisco, CA. Available on *WWW: www.cs.waikato.ac.nz/ml/weka.*

[12] 김진용, 유재휘, "아바타 통신에서의 얼굴 표정의 생성 방법," 한국컴퓨터정보학회, 제12권 제5권, 55-64쪽, 2007년 11월

[13] B. Pang, L. Lee and S. Vaithyanathan, "Thumbs up? Sentiment Classification USing Machine Learning Techniques," EMNLP, pp.79-86, 2002.

[14] K. Dave, S. Lawrence, D. M. Pennock, "Mining the Peanut Callery: Opinion Extraction and Sematic Classification of Product Reviews," In Proceedings of WWW 2003, Budapest, Hungry, 2003.

저 자 소 개

윤 영 미(Youngmi Yoon)

1981년 2월 : 서울대학교 자연과학대학 졸업(학사).

1983년 6월 : 오하이오 주립대학 수학과(학사수료)

1987년 3월 : 스탠포드대학교 컴퓨터과학과 졸업(이학석사)

2008년 8월 : 연세대학교 컴퓨터과학과 졸업(공학박사)

1987년 5월 ~ 1993년 5월 : IntelliGenetics Inc., California, USA, Software Engineer

1995년 2월 ~ 현재 :가천의과학대학교 부교수

〈관심분야〉 : 데이터베이스 시스템, 데이터 마이닝, 바이오인포매틱스

이영호(Young-HoLee)

2001년 2월 : 한국외국어대학교 응용전산학과(이학석사)

2005년 8월 : 아주대학교 의과대학 의료정보학과(이학박사)

2000년 ~ 2002년 : 한국IBM BI & CRM EM

2002년 ~ 현재 : 가천의과학대학교 의료공학부 교수

2007년 ~ 현재 : ISO/TC215전문위원

〈관심분야〉 : 데이터마이닝, 의료정보, u-헬스케어