Journal of Korea Multimedia Society Vol. 17, No. 2, February 2014(pp. 232-239)
http://dx.doi.org/10.9717/kmms.2014.17.2.232

한국어 트위터의 감정 분류를 위한 기계학습의 실증적 비교

임좌상[†], 김진만[††]

요 약

온라인에서의 글쓰기가 늘어나면서, 기계학습을 통해 이를 분류하는 연구가 늘고 있다. 그럼에도 불구하고 한국어로 작성된 마이크로블로그를 대상으로 한 연구는 많지 않다. 또한 통계적으로 기계학습을 평가한 연구를 찾아보기 힘들다. 본 논문에서는 트위터를 대상으로, 표본을 추출하고, 형태소와 음절을 자질로 사용하여 기계학습에 따라 감정을 분류하였다. 그 결과 약 76%정도 트위터에 포함된 감정이 분류되었다. Support Vector Machine이 Naïve Bayes보다 정확했고, 선형모델도 비구조적인 텍스트 처리에 비선형모델에 상응하는 정확성을 보였다. 또한 형태소가 음절 자질에 비해 높은 정확성을 보이지 않았다.

An Empirical Comparison of Machine Learning Models for Classifying Emotions in Korean Twitter

Joa-Sang Lim[†], Jin-Man Kim[††]

ABSTRACT

As online texts have been rapidly growing, their automatic classification gains more interest with machine learning methods. Nevertheless, comparatively few research could be found, aiming for Korean texts. Evaluating them with statistical methods are also rare. This study took a sample of tweets and used machine learning methods to classify emotions with features of morphemes and n-grams. As a result, about 76% of emotions contained in tweets was correctly classified. Of the two methods compared in this study, Support Vector Machines were found more accurate than Naïve Bayes. The linear model of SVM was not inferior to the non-linear one. Morphological features did not contribute to accuracy more than did the n-grams.

Key words: Machine Learning(기계학습), Support Vector Machine(서포트벡터머신), Naïve Bayes(나이브베이즈), Twitter emotion classification(트위터감성분류)

1. 서 론

날씨와 같이, 사람들의 감정을 알 수 있다면 어떨까? 날씨 예보를 듣고 우산을 준비하는 것처럼, 사람들의 감정을 알 수 있다면 우리는 생각과 행동을 조절할 수 있다. 감정에 대한 반응은 주관적으로, 생리적으로 또는 행동으로 나타나게 된다. 따라서 감정은 주관적으로 묻거나, 생리신호를 측정하거나 표정, 몸짓 등에서 알 수 있다. 최근에는 온라인에서 서로 만나지 않고도 상호작용이 늘고 있다. 이런 경우 그들

※ 교신저자(Corresponding Author) : 김진만, 주소 : 서울특별시 종로구 홍지동 상명대학교 소프트웨어대학관 G519호(110-743), 전화 : 010-2007-5645, E-mail : hansumo81@gmail.com
접수일 : 2014년 1월 15일, 수정일 : 2014년 2월 5일
완료일 : 2014년 2월 19일

[†] 정회원, 상명대학교 미디어소프트웨어학과
(E-mail : jslim@smu.ac.kr)
[††] 상명대학교 일반대학원 컴퓨터과학과
※ 본 연구는 2012학년도 상명대학교 교내연구 지원으로 수행되었음.

이 남긴 글에서 감정을 유추하게 된다. 요즘 인기를 끌고 있는 트위터가 그 한 예이다. 트위터는 140자의 짧은 글에 자신의 감정이나 의견을 표현한다. 기업은 이러한 트위터에서 브랜드 이미지와 상품에 대한 반응을 살펴보고 있다. 국가에서도 정책에 대한 의견을 트위터를 활용하여 분석하고 있다.

종전의 방식, 즉 주관적 설문, 생리적 반응, 행동관찰에서 감정을 측정하는 것은 노력과 비용이 많이 수반되어 기민한 대응이 어렵다. 하지만 트위터와 같은 마이크로블로그는 수집 및 분석 시스템을 구축하여 많은 사람들의 감정을 빠른 시간에 측정할 수 있다. 최근에는 마이크로블로그의 감정을 분석하여 의미 있는 정보를 발견하려는 연구가 해외에서 활발히 이루어지고 있다. 하지만 한국어를 대상으로 분석하는 연구는 매우 부족하다.

마이크로블로그는 신문기사와 같은 구조적인 텍스트와는 많이 다르다. 신조어나 줄임말의 형태가 자주 등장하고, 이모티콘의 사용률이 높으며, 띄어쓰기를 하지 않거나 주어가 생략된 형태가 많다. 특히 한국어는 영어로 작성된 것과 다르다. 영어는 고립어(isolating language)인 반면 한국어는 교착어(agglutinative language)의 특징을 가진다. 즉, 영어는 단어의 변형이 없어 조사나 접미사가 필요 없지만 한국어는 단어에 이들이 교착된다. 이 같은 특성은 한국어로 작성된 마이크로블로그에서의 감정분석을 어렵게 한다.

본 연구에서는 마이크로블로그 서비스 중 트위터를 대상으로 한국어로 작성된 트윗의 감정 분류에 효과적인 기계학습을 살펴보고, 형태소와 음절 방식을 사용하여 한국어 트윗의 감정 분류에 적합한 자질 추출 방식을 확인하는 것을 목적으로 했다. 이를 위해 감정이 담긴 트윗을 표본으로, 긍정과 부정 감정을 분류하였다. 그 결과 기계학습에 따라 분류 정확성에 차이를 보였지만 평균적으로 76%이상 정확하였으며, 형태소가 음절 자질에 비해 높은 정확성을 보이지 않았다.

2. 관련 연구

감정은 인간의 생각과 행동에 많은 영향을 미치고 있다[1,2]. 최근에는 온라인의 사용이 늘어나면서 정서분석이 관심을 끌고 있다. 감정은 심리학에서는 개인적인 반응으로 보고 있는데 반해, 정서는 보다 사회적인 것으로 정의되고 있다[3]. 즉 정서분석이란 어떤 사건, 대상에 대해 느끼는 의견을 보통 극성으로 나누어서 긍정, 부정으로 분류한다. 그러나 감정이 차원에서 극성을 가진다는 점에서 그 경계가 모호하다. 특히 온라인에서는 매우 많은 다수의 감정(또는 정서)을 분석해야 하므로, 이를 집계하는 것이 필요하다.

이러한 감정의 분류 연구는 외국에서 많이 등장하고 있다. 온라인에 작성된 상품평과 영화평에 나타나는 감정을 극성으로 분류하는 연구가 그렇다. 최근에는 마이크로블로그 중 트위터를 대상으로 감정을 분류하는 연구가 활발히 진행 중이다. 반면 한국어를 대상으로 한 연구는 상대적으로 적다. 국내 학술 DB인 DBPIA에서 '트위터 분석'으로 검색한 결과 총 47편이 나왔고(검색일: 2013.11.10), 이 중 분류와 관련된 연구는 3편[4-6](학술대회논문 제외)이었다.

기계가 문장에서 감정을 분류하는 데는 어휘집(lexicon) 또는 기계학습을 사용한다. 외국에서는 WordNet-Affect, SentiWordnet과 같은 사전을 어휘집으로 이용하고 있다[7].

사전 기반 감정 분류는 보통 그림 1과 같이, 우선 감정 분류 대상인 텍스트 문장($sentence$)에서 단어(w_i)를 추출한다. 다음으로 그 단어가 사전($dictionary$)에 포함되어 있는지 확인하고 사전에 정의된 감정값을 추출($score(w_i)$)하여 문장의 감정 정도($emotion$)를 계산하고 그 결과값으로 감정을 분류한다[8-10]. 이 방법은 높은 분류 정밀도(precision)를 보이는 반면에 재현율(recall)은 상대적으로 낮은 경향을 나타낸다[11].

기계학습 기반 감정 분류는 그림 2와 같이 학습단계와 테스트단계로 나뉜다. 학습단계에서는 학습 집합($trainset$)으로부터 텍스트 문장(s_i)과 라벨($label_i$)을 추출한다. 문장에서 감정 분류 자질이 추출

Algorithm 1 Lexicon Classification

```
for w_i ∈ sentence do
  while w_i ∈ dictionary do
    emotion = ∑(score(w_i))
  end while
end for
```

그림 1. 사전기반 분류 알고리즘

Algorithm 2 Machine Learning Classification

for $s_i, label_i \in trainset$ **do**
  $vec_i = feature(s_i)$
  $classifier = vec_i, label_i$
end for
for $s_i \in testset$ **do**
  $vec_i = feature(s_i)$
  $emotion = classifier(vec_i)$
end for

그림 2. 기계 학습기반 분류 알고리즘

($feature(s_i)$)되고 그 결과로 벡터(vec_i)가 생성 된다. 이 벡터와 라벨을 이용하여 분류 모델($dassifier$)이 만들어 진다. 다음으로 테스트 단계에서는 테스트 집합($testset$)의 각 문장에서 추출된 자질을 벡터화 하고 이를 앞서 생성된 분류 모델에 입력하여 문장의 감정을 분류($emotion$) 한다. 즉, 수집된 데이터를 분류하기 위해 알고리즘을 훈련시키고, 훈련에 사용된 자질이 실제 분류 시 가중치를 갖게 되어 문장의 감정이 특정 극성으로 치우치게 된다[5,12].

[13]은 영화평을 긍정과 부정으로 이분화 하는 연구를 진행했다. 그 결과 자질에 따라 차이는 보였지만 Naïve Bayes(NB), Maximum Entropy 보다 Support Vector Machine(SVM)이 평균적으로 분류 정확성이 높게 나왔다. 또한, [14]는 한국어 뉴스 댓글을 대상으로 자질 가중치 조절 없이 NB, k-Nearest Neighbor, SVM의 분류 정확성을 비교한 결과 SVM이 우세한 것으로 나왔다.

일반적으로 기계학습은 구조적인 텍스트의 분류 정확성을 입증하고 있으나 마이크로블로그와 같은 비구조적인 경우에는 어려움이 뒤 따른다. 특히, [15]의 연구와 같이 트위터와 같은 비구조적 텍스트인 경우에도 기계학습이 우월한 것으로 보였지만, 자질의 특성과 기계학습의 특성에 따라 정확성에 차이가 있는 것으로 나타났다.

따라서 본 논문에서는 비구조적 텍스트의 감정을 분류하기 위해 적합한 기계학습과 분류에 영향을 미치는 자질의 정확성을 비교하였다.

3. 연구 방법

3.1 연구 설계

본 논문에서는 한국어 트위터 감정 분류에 적합한 기계학습과 자질을 확인하기 위해 '2 기계학습 × 3 자질'((NB-BD(Bernoulli Distribution), LSVC(Linear Support Vector Classification)) × (형태소, 2음절, 3음절)) 실험이 설계되었다.

3.2 데이터

본 연구에서는 실험을 위해 마이크로블로그인 트위터를 선택하고 한국어로 작성된 트윗으로 한정하여 2,759명 트위터 사용자가 2011년 1월~2012년 4월 사이 작성한 트윗을 수집하였다. 총 1,563,944개가 수집되었고 이 중 감정이 포함된 트윗을 무작위로 1,333개 선별하였다. 다음으로 이를 긍정, 부정 감정으로 분류하였다. 분류에는 3명이 참여하였고, 이들은 현재 트위터를 사용해오고 있으며, 감정의 인지에 문제가 없었다. 2명 이상 동일한 감정으로 판정된 트윗만을 선별한 결과, 실험데이터로 총 1067개(긍정 509개, 부정 558개) 트윗이 추출되었다.

3.3 전처리

실험데이터는 멘션 트윗, RT 트윗, URL을 포함한 트윗이 각각 535개, 48개, 85개로 전체의 약 63%(668개)를 차지하고 있고, 'ㅠㅠ', 'ㅡㅡ', '♡', '♥', '^^', 'ㅋ', 'ㅎ'와 같은 이모티콘을 포함한 트윗이 647개로 약 61%를 차지하고 있다. 또한, 이 데이터에서 URL, 'RT'문자, 아이디(예: @hansumo81)를 제외한 트윗의 평균 글자 수는 약 49개로, 최대, 최소 글자 수는 각각 131개, 3개로 나타났다. 위와 같이 트위터는 구조적인 텍스트와 달리 웹에서 쓰이는 기호, 이모티콘 등의 처리하기 어려운 데이터를 포함하고 있다. 따라서 본 연구에서는 다음과 같은 처리를 수행하였다.

(1) 사용자명 - 트윗은 상대방 트윗에 응답하기 위해 상대방의 아이디를 포함한다. 이것은 사용자명 앞에 '@' 기호가 포함되어 작성된다(예: @hansumo81). 우리는 '@' 기호가 나타난 위치를 시작으로 띄어쓰기가 나오는 위치까지를 추출하여 'AT_USER'로 변환한 뒤 제거하였다.

(2) 웹링크 - 트윗에는 신문기사나 웹페이지 혹은 사진과 관련된 웹링크를 달아 해당 정보에 대한 자신의 감정을 표현할 수 있다. 우리는 'http' 문자열이 나타난 위치를 추출하고 웹링크 주소를 'URL'로 변환하여 제거하였다.

(3) 이모티콘 - 이모티콘은 사용자의 감정을 빠르게 파악하는데 도움이 된다. 하지만 텍스트의 감정에 상관없이 혹은 강조하기 위해 사용되는 경우도 많다 (예: '정말 기분이 나쁘네요 ㅎㅎㅎ', '우와 너무 감사해요 ㅠㅠ'). 우리는 변환 없이 이모티콘을 트윗에서 제거하였다.

3.4 트윗 자질 추출

영어는 고립어 특징을 가지고 있어 비구조적 텍스트인 트윗에서도 비교적 단어의 추출이 쉽다. 반면 한국어 트윗은 그렇지 못하다. 따라서 우리는 한국어 트위터 감정 분류에 적합한 자질을 확인하기 위해 (1) 형태소 분석과 (2) 음절 방식을 사용하였다.

형태소 분석은 어절에서 형태소 원형을 복원하고 형태소 단위로 분리하여 어절에 포함된 형태소를 찾아내는 것이다. 우리는 형태소 분석기[16]를 사용하여 트윗의 품사를 추출하였고, 이 중에서 명사, 형용사, 동사를 트윗 감정 분류 자질로 사용하였다.

음절 방식은 문장에서 n개의 연속된 단어나 음절을 추출하는 것으로 영어권 계산 언어학(Computational Linguistics)에서는 단어 단위로 이를 구분한다. 한국어는 이와 다르게 보통 문장을 어절 단위로 분리한 후 그 결과로 생성된 분절을 음절 단위로 추출한다. 우리는 2음절, 3음절을 분류 자질로 사용했다. 이 두 개는 한국어 정보검색 연구에서 주 자질을 추출할 때 사용되어 왔고[17], 특히 한국어 단어의 약 80%가 이들로 구성되어 있다[18]. 또한 우리는 트윗이 띄어쓰기가 되어 있지 않은 경우를 감안하여 어절 단위 분리를 하지 않고 음절을 추출하였다. 즉, '트윗의 분류'라는 문자열의 경우 2음절은 '트윗', '윗의', '의 ', ' 분', '분류'로 3음절은 '트윗의', '윗의 ', '의 분', ' 분류'로 추출된다.

3.5 기계 학습

3.5.1 Naïve Bayes

NB는 베이즈 이론[19]에 기반한 것으로 클래스 조건부 독립을 가지고 데이터가 어느 한 클래스에 속할 확률을 예측한다.

트윗의 감정을 클래스 C라고 했을 때, $p(C)$는 클래스 C의 사전확률(prior probability)로 임의의 트윗이 클래스 C가 될 확률을 나타내고 $p(F_1,...,F_n|C)$는 자질 $F_1,...,F_n$의 사후확률(posterior probability)로 특정 트윗이 클래스 C일 때 그 트윗이 자질 $F_1,...,F_n$을 포함할 확률이다. 이를 식으로 나타내면 아래와 같다. 여기서 Z는 $F_1,...,F_n$에만 의존하는 계수이며, 자질값을 알고 있으면 상수가 된다.

$$p(C|F_1,...,F_n)=\frac{1}{Z}p(C)\prod_{i=1}^{n}p(F_i|C)$$

우리는 확률 분포에 따라 BD와 Multinomial Distribution(MD)을, 우도(likelihood)에 따른 Maximum-Likelihood Estimator(MLE)와 Expected-Likekihood Estimator(ELE)를 적용한 NB를 트윗의 감정 분류에 사용하였다.

3.5.2 Support Vector Machine

1992년에 발표된 SVM[20]은 서로 다른 클래스에 속한 벡터들 간에 거리의 최대마진을 구하는 초평면을 찾는 것으로 느린 훈련 시간에도 불구하고, 비선형 의사결정 영역을 모형화할 수 있고, 다른 기계학습에 비해 과대적합 되는 경향이 매우 낮아 많은 분류 문제 연구에 사용되고 있다[21].

SVM은 기본적으로 이진 분류기로써 아래와 같이 이진 분류를 위한 초평면을 나타낼 수 있다. 여기에서 $\boldsymbol{x}$는 트윗의 자질 벡터로서 $\boldsymbol{x}=(x_1,...,x_n)$이다. $\boldsymbol{w}$와 b는 벡터들을 분류하는 초평면을 정의하는 매개 변수이다. 따라서 $d(\boldsymbol{x})$는 벡터를 특정 영역으로 분류하는 값으로 트윗의 자질 $\boldsymbol{x}$가 $d(\boldsymbol{x})>0$ 또는 $d(\boldsymbol{x})<0$의 값을 가짐으로 특정 영역 즉, 특정 감정으로 분류될 수 있다.

$$d(\boldsymbol{x})=\boldsymbol{w}\cdot\boldsymbol{x}+b=0$$

우리는 분류관점에서 LSVC와 Non-Linear SVC (NLSVC)를 사용하였다. 이 때, NLSVC는 3개 커널(linear, polynomial(poly), radial basis function (rbf))이 사용되었다.

3.6 정확성

본 연구에서는 트윗 감정 분류의 정확성 척도로 F 값을 사용하였고 분류 감정을 묶어서(macroaveraging) 이를 분석하였다. F 값은 아래 식과 같이 precision과 recall의 조화평균으로 정의된다.

$$F=2\cdot\frac{preision\cdot recall}{preision+recall}$$

표 1. 기계학습과 자질의 F-measure 평균

	Morpheme	bi-gram	tri-gram	Average
NB-MD	.7218	.6948	.8281	.7376
NB-BD	.7464	.6905	.8096	.7479
NB-ELE	.7152	.7006	.8316	.7356
NB-MLE	.7230	.7115	.8365	.7434
NB Average	.7266	.6993	.8265	.7411
LSVC	.7960	.7006	.8316	.7840
NLSVC-linear	.7710	.7006	.8316	.7691
NLSVC-poly	.8117	.7006	.8316	.7935
NLSVC-rbf	.8049	.7006	.8316	.7894
SVM Average	.7059	.7006	.8316	.7840
All Average	.7612	.7000	.8290	.7625

3.7 기계학습의 훈련

기계학습의 훈련을 위해 본 연구에서는 1067개 실험데이터를 무작위로 섞고, n-fold 방식으로 80%를 훈련데이터로 사용하였다.

4. 실 험

4.1 실험 준비

앞서 연구방법에 정의한 대로 각 기계학습을 훈련한 후 실제 데이터에 적용하여 분류 정확성을 산출하였다. 실험에 사용할 요인을 찾기 위해 각 4개의 NB(MD, BD, ELE, MLE)와 SVM(LSVC, NLSVC-linear, NLSVC-poly, NLSVC-rbf)이 비교되었다. 또한 한국어 분류에서 많이 사용되는 방식으로 자질을 추출하여(형태소, 2음절, 3음절) 비교하였다.

기계학습 요인은 NB 4개와 SVM 4개에 대한 사전 실험을 통해 결정되었다. 각 기계학습 간에 정확성에 다소간의 차이가 있었다. NB 중에서 NB-BD(.7479)가 가장 높은 분류 정확성을 보였고 가장 낮은 것은 NB-MLE(.7434)로 나타났다(표 1 참조). 그러나 이 차이는 통계적으로 유의적이지 않았다. 즉, 일원분산분석(One-Way ANOVA)을 수행하여 NB 4개에 대해 분류 정확성 차이가 있는지를 살펴보았고 그 결과 유의수준 .05에서 NB들 사이에 차이가 없는 것으로 나왔다($F(3, 399)=.1773$, $p=.9118$). SVM 4개를 분석한 결과도 마찬가지로 유의수준 .05에서 SVM 사이에 차이가 없었다($F(3, 399)=.6112$, $p=.6081$). 이처럼, 기계학습 사이에 유의한 차이가 없으므로 우리는 기계학습 요인을 NB 4개와 SVM 4개에서 하나씩 무작위로 추출하였다. 그 결과 LSVC와 NB-BD가 선택되었다.

자질 요인은 형태소와 2음절, 3음절로 설정하였다. 음절 자질 추출 결과는 그 규모가 형태소에 비하여 과도하여 자질에 따른 차이가 기계학습에 따른 차이를 상쇄할 수 있다고 판단하여 형태소는 품사별로 정확성을 비교하여, 명사, 형용사, 동사를 사용하였다.

4.2 실험 결과와 분석

우리는 기계학습, 자질에 따른 정확성의 차이를 분석하기 위해 이원분산분석(Two-Way ANOVA)을 실시하였다. (1) 기계학습, 자질의 주효과와 (2) 기계학습과 자질의 상호작용 효과에 대한 통계적 유의성을 검정한 결과는 주효과는 있었지만, 상호작용 효과는 없었다(표 2).

4.2.1 기계 학습에 따른 정확성

기계학습이 트윗 감정 분류 정확성에 영향을 주는지에 대한 검정결과 유의수준 .05에서 기계학습 간

표 2. 기계학습과 자질의 이원분산분석

		df	F	Sig.
Main Effect	Machine Learning(A)	1	4.9465	.0281*
	Features(B)	2	35.3553	.0000**
Interaction	A × B	2	.9130	.4042

*p<.05 **p<.01

유의한 차이가 있는 것으로 분석되었다($F(1, 119)=4.9465$, $p=.0281$). 기계학습에 따른 분류 정확성은 평균적으로 SVM이 NB보다 높은 것으로 나타났다(표 1 참조). 따라서 한국어 트위터 감정 분류에 SVM이 NB에 비해 효과적이었다.

4.2.2. 자질에 따른 정확성

자질이 트윗 감정 분류에 영향을 주는지에 대한 분석결과 유의수준 .01에서 자질에 따라 분류 정확성에 유의한 차이가 있는 것으로 나타났다($F(2, 119)=35.3554$, $p=.0000$). 자질의 분류 정확성 평균은 3음절이 가장 높고, 형태소, 2음절 순으로 나타났다(표 1 참조). 따라서 3음절은 다른 자질에 비해 한국어 트위터 감정 분류에 효과적이었다. 이러한 자질 효과는 기계학습과 무관하였다.

5. 결 론

트위터에 작성된 글은 비구조적인 특성을 갖고 있다. 본 연구에서는 기계학습을 이용하여 비구조적인 트윗에 포함된 감정을 분류하였고, 기계학습간 또는 자질에 따른 정확성의 차이를 비교하는 것에 초점을 맞췄다.

본 논문에서 사용된 8개 기계학습 모두 트윗을 상당히 정확하게 분류해냈다. 가장 정확한 기계학습은 약 84%에 가까운 정확성을 보였다(NLSVC-poly). 즉 트위터 트윗과 같은 온라인 문서에는 띄어쓰기가 불규칙적이고, 신조어가 포함되어 있는 비구조적인 특성이 포함되었음에도 불구하고, 텍스트 감정 분류에 기계학습이 유용함을 다시 입증하였다.

검증된 2개의 기계학습 가운데 NB보다는 SVM이 분류 정확성이 높았다. 텍스트 분류에 SVM이 보다 우수하다는 것은 여러 연구에서 보고되고 있다[13, 22-24]. NB보다 SVM이 그 정확성이 높은 이유는 데이터가 많을 경우, SVM은 과도하게 에러를 줄이려 하지 않고, 불필요한 자질을 걸러내면서 학습하는 능력이 뛰어나기 때문이다[25]. 반면 NB는 서로 자질이 독립적이라는 가정을 갖고 실행이 되기 때문에 자질이 서로 관련이 있는 경우 정확성이 떨어진다. 또한 관련성이 없거나 중복이 되는 자질이 포함되는 경우에도 그 정확성이 떨어지는 것으로 나타나고 있다[26].

자질에서는 3음절로 추출된 자질이 분류 정확성이 높았다. 이러한 결과는 사전 연구와 일치한다[17]. 트윗에는 신조어나 의도적 오타가 많이 포함되고, 한국어도 예외가 아니다(예: '개드립', '좋아욬ㅋㅋㅋ'). 즉 이러한 것들은 형태소 분석에서 제외되지만 음절 방식에서는 이런 형태의 음절을 인식할 수 있다. 또한 형태소는 단어벡터의 각 요소가 서로 독립적인 반면, 긴 음절단위의 자질은 앞 뒤 단어의 연관성을 포함하는 특성을 가지고 있으며, 이는 정확성에 영향을 미치게 된다.

본 연구는 몇 가지 한계점을 지닌다. 한국어는 어미의 변형이 많아 분류에 많은 자질이 포함될 수밖에 없다. 기계학습은 어떤 자질을 얼마나 포함할 것인지에 따라 그 정확성이 영향을 받게 된다[27]. 또한 본 연구에서는 SVM과 NB만을 포함하였지만, 한국어 특성에 맞는 기계학습이나 대량의 데이터를 실시간으로 처리해야 하는 온라인 특성을 고려하여 훈련이 필요 없는 기계학습을 포함하여 특성화하는 연구가 필요하다.

참 고 문 헌

[1] Gerald L Clore, Norbert Schwarz, and Michael Conway, *Handbook of Social Cognition*, Psychology Press, New York, pp. 323-417, 1994.

[2] Michael W Morris and Dacher Keltner, "How Emotions Work: the Social Functions of Emotional Expression in Negotiations," *Research in Organizational Behavior*, Vol. 22, pp. 1-50, 2000.

[3] Peggy A Thoits, "The Sociology of Emotions," *Annual Review of Sociology*, Vol. 15, pp. 317-342, 1989.

[4] 홍초희, 김학수, "트윗 감정 분류를 위한 다양한 기계학습 자질에 대한 비교 연구," 한국콘텐츠학회논문지, 제12권, 제12호, pp. 471-478, 2012.

[5] 이철성, 최동희, 김성순, 강재우, "한글 마이크로블로그 텍스트의 감정 분류 및 분석," 정보과학회논문지:데이타베이스, 제40권, 제3호, pp. 159-167, 2013.

[6] 김민철, 심규승, 한남기, 김예은, 송민, "트위터

상의 악의적 이용 자동분류," 한국문헌정보학회지, 제47권, 제1호, pp. 269-286, 2013.

[7] Angela Fahrni and Manfred Klenner, "Old Wine or Warm Beer: Target-specific Sentiment Analysis of Adjectives," *Proc. The Symposium on Affective Language in Human and Machine*, pp. 60-63, 2008.

[8] Minqing Hu and Bing Liu, "Mining and Summarizing Customer Reviews," *Proc. The Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 168-177, 2004.

[9] Xiaowen Ding, Bing Liu, and Philip S Yu, "A Holistic Lexicon-based Approach to Opinion Mining," *Proc. The International Conference on Web Search and Web Data Mining*, pp. 231-240, 2008.

[10] Maite Taboada, Julian Brroke, Milan Tofiloski, Kimberly Voll, and Manfred Stede, "Lexicon-based Methods for Sentiment Analysis," *Computational Linguistics*, Vol. 37, No. 2, pp. 267-307, 2011.

[11] Ley Zhang, Riddhiman Ghosh, Mohamed Dekhil, Meichun Hsu, and Bing Liu, *Combining Lexiconbased and Learning-based Methods for Twitter Sentiment Analysis*, HP Laboratories, Technical Report HPL-2011, Vol. 89, 2011.

[12] Bo Pang and Lillian Lee, "A Sentimental Education: Sentiment Analysis using Subjectivity Summarization based on Minimum Cuts," *Proc. The 42nd Annual Meeting on Association for Computational Linguistics*, pp. 271, 2004.

[13] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, "Thumbs Up? Sentiment Classification using Machine Learning Techniques," *Proc. Emnlp 2002*, pp. 79-86, 2002.

[14] 이공주, 김재훈, 서형원, 류길수, "뉴스 댓글의 감정 분류를 위한 자질 가중치 설정," 한국마린엔지니어링학회지, 제34권, 제6호, pp. 871-879, 2010.

[15] Alec Go, Richa Bhayani, and Lei Huang, *Twitter Sentiment Classification using Distant Supervision*, CS224N Project Report, Stanford, pp. 1-12, 2009.

[16] Taku Kudo, MeCab. version 0.996, 2013.

[17] 이준호, 안정수, 박현주, 김명호, "한글 문서의 효과적인 검색을 위한 n-Gram 기반의 색인 방법," 정보관리학회지, 제13권, 제1호, pp. 47-63, 1996.

[18] 김철수, 김양범, "대용량 전자사전 구축을 위한 국어 대사전의 통계 정보," 한국콘텐츠학회논문지, 제7권, 제6호, pp. 60-68, 2007.

[19] J Susan Milton and Jesse C Arnold, *Introduction to Probability and Statistics: Principles and Applications for Engineering and the Computing Sciences*, McGraw-Hill, Inc., New York, 2002.

[20] Bernhard E Boser, Isabelle M Guyon, and Vladimir N Vapnik, "A Training Algorithm for Optimal Margin Classifiers," *Proc. The Fifth Annual Workshop on Computational Learning Theory*, pp. 144-152, 1992.

[21] Jiawei Han, Micheline Kamber, and Jian Pei, *Data Mining: Concepts and Techniques*, Morgan kaufmann, San Francisco, California, 2006.

[22] Yiming Yang and Xin Liu, "A Re-examination of Text Categorization Methods," *Proc. The 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 42-49, 1999.

[23] Jason DM Rennie and Ryan Rifkin, *Improving Multi Class Text Classification with the Support Vector Machine*, Technical Report 2001-026, MIT. 2001.

[24] 황두성, "지지벡터기계를 이용한 다중 분류 문제의 학습과 성능 비교," 멀티미디어학회논문지, 제11권, 제7호, pp. 1035-1042, 2008.

[25] Thorsten Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," 1998.

[26] Sotiris B Kotsiantis, “Supervised Machine Learning: a Review of Classification Techniques,” *Informatica*, Vol. 31, No. 3, pp. 249-268, 2007.

[27] Fabrice Colas and Pavel. Brazdil, “Comparison of Svm and Some Older Classification Algorithms in Text Classification Tasks,” *In Artificial Intelligence in Theory and Practice*, Vol. 217, pp. 169-178, 2006.

임 좌 상

1991년 New South Wales University MIS전공 석사
1994년 New South Wales University MIS전공 박사
1997년~현재 상명대학교 미디어소프트웨어학과 교수

관심분야 : 소프트웨어공학, 빅데이터, 텍스트마이닝, 위치기반서비스

김 진 만

2003년 한서대학교 수학과(컴퓨터과학 복수) 학사
2006년 상명대학교 디지털미디어대학원 정보통신학과 석사
2008년~현재 상명대학교 일반대학원 컴퓨터과학과 박사과정

관심분야 : 소프트웨어공학, 빅데이터, 텍스트마이닝, 시각화, 위치기반서비스