# 2-Phase CNN을 이용한 SNS 글의 논쟁 유발성 판별

허상민[1], 이연수[2], 이호엽[2]
(주)엔씨소프트[1,2]
leo6104@gmail.com, {yeonsoo, hoyeoplee}@ncsoft.com

# Debatable SNS Post Detection using 2-Phase Convolutional Neural Network

Sang-Min Heo[1], Yeon-soo Lee[2], Ho-Yeop Lee[2]
NCSOFT Corp.[1,2]

**요 약**

본 연구는 SNS 문서의 논쟁 유발성을 자동으로 감지하기 위한 연구이다. 논쟁 유발성 분류는 글의 주제와 문체, 뉘앙스 등 추상화된 자질로서 인지되기 때문에 단순히 n-gram을 보는 기존의 어휘적 자질을 이용한 문서 분류 기법으로 해결하기가 어렵다. 본 연구에서는 문서 전체에서 전역적으로 나타난 추상화된 자질을 학습하기 위해 2-phase CNN 기반 논쟁 유발성 판별 모델을 제안한다. SNS에서 수집한 글을 바탕으로 실험을 진행한 결과, 제안하는 모델은 기존의 문서 분류에서 가장 많이 사용된 SVM에 비해 월등한 성능 향상을, 단순한 CNN에 비해 상당한 성능 향상을 보였다.

주제어 : 문서 분류, 논쟁 유발성, SNS 문서 분류, CNN

## 1. 서론

논쟁은 어떤 문제에 대하여 각자의 의견을 주고받으며 토론을 통해 적절한 절충안을 찾아가는 과정을 의미한다. 특히 SNS에서 이루어지는 논쟁은 누구라도 자유롭게 발의할 수 있고, 직접적으로 찬반 의견을 피력할 수 있으며, 기존 매체에서 다룰 수 없는 다양한 사안에 대해 논의할 수 있다는 장점이 있다. 반면, 이러한 자유로움으로 인해 어떤 글은 사실을 왜곡하거나 악의적인 폭로와 비방으로 인해 사회적 문제를 일으키기도 한다. 논쟁을 유발하는 글은 긍정적으로든 부정적으로든 사회적 파급력이 크다고 할 수 있다.

최근에는 SNS에서 발생한 문서를 다양한 관점에서 분류하는 연구가 진행되었다. 대표적인 연구로는 트위터(twitter)에서 트윗(tweet)의 감정을 분류하는 연구를를 들 수 있다[1]. 해당 연구에서는 트윗에서 정형어휘(formal vocabulary) 추출하였으며, 사전 기반의 분류기를 통해 감정을 구분하였다. 이보다 확장된 개념으로 트윗에서 정형어휘(formal vocabulary) 뿐만 아니라 비정형어휘(informal vocabulary)도 함께 사용하여 특징(feature)을 추출한 연구[2]도 진행되었는데, 그 연구에서는 SVM(support vector machine)을 통해 문서의 감정 분류를 수행하였다. 중국어 SNS 블로그(blog) 사용자의 기호를 파악하기 위한 관점에서 SNS 분류 문제를 접근한 연구도 진행되었으며[3], 앙상블(ensemble) 모형을 응용하여 트위터, 마이스페이스(MySpace), 슬래시닷(Slashdot) 등 다양한 SNS에서 수집된 데이터를 바탕으로 인터넷 집단 따돌림(cyberbullying)을 감지하는 시스템이 제안되기도 했다[4].

글의 논쟁 유발성을 자동으로 판별하는 것은 하나 이상의 문장으로 구성된 문서를 대상으로 특정 기준에 의해 가부를 판단한다는 점에서 기존의 문서 분류 테스크들과 유사하다고 할 수 있다. 그러나 기존 문제 영역과는 달리 특정 어휘나 문체에서 직접적으로 드러나지 않고, 주제가 다양하며 구어체를 비롯한 다양한 SNS 언어 행태가 나타나기도 한다. 아래는 최근 페이스북 대학생 커뮤니티에서 발생한 논쟁적인 글의 예이다.

> 총학생회 투표 기간입니다. 그런데 투표 마지막 날에 접어든 오늘까지도 아직도 투표율이 30%밖에 되지 않습니다. 어떻게 지성인들이 모인 대학교에서 이렇게 투표율이 낮은지 도무지 이해가 되지 않습니다. 건물마다 투표소가 설치되어있고 문자로 온라인 투표 안내도 되고 있는데... 여러분이 어느 선본을 지지하는지는 중요하지 않습니다. 학생 정치에 환멸을 느꼈으면 지지하는 선본없음이라도 체크 해 주세요. ... (생략) 제발 우리들의 권리를 썩히지 맙시다 ...

최근에는 이처럼 사람이 판단하기 힘든 자질을 학습하기 위해 이미지 인식에서 주로 사용하는 CNN (Convolutional Neural Network) 기법을 사용한 문서 분류 연구가 수행되고 있다. 대표적으로 Kim은 단어 표현(word embedding)과 CNN 모형을 활용하여 문장의 극성(polarity), 주체성(subjectivity) 등에 대한 분류를 수행하였다[5]. 여기서 사용된 CNN 모델은 매우 단순하면서도 좋은 성능을 보였다.

우리는 논쟁 유발성 판단을 위해서는 보다 추상화된

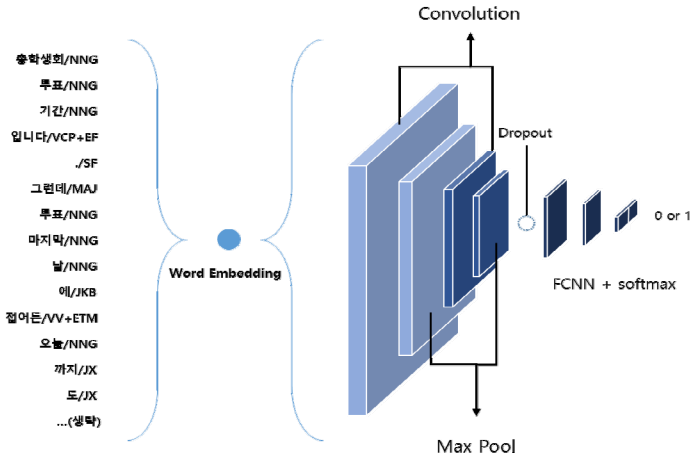


그림 2. 2단계 CNN이 적용 된 논쟁 유발성 판별 방법

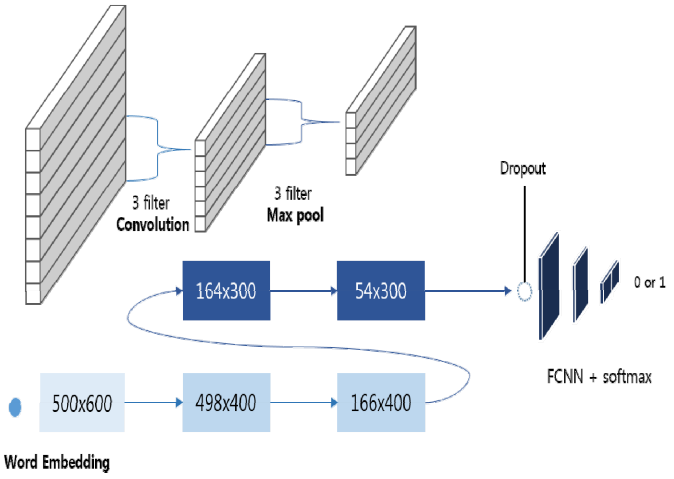


그림 3. Convolution, Max Pooling 과정

자질이 학습될 수 있어야 하는 가정하에 기존 [5]의 CNN을 확장한 2-phase CNN 모델을 제안한다. 그리고 이를 학습하기 위해 한국어 SNS 데이터의 메타 정보를 이용하여 자동으로 학습데이터를 구축하고, [5]와 제안하는 모델을 비교 실험하였다. 2장에서는 제안하는 모델을 자세히 설명하고, 3장에서는 SNS에서 수집한 데이터를 바탕으로 실험한 결과를 분석한다. 마지막으로 4장에서는 연구의 결론과 향후 연구에 대해서 다룬다.

## 2. 2-Phase CNN

그림 2는 본 논문이 제안하는 2-Phase CNN 모형을 보여준다. 제안 모형은 SNS의 글을 입력으로 받아, 2 단계로 구성된 CNN 층을 거친다. 2단계의 CNN을 통과한 이후에 입력된 글은 주제, 문체, 뉘앙스 등 추상화된 자질이 추출된다. 추상화 된 자질을 바탕으로 글의 논쟁 유발성을 판별하기 위해 3번의 FCNN(Fully Connected Neural Network) 층을 거치도록 한다. 이때, 마지막 FCNN은 softmax 층으로 구성하여 입력된 글이 논쟁 유발 글인지 아닌지를 판별할 수 있도록 한다.

CNN 층은 크게 Convolution 층과 Pooling 층으로 세분화 할 수 있는데, 그림 3은 이를 시각적으로 표현하고 있다. 예를 들어, SNS 글이 $n$개의 단어로 구성되어 있고 단어를 $k$ 차원으로 임베딩 시켰을 때, 입력 층은 $n \times k$의 행렬로 구성할 수 있다. 본 연구에서는 3개의 단어 단위로 SNS 글이 추상화 될 수 있도록 Convolution 층을 구성하며, 이를 다시 한 번 추상화시키기 위해 3-Max Pooling 층으로 통과시킨다. 2단계 CNN는 1단계 CNN의 출력을 입력으로 넣어 같은 과정을 반복하도록 한다.

### 2.1. 단어 표현 (Word Embedding)

본 논문이 제안한 CNN 모형의 입력은 문장의 각 단어를 one-hot 벡터로 표현한다. 단어 $w$의 one-hot 벡터 $v^w$는 그 단어를 나타내는 특정 위치의 값만 1이고 나머지가 모두 0인 1차원 벡터를 의미한다. 학습 말뭉치에서 나타난 단어 수를 $m$개라 할 때, 각 단어는 $1 \times m$로 구성된 벡터 대응된다.

이와 같이 각 단어에 대한 one-hot 벡터를 $k$차원으로 임베딩 시키기 위해, 우리는 $m \times k$ 차원의 표현 행렬 (embedding matrix) $W^{wrd}$를 구성한다. 글의 $i$번째에 단어 $w$가 있는 경우, $i$번째 단어의 단어 표현 벡터(word embedding vector) $x_i$는 다음 식을 통해 계산할 수 있다.

$$x_i = W^{wrd} v^w$$

이를 바탕으로 길이가 $n$인 글 $s$는 아래 수식과 같이 단어 표현 벡터 $x_i$ $n$개를 이어붙인 $n \times k$ 크기를 갖는 행렬 $s$로 표현할 수 있다.

$$s = x_0 x_1 x_2 \ldots x_n$$

표현 행렬 $W^{wrd}$를 오류 역전파 알고리즘(backpro-pagation algorithm)을 통해 학습할 수 있도록 한다.

### 2.2. CNN

CNN의 주요 층은 convolution 계층와 pooling 계층으로 구성되어있다. CNN은 하위 계층부터 상위 계층을 지나면서 점차 추상화된 특징들이 추출된다.

$x_{i:j}$를 글의 i번째부터 j번째까지를 이어 붙인 벡터 $x_i, x_{i+1} \ldots, x_j$ 라 할 때, convolution 계층에서는 다음과 같이 h개의 인접한 단어로 이루어진 행렬 $x_{i:i+h-1}$에 대해 convolution 연산을 수행한다. convolution 연산은 다음과 같다.

$$c_i = f(w \bullet x_{i:i+h-1} + b)$$

이 식에서 $f$는 비선형 함수로써 tanh나 ReLU와 같은 함수들이 사용될 수 있으며. 본 연구에서는 ReLU를 사용

한다. $w$는 가중치 행렬, $b \in R$은 bias에 해당하며, $w$와 $b$는 역전파 과정에서 결정된다. 길이가 $n$인 글에서 $h$개의 인접한 단어로 만들 수 있는 조합은 다음과 같이 나타날 수 있다.

$$\{x_{1:h}, x_{2:h+1} \cdots x_{n-h+1:n}\}$$

convolution 연산 결과로 특징 벡터 $c \in R^{n-h+1}$를 얻는다.

$$c = [c_1, c_2 \cdots, c_{n-h+1}]$$

그 후에는 Pooling 계층에서 max pooling을 사용하여 $c' = \max(c)$ 인 $c'$를 구한다. 이는 가장 중요한 특징들만 남겨놓는 과정이라고 볼 수 있다[6].

본 연구에서는 복수의 문장이 들어간 경우에 대해서 분류하는 모델이 필요하기 때문에 [5]과 달리 CNN을 2단계로 구성하였다. [5]에서는 100글자 내로 구성 된 문장에 대해 감정 분류하는 CNN 분류 모형을 제안하였는데 본 연구에서는 다수의 문장으로 구성 된 글에 대해 논쟁 여부를 분류하는 모형이기 때문에 2단계로 구성하였다.

1단계 CNN에서는 3개의 단어를 나타내는 단어 표현 벡터 $x_{i:i+2}$에 대해 Convolution, Max Pooling 계층을 통과시키기 때문에 출력으로 각 문장에서 중요한 특징이 되는 단어 또는 추상적 문법을 표현하는 벡터가 나타나고, 2단계 CNN에서는 1단계 CNN을 거친 $c'_{i:i+2}$에 대해 Convolution, Max Pooling 계층을 통과시키기 때문에 논쟁이 될 만한 특징이 추출 될 것이라고 판단하였다.

1단계 CNN의 Convolution 특징에서는 특징 벡터의 크기를 500으로 설정하였고, 2단계 CNN의 Convolution 계층에서는 특징 벡터의 크기를 300으로 설정하였다. 2단계 CNN을 통과한 후에는 단조화(Flatten) 한 다음 Dropout을 적용하였다. Dropout은 CNN 신경망 학습 과정에서 가장 성능에 영향을 주는 과적합(Overfitting) 현상을 막아주는 방법이다[7]. 그 후, FCNN을 3번 통과시켜 softmax 값을 계산한다. 첫 번째 통과하는 FCNN에는 입력이 $1 \times 19500$인 크기의 벡터로 들어가고, $1 \times 500$ 크기의 벡터가 나온다. 두 번째 통과하는 FCNN에서는 $1 \times 200$ 벡터가 나온다. 마지막 FCNN에서는 $1 \times 2$ 벡터가 나오고 Softmax 값을 계산한다.

### 2.3. 경사 하강법

신경망 네트워크에서는 에러(예측한 값과 실제 값의 차이)를 줄이기 위해 지속적으로 각 계층에 들어가는 가중치 벡터 $w$ 값을 학습한다. 이러한 학습은 오류 역전파 알고리즘에 의해 이루어지는데 그 방법으로 경사 하강 방법을 주로 사용한다. 일반적으로 고급 경사 하강법에는 Adadelta, Adagrad. RMSProp, Adam 등이 있으며, 본 연구에서는 Adam 기법을 사용하였다[8].

## 3. 실험 및 결과

### 3.1. 학습 말뭉치

SNS 논쟁 유발 글 말뭉치는 페이스북으로부터 수집하였다. 페이스북을 선택한 이유는, 페이스북에는 대학교별로 학생들이 자신의 고민을 익명으로 털어놓을 수 있는 '대나무숲' 이라는 소통 공간이 존재한다. 이 공간에서는 익명성을 빌려 소통하는 공간이기에 선·후배 사이의 악습, 대학 총장 관련 논란 등과 같이 사회적으로 이목이 집중되는 글이 제보되고 있다. 더 나아가, 해당 공간은 대학생뿐만 아니라 일반인도 쉽게 접근이 가능하기 때문에, 학생뿐만 아니라 대학교 주변 상인들이 허위 사실을 유포하는 경우도 있다. 이로 인해 경쟁 업체와 소송을 진행하는 등 SNS 내에서는 다양한 주제로 논쟁이 진행되고 있다. 다양하면서도 대학생 또는 대학교 주변 사람들 관련 논쟁으로만 구성되어있는 대나무숲 페이지 글을 수집하여 논쟁 유발 글, 논쟁을 유발하지 않는 글을 구성하였다. 본 연구에서는 논쟁 유발성 SNS 글은 그 글에 대해서 다양한 의견이 오가는 것으로 삼았으며, 이와 같은 글은 댓글 및 공감(예: 좋아요, 싫어요)이 많은 특징을 보인다.

우리는 페이스북의 77개 대학 대나무숲 페이지로부터 91만개의 글(내용, 좋아요 수, 댓글)을 수집하였다. 수집 된 글은 페이지 생성 시점부터 작성 된 모든 글을 포함하고 있다. 91만개 글 중 댓글 10개 이상, 좋아요 100개 이상, 댓글의 좋아요 수가 10 이상인 글을 논쟁이 나타난 글로 간주하여 논쟁 글 학습 말뭉치 (8135개)를 구성하였다. 논쟁 아닌 글 말뭉치는 논쟁 글 8,135개를 제외한 나머지 글에 대해 임의로 과소 표본추출(Random Under. Sampling)을 적용하여 15,000개를 추출하였다.

추출한 글에 대해 내용의 HTML 태그를 제거한 후 KoNLPy mecab 형태소 분석을 수행하였다. CNN의 입력으로 들어갈 때는 각 단어마다 품사태그가 붙어서 입력되도록 하였다. [9,10]

본 연구에서는 각 글의 최대 길이를 단어 500개로 하였다. 논쟁 유발 글은 많은 사람들의 눈길을 끌고, 좋아요 클릭, 댓글 입력 등의 활동을 유도하는 글이다. 페이스북에서는 일정 길이 이상의 게시글에 대해 "계속 읽기(See More)" 버튼을 두어 모든 내용이 처음부터 보이지 않는다. 또한, 긴 글은 서두에서 논쟁에 대해 언급을 하는 경우가 많다. 이런 특성을 고려할 때, 먼저 출현한 500 단어만을 보고도 이 글이 충분히 SNS 상에서 논쟁이 될지 안 될지 판단할 수 있어야 한다고 가정하였다. CNN의 입력 크기가 고정되어 있어야 하므로, 단어 개수가 500개 미만인 글은 임의의 단어 <PAD>를 붙여서 500개가 되도록 하였다.

### 3.2. 실험 설정

본 연구에서는 하이퍼파라미터에 해당하는 learning rate를 1e-5, dropout rate를 0.5, batch size는 128로

구성하였다. epoch은 15까지 진행하였다.

### 3.3. 결과 및 분석

본 연구에서 제안한 모델의 성능을 검증해보기 위해 SVM과 [5]의 연구에서 쓰인 non-static CNN-rand 모델을 비교 실험하였다. 비교 실험에 쓰인 SVM은 N-gram 기반으로 단어를 추출하고 TF-IDF를 계산한 값을 입력으로 사용하여 Linear Support Vector Machine에 돌렸다. SVM 모형에선 10-겹 교차 검증을 통해 76.30%의 Accuracy, 61.13%의 Recall, 68.17%의 Precision, F1 점수는 0.64가 나왔다. SVM에 비해 본 연구에서 제안한 2-Phase CNN 모형은 F1 점수가 0.0555 올라갔다.

표 1. 3개 모형 실험 결과

| | Acc. | Recall | Prec. | F1 |
|---|---|---|---|---|
| SVM | 0.7630 | 0.6113 | 0.6817 | 0.6446 |
| CNN-rand (Kim) | 0.7933 | 0.6420 | **0.7362** | 0.6859 |
| 2-Phase CNN | **0.7964** | **0.6765** | 0.7255 | **0.7001** |

[5]의 연구에서 제안한 모델은 우리가 제안한 2-Phase CNN 모델보다 Precision은 1.1%P 높게 나왔으나 Recall은 3.45%P 낮게 나왔다.

교내 포교에 대해서 한마디 하겠습니다. 특정종교를 꼭 찝어서 비하하려는 의도는 아닙니다. ...(생략) 그런데 왜 교내에서까지 갈길 바쁜 사람을 붙들고 노상 전도를 하려는 것인지요? ...(생략) 당신이 믿으시는 신이 타인에게는 중요하지 않다는 것을 정녕 모르는 것인지. 개인이 느끼는 하나하나의 사례들이 당신이 믿는 종교의 이미지를 결정짓습니다. ... (생략)

어떤 사람이 아카라카 표를 산다고 글을 올렸길래 판다고해서 나갔는데 자기가 응원단이라면서 기존 가격 이상으로 표를 판 것에 대해서 불이익이 있다는 점이 표 뒤에 있다며 표를 몰수해야겠다고해서 표를 압수당했습니다. 표 매매 행위의 옳고 그름을 떠나 저는 몇가지 문제가 있다고 생각합니다. 첫번째 법리적 해석에 따라 함정수사는 위법행위가 될 수 있습니다. ...(생략) 응원단의 지침은 권력의 남용이 아닌가 생각합니다. 여러분 어떻게 생각하시나요?? 여러분의 생각이 궁금합니다.

그림 4. 논쟁 유발성이 있는 글(True Positive)로 판별한 예 .

아니 우리학교는 도대체 왜 아직도 운동장이 흙바닥인거죠 ;; 초중고도 인조잔디 깔아주는데.... 대학교가.... 하..... 심지어 덕성여대도 인조잔디풋살장이 있음; 여대도 있는데!!! 왜!!! 우리는 흙바닥인가!!!!!!! 총장님 제발..... 이정도는 기본 아닙니까? 하 내 등록금;; 09년에 입학할때부터 잔디구장 만든다 만든다 카더만 벌써 몇년째.. (생략)

전통이랍시고 16이 15에게 성년의날 선물을 꼭 줘야한다고 강요했던 ㅁㅁ학과, 대나무숲에 글 올라오니 해결은 무슨 ㅋㅋㅋㅋ 16 단체 집합시켜서 블랙오티니 뭐니 하면서 기합주네요 세상에 ㅎ 몇년되지도 않은 학과에 이런 이상한 전통은 누가 만들었으며 왜 하는 겁니까 ㅋㅋㅋㅋㅋ 선배들 실습가면 실습물품 사다바쳐, 졸업하면 졸업반지 사다바쳐, 무슨 특별한 날이기만 하면 교수들한테 선물 사다바쳐... 악습 언제끝납니까? 대나무숲 올라오면 또 기합주시나요 선배님들? 이번에는 책상에 무릎꿇고 올라가서 누가 올렸는지 자수할 때 까지 벌세울겁니까?

그림 5. 논쟁 유발성이 없는 글(False Positive)로 판별된 예.

그림 4에서 본 논문이 제안한 2-Phase CNN이 논쟁 유발 글로 판별한 예시를 보면, 문장 구성이 문어체에 가까운 문장임을 볼 수 있다. 반면, 그림 5에서 논쟁 유발 글 중 논쟁 유발 글로 판별하지 못 한 문장들이다. 이 문장들은 구어체에 가까운 단어들이 들어가 있는 경우가 많았다. 이처럼 본 논문이 제안한 CNN 모형에서 문장의 완성도가 높은 문장 구성일수록, 구어체가 적게 들어간 문장일수록 더 잘 판단하는 현상을 볼 수 있었다.

## 4. 결론

본 연구에서는 논쟁적 글을 분류하는데 있어 CNN기반의 새로운 모형을 제안하였다. 그리고 제안하는 모형을 검증하기 위해, 대학생들의 페이스북 소통공간인 대나무숲에서 코퍼스를 수집하고, 댓글이나 대댓글 정보를 이용하여 자동으로 논쟁 여부에 대한 학습 데이터를 구축하였다. 제안하는 모델은 SVM 기반의 기존 문서 분류 모델과 성능 비교를 하였을 때 상당한 성능향상이 있었고 기존의 단순한 CNN에 비해서도 성능향상을 보였다. 이를 통해 우리는 제안 모형을 바탕으로 사전에 SNS 내에서 논쟁을 일으키는 글을 사전에 탐지할 수 있음을 파악했다.

추후에는 SNS 상에서는 대학생들의 소통 공간에서뿐만 아니라 다양한 공간에서 논쟁 데이터를 대상으로 실험하여 제안 논쟁 탐지 모형이 잘 작동하지는 검증해야 할 필요가 있다. 또한 2-phase 이상의 다양한 CNN구조와 데이터의 길이가 매우 긴 경우 어떠한 구조가 적절한지에 대한 연구가 필요하다고 할 수 있다. 추가적으로 논쟁 유발성을 판별하는 것은 SNS의 글뿐만 아니라 각 글에 달린 댓글을 분석을 함께 고려할 필요가 있다. 따라서 향후 이러한 요소들을 고려한 모형으로 발전되어야 한다.

## 참고문헌

[1] 김재영, 김명관, “감정요소를 이용한 SNS 메시지 분류기 구현에 대한 연구.” 한국인터넷방송통신학회논문지, 제11권, 제4호, pp. 217-222, 2012,

[2] Cho, Sang-Hyun, and Kang, Hang-Bong. "Text sentiment classification for SNS-based marketing using domain sentiment dictionary." 2012 IEEE International Conference on Consumer Electronics (ICCE). IEEE, 2012.
[3] Shi, Chenye, Li, Jianhua, Chen, Jieyuan , and Chen, Xiuzhen, "Chinese SNS blog classification using semantic similarity." Computational Aspects of Social Networks (CASoN), 2013 Fifth International Conference on. IEEE, 2013.
[4] Nahar, Vinita, Li, Xue, Pang, Chaoyi, Zhang, Yang, "Cyberbullying detection based on text-stream classification." The 11th Australasian Data Mining Conference (AusDM 2013). 2013.
[5] Kim, Yoon. “Convolutional neural networks for sentence classification.” arXiv preprint arXiv:1408.5882, 2014.
[6] Collobert, J. Weston, L. Bottou, M. Karlen, K.Kavukcuglu, P. Kuksa, “Natural Language Processing (Almost) from Scratch. Journal of Machine Learning Research” 12:2493-2537, 2011.
[7] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, Ruslan Salakhutdinov . “Dropout: A Simple Way to Prevent Neural Networks from Overfitting.” 2014 Journal of Machine Learning Research 15, 2014
[8] Diederik Kingma, Jimmy Ba, “Adam: A Method for Stochastic Optimization”, arXiv preprint arXiv:1412.6980, 2014.
[9] 강승식, "다층 형태론과 한국어 형태소 분석 모델", 제6회 한글 및 한국어 정보처리 학술발표 논문집, pp.140-145, 1994.
[10] 최재혁, "형태소 분석을 통한 한영 자동 색인어 추출", 정보과학회논문지(B), 제23권, 제12호, pp. 1279-1288, 1996.