Multi-channel CNN을 이용한 한국어 감성분석

김민†, 변증현O,‡, 이충희‡, 이연수‡

스탠포드대학교†, ㈜엔씨소프트‡

tomas76@stanford.edu, {jhbyun, forever73, yeonsoo}@ncsoft.com

Multi-channel CNN for Korean Sentiment Analysis

Min Kim†, Jeunghyun ByunO,‡, Chunghee Lee‡, Yeonsoo Lee‡

Stanford University†, NCSOFT Corp.‡

요 약

본 논문은 한국어 문장의 형태소, 음절, 자소를 동시에 각자 다른 합성곱층을 통과시켜 문장의 감성을 분류하는 Multi-channel CNN을 제안한다. 오타를 포함하는 구어체 문장들의 경우에 형태소 기반 CNN으로 추출 할 수 없는 특징들을 음절이나 자소에서 추출 할 수 있다. 한국어 감성분석에 형태소 기반 CNN이 많이 쓰이지만, 본 논문의 Multi-channel CNN 모델은 형태소, 음절, 자소를 동시에 고려하여 더 정확하게 문장의 감성을 분류한다. 본 논문이 제안하는 모델이 형태소 기반 CNN보다 야구 댓글 데이터에서는 약 4.8%, 영화 리뷰 데이터에서는 약 1.3% 더 정확하게 문장의 감성을 분류하였다.

주제어: Multi-channel CNN, 감성 분석, Text Classification, CNN

1. 서론

감성분석(sentiment analysis)은 자연어 분석 기술 중 하나로, 텍스트에 포함된 감성의 극성을 알아내는 기술이다. 인터넷에 올라오는 방대한 양의 댓글, 트윗, 상품평 등을 자동으로 분류하고 필요한 정보를 추출하기 위해서는 텍스트의 감성분석이 필요하고, 감성분석은 사용자 의견 파악이나 미래 선거 결과 예측 등에 쓰이고 있다[1]. 기존의 감성분석에는 Naive Bayes 와 Logistic Regression 등의 전통적인 기계학습 방법이 많이 사용되었으나, 최근에는 딥러닝 기반의 기술들이 많은 분야에서 높은 성능을 기록했다[2].

오피니언 마이닝과 감성분석 기술에 대한 관심이 커지고 있지만, 대부분의 감성분석 연구들은 영어에 초점이 맞추어져 있다[3]. 언어마다 문법 규칙이 다르고, 한국어와 영어는 문법적으로 차이가 있기에 영어 감성분석을 할 때에 사용하는 기술과 모델을 그대로 사용 할 시에 한국어 감성분석에서 부정확한 결과가 나올 수 있다. 한국어의 언어적 특성에 맞는 모델과 기술을 사용해 감성분석을 해야 높은 성능을 기대 할 수 있다[4].

본 논문에서는 한국어 문장의 감성 분류에 효과적인 Multi-channel CNN(Convolutional Neural Network) 모델을 제안한다. 제안하는 모델은 세 개의 다른 합성곱층을 이용하여 형태소, 음절, 자소를 모두 입력 값으로 받는다. 특히 온라인 상의 텍스트의 경우 줄임말이나 맞춤법 오류가 많아서 형태소 단위에서의 정보 손실이 많을 수 있다. 자소와 음절에서 특징 벡터를 추출 함으로서 자소 단위의 오타 및 문법적 오류와 음절 단위의 합성어와 줄임 말에서 손실된 정보를 추출 할 수 있다. 그리고, 학습하지 못한 새로운 단어에 대해서 형태소가 추출하지 못하는 특징을 자소와 음절에서 추출 할 수 있다. 형태소, 자소, 음절에서 찾은 특징벡터를 모두 고려하여 문장의 감성을 분류하는 본 논문의 모델이 기존의 다른 CNN 모델들보다 더 정확하게 문장의 감성을 분류 해 낼 수 있다. 이를 두 종류의 다른 온라인 데이터 셋에서 확인하였다.

2. 관련 연구

Y. Kim 의 CNN 을 이용한 텍스트 분류 연구[5]를 확장하여 한국어에 적용한 연구들이 있다[6,7]. Y. Kim은 논문[5]에서 어절 기반의 CNN 을 제안 하였으나, 이는 한국어에서는 심각한 OOV(Out of Vocabulary) 문제가 나타난다. 왜냐하면 교착어인 한국어의 특성상 무수한 어절이 생성 될 수 있기 때문이다. 어절 대신 형태소나 음절이 입력으로 사용되는 경우가 많다.

한국어에서 의미를 가지는 가장 작은 단위인 형태소가 한국어 텍스트 분류 연구에서는 활발히 사용된다. [6]에서는 네이버 영화 리뷰를 Konlpy 의 Twitter 형태소 분석기를 사용하여 형태소로 나누어 word2vec 을 학습시킨 후 CNN 의 입력 값으로 사용하였다. 전체적인 모델의 구조는 [5]의 모델과 유사한데, 이 모델이 Naive Bayes 모델보다 약 8%정도 더 정확하게 네이버 영화 리뷰 감성을 분류한다.

한국어에서 음절 기반의 CNN 모델을 제안한 연구로는 [7]이 있다. [7]에서는 학습되지 않은 새로운 단어들에 대해서도 예측을 할 수 있도록 형태소 기반이 아닌 음절 기반의 CNN 모델을 사용하였다. 이를 통해서 '구글' 혹은 '갤럭시노트'를 학습한 음절 기반 CNN 모델이 '구글신' 과 '갤노트' 등의 합성어 및 줄임 말에 대해서도 비슷한 예측을 할 수 있어 OOV(Out of Vocabulary) 문제가 형태소 기반 CNN 보다 개선 되었다[7].

본 논문의 기반이 된 Multi-channel CNN 을 이용한 감성 분석 연구로는 [8,9,10]이 있다. [8]의 연구의

경우에는 영어 구어체 문장을 분류 할 시에 단어와 글자(character)를 동시에 사용하는 Multi-channel CNN 이 단어 기반 CNN 혹은 글자 기반 CNN 보다 뛰어남을 확인하였다. [9]에서는 감성분류에서 Word2vec, Glove, Syntactic 등의 word embedding 을 Multi-channel CNN 을 통하여 동시에 사용 할 시에 하나의 word embedding 을 사용하는 것 보다 성능이 향상하였다. 또한 최근에는 [10]에서 한국어 감성분석에서 문장의 자소와 형태소를 동시에 사용하여 영화 리뷰의 감성을 분류하였다. 그러나 [10]에서는 문장의 음절을 고려하지 않았고 온라인 댓글 중에서 비교적으로 정형화가 되어 있고 문장의 길이가 긴 편인 네이버 영화 리뷰 데이터에서만 실험을 하였다. 반면 본 논문에서는 형태소, 자소, 음절 세가지 모두를 이용하는 Multi-channel 모델을 통하여 감성분류 성능을 더 높였을 뿐만 아니라 특징이 다른 두 개의 데이터 셋에 실험을 하였다. 실험을 통하여 본 논문의 최종 모델이 음절을 고려하지 않은 Multi-channel 모델보다 야구 댓글 데이터에서는 약 2.4% 더 높은 성능을 기록하고 네이버 영화 리뷰 데이터에서도 약 1.12% 더 높은 성능을 기록하였다. 또한 본 연구에서 여러 한국어 단위를 조합 한 CNN 에 대한 실험을 통하여 최적화 된 CNN 의 한국어 문장 입력 단위 조합을 제시한다.

3. Multi-channel CNN을 이용한 한국어 문장 분류

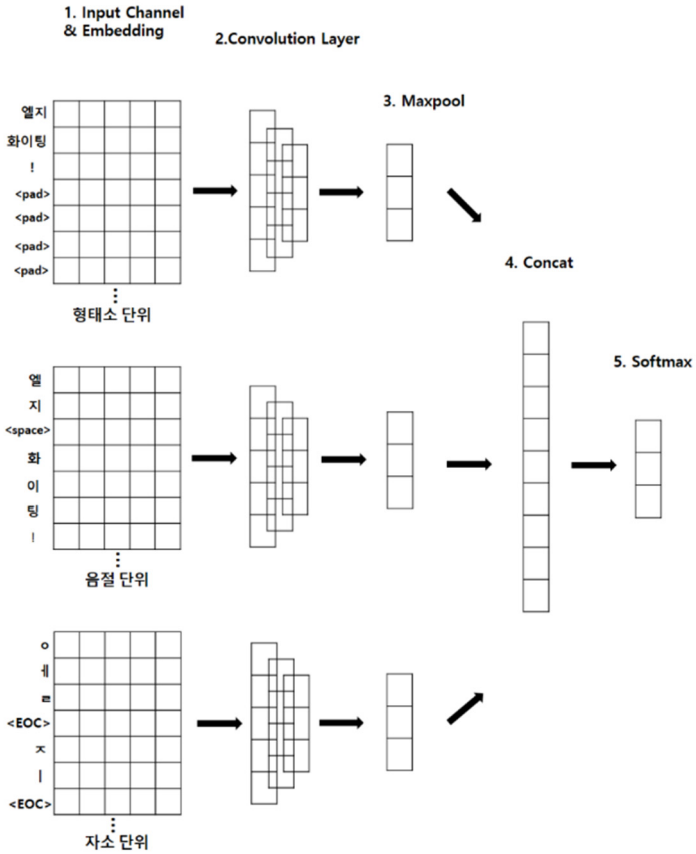

그림 1. Multi-channel CNN 모델 구조

본 논문이 제안하는 Multi-channel CNN 모델은 그림 1 로 형태소, 음절, 자소를 모두 입력값으로 사용하는 형태이다. Multi-channel CNN 모델은 Y. Kim 의 하나의 subword-level 만 입력하는 CNN 모델[5]의 변형이며 형태소, 음절, 자소를 동시에 입력 할 수 있다. 이를 통하여 하나의 subword-level 만 고려하는 게 아니라 여러 가지 subword-level 을 고려 하여 문서를 분류 할 수 있다. 형태소, 음절, 자소로 부터 추출한 특징벡터를 동시에 사용하고자 위의 Multi-channel CNN 모델을 구축하였다.

그림 1 과 같이 Multi-channel CNN 모델은 세 개의 입력 채널을 가지고 있으며 하나의 문장이 형태소, 음절, 자소로 나눠져서 세 개의 채널의 입력으로 사용된다. 그림 1 에서와 같이 “엘지 화이팅!” 이라는 문장은 형태소로 나누면 [“엘지”, “화이팅”, “!”]이 되고 이는 첫 번째 채널로 들어가게 된다. 또 위의 문장을 음절로 나눌 시에는 [“엘”, “지”, “<space>”, “화”, “이”,”팅”, “!”] 이 되며 이를 두 번째 채널의 입력으로 사용한다. 단어와 단어 사이에는 <space> 토큰을 넣어 주었다. 마찬가지로 위의 문장을 자소로 나누어 [“ㅇ”,”ㅔ”,”ㄹ”,”<eoc>”, “ㅈ”,”ㅣ”, ……]를 마지막 채널의 입력으로 사용한다. 음절과 음절 사이에는 <eoc> 토큰을 넣어 주었다. 형태소 기반 채널과 음절 기반 채널은 문장의 최대 길이를 50 토큰으로 제한하여서 더 긴 문장은 잘라내고 짧은 문장은 padding 으로 길이를 채워주었다. 자소 기반 채널의 경우에는 최대 길이를 150 토큰으로 제한하여서 위와 같이 잘라내거나 padding 을 입력하였다. 각각의 채널에서 embedding 층을 거쳐서 입력 값들의 embedding 을 학습하게 된다. Facebook Research 에서 제공한 pre-trained 된 한국어 FastText embedding 을 사용하여서 실험해 보았으나 큰 성능의 향상이 없어서 이 모델에서는 300 차원으로 랜덤 초기화된 embedding 들을 사용하였다. Embedding 층을 거친 결과 값은 세 개의 다른 크기의 filter window 로 구성된 합성곱층의 입력으로 사용된다. 합성곱층의 출력은 Max-Pooling 을 거쳐서 하나의 행렬로 연결되어 Softmax 층을 거쳐서 문장을 분류한다.

4. 데이터 및 실험 방법

4.1 데이터

표 1: 데이터 셋 통계

	야구 댓글	영화 리뷰
평균 문장 길이(글자)	25.2	35.24
감성 개수	3	2
학습 단위 크기	15846	150000
테스트 단위 크기	1761	50000
형태소 Vocab 크기	17313	94747
음절 Vocab 크기	1777	3006
자소 Vocab 크기	156	162

표 2: 야구 댓글 감성 예제

문장	감성
“기아우승 확신 100%”	긍정
“시즌끝나고 걍 푹쉬어”	부정
“잘한다”	중립
“잘했다 현식아!”	긍정
“웃돈주고 21 만원에 샀다. 개비오야, 3 시간을매달려도 티켓성공못하겠다.”	부정
“부상부위도 특이하네”	중립

야구 댓글: 실험 데이터로 다음 야구 기사 댓글 감성 말뭉치 데이터를 사용하였다. 이 데이터는 사내에서 제작한 데이터 셋이며 댓글들에 ‘긍정’, ‘부정’, ‘중립’의 감성 태깅이 되어 있다. 긍정, 부정, 중립은 1:1:1 로 학습 단위와 테스트 단위에 균등하게 분포되어 있다. 표 2 의 예제를 보면 알 수 있듯이 문장들은 구어체로 이루어져 있으며 많은 경우에 띄어쓰기와 문법적인 오류를 포함하고 있고 문장의 길이 또한 길지 않은 경우가 많다. ”잘했다 현식아!” 같은 경우에는 긍정으로 태깅이 되어 있지만 비슷한 “잘한다”는 중립으로 태깅이 되어 있다. “잘한다” 처럼 긍정으로 쓰일 수도 있지만 어떤 상황에서는 비꼬는 표현으로 부정 혹은 중립으로도 쓰일 수 있는 표현들이 있어서 감성 분석에 어려움이 있을 수 있다. 또한 “부상부위도 특이하네”와 같이 사람이 판단하기에도 중립인지 부정인지 애매한 감성을 가지고 있는 문장들도 다수 존재한다.

표 3: 영화 리뷰 감성 예제

문장	감성
“아 더빙.. 진짜 짜증나네요 목소리”	부정
“너무재밋었다그래서보는것을추천한다”	긍정
“교도소 이야기구먼 ..솔직히 재미는 없다..평점 조정”	부정
“액션이 없는데도 재미 있는 몇안되는 영화”	긍정

영화 리뷰: 공개된 네이버 영화 리뷰 코퍼스 v1.0(https://github.com/e9t/nsmc) 데이터를 이용하여서 모델들을 학습시키고 테스팅 하였다. 감성 태깅의 경우 긍정 부정 두 가지로 균등하게 학습 단위와 테스트 단위에 분포되어 있다. 이 코퍼스는 전처리 과정에서 평점 1-4 사이의 리뷰들은 부정으로 태깅이 되었고, 평점 9-10 사이의 리뷰들은 긍정으로 태깅이 되어서 감성 분류가 상대적으로 명확하다. 그리고, 야구 댓글 데이터에 비해서 문장의 길이가 더 길고 학습 단위가 훨씬 크다. 또한, 오픈소스로 공개된 코퍼스라서 다른 연구들과의 성능 비교가 가능하다.

4.2 실험 모델

어절 기반 CNN(Baseline): Y. Kim 의 CNN 모델[5]을 사용하였다. 어절 단위로 문장을 분리하여서 CNN 의 입력 값으로 사용하였다.

형태소 기반 CNN: 어절 기반 모델과 동일하며 어절이 아닌 형태소를 입력 값으로 사용하였다. 사내의 Bidirectional LSTM 과 CRF 를 이용한 음절 단위 형태소 분석기를 사용하였다[11]. 여러 형태소 분석기를 실험해본 결과 사내 형태소 분석기를 사용하면 Konlpy 의 Twitter 혹은 Kkma 를 사용한 형태소 기반 CNN 에 비해 약 2%정도 더 정확하게 야구 댓글 데이터의 감성을 구별 할 수 있었다.

음절 기반 CNN: 위의 모델들과 동일하며 음절을 입력 값으로 사용하였다. 전처리 과정에서 어절과 어절 사이의 경우 <space> 토큰을 입력하였다.

자소 기반 CNN: 위의 모델들과 동일하며 자소를 입력 값으로 사용하였다. 전처리 과정에서 문장을 자소로 분해 할 시에 음절과 음절 사이에는 <eoc> 토큰을 입력하고 어절과 어절 사이에는 <space> 토큰을 입력하였다.

Multichannel CNN(형태소 + 음절 + 자소): 본 논문에서 제안하는 최종 모델이며 형태소, 음절, 자소를 모두 사용하는 Multi-channel CNN 모델이다. 전처리는 sub-word 별로 위의 모델들과 동일하다. 이 최종 모델 이외에도 형태소, 음절, 자소 중 2 개의 조합으로 이루어진 Multi-channel CNN 들도 비교를 위해서 실험하였다

4.3 실험 환경 및 모델 파라미터

표 4: 모델 별 최종 파라미터

파라미터	형태소 기반 CNN	음절 기반 CNN	자소 기반 CNN	Multi-channel CNN
Feature map size	100	300	300	100,300,300 (형태소,음절,자소)
L2 constraint	1	0	0	0
Batch size	30	128	30	30
Learning Rate	1e-5	1e-5	1e-4	1e-4

실험 모델들의 파라미터는 야구 댓글 데이터 검증 셋에 대한 그리드 탐색(Grid Search)를 사용하여서 정하였다. 그리드 탐색에 사용한 파라미터의 범위는 다음과 같다.

Feature map size: [50, 100, 300]
L2 constraint: [0, 0.1, 1, 3]
Batch size: [30, 60, 128, 256]
Learning rate: [1e-5, 1e-4, 1e-03, 1e-02]

표 4 에 나와있는 값들이 그리드 탐색에서 찾은 최종 파라미터이다. Epoch 은 100 으로 두고 early-stopping 을 해주었다. Dropout 은 공통적으로 0.5 로 적용해 주었고, Adam optimizer 를 사용하였다. 모든 모델에 Filter window 는 3, 4, 5 의 세 개의 다른 크기를 사용하였다.

4.4 실험 평가 기준

야구 댓글 데이터와 영화 리뷰 데이터에서 감성이 균등하게 분포되어 있어서 전체적인 모델의 성능을 평가할 때에는 정확성(Accuracy)를 사용하였다. 또한 각 감성에 대한 분류 F1 score 도 실험 결과에 포함하여 감성 별 분류 성능도 나타내었다.

표 5: 감성분석 실험 결과

Model	야구 댓글				영화 리뷰		
	Accuracy	긍정(f1)	중립(f1)	부정(f1)	Accuracy	긍정(f1)	부정(f1)
CNN (어절) Baseline	45.03%	55.85%	32.41%	39.10%	79.17%	80.68%	77.40%
CNN (형태소)	62.12%	67.56%	57.86%	61.67%	85.02%	85.14%	84.90%
CNN (음절)	59.51%	66.03%	49.48%	62.77%	84.92%	84.83%	85.01%
CNN (자소)	59.23%	63.62%	53.95%	61.33%	84.4%	84.18%	84.63%
Multi-channel CNN (음절+자소)	60.25%	66.43%	50.48%	63.64%	84.96%	85.02%	84.90%
Multi-channel CNN (형태소+음절)	63.14%	69.94%	57.79%	62.48%	85.33%	85.18%	85.48%
Multi-channel CNN (형태소+자소)	64.56%	70.08%	58.31%	65.37%	85.15%	85.38%	84.92%
Multi-channel CNN (형태소+음절+자소)	**66.95%**	71.46%	60.88%	68.91%	**86.27%**	86.42%	86.11%

표 6: 문장 분류 예시

Sentence	정답 Label	Multi-channel CNN (형태소+음절+ 자소)	CNN (형태소)	CNN (음절)	CNN (자소)
“기다된다”	긍정	긍정	중립	긍정	긍정
“우리팀은 연패해도 선수들 분위기 너무 좋당께....”	긍정	긍정	부정	긍정	긍정
“역대 최고의 먹튀 이)돼호,,”	부정	부정	중립	부정	긍정
“악플이 없네...역시 성공한 인생이시네요”	긍정	긍정	긍정	부정	부정
“국내용”	부정	중립	중립	중립	중립
“축구에 수아레즈”	부정	중립	중립	긍정	긍정

5. 실험 결과

모델 별 CNN의 감성분석 결과가 표 5에 정리되어 있다. Baseline 인 어절 기반 CNN은 다른 모델들에 비하여 성능이 떨어진다. 본 논문이 제안 하는 형태소, 음절, 자소를 동시에 사용하는 Multi-channel CNN 모델이 야구 댓글 데이터와 영화 리뷰 데이터 모두에서 제일 높은 정확도로 문장의 감성을 분류하였다.

형태소가 의미를 가지는 최소 단위이기 때문에 음절이나 자소 기반 CNN보다 감성분류 성능이 더 높았다. 그러나 형태소 기반 CNN이 올바르게 분류하지 못한 문장들을 음절 기반 CNN이나 자소 기반 CNN이 올바르게 분류 한 경우도 있다. 표6 에는 실제로 모델들이 분류한 야구 댓글 문장들의 예시이다. “기다된다” 같은 경우에는 형태소로 나누면 “기다되, ㄴ다” 로 나눠질 수 있다. 그러나 “기다된다”는 “기대된다”의 오타 이므로 “기다되” 라는 형태소에는 큰 의미가 없다. 하지만 음절 기반 및 자소 기반 CNN은 오타를 포함한 문장에서도 특징 벡터를 추출 할 수 있기 때문에 “기다된다” 라는 문장을 “기대된다” 혹은 “기대”와 연관 지어 올바르게 분류한다. “우리팀은 연패해도 선수들 분위기 너무 좋당께…” 라는 문장에서는 ‘연패하’ 라는 부정적인 감성을 가진 형태소가 있어서 형태소 기반 CNN은 문장을 부정이라고 감성을 분류한다. ‘좋당께’ 라는 단어를 형태소로 분해 할 수 있었다면 이 문장을 긍정으로 구별 할 수도 있었을 것 같으나, 너무 구어체적인 문장이어서 형태소로 올바르게 분해하지 못한다. 그러나 음절과 자소 기반 CNN에서는 ‘좋당께’ 라는 단어를 ‘좋다’ 혹은 ‘좋당’ 이랑 연관 지어서 올바르게 이 문장을 긍정으로 구분하였다. Multi-channel CNN은 형태소뿐만 아니라 음절과 자소를 동시에 이용하기에 위의 문장들의 감성을 정확하게 분류한다. 음절과 자소를 같이 이용함으로써 형태소 기반 CNN의 큰 문제점인 OOV(Out of Vocabulary)문제가 많이 해결 되었다. 특히 구어체 문장이나 오타를 포함하는 문장에서 Multi-channel CNN이 효과적임을 분류 예시들을 통해서 확인 할 수 있다.

감성 별 F1 score 실험 결과를 비교해 보면 형태소 기반 CNN에 비해서 최종 Multi-channel CNN의 부정 감성 분류 성능이 많이 높아졌다. 이를 야구 댓글 데이터와 영화 리뷰 데이터 모두에서 확인 할 수 있는데, 이는 부정적인 댓글들에 형태소로 분해하기 힘든 구어체 및 합성어가 다른 댓글들보다 더 많이 포함되어서라고 판단된다.

형태소, 음절, 자소 세가지를 동시에 사용하는 Multi-

channel CNN이 형태소와 음절 혹은 형태소와 자소 만을 사용하는 Multi-channel CNN 보다 성능이 높은 것으로 보아 세가지 subword-level을 동시에 사용 하였을 시에 더 다양한 특징벡터를 고려하면서 더 정확하게 문장의 감성을 분류 할 수 있다. “역대 최고의 먹튀 이)돼호,,” 라는 문장에서 형태소 기반이나 자소 기반 CNN 에서는 부정적인 특징벡터를 추출하지 못하여서 잘못 분류하였으나 음절 기반 CNN 에서는 음절들 사이의 특징을 고려하여 문장을 부정으로 올바르게 분류하였다. 또한 형태소와 음절 기반 CNN에서 추출하지 못한 특징들을 자소 기반 CNN에서 추출 할 수도 있기에 세 종류의 subword-level을 전부 사용하였을 시에 감성분류 성능이 제일 높다.

야구 댓글에서 문장의 문맥을 모르면 정확하게 감성을 분류하기 힘들 “국내용” 같이 함축적이거나 모호한 문장들은 실험 모델들이 중립으로 분류하는 경우가 많아서 중립 감성에 대한 성능이 대체적으로 긍정이나 부정보다 낮다. 중립 감성이 없는 영화 리뷰 데이터에 대해서는 최종 Multi-channel CNN 모델은 86.7%까지 정확하게 리뷰의 감성을 분류한다. Multi-channel CNN을 사용 하였을 시에 감성분류 성능의 향상 폭이 영화 리뷰 데이터에서 보다 야구 댓글 데이터에서 더 크다. 영화 리뷰 데이터에서는 문장의 길이도 길고 학습 단위도 크기 때문에 형태소로도 대부분의 문장의 감성을 분류 할 수 있으나, 야구 댓글 데이터에서는 짧고 구어체적인 문장이 많고 중립 감성을 포함하여 형태소만 사용해서는 분류하기 힘든 문장들이 많아서 인 것으로 판단된다. 두 가지의 다른 특징을 가진 데이터 셋에 모델을 검증함으로써 본 논문이 제시하는 모델이 온라인 데이터 감성분류에 효과적임을 확인하였다.

6. 결론

본 논문에서 여러 실험을 통해서 한국어 구어체 감성분석에 효과적인 형태소, 자소, 음절 기반 Multi-channel CNN을 제안하였다. OOV(Out of vocabulary) 문제를 가지는 형태소 기반 CNN의 문제점을 해결하며 음절과 자소에서 추출한 특징벡터를 형태소 기반의 특징벡터와 상호보완적으로 사용할 수 있음을 확인하였다. 기존에 많이 쓰이는 형태소 및 음절 기반 CNN보다 높은 감성분석 정확도를 보여주어 한국어 구어체를 분류하는 다른 연구에도 활용 가능성이 있음을 확인하였다.

참고문헌

[1] H. Lee and S. Lee, 감성 분석 및 감성 정보 부착 시스템 구현, KIPS Tr. Software and Data Eng, Vol.5, No.8, pp. 377-384, 2016.

[2] X. Glorot, A. Bordes and Y. Bengio, Domain adaptation for large-scale sentiment classification: a deep learning approach, ICML'11 Proceedings of the 28th International Conference on International Conference on Machine Learning, pp. 513-520, 2011.

[3] M. Bautin, L. Vijayarenu and S. Skiena, International Sentiment Analysis for News and Blogs, ICWSM, 2008.

[4] H. Jang and H. Shin, Language-Specific Sentiment Analysis in Morphologically Rich Languages, COLING '10 Proceedings of the 23rd International Conference on Computational Linguistics: Posters, pp. 498-506, 2010.

[5] Y. Kim, Convolutional Neural Networks for Sentence Classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, pp. 1746-1751, 2014.

[6] W. Kim and K. Park, 합성 곱 신경망을 이용한 한글 텍스트 감성 분류기 설계, 한국정보과학회 2017 년 한국컴퓨터종합학술대회 논문집, pp. 642-644, 2017.

[7] S. Choi et al., A Syllable-based Technique for Word Embeddings of Korean Words, Proceedings of the First Workshop on Subword and Character Level Models in NLP, pp. 36-40, 2017.

[8] J. Park and P. Fung, One-step and Two-step Classification for Abusive Language Detection on Twitter, Proceedings of the First Workshop on Abusive Language Online, pp. 41-45, 2017.

[9] Y. Zhang, S. Roller and B. Wallace, MGNC-CNN: A Simple Approach to Exploiting Multiple Word Embeddings for Sentence Classification, Proceedings of NAACL-HLT 2016, pp. 1522-1527, 2016.

[10] K. Mo et al., Text Classification based on Convolutional Neural Network with Word and Character Level, Journal of Korean Institute of Industrial Engineers Online, 2018.

[11] H. Kim et al., 품사 분포와 Bidirectional LSTM CRFs 를 이용한 음절 단위 형태소 분석기, 제 28 회 한글 및 한국어 정보처리 학술대회 논문집, 2016.