5조 최종발표

Convolutional Neural Networks와 Weakly Supervised Learning을 활용한 문장 Attention 모델 연구

산업경영공학부 DSBA연구실
김창엽, 이기창, 서승완, 정재윤

목차

- 프레임워크
- 데이터 소개
- 실험 설계
- 실험 결과
- 요약 및 향후계획

프레임워크

01

02

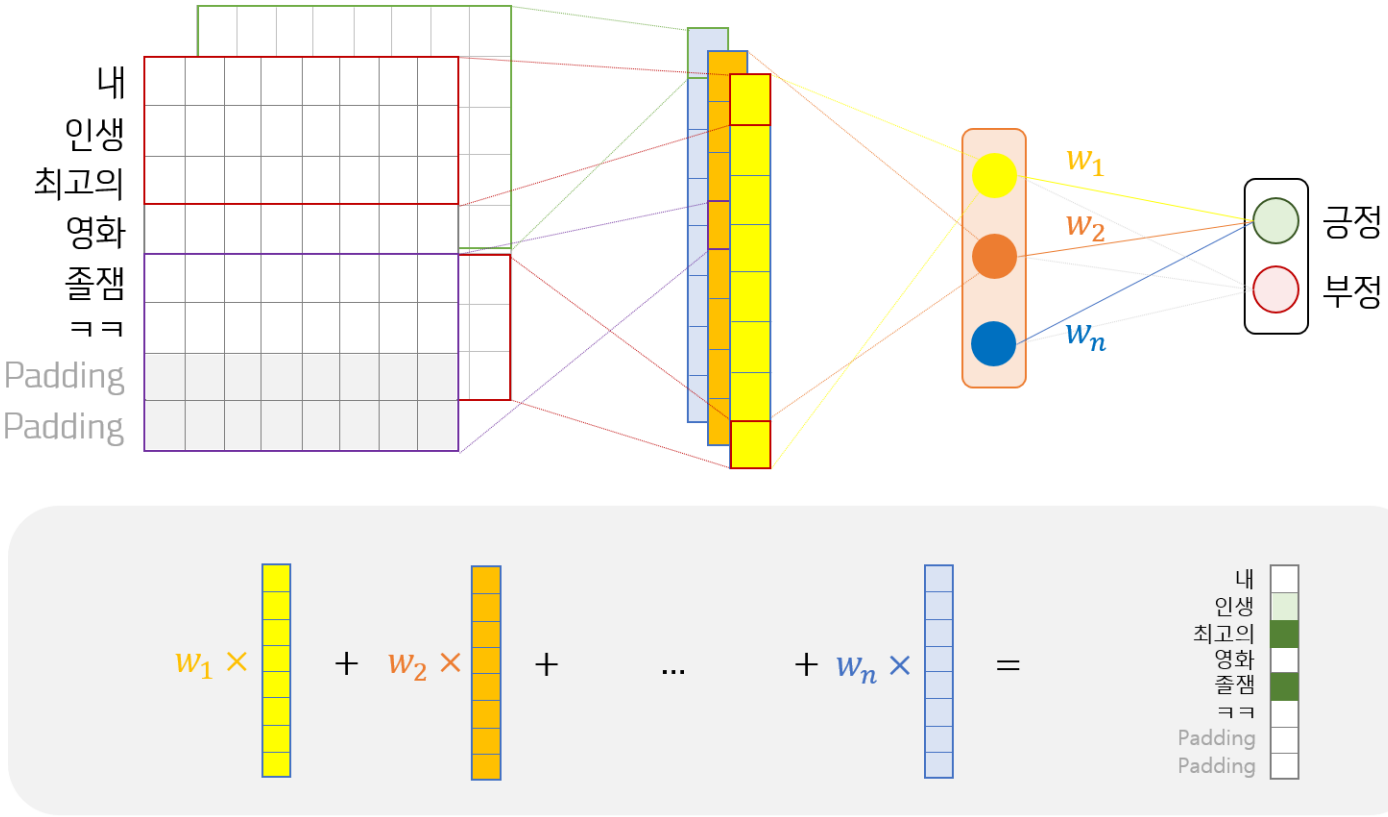
03

04

05

06

❖ 연구목적 및 모델 개요



데이터 소개

01

02

03

04

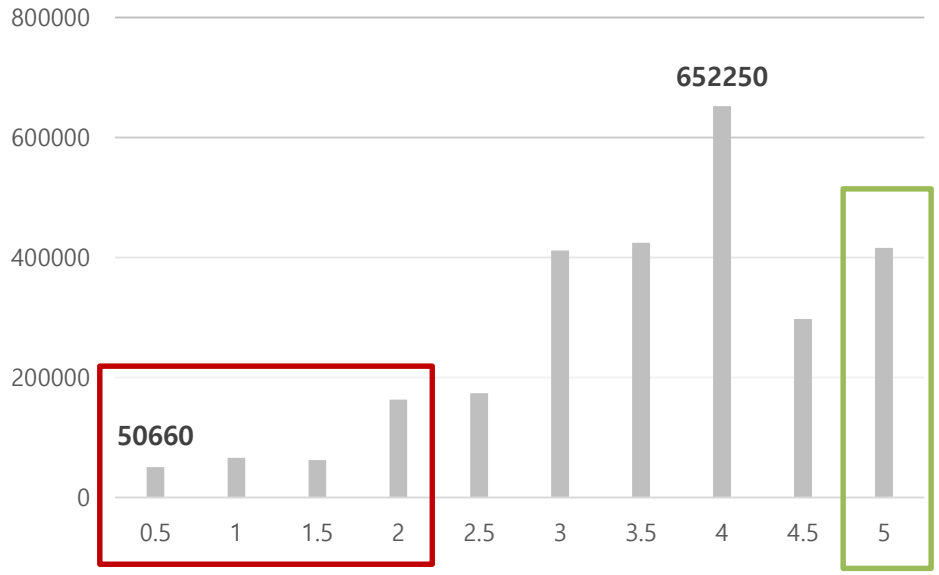
05

06

❖ 한국어 영화 리뷰

- 영화 사이트 '왓챠'에서 수집 (2012. 11~2016. 7)
- 전체 271만7668건 → 72만2813건(중간 범주 제거)
- 학습/검증데이터를 7:3 비율로 분리

❖ 평점 분포 및 전처리



데이터 소개

01

02

03

04

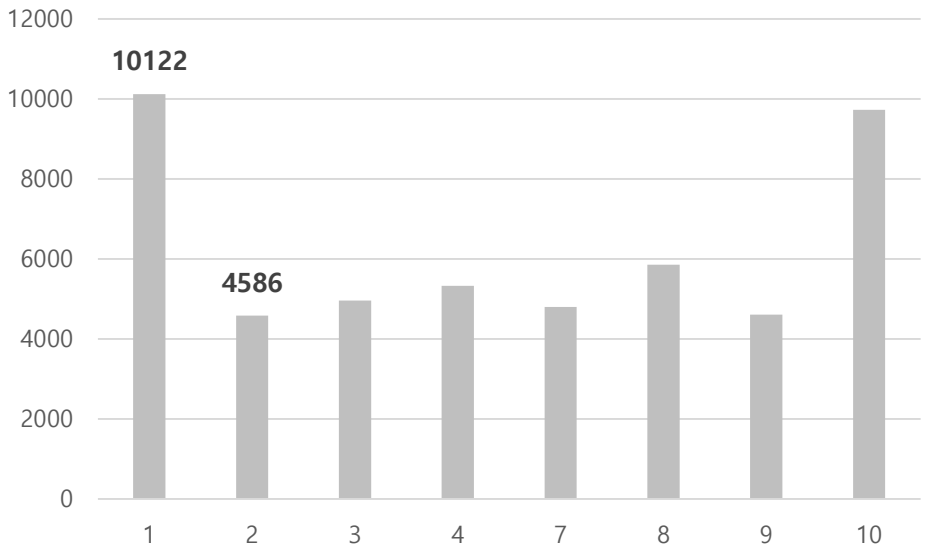
05

06

❖ 영어 영화 리뷰

- IMDB 데이터셋
- 전체 5만건
- 학습/검증데이터를 7:3 비율로 분리

❖ 평점 분포



데이터 소개

01

02

03

04

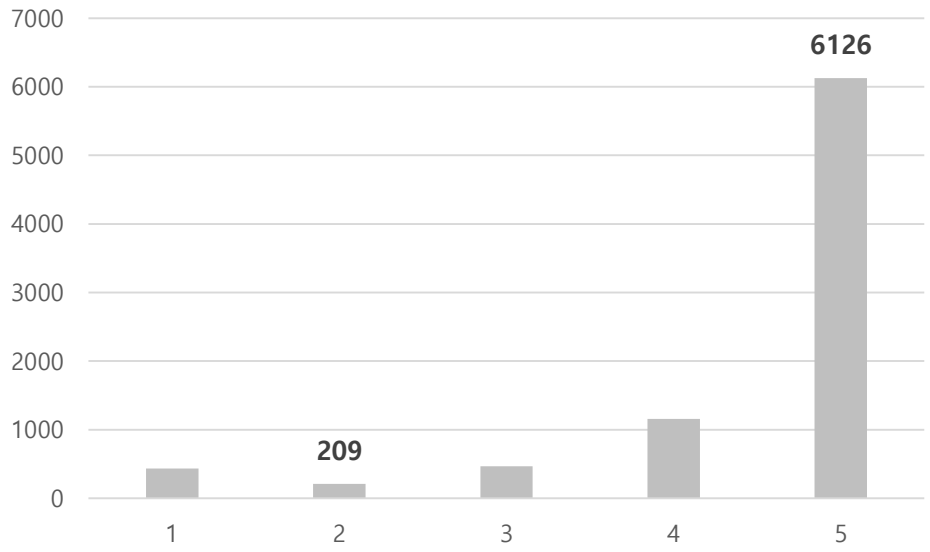
05

06

❖ 한국어 상품 평가

- 소셜 커머스 '쿠팡'에서 수집
- 전체 8393건
- 학습/검증데이터를 7:3 비율로 분리

❖ 평점 분포

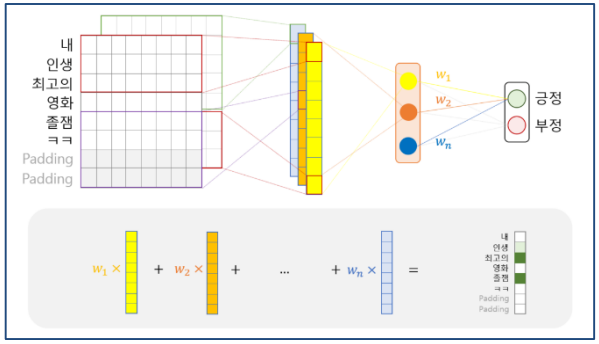


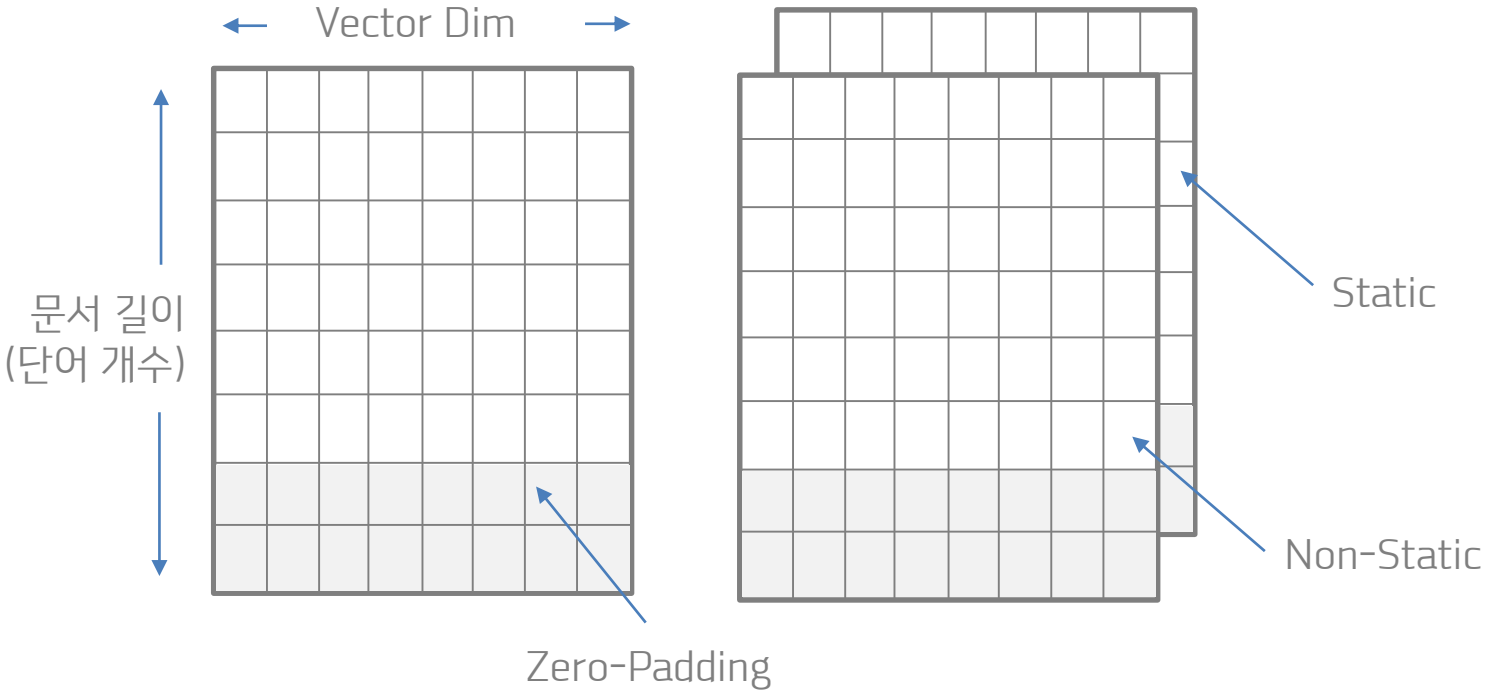
실험 설계

01
02
03
04
05
06

❖ Input

- Word2Vec vs Random
- Static vs Non-Static
- Single Channel vs Multi-Channel

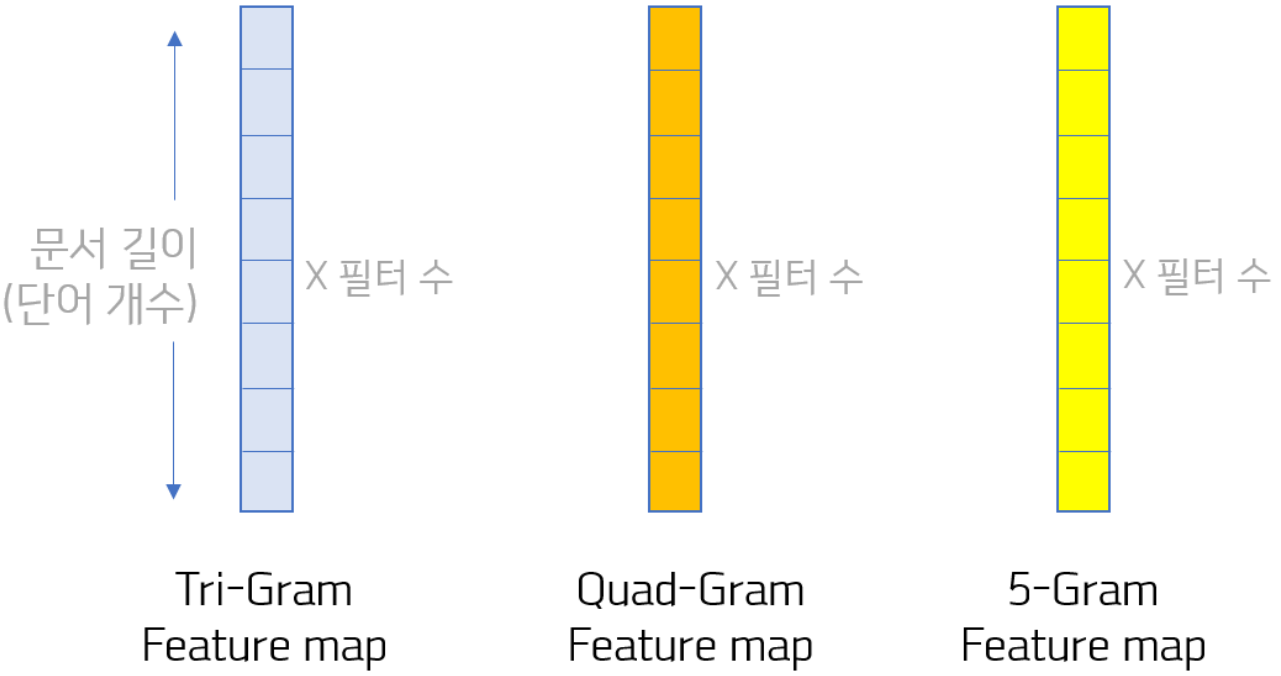


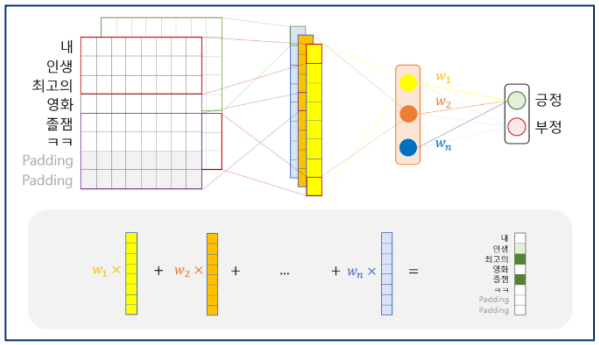


실험 설계

01
02
03
04
05
06

❖ Filter & Feature Map

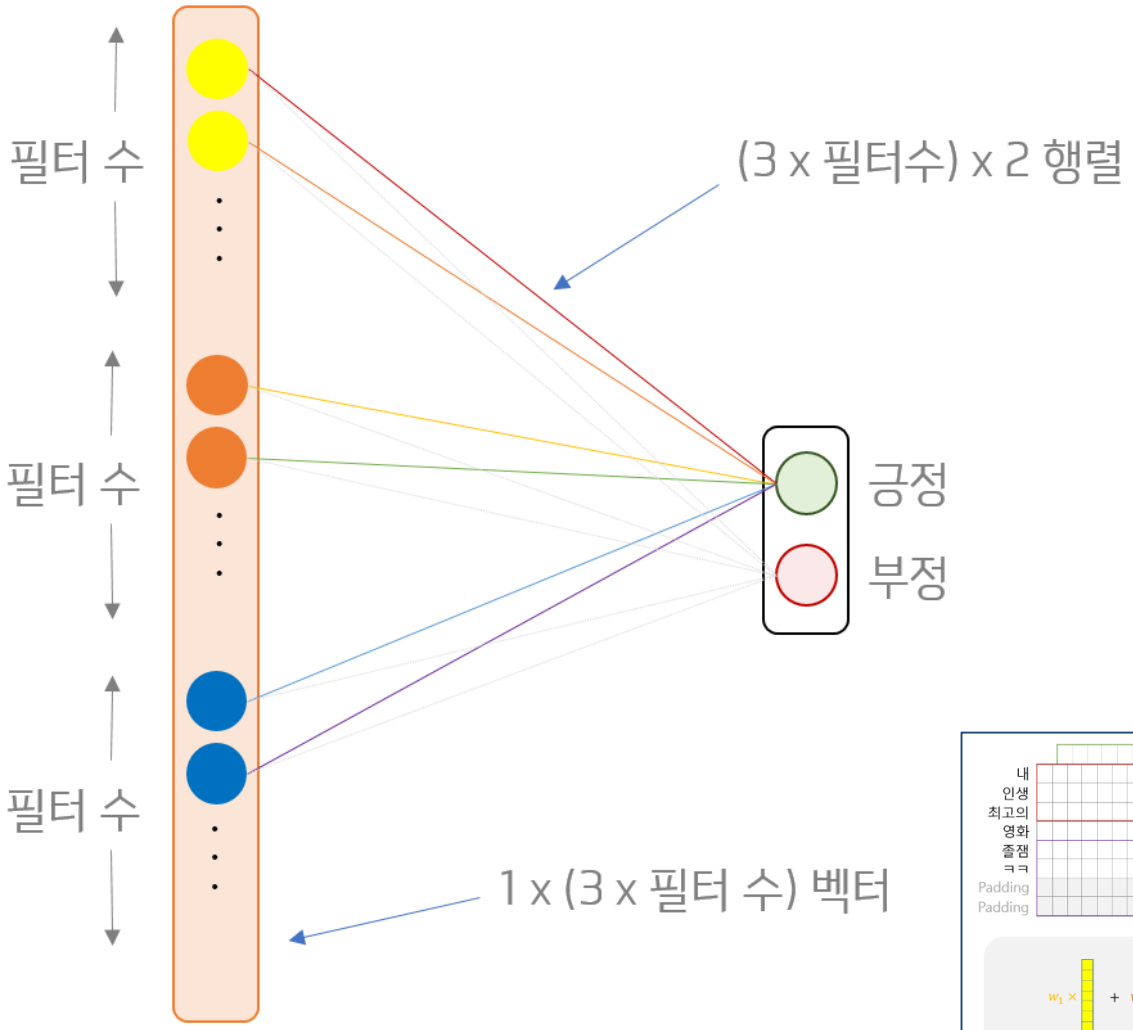


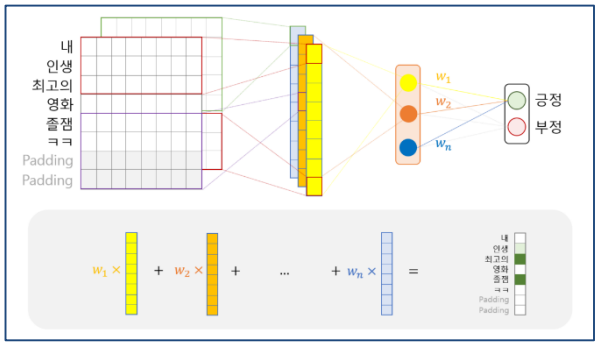


실험 설계

01
02
03
04
05
06

❖ FC Layer

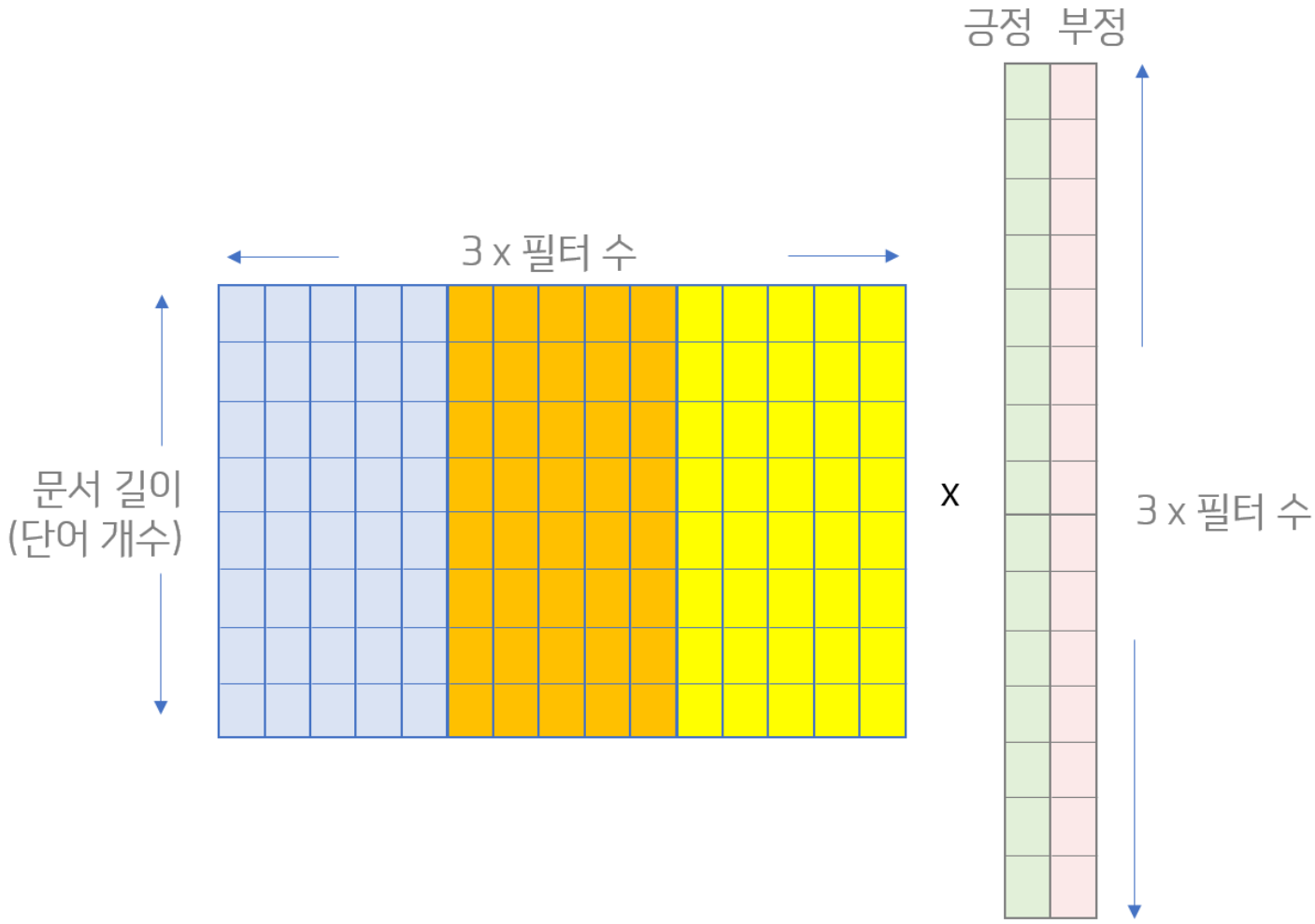




실험 설계

01
02
03
04
05
06

❖ 결과 산출



실험 설계

01

02

03

04

05

06

❖ 실험 환경

- MS Windows 10, Tensorflow
- Intel i5-4690 @3.5GHz
- Nvidia GTX-1080 (8Gb)

❖ 하이퍼 파라메터

- Word2Vec : SG, 100차원, window=3, min_count=100
- 필터 종류 : 3, 4, 5, 필터 수 : 128개씩 총 384개
- 문서 길이(단어 개수) : 100
- Dropout Rate : 0.5, L2 regularization Lambda : 0.1
- Batch size : 64, 에폭 수 : 10

❖ 실험 종류

- 2-channel : Word2Vec Static + Dynamic
- 1-chaneel : Word2Vec Static, Word2Vec Dynamic, One-hot-Vector

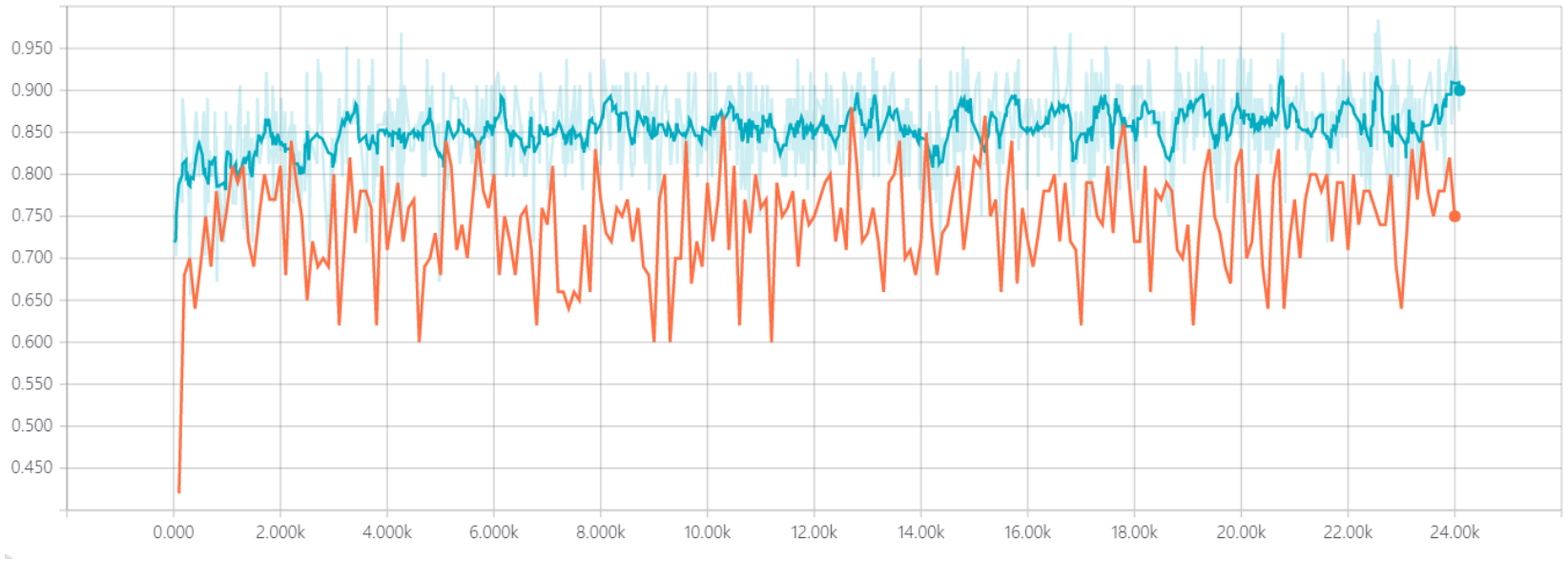
❖ 데이터

- 한국어 영화 리뷰
- 영어 영화 리뷰
- 한국어 상품 평가

실험 결과

01
02
03
04
05
06

❖ 학습 그래프 (WATCHA, MultiChnnel)



실험 결과

01
02
03
04
05
06

❖ 분류 정확도

Model	MR	SST-1	SST-2	Subj	TREC	CR	MPQA
CNN-rand	76.1	45.0	82.7	89.6	91.2	79.8	83.4
CNN-static	81.0	45.5	86.8	93.0	92.8	84.7	**89.6**
CNN-non-static	**81.5**	48.0	87.2	93.4	93.6	84.3	89.5
CNN-multichannel	81.1	47.4	**88.1**	93.2	92.2	**85.0**	89.4

Yoon Kim(2014). "Convolutional Neural Networks for Sentence Classification"

Training ACC	WATCHA	IMDB	COUPANG
CNN-rand	0.97800	0.98600	0.93400
CNN-static	0.83106	0.80468	0.77129
CNN-non-static	0.86576	0.89414	0.91243
CNN-multichannel	0.86570	0.89149	0.90161

Test ACC	WATCHA	IMDB	COUPANG
CNN-rand	0.77900	0.86100	0.71000
CNN-static	0.72033	0.77929	0.70000
CNN-non-static	0.75250	0.81714	0.75000
CNN-multichannel	0.75283	0.80571	0.79000

실험 결과

01
02
03
04
05
06

❖ 스코어 상위 단어 (WATCHA)

긍정

Static	Non-static	Multi	Rand
영화	영화	영화	영화
수	수	수	이
최고의	최고의	최고의	너무
정말	그	그	스토리
그	정말	정말	왜
다시	가장	다시	수
가장	다시	가장	그냥
내	있는	잘	영화를
있는	잘	있는	더
그리고	이	너무	없다
이	너무	이	없는
잘	그리고	내	이런
더	더	더	다
진짜	내	그리고	이야기
모든	진짜	진짜	영화가
너무	모든	모든	영화는
좋다	최고	한	그
아름다운	한	최고	것
또	좋다	좋다	내가
마지막	영화를	마지막	진짜
최고	마지막	영화를	본
본	보고	보고	내
보고	또	또	이렇게
내가	아름다운	아름다운	정말
꼭	본	꼭	재미없
봐도	꼭	본	잘
완벽한	완벽한	완벽한	배우들
영화를	봐도	좋았다	좀
나의	좋았다	봐도	보는
인생	영화가	것	캐릭터

부정

Static	Non-static	Multi	Rand
너무	너무	너무	영화
영화	영화	영화	이
왜	그냥	그냥	너무
그냥	왜	왜	스토리
좀	영화는	영화는	더
없고	좀	좀	왜
영화는	이런	이런	수
없다	없는	이	그냥
다	이	없다	영화를
이	없다	없고	없다
없는	없고	없는	내가
느낌	차라리	차라리	없는
이런	그나마	안	이런
안	느낌	별	다
스토리	그래도	그나마	그
뻔한	별	그래도	영화는
영화를	안	이건	영화가
뭐	다	다	것
별	이건	느낌	이야기
모르겠다	전혀	전혀	정말
차라리	모르겠다	뻔한	이렇게
영화가	뻔한	영화를	진짜
이건	별로	모르겠다	좀
건	스토리	스토리	보는
그나마	최악의	별로	재미없
많이	뭐	이렇게	시리즈
무슨	굳이	최악의	내
재미도	더	뭔가	주인공
별로	뭔가	뭐	모르겠
전혀	아깝다	굳이	한

Model	ACC
CNN-static	0.72033
CNN-non-static	0.75250
CNN-multichannel	0.75283
CNN-rand	0.77900

실험 결과

01 02 03 04 05 06

❖ 스코어 상위 단어 (IMDB)

긍정

Static	Non-static	Multi	Rand
is	is	is	the
great	and	and	and
and	a	a	of
a	great	very	this
very	very	It	a
well	it	it	to
it	It	great	is
It	excellent	s	that
best	best	best	film
of	s	excellent	movie
I	an	This	in
was	the	an	The
s	most	most	with
excellent	was	was	was
film	in	I	for
most	I	this	I
This	this	of	it
good	really	the	comedy
wonderful	This	really	you
loved	of	in	see
movie	to	movie	not
an	movie	enjoyed	but
also	The	The	even
as	enjoyed	A	one
this	with	with	act
favorite	amazing	to	really
story	A	film	have
A	film	also	as
still	loved	loved	on
my	but	well	show

부정

Static	Non-static	Multi	Rand
bad	worst	worst	the
The	is	is	this
is	The	The	movie
acting	bad	bad	of
plot	a	This	that
this	this	A	to
And	awful	awful	and
was	of	of	a
Just	and	and	in
awful	was	to	is
So	to	was	The
Movie	the	movie	was
boring	I	the	film
of	boring	terrible	for
a	terrible	boring	it
terrible	movie	I	I
script	just	just	with
no	poor	It	you
worst	It	poor	not
horrible	waste	so	act
stupid	so	waste	see
waste	horrible	are	even
poor	but	horrible	have
an	very	acting	on
to	are	very	as
the	for	for	completely
that	s	but	one
I	really	s	really
t	film	it	some
low	it	script	but

Model	ACC
CNN-static	0.77929
CNN-non-static	0.81714
CNN-multichannel	0.80571
CNN-rand	0.86100

실험 결과

❖ 스코어 상위 단어 (COUPANG)

긍정

Static	Non-static	Multi	Rand
잘	좋아요	좋아요	로켓배송
항상	항상	항상	쿠팡맨
좋아요	너무	너무	너무
너무	잘	빠르고	물
가격도	배송도	좋네요	잘
배송도	가격도	가격도	동원샘
빠르고	빠르고	물	찌그러
좋네요	쿠팡맨	배송도	물을
감사합니다	감사합니다	감사합니다	좋아요
늘	늘	쿠팡맨	무거운
쿠팡맨	좋네요	또	저렴하
물	또	잘	늘
많이	물	쿠팡	생수
저렴하고	매번	수	유통기
수	저렴하고	생수	빠르고
자주	수	매번	항상
저렴한	많이	물은	저렴한
물은	생수	정말	가격도
배송	쿠팡	늘	주문하
또	물은	저렴하고	계속
로켓배송	정말	있어서	더
매번	더	자주	배송도
동원샘물	있어서	좋고	빠른
더	좋고	많이	그냥
마시고	없이	더	매번
가격에	자주	없이	주문했
생수	싸고	집앞까지	배송해
가격	친절하게	저렴해서	친절하
물맛도	로켓배송으로	싸고	다
있어서	집앞까지	역시	쿠팡

부정

Static	Non-static	Multi	Rand
물이	물이	물이	로켓배송
개	그냥	그냥	물이
다	좀	좀	물
좀	조금	조금	좀
얇아서	개	터져서	주문하
개가	다	다	배송이
이번에	터져서	얇아서	좋아요
터져서	얇아서	배송이	저렴하
찌그러져서	통이	다른	유통기
뚜껑	잡으면	통이	기사님
잡으면	왔는데	왔는데	항상
처음에	개가	잡으면	찌그러
조금씩	물통이	물통	빠르고
처음	배송이	줄줄	다른
새서	물통	개	배송
살짝	조금씩	조금씩	물은
물병이	주문	주문	매번
물	찌그러져서	개가	무거운
줄줄	총	물병이	구매해
얇아	했는데	물통이	하고
한	한손으로	생수가	이번에
왔네요	생수가	힘이	친절하
물통이	일	이번에는	너무
눌려서	다른	눌려서	저렴한
근데	줄줄	너무	가격도
조금	물병이	물병	개
왈칵	너무	자꾸	마시는
시켰는데	힘이	하나는	먹어봤
그런지	자꾸	한손으로	얇아서
왜	이번에는	가격이	물맛도

Model	ACC
CNN-static	0.70000
CNN-non-static	0.75000
CNN-multichannel	0.79000
CNN-rand	0.71000

실험 결과

01

02

03

04

05

06

❖ Format of results

16452,"{'예측값': 1, '실제값': array([0., 1.])}","OrderedDict([('아무리', 7.5705439980284677), ('에로영화라지만', 5.4930844795135112), ('이렇게', -3.4343334494547237), ('시대착오적일', -7.8837150512679459), ('수', -8.1239408364428449), ('있을까', -10.026943474971068), ('대세에서', -6.6812849427489667), ('한참', 6.0800347052073), ('빠지는', 4.8918806475680334), ('촌발날리는', -2.0154619328572436), ('러브씬들과', -3.086288236768012), ('배우들의', -0.53468614721551333), ('연기는', 4.5795720389206789), ('진짜', 3.606055154268927), ('국어책', -1.5018738568921131), ('읽는', -0.42479367135799412), ('우리', -3.1149122937831866), ('초딩들한테', -7.8916166709380455), ('미안해지는거다', -3.1103720690219028), ('특히', -3.0767151289084502), ('여주와', -2.4799685140713734), ('사장', -2.1679664904300449), ('두', -4.7167526707194067), ('사람이', -6.881994227561246), ('함께', -5.9618903076589955), ('대사를', 3.1944460656223419), ('주고', 7.336144789910648), ('받을때는', 1.6721463593003698), ('이', -1.5173537002015269), ('사람들이', -3.3757389634858601), ('어떤', -3.6819764558203625), ('감정인지도', -1.9408353708833022), ('알아내기', 6.6483964827342037), ('힘들', 11.628917735462213), ('지경이다', 15.512225799886497), ('차라리', 18.994444636317137), ('연기를', 20.898081628949491), ('알파고한테', 2.3935199949249193), ('시키면', 0.77416766908018231), ('더', 6.5241908848433923), ('잘할듯', 11.424342191878592), ('도무지', 24.334843770775699), ('손발이', 30.196296548196727), ('오그라들다', 22.069245423663752), ('못해', 8.4805071965512084), ('남아나지', 3.1873114013184516), ('않을', -2.8741088689097176), ('것', -0.15414958052426164), ('같아', 1.4482281627021383), ('보다가', 17.507393952248904), ('꺼버린', 12.044091515252539), ('영화', -2.7895662007681263), ('아', -13.283676954450346), ('내', -21.472688753948681), ('눈과', -26.475651898363886), ('귀가', -26.392057572147465), ('다', -10.173891120293645), ('썩어버린', 1.3757933689882895), (60, 3.4416249133705041), (61, -0.099599579684421347), (62, -0.099599579684421347), (63, -0.099599579684421347), (64, -0.099599579684421347), (65, -0.099599579684421347), (66, -0.099599579684421347), (67, -0.099599579684421347), (68, -0.099599579684421347), (69, -0.099599579684421347), (70, -0.099599579684421347), (71, -0.099599579684421347), (72, -0.099599579684421347), (73, -0.099599579684421347), (74, -0.099599579684421347), (75, -0.099599579684421347), (76, -0.099599579684421347), (77, -0.099599579684421347), (78, -0.099599579684421347), (79, -0.099599579684421347), (80, -0.099599579684421347), (81, -0.099599579684421347), (82, -0.099599579684421347), (83, -0.099599579684421347), (84, -0.099599579684421347), (85, -0.099599579684421347), (86, -0.099599579684421347), (87, -0.099599579684421347), (88, -0.099599579684421347), (89, -0.099599579684421347), (90, -0.099599579684421347), (91, -0.099599579684421347), (92, -0.099599579684421347), (93, -0.099599579684421347), (94, -0.099599579684421347), (95, -0.099599579684421347), (96, -0.099599579684421347), (97, -0.099599579684421347), (98, -0.099599579684421347), (99, -0.099599579684421347)])"

16452. 아무리 에로영화라지만 이렇게 시대착오적일 수 있을까 대세에서 한참 빠지는 촌발날리는 러브씬들과 배우들의 연기는 진짜 국어책 읽는 우리 초딩들한테 미안해지는거다 특히 여주와 사장 두 사람이 함께 대사를 주고 받을때는 이 사람들이 어떤 감정인지도 알아내기 힘들 지경이다 차라리 연기를 알파고한테 시키면 더 잘할듯 도무지 손발이 오그라들다 못해 남아나지 않을 것 같아 보다가 꺼버린 영화 아 내 눈과 귀가 다 썩어버린 Negative

실험 결과

❖ highlight on words which has high scores (WATCHA)

Model	Contents
CNN-static	2025. 이 영화는 뭘까 보면서 한참을 생각했다 뭐지 아니 보긴 봤으니까 무슨 말을 적긴 적어야 할텐데 내용인지 하나도 모르겠다 특이한 사건 없이 흘러가는 영화였고 갑자기 사랑에 빠졌듯 그래서 정말 너무 당황스러웠다 그그 새로운 여자가 나올 때부터 음 싶었고 결국 막판에 졸았다 진짜 보는 내내 지루해서 결말에 와아아아 드디어 끝난다 이런 생각이 들어서 기뻤다 뭔가 사랑이랑 뭐시기에 대해 말하는 것 같았지만 내가 빠가사리라 알아먹지 못했다 저들은 죽을까봐 무서워해 네가 여기 있는데 왜 죽어 어때 논리적이지 대사가 기억에 남는다 와이퍼의 법칙이랑 아 맞다 미친듯한 인물 클로즈업 감독 특징이라는데 그닥 다른 작품을 보고 싶지 않다 Negative
CNN-non-static	558. 허술한 스토리 그래픽 간만에 극장에서 본 건질게 없는 영화 상영시간은 또 왤케 길어 이걸 보느라 소비된 내 공짜 쿠폰도 아까움 ㅜㅜ Negative
CNN-multi channel	172. 솔직히 여태 봤던 시리즈가 재밌어서 이 정도의 평점을 유지하는 듯 모든 이야기를 끝 맺는 영화 결말이 궁금해 미치겠다면 보고 상관없다면 영원히 안 보는게 더 재미있을 것 같다 시리즈 중 유일하게 망작 Negative
CNN-rand	17257. 화담의 의식의 흐름 과 임수정의 존재 이유 를 납득할 수 없었으나 강동원이 진짜진짜 귀여워서 너무 행복했다 positive

실험 결과

❖ highlight on words which has high scores (IMDB)

Model	Contents
CNN-static	5703. First off I saw another reviewer said this movie was fantastic Well nothing could be further from the truth This is complete garbage A moronic horror comedy that NOT even slightly funny Don t take mean it s so bad good because not It a total waste of time and money Here what see in DVD group friends get together on weekend drunk then decide to make backyard video They grab Mom Dad camera Negative
CNN-non-static	673. This is without a doubt one of my favorite Columbo episodes ever The acting very well done the music catchy script ingenious and direction fabulous Peter Falk who acts brilliantly in every particularly this episode Also great performances from Stephen Caffrey Gary Hershberger Alan Fudge Robert Culp ending absolutely brilliant I love way describes it movie that WON T go amiss Positive
CNN-multi channel	818. This is one of Bruce s most underrated films in my opinion its an awesome heartwarming film with a neat story and amazing performance from Willis All the characters are great I thought Spencer Breslin were just together plus simply this definitely best comedic performances The waaaaaaaaamabulance thing was it very well written made as finale especially cool It good natured how you can see Russell Positive
CNN-rand	8289. I have seen this movie more than several times on TV. ALWAYS watch it again...NEVER turning the channel. This is full of chilling surprises and absolutely edge-of-your-seat suspenseful without being overbearing or stupid. Helen Hunt's talent magnificently shown in movie! recommend to anyone!!! Positive

실험 결과

❖ highlight on words which has high scores (COUPANG)

Model	Contents
CNN-static	430 물병이 약해요 찌그러지네용온도차때매그런가 개중에 개는 입구바로아래부분이 휘어있어요 글고 집을때 꽉집으면 패트가 꿀렁이면서 휙 쭈그러들어서물 울컥하고 나오는 경향이 물맛은 좋으나 음용하기에 불편이 따르기에 원래먹던 제주 삼다수로 갈아탑니다 동원 쏘리 Negative
CNN-non-static	228. 저렴하게 잘샀어요 찌그러짐없이 깨끗히 왔구요 네모난 모양일줄 알았는데 동그란 모양이네요 ㅠㅠ 쌀통할겸 샀는데 그래도 저렴한가격에 잘샀구요 택배아저씨도 엘베없는 층인데 힘든기색없이 친절하게 갖다주셧네요 ㅎㅎ 쿠팡짱 Positive
CNN-multi channel	784. 물을워낙 많이먹어서 다소저렴한 동원을 주문하는데 병이너무 얇아 불편해요 처음엔 자꾸 쏟게되네요 Negative 712. 젤 싸서 먹고있는중이에요 용기가 너무 말랑해서 잡을때 불편한거 빼고는 물맛도 괜찮고 무엇보다 가격이 착해서 좋아요 Positive
CNN-rand	493. 배송은 빨랐어요근데 물병이 이상해요 비닐도아니구 물이들었는데 넘 얇아서 근방이라도찢어질것같아요한손으로 물병을들면 재질이얇아서 흐느적거려서 물이쏟아져요 컵에딸아먹을수가없네요 동원샘물 실망입니다 Negative

결과 요약

01

02

03

04

05

06

- Convolutional Neural Networks와 Weakly Supervised Learning을 활용해 WATCHA, IMDB, COUPANG 데이터에 Sentence attention task 수행
- 스코어 상위 단어들을 살펴본 결과 극성 단어들이 비교적 많은 것으로 나타남
- 문장별 정성 평가 결과 제안 모델의 성능이 비교적 우수함을 확인
- 분류 정확도 면에서는 Kim(2014) 결과처럼 데이터마다 우수 모델이 다르나, 연구목적인 Sentence attention task에 있어서는 기법 간 차이가 크지 않은 경향 확인
- 제안 모델은 앞으로 질의/응답(Q&A), 요약 등에 활용될 수 있을 것으로 기대