

데이터 분석 캡스톤 디자인 중간 발표

**2016100946
산업경영공학과
김효준**

목차

- 주제
- 데이터
- 연구 방법론
- 분석 과정
- 결과
- 향후 진행 방향

주제

■ 악성 댓글 분류 모델 성능 개선

- 최근, SNS가 활발해지면서 영화, 쇼핑, 뉴스 등 다양한 산업 분야에서 많은 익명의 리뷰들이 등장
- 익명의 특성으로 인해 단순 욕설, 차별적인 말, 타인 비하 등 불쾌감을 주는 리뷰로 인해 피해 받고 있는 실정
- 현재, 네이버 AI 악플 감지기 서비스가 있지만, 반어적인 어투 또는 비꼬는 말 등의 여러 악성 리뷰를 판 단하지 못하는 경우가 있기에 더 높은 정확도가 요구되는 모델 필요
- 더구나, 크롤링 기법을 통해 악성 데이터를 수집하는데 있어 진입장벽이 낮지만, 이러한 데이터를 일일이 악성 라벨링 해주는데 들어가는 시간적 비용이 막대한 실정

궁극적 목적



- 한국어 전처리 기법 적용
- Text Classification 모델 적용
- Data Augmentation 도입하여 기존 연구와의 차별성 부여

데이터셋

■ 사용 데이터셋

욕설 감지 데이터셋

커뮤니티 사이트

→ 단순 욕설, 인종 차별적인 말, 정치적 갈등을 조장하는 말, 성적·성차별적인 말, 타인을 비하하는 말, 그 외에 불쾌감을 주거나 욕설로 판단되는 말

	document	label
0	좌배 까는건 ㅇㅈ	1\n
1	집에 롱 패딩만 세 개다. 10년 더 입어야지 ㅋㅋ	0\n
2	개소리야 니가 빨갱이를 옹호하고 드루킹을 ㅇㅇ짓이라고 잘못해서 빠진거야 빨갱아	1\n
3	세탁이라고 봐도 된다	0\n
4	애새끼가 초딩도 아니고 ㅋㅋㅋㅋ	1\n
...
5820	좌우 헬파이어 3개씩 6개 장착에 아파치보다 약하지만 20mm 기관포 장착임	0\n
5821	세금 내놓으라고 데모질 중 ㅋㅋ간첩, 도둑놈 새끼들이 대통령 해쳐먹으니까 나도 같...	1\n
5822	너가 한 말 중에	0\n
5823	제갈대중 ㅇㅈ	0\n
5824	우리나라교회는 악마들이모여 주영이 저벌리고	1\n

출처 : <https://github.com/2runo/Curse-detection-data>

→ Text classification

AI 허브 데이터셋

인공지능 윤리 연구를 위한 비정형 텍스트 데이터

	네놈이 죽었으니 영원히 못잊겠지 돼지새끼 뽀뽀함 지리구요. 인간성 더러운거 오지구요.
0	김양건이 영웅대접 받아서 또 질투났나보네.자기 위치를 조금 불안하게 할 것 같은 사...
1	암살하고 조문... 짜속 많이 늘었군
2	교통사고인데 되게 말짱하네요
3	아니 스.벌. 과거에 독립군들 그리 많았던 민족 맞냐..??어떻게 정은이 한명 암살...
4	차도 없는 나라에서 어떻게 하면 교통사고로 죽을 수 있지
...	...
9999994	정희 재연이 다쳐볼어라
9999995	좌빨 운운하는새끼들봐라. 일의 시작이 나쁘다는거냐?과정이 더럽다는거다. 그리고 그놈...
9999996	아니근데 뉴스 기레기놈들은 왜 꼬박꼬박 돼지한테 위원장이라고 존칭을 붙이냐??
9999997	가족을 벗기고 눈알을 파내고 코와 귀를 베어내는 등 빨갱이는 빨갱이들이 쓰는 형벌로...
9999998	분명 정상은 아니라고 봤었는데 역시나ㅋㅋ 공지영, 한상열,이정희... 같은 애들도 ...

9999999 rows × 1 columns

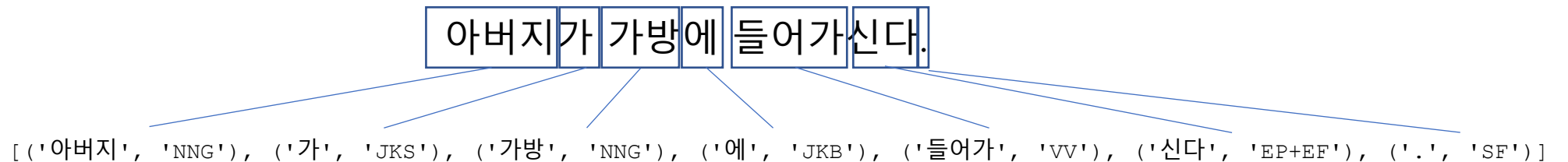
출처 : <https://aihub.or.kr/node/276>

→ Word2Vec 임베딩 학습, 준지도 학습

연구 방법론

1. 텍스트 전처리

- 형태소 분석



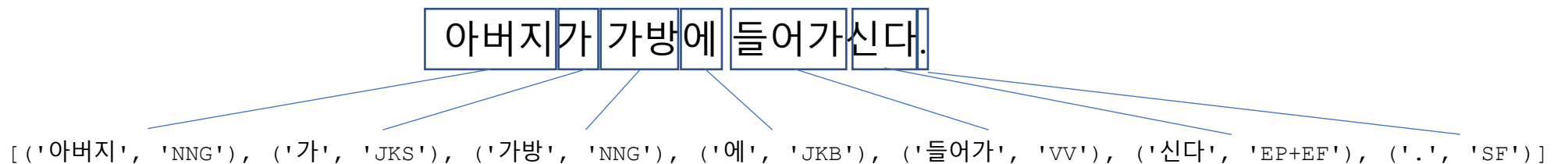
- 형태소 분석 라이브러리

- ① Hannanum : KAIST의 SWRC (Semantic Web Research Center)에서 Java로 작성된 형태소 분석기 및 POS 태거
- ② Kkma (꼬꼬마 형태소 분석기) : 서울대 지능형 데이터 시스템 (IDS) 연구실에서 개발 한 자바로 작성된 형태소 분석기 및 자연어 처리 시스템
- ③ Komoran : 2013 년부터 Shineware에서 개발한 Java로 작성된 비교적 새로운 오픈 소스 한국어 형태소 분석기
- ④ Mecab : 일본 형태소 분석가에 의해 개발되었지만 추후 한국어에 적합하게 수정됨
- ⑤ Okt : 과거 Twitter()와 동일. 한국어 트위터 기반으로 작성된 라이브러리

연구 방법론

1. 텍스트 전처리

- 정수 인코딩



`tf.keras.preprocessing.text.Tokenizer()`

: `fit_on_texts()`안에 코퍼스를 입력으로 하면 **빈도수를 기준으로** 단어 집합을 생성한다.

아버지가 가방에 들어가신다.



[13, 1, 10, 2, 30, 2, 3]

연구 방법론

2. 워드 임베딩

■ 희소 표현(Sparse Representation)

아버지가 가방에 들어가신다.

아버지	[1, 0, 0, 0, 0, 0, 0]
가	[0, 1, 0, 0, 0, 0, 0]
가방	[0, 0, 1, 0, 0, 0, 0]
에	[0, 0, 0, 1, 0, 0, 0]
들어가	[0, 0, 0, 0, 1, 0, 0]
신다	[0, 0, 0, 0, 0, 1, 0]
.	[0, 0, 0, 0, 0, 0, 1]

13
:단어 수

아버지	[1, 0, 0, 0, 0, 0, 0, 0, 0, 0]
가	[0, 1, 0, 0, 0, 0, 0, 0, 0, 0]
가방	[0, 0, 1, 0, 0, 0, 0, 0, 0, 0]
에	[0, 0, 0, 1, 0, 0, 0, 0, 0, 0]
들어가	[0, 0, 0, 0, 1, 0, 0, 0, 0, 0]
시고	[0, 0, 0, 0, 0, 1, 0, 0, 0, 0]
어머니	[0, 0, 0, 0, 0, 0, 1, 0, 0, 0]
가	[0, 1, 0, 0, 0, 0, 0, 0, 0, 0]
방	[0, 0, 0, 0, 0, 0, 0, 1, 0, 0]
에	[0, 0, 0, 1, 0, 0, 0, 0, 0, 0]
들어가	[0, 0, 0, 0, 1, 0, 0, 0, 0, 0]
신다	[0, 0, 0, 0, 0, 0, 0, 0, 1, 0]
.	[0, 0, 0, 0, 0, 0, 0, 0, 0, 1]

10 : 단어 사전 개수

Shape : (10, 13)
→ N개 문장
→ (N, 10, 13)

연구 방법론

2. 워드 임베딩

10 : 단어 사전 개수

■ 희소 표현(Sparse Representation)

문장 수가 많아진다면??

- 단어 수도 증가
- Sparse Vector 형성

```
[1, 0, 0, 0, 0, 0, 0, 0, 0, ..., 0]  
[0, 1, 0, 0, 0, 0, 0, 0, 0, ..., 0]  
[0, 0, 1, 0, 0, 0, 0, 0, 0, ..., 0]  
[0, 0, 0, 1, 0, 0, 0, 0, 0, ..., 0]  
...  
...  
[0, 0, 0, 0, 0, 0, 0, 0, 1, ..., 0]  
[0, 0, 0, 1, 0, 0, 0, 0, 0, ..., 0]  
[0, 0, 0, 0, 1, 0, 0, 0, 0, ..., 0]  
[0, 0, 0, 0, 0, 0, 0, 0, 0, ..., 0]  
[0, 0, 0, 0, 0, 0, 0, 0, 0, ..., 1]
```

한 문장 Shape : (20, 100000)

연구 방법론

2. 워드 임베딩

■ 밀집 표현(Dense Representation)

희소 표현

```
[1, 0, 0, 0, 0, 0, 0, 0, ..., 0]
[0, 1, 0, 0, 0, 0, 0, 0, ..., 0]
...
...
[0, 0, 0, 0, 0, 0, 0, 0, ..., 0]
[0, 0, 0, 0, 0, 0, 0, 0, ..., 1]
```

한 문장 Shape : (20, **10000**)

Embedding(input_dim = 10000,
output_dim = 32,
input_length=None)



```
[0.54, 1.24, -0.1, ..., 0.46]
[0.11, -1.34, -1.21, ..., -0.52]
...
...
[1.01, -0.94, -0.61, ..., 0.06]
[0.21, -1.12, -0.27, ..., 0.38]
```

한 문장 Shape : (20, **32**)

연구 방법론

2. 워드 임베딩

■ 분산 표현(Distributed Representaion)

'비슷한 위치에서 등장하는 단어들은 비슷한 의미를 가진다'

: 저차원에 **단어의 의미를 여러 차원에다가 분산**하여 표현

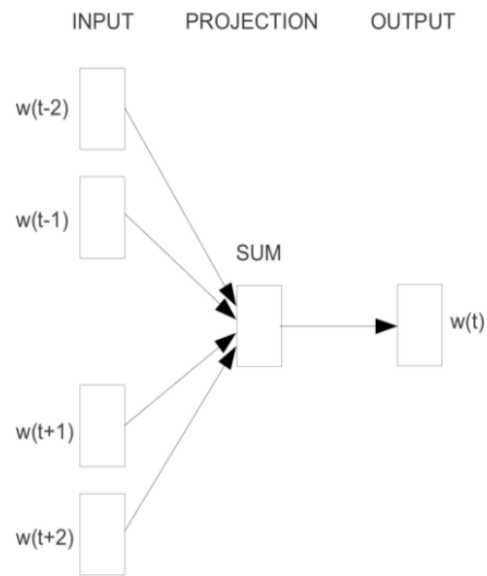
→ 그 단어를 표현하고자 주변을 참고하여 단어를 표현하는 방법

대표적인 워드 임베딩 방법론 : LSA, **Word2Vec**, FastText, Glove 등

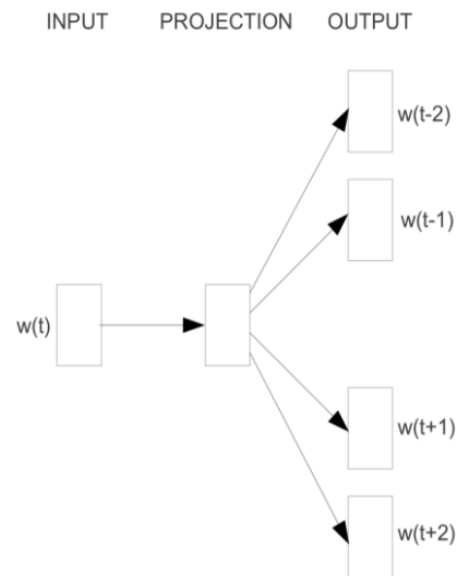
연구 방법론

3. Word2Vec (워드투벡)

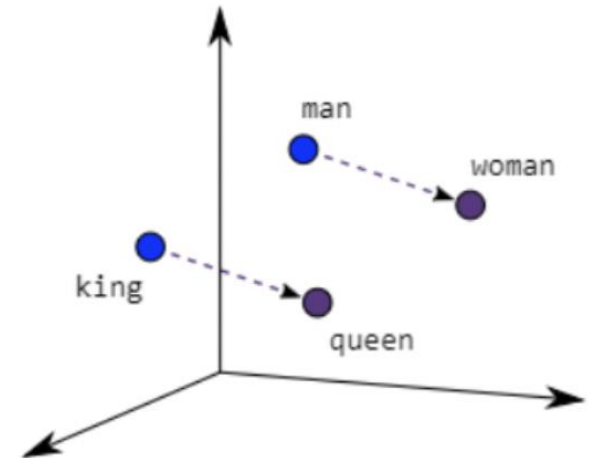
“단어의 주변을 보면 그 단어를 안다”



CBOW

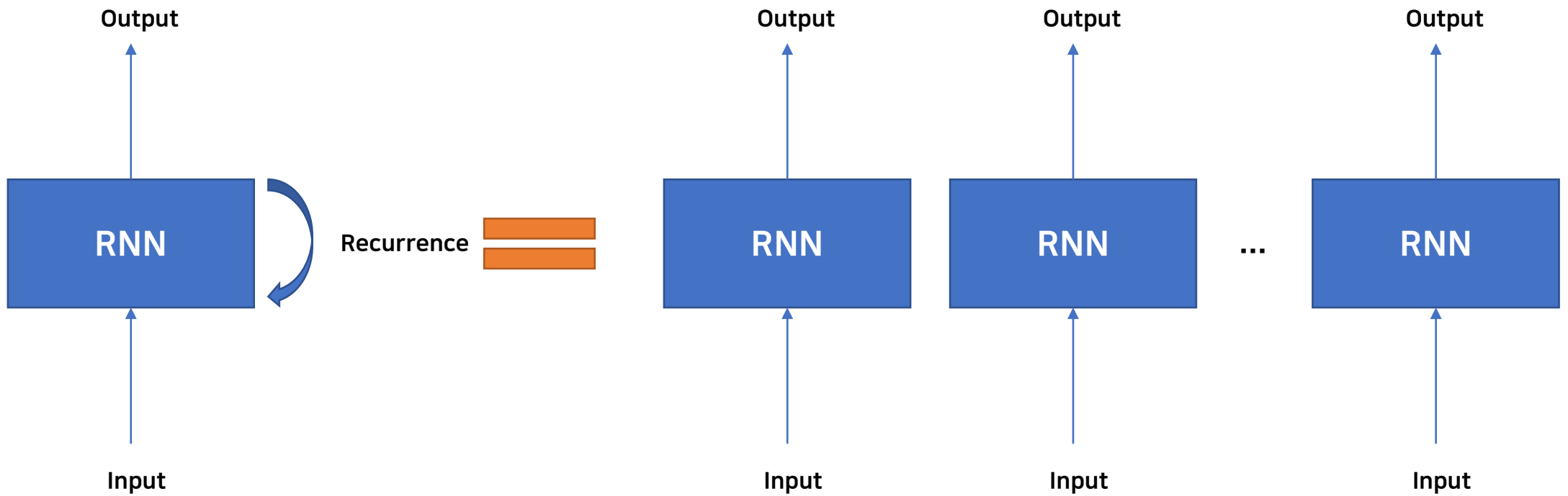


Skip-gram



4. Text Classification Model

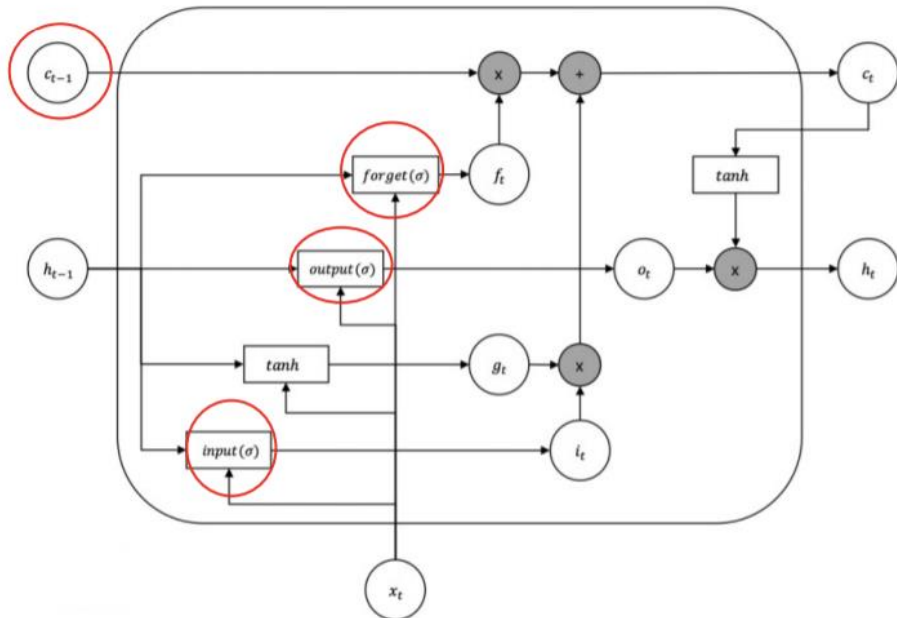
■ RNN



연구 방법론

4. Text Classification Model

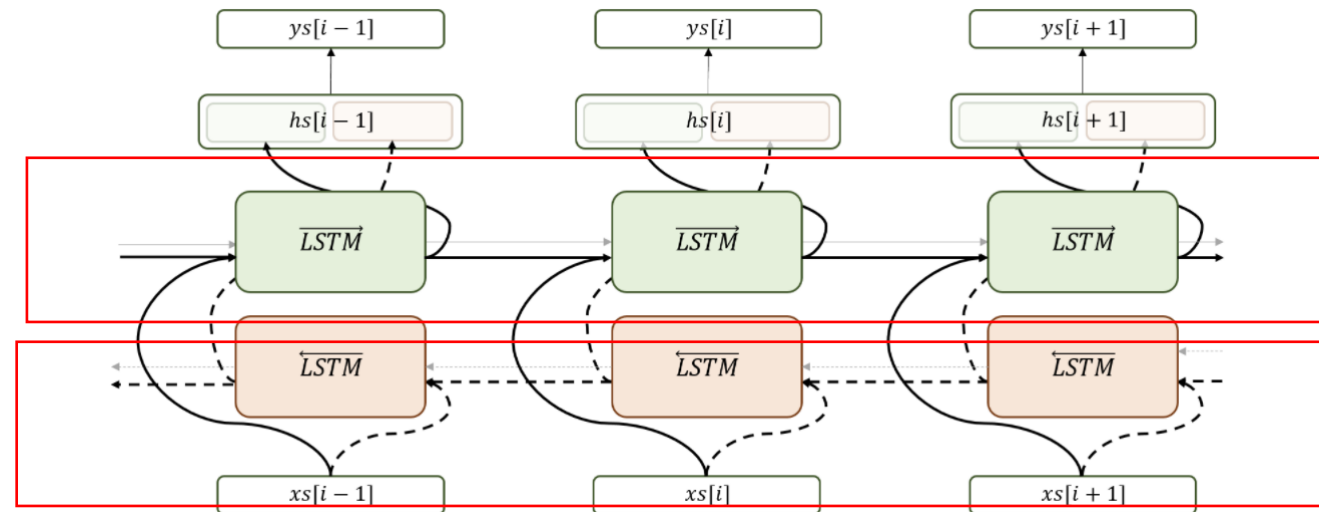
■ LSTM Architecture



■ Bidirectional LSTM

나는 ____를 뒤집어 쓰고 펑펑 울었다.

$P(\text{이불/나는} + \text{____}) \lll P(\text{이불/를 뒤집어 쓰고 펑펑 울었다.})$

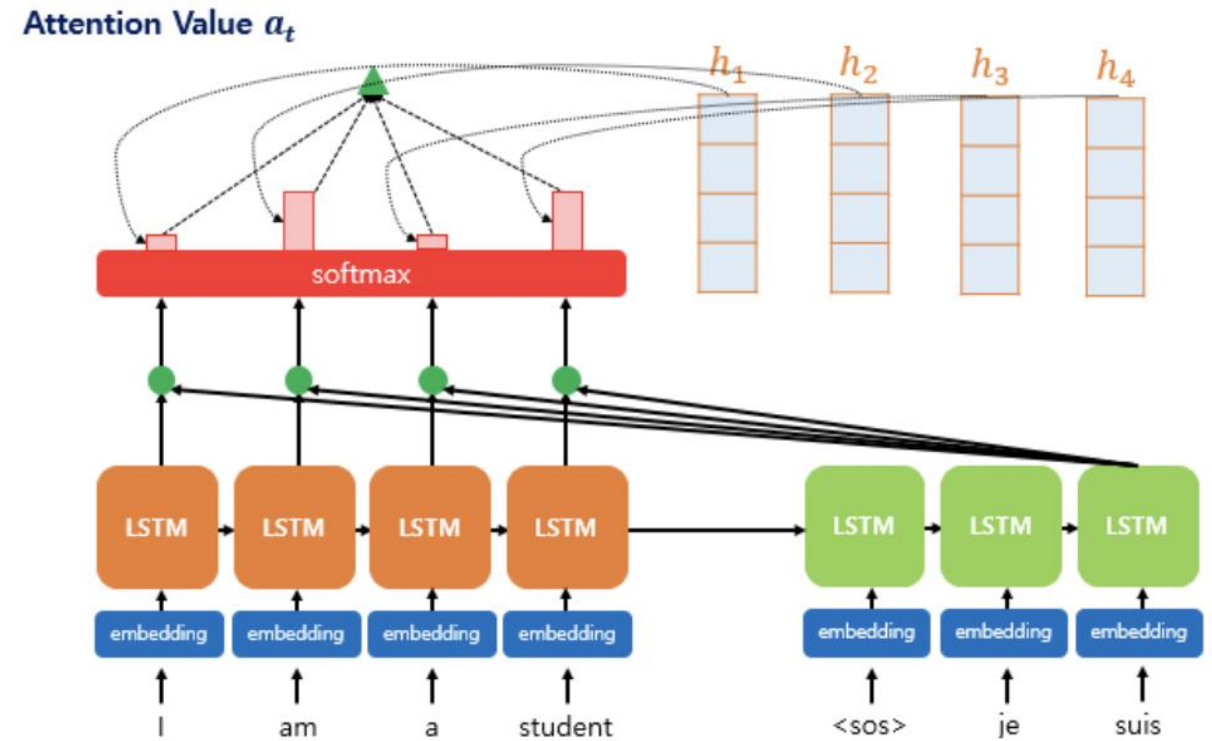


연구 방법론

5. 어텐션 메커니즘 (Attention Mechanism)

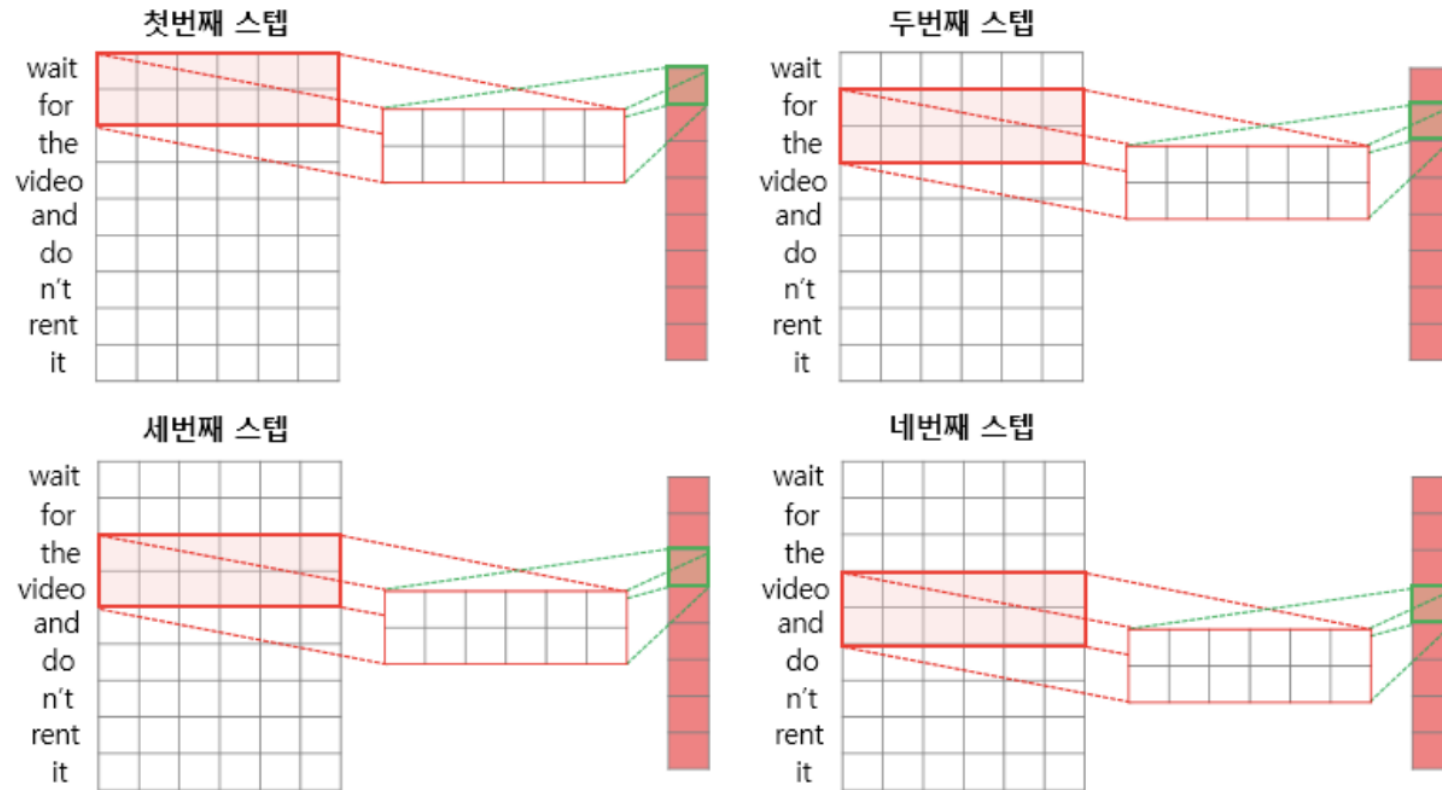
■ 문제점

1. 하나의 고정된 크기 벡터에 모든 정보를 압축 시, 정보 손실 발생
2. RNN의 고질적인 문제인 기울기 소실 문제



연구 방법론

6. 1D CNN



분석 과정

1. 텍스트 형태소 분석

형태소 분석기	한국어 욕설 데이터셋	시간(초)
Hannanum	7896건	24.760
Kkma	7896건	248.536
Komoran	7896건	5.726
Mecab	7896건	1.049
Okt(twitter)	7896건	17.44



Mecab 라이브러리 사용

분석 과정

1. 텍스트 형태소 분석- 결과

	document	label	tokenized
2328	안그래도 요즘 숙박업도 올상일텐데.....33오오 모이지말고. 시키들아.	1	[안, 그래도, 요즘, 숙박업, 올상, 일텐데, .,, 33, 오, 오,...
51	그냥 흘딩하세요	0	[그냥, 흘딩, 세요]
4897	빙신들!!!!!! 저 와꾸에 저 피지컬이면 한국어 좀 배워서 김치만 쳐먹어도 될 1...	1	[빙신, I, !!!!!, 저, 꾸, 저, 피지컬, 한국어, 좀, 배워서, 김치, ...
2133	정신 피폐해지고, 돈도 나눠 줘야 하고, 애까지 낳으면 어쩔	0	[정신, 피폐, 해, ,, 돈, 나눠, 줘야, ,, 애, 까지, 낳, 으면, 어쩔]
5773	예를들면 조선이 -100 -> 일제가 -50라는 거임. 둘 다 지금처럼 완전히 자...	0	[예, 조선, -, 100, ->, 일제, -, 50, 라는, 거, ,, 둘, 지금...
...
3772	한민관 존잘이네	0	[한민관, 존, 잘]
5191	한달에 한두번 집에 갈 수 있는 그런 미개한 근로 조건이라면 누가 하겠음?	0	[달, 한두, 번, 집, 갈, 수, 있, 그런, 미개, 근로, 조건, 라면, 누가,...
5226	힘있어도 저지랄하는건 똑같은듯?	0	[힘, 있, 어도, 저, 지랄, 건, 똑같, ?]
5390	와 씹고수 ㅋㅋㅋ	1	[씹, 수, ㅋㅋㅋ]
860	ㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋㅋ표정	0	[ㅋㅋㅋ, ㅋㅋㅋ, ㅋㅋㅋ, ㅋㅋㅋ, ㅋㅋㅋ, ㅋㅋㅋ, 표정]

분석 과정

2. 정수 인코딩 + 패딩 적용

```
array(['인민', '감시', '시스템', '구축', '위해서', '안면', '인식', '기술', '데이터', '로', '쓰',  
      '려고', '중국', '공산당', '큰', '그림', '그린', '거'], dtype='<U3')
```

정수 인코딩

```
array([[2875, 2876, 1237, 2145, 1082, 4455, 604, 298, 2146, 19, 133,  
       357, 168, 2877, 412, 4456, 4457, 4]])
```

패딩

```
array([[ 0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  
        0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  
        0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  
        0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  
        0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  
        0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  0,  
        0,  0,  0,  0,  0, 2876, 2877, 1238, 2146, 1083, 1,  
605, 299, 2147, 20, 134, 358, 169, 2878, 413, 1, 1,  
5])
```

분석 과정

3. 최종 데이터

Train Data

```
array([[ 0,  0,  0, ..., 431,  12,   2],
       [ 0,  0,  0, ...,  61,   1, 356],
       [ 0,  0,  0, ...,  29,  55, 1737],
       ...,
       [ 0,  0,  0, ...,  67, 535,   4],
       [ 0,  0,  0, ..., 123,  45,   3],
       [ 0,  0,  0, ...,   3,   3, 1042]])
```

(4368, 100)

Test Data

```
array([[ 0,  0,  0, ..., 124,  54,   8],
       [ 0,  0,  0, ...,   0,   0, 394],
       [ 0,  0,  0, ...,  60,  60,  60],
       ...,
       [ 0,  0,  0, ..., 3999, 714, 1609],
       [ 0,  0,  0, ..., 1177, 298,  459],
       [ 0,  0,  0, ...,   0, 294,  101]])
```

(1456, 100)

분석 과정

4. Word2Vec 임베딩 벡터

한국어 Pretrained 된 모델을 임베딩 가중치로 사용하였지만, 성능 개선이 보이지 않음

왜 그럴까?

도메인이 다르기에 댓글에 국한된 W2V 임베딩 벡터가 더 문맥 내용을 잘 담고 있을 것이다.



W2V 모델을 직접 학습시켜 저장시킨 모델을 임베딩 가중치로 사용하자!

분석 과정

4. Word2Vec 임베딩 벡터

네놈이 죽었으니 영원히 못잊겠지 돼지새끼 뽀뽀함 지리구요. 인간성 더러운거 오지구요.	
0	김양건이 영웅대접 받아서 또 질투났나보네.자기 위치를 조금 불안하게 할 것 같은 사...
1	암살하고 조문... 짜속 많이 늘었군
2	교통사고인데 되게 말짱하네요
3	아니 스.벌. 과거에 독립군들 그리 많았던 민족 맞냐..??어떻게 정은이 한명 암살...
4	차도 없는 나라에서 어떻게 하면 교통사고로 죽을 수 있지
...	...
9999994	정희 재연이 다쳐넣어라
9999995	좌빨 운운하는새끼들봐라. 일의 시작이 나쁘다는거냐?과정이 더럽다는거다. 그리고 그놈...
9999996	아니근데 뉴스 기레기놈들은 왜 꼬박꼬박 돼지한테 위원장이라고 존칭을 붙이냐??
9999997	가죽을 벗기고 눈알을 파내고 코와 귀를 베어내는 등 빨갱이는 빨갱이들이 쓰는 형벌로...
9999998	분명 정상은 아니라고 봤었는데 역시나ㅋㅋ 공지영, 한상열,이정희... 같은 애들도 ...

9999999 rows × 1 columns

리소스 한계로 6백만건에 대해 W2V 학습 진행

➔ Skipgram 방법론으로 진행

분석 과정

4. Word2Vec 임베딩 벡터

`model_sg.similar_by_vector('사랑', topn=100)`

```
[('존경', 0.7059704661369324),  
 ('~♪~♪~♡', 0.648003876209259),  
 ('애틀', 0.6235745549201965),  
 ('끔찍히', 0.6187503337860107),  
 ('진정코', 0.6181071996688843),  
 ('경외', 0.6144888401031494),  
 ('로우신', 0.6112179160118103),  
 ('~~^♡^', 0.6105014085769653),  
 ('부디부디', 0.6011748909950256),
```

`model_sg.similar_by_vector('ㅅㅅㅅ', topn=1000)`

```
[('시바', 0.8410485982894897),  
 ('시발', 0.8202505111694336),  
 ('ㅅㅅ발', 0.7903104424476624),  
 ('ㅅㅅ표', 0.7833770513534546),  
 ('아오', 0.782518208026886),  
 ('슈바', 0.7626177072525024),  
 ('시밤', 0.7533204555511475),  
 ('젠장', 0.7490739226341248),  
 ('어휴', 0.7402772903442383),  
 ('ㅅㅅㅅ', 0.7374851703643799),  
 ('ㅅㅅㅅ', 0.7353333333333333),
```



Word2Vec 학습을 통해 악플 댓글 데이터라는 도메인의 문맥을 잘 반영

결과

모델	Test 정확도
DNN	0.6452
LSTM	0.8296
Bidirectional-LSTM	0.833
Bidirectional-LSTM-2	0.830
1D-CNN	0.843
Bidirectional-LSTM + Attention	0.863

여러가지 실험 중 가장 최고의 성능을 기술

→ 그 중, Word2Vec 임베딩 가중치를 사용하였을 때 성능이 개선됨을 확인

→ 활성화 함수, optimizer 변경 등을 통한 파라미터 개선이 필요

향후 진행 방향

1. 텍스트 전처리

- 중복 단어의 정규화
Ex) .. , ...
- 토큰화 이루어지지 않은 단어 처리
Ex) 비트, 코인 → 비트코인



모델 성능 개선

향후 진행 방향

2. 모델 학습 원리 공부 및 적용

- 모델의 학습 파라미터 개념 미숙
- Optimizer, node 등 성능 개선에 투자



모델 성능 개선

향후 진행 방향

3. Data Augmentation 적용

EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks

Jason Wei^{1,2} Kai Zou³

¹Protago Labs Research, Tysons Corner, Virginia, USA

²Department of Computer Science, Dartmouth College

³Department of Mathematics and Statistics, Georgetown University

jason.20@dartmouth.edu kz56@georgetown.edu

Model	Training Set Size			
	500	2,000	5,000	full set
RNN	75.3	83.7	86.1	87.4
+EDA	79.1	84.4	87.3	88.3
CNN	78.6	85.6	87.7	88.3
+EDA	80.7	86.4	88.3	88.8
Average	76.9	84.6	86.9	87.8
+EDA	79.9	85.4	87.8	88.6

Table 2: Average performances (%) across five text classification tasks for models with and without EDA on different training set sizes.

Data Augmentation Using Pre-trained Transformer Models

Varun Kumar

Alexa AI

kuvrun@amazon.com

Ashutosh Choudhary

Alexa AI

ashutoch@amazon.com

Eunah Cho

Alexa AI


eunahch@amazon.com

Understanding Back-Translation at Scale

Sergey Edunov[△] Myle Ott[△] Michael Auli[△] David Grangier^{▽*}

[△]Facebook AI Research, Menlo Park, CA & New York, NY.

[▽]Google Brain, Mountain View, CA.

- 
- 한국어 적용
 - 새로운 Data Augmentation 제시
 - 준지도 학습 ?
 - 궁극적으로 모델 성능 개선

감사합니다