

데이터 분석 사례 연구

산업경영공학과
2016100946 김효준


1. 악플 데이터셋 탐지 분류기

1> 관련 데이터셋

1) Korean HateSpeech Dataset : from Korean entertainment news aggregation platform

kocohub/korean-hate-speech

We provide the first human-annotated Korean corpus for toxic speech detection and the large unlabeled corpus. The data is comments from the Korean entertainment news aggregation

 <https://github.com/kocohub/korean-hate-speech>

KOCO

- 9,381개의 human-labeled comments
- 2,033,893개의 unlabeled comment

```
comments contain_gender_bias bias hate
송중기 시대극은 믿고본다. 첫회 신선하고 좋았다. False none none
지현우 나쁜놈 False none offensive
알바쓰고많이만들면되지 돈욕심없으면골목식당왜나온겨 기뻐기게나하고 산에가서팔어라 False none hate
설마 ㅈ 현정 작가 아니지?? True gender hate
```

2) 욕설 감지 데이터셋

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

- 일간베스트, 오늘의 유머 등 각종 커뮤니티 사이트의 댓글에 대해 분류

- 단순 욕설, 인종 차별적인 말, 정치적 갈등을 조장하는 말, 성적·성차별적인 말, 타인을 비하하는 말, 그 외에 불쾌감을 주거나 욕설로 판단되는 말
- 5825 개의 labeled comments

좌배 까는건 ㅇㅂ|1
 집에 룬 패딩만 세 개다. 10년 더 입어야지 ㅋㅋ|0
 개소리야 니가 빨갱이를 옹호하고 드루킹을 ㅇㅇ짓이라고 잘못해서 빠진거야 빨갱아|1
 세탁이라고 봐도 된다|0
 애새끼가 초딩도 아니고 ㅋㅋㅋㅋ |1

3) AI 허브 데이터셋 (라벨링 X) - 약 백만건

| | 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?과정의 더럽다는거다. 그리고 그놈... |
| 9999996 | 아니근데 뉴스 기레기놈들은 왜 꼬박꼬박 돼지한테 위원장이라고 존칭을 붙이냐?? |
| 9999997 | 가족을 버리고 눈알을 파내고 코와 귀를 베어내는 등 빨갱이는 빨갱이들이 쓰는 형벌로... |
| 9999998 | 분명 정상은 아니라고 봤었는데 역시나ㅋㅋ 공지영, 한상열,이정희... 같은 애들도 ... |

- 인공지능 윤리 연구를 위한 비정형 텍스트 데이터셋
 - 1차년도: 뉴스기사 댓글 7,000만 건, 트위터 3,000만 건
 - 2차년도: 온라인커뮤니티 (일베저장소) 댓글4,500만건
 - 3차년도: 온라인커뮤니티 (일베저장소) 댓글2,000만건
- 욕설이 포함되지 않은 비윤리 데이터는 문맥추출을 위한 딥러닝 모델 개발에 주요 데이터로 사용됨

2> 분석 방향

1. LDA 토픽 모델링

- LDA(잠재 디리클레 할당) 방법론 : 추출한 문서에 담긴 단어들의 주제(토픽)을 추출하는 '토픽모델링' 기법 중 하나
- 기존 논문

에너지공학 ISSN 1598-7981, 제29권 제2호(2020)
Journal of Energy Engineering, Vol. 29, No. 2, pp.23-29(2020)
<https://doi.org/10.5855/ENERGY.2020.29.2.023>

LDA 기법을 이용한 미세먼지 이슈의 토픽모델링 분석

윤순옥* · 김민철**†

*녹색기술센터 정책연구부 Post-doc., **녹색기술센터 정책연구부 선임연구원
(2020년 4월 3일 접수, 2020년 5월 15일 수정, 2020년 5월 20일 채택)

Topic Modeling on Fine Dust Issues Using LDA Analysis

Yoon soonuk* · Kim Minchul**†

Green Technology Center

(Received 3 April 2020, Revised 15 May 2020, Accepted 20 May 2020)

요 약

본 연구에서는 최근 10년간의 미세먼지 관련 뉴스 데이터를 수집하여 LDA 분석을 통해 최적 토픽을 도출하였다. 최적 토픽으로 선별된 80개의 이슈를 미세먼지 정책의 시각에서 해석하였다. 연구결과, 기온과 같은 날씨와 관련된 정보와 미세먼지 농도가 관련되어서 이슈화되는 경향이 있었다. 다음으로 미세먼지 저감 대책의 일환으로 노후경유차 운행 제한 제도와 저감 장치 부착과 같은 이슈의 빈도수가 높았다. 국민에 대한 제도 변경 안내를 포함하여 시민과 운수업자와의 갈등도 주요한 토픽으로 나타났다. 미세먼지 문제의 해결을 위한 수소차 보급과 같은 대안도 주요 토픽으로 분석되었다. 또한 미세먼지 관련 공기청정기 등 제품 관련 주제, 취약계층을 미세먼지로부터 보호하는 정책과 관련된 주제, 연구개발을 통한 미세먼지 저감 관련 주제가 주요 화두로 제기되었다. 미세먼지 대책은 사회 이슈로 정부 정책과 밀접한 관련이 있다고 볼 수 있다. 또한 본 연구를 통해 토픽 상에서는 거시적인 정부정책 자체보다는 시민의 안전, 시혜적인 정책이나 이해관계자간의 갈등이 정부정책 변화와 연동하여 중요한 의미를 지니는 것으로 나타났다.

주요어 : 미세먼지, 잠재적 디리클레 할당모형(LDA), 토픽 모델링, 정부 정책, 공기청정기, 취약계층, 연구 개발

: 각 토픽의 주요 단어를 분석해서 각 토픽이 어떤 기사를 의미하는지를 분석

- 주요 결과
 - 날씨관련 정보가 토픽의 주요 단어를 구성하고 있었고, "5도", "6도" 등의 10도 이하 낮은 기온 단어가 많이 등장하였음
 - ⇒ 미세먼지의 농도가 높음을 파악할 수 있었다.

• 본 연구 : topic별 악플 분석

⇒ 악플의 유형(즉, 토픽)을 파악하는데 기여하지 않을까 고안

1) Hatespeech dataset: contain_gender_bias, hate

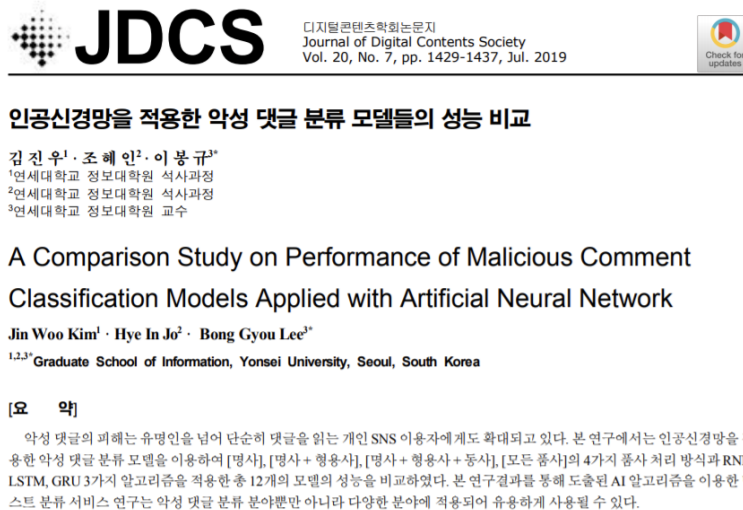
2) 욕설 감지 데이터셋: 단순 욕설, 인종 차별적인 말, 정치적 갈등을 조장하는 말, 성적·성 차별적인 말, 타인을 비하하는 말, 그 외에 불쾌감을 주거나 욕설로 판단되는 말

3) AI 허브 데이터셋 : 욕설이 포함되지 않은 비윤리 데이터

⇒ 추후 시각화 활용 PyLDAvis

2. 악성 댓글 분류 모델

- 기존 논문 : 12개의 task 성능 비교



- 4가지 품사 처리 (명사, 명사+형용사, 명사+형용사+동사, 모든 품사)* 3가지 모델 (RNN, LSTM, GRU)
- 본 연구: BiLSTM, CNN+BiLSTM, BERT, .. 등 모델 추후 모색
 - **Semi supervised Learning 방법을 추가**
 - ⇒ 입력변수만 존재하는 악플 데이터들도 학습 과정에 포함시켜 예측 모델의 성능을 향상
 - 1) Pseudo-Label (수도 레이블링)
 - *Labeled Data*로 *Model*을 먼저 학습
 - 학습된 모델을 사용하여, *Unlabeled Data*를 예측하고 그 결과를 *Label*로 사용하는 *Pseudo-labeled data* 생성
 - *Pseudo-labeled data*와 *Labeled data*를 모두 사용하여 다시 그 모델을 학습
 - 2) Graph-based

3> 웹 구현

- Flask 기반
 - 악플 탐지 서비스
 - 댓글 분석 (특정 단어 탐지)